The Use of Simplified or Misspecified Models: Linear Case

Shaohua Wu, T. J. Harris*, K. B. McAuley

Department of Chemical Engineering, Queen's University, 19 Division Street, Kingston, ON Canada K7L 3N6

Simplified models have many appealing properties and sometimes give better parameter estimates and model predictions, in sense of mean-squared-error, than extended models, especially when the data are not informative. In this paper, we summarize extensive quantitative and qualitative results in the literature concerned with using simplified or misspecified models. Based on confidence intervals and hypothesis tests, we develop a practical strategy to help modellers decide whether a simplified model should be used, and point out the difficulty in making such a decision. We also evaluate several methods for statistical inference for simplified or misspecified models.

Keywords: simplified/misspecified models, mean-squared-error, non-central F distribution, non-centrality parameter

INTRODUCTION

Chemical engineers develop simplified models (SMs) and use them for simulating, designing, controlling and optimizing many different types of processes (Brendel et al., 2006; Chang et al., 2005; Romdhane and Tizaoui, 2005; Golbert and Lewin, 2004; Lv et al., 2004; Maria, 2004; Mchaweh et al., 2004; Bagajewicz and Cabrera, 2003; Yoshida et al., 2003; Perregaard, 1993). SMs are developed when the modeller lacks an understanding of some of the underlying phenomena in complex processes, or there are insufficient data to adequately calibrate or validate an extended model (EM). An extended model is a mathematical description of the process that contains sufficient phenomenological detail to provide good predictions over the range of conditions of interest, including those experimental conditions where data may not have been collected. SMs often have reduced input and computational requirements compared with EMs, making them more portable and less expensive to use and maintain (Brooks and Tobias, 1996; Rexstad and Innis, 1985; Innis and Rexstad, 1983). SMs usually contain fewer unknown parameters, which are more readily and precisely estimated than the numerous parameters that appear in EMs. Problems of parameter estimability in EMs are more pronounced when limited experimental data are available. However, SM may fail to account for important phenomena, resulting in poor operating decisions and ill-conceived process designs. Although formal techniques have been developed for model simplification (Sun and Hahn, 2006; Kou et al., 2005a; Kou et al., 2005b; Maria, 2004; Rexstad and Innis, 1985; Innis and Rexstad, 1983), finding the right balance between simplicity and complexity usually involves more engineering judgment than science (Brooks and Tobias, 1996).

There has been considerable research in the statistics literature on the statistical consequences of using simplified or misspecified models (O'Brien et al., 2006; Waldorp et al., 2006; Waldorp et al., 2005; Rao and Wu, 2001; Bera, 2000; Golden, 1995; Miller, 1990; White, 1982, 1981; Hocking, 1976; Rosenberg and Levy, 1972; Rao, 1971; Toro-Vizcarrondo and Wallace, 1968; Wallace, 1964; Kabe, 1963; Freund et al., 1961; Goldberg, 1961; Goldberg and Jochems, 1961), particularly for models that are linear in the parameters. Three general situations have been considered: (1) input variables that belong in the true model are

* *Author to whom correspondence may be addressed.*
E-mail address: harrist@post.queensu.ca

mistakenly omitted (this is sometimes known as undermodelling); (2) variables that have no real influence on the output variables are mistakenly included (this is sometimes known as overfitting); and (3) errors are made in the distributional assumptions of the stochastic or random component of the model (Seber and Wild, 2003; Hocking, 1976; Rao, 1971). Issues related to the third item are more difficult to generalize and are not discussed further in this paper.

Important quantitative and qualitative results have been derived to compare the parameter estimates and model predictions from misspecified models with those from correctly structured models (Miller, 1990; Abdullaev and Geidarov, 1985; Hocking, 1976; Rosenberg and Levy, 1972; Rao, 1971; Wallace, 1964; Kabe, 1963; Freund et al., 1961; Goldberg, 1961; Goldberg and Jochems, 1961). While it might seem that the use of misspecified models will always lead to inferior model predictions, and parameter estimates that are biased, this is not always true (Waldorp et al., 2006; Hocking, 1976; Rao, 1971).

Like many chemical engineers, we are particularly interested in developing phenomenological models based on material and energy balances and constitutive equations. The usual objective of the statistical approach to model building (for either empirical or mechanistic models) is to develop models that are of sufficient complexity so that the model passes statistical adequacy tests (Montgomery and Runger, 2003). When this objective has been achieved, probability statements can be assigned to the precision of the estimated parameters and model predictions (Montgomery et al., 2001; Draper and Smith, 1998). There has been far less research on providing similar information for simplified or misspecified models (Bera, 2000; Golden, 1995; White, 1982, 1981). Recent research in the use of more computationally intensive methods for statistical analysis of misspecified models (e.g. nonparametric bootstrapping) has revived interest in this topic (Waldorp et al., 2006; Fushiki, 2005; Aerts and Claeskens, 2001; Velilla, 2001; Davison and Hinkley, 1997).

In this paper we: (1) summarize the quantitative and qualitative results in the literature concerned with using simplified or misspecified models; (2) provide new insights into the conditions under which simplified or misspecified models give superior predictions compared with the correctly structured EM; and (3) evaluate methods that can be used for statistical inference for models that are simplified or misspecified. A new practical strategy, based on confidence intervals and hypothesis tests, is developed to help modellers decide whether a SM will give better predictions than the truly structured EM. However, there are considerable challenges in making such a decision. The resulting confidence intervals are quite large and the statistical tests, while exact in their construction, have poor discrimination properties for the alternative hypothesis that the SM is better or that the EM is better.

We focus on models that are linear in the parameters. While this choice might at first seem restrictive because chemical engineers tend to use non-linear models, we note that, the statistical analysis of non-linear models usually involves a linearization of the model around the nominal parameter values (Seber and Wild, 2003). Thus, the results of this paper can assist model developers in the analysis of phenomenologically based models that are non-linear in the parameters.

The paper is organized as follows. A general description of model misspecification is given in the second section. In the third section, misspecification in models that are linear in the parameters is analyzed theoretically. There are extensive results in the literature for this topic, but an important unresolved issue is the lack of practical tests for determining whether a SM will give better predictions than the truly structured EM. The new practical strategy that uses the model structure and the data available for parameter estimation is provided in the fourth section. This is followed in the fifth section by an analysis of constructive numerical methods that can be used to make statistical statements regarding the uncertainty in parameters and model predictions when the SM is used. In the sixth section, analytical results and Monte Carlo simulations from a simple example are used to provide insights into the most important results from the previous sections. The paper concludes with a brief discussion on the applicability of these methods to models that are non-linear in the parameters.

MISSPECIFIED MODELS—THE GENERAL CASE

We assume that a process can be truly described by

$$y_i = f(x_i, \beta) + \varepsilon_i, \quad i = 1, 2, \ldots, n \tag{1}$$

where x_i is a k dimensional vector of explanatory variables for the i^{th} observation, β is an m dimensional vector of model parameters, and random variable ε_i is uncorrelated and is identically distributed with mean 0 and variance σ^2.

Typically, a modeller supposes that the observed response variable can be represented as

$$y_i = g(z_i, \theta) + e_i, \quad i = 1, 2, \ldots, n \tag{2}$$

where z_i is a vector of explanatory variables, θ is vector of parameters and $g(z_i, \theta)$ is the function that the modeller believes (or hopes) relates (z_i, θ) to the response. The term e_i encompasses the stochastic component and any deterministic part that is not captured by the model. The functional form of $g(z_i, \theta)$ may be specified from a fundamental understanding of the process or a desire to find a purely empirical representation. In either case, the parameters are often estimated as the solution to the least-squares problem

$$\hat{\theta} = \arg\min_{\theta} S(\theta) = \arg\min_{\theta} \sum_{i=1}^{n} \left(y_i - g(z_i, \theta)\right)^2 \tag{3}$$

When the parameters enter the model non-linearly, a non-linear optimization algorithm is required to determine $\hat{\theta}$. When the parameters enter linearly, ordinary least squares (OLS) is commonly used (Montgomery and Runger, 2003).

A model-building strategy is typically iterative. A model form is postulated, and the parameters are estimated. Typically the residuals $\hat{e}_i = y_i - g(z_i, \hat{\theta})$ are examined for systematic patterns that suggest omitted variables. The model may also by "pruned" by eliminating parameters that are not statistically different from zero. While no assumptions on the probability structure of the stochastic components are required to determine the least-squares estimates, the validity of the statistical analysis requires several assumptions (Montgomery et al., 2001). Typically it is assumed that the model structure is correct ($g(z_i, \theta) = f(x_i, \beta)$), that the explanatory variables are deterministic, and that ε_i, in addition to being uncorrelated and identically distributed, follows a normal distribution. When these assumptions are satisfied, there is a rich body of knowledge related to model building and statistical assessment of the model (Montgomery et al., 2001; Draper and Smith, 1998).

However, there are many instances when one deliberately chooses a structural form that does not match the true process.

In these instances the model may be acceptable (in an intended end-use sense), but may fail a statistical test for adequacy (Chang et al., 2005; Golbert and Lewin, 2004; Bagajewicz and Cabrera, 2003; Yoshida et al., 2003). This is particularly true when fundamental models are used. In the case of empirical models, it may happen that some of the predictor variables are deleted from the model because they are inaccessible, in which case the model is incorrect. We will refer to these structurally imperfect models as simplified or misspecified models. Several interesting questions arise:

1. Can misspecified models give better parameter estimates and model predictions than the correctly structured extended model?
2. What statistical methods can be used to analyze these models and to make statements about the quality of their predictions?

There is a rich literature that addresses question 1 (Miller, 1990; Hocking, 1976; Rao, 1971). Not surprisingly, almost all of this work relates to models that are linear in parameters. In these instances, closed-form solutions can be obtained after a definition for "better" is specified. Some results are also available to address question 2 (Waldorp et al., 2006; Waldorp et al., 2005; Bera, 2000; Golden, 1995).

MISSPECIFIED MODELS—THE LINEAR CASE

Assume that the true process is described by

$$Y = X_1\beta_1 + X_2\beta_2 + \varepsilon \tag{4}$$

where $Y \in \mathbb{R}^n$, $X_1 \in \mathbb{R}^{n\times p}$, $X_2 \in \mathbb{R}^{n\times q}$, $\beta_1 \in \mathbb{R}^p$, $\beta_2 \in \mathbb{R}^q$, and $\varepsilon \in \mathbb{R}^n$. We will refer to this correctly structured model as the extended model (EM).

The following assumptions are usually made (Beck and Arnold, 1977):

1. X_1 and X_2 have full column rank, and are deterministic;
2. The stochastic component ε is a mean zero, uncorrelated random sequence with constant variance σ^2.

Define $\beta = (\beta_1\ \beta_2)^T$, $X = (X_1\ X_2)$. Then the OLS estimates are given by

$$\hat{\beta}_E = \left(X^T X\right)^{-1} X^T Y \tag{5}$$

where the subscript "*E*" indicates the use of the truly structured extended model.

When the OLS assumptions are satisfied,

$$\begin{aligned} E\left(\hat{\beta}_E\right) &= \beta \\ Cov\left(\hat{\beta}_E\right) &= \sigma^2\left(X^T X\right)^{-1} \end{aligned} \tag{6}$$

where $E(\bullet)$ and $Cov(\bullet)$ denote the mathematical expectation and covariance of the $(\bullet)$.

It is convenient to write (6) in the form of $\hat{\beta}_E \sim (\beta, \sigma^2 (X^T X)^{-1})$. The notation $(\bullet)\sim(\mu, \Sigma)$ denotes that $\mu = E(\bullet)$ and $\Sigma = Cov(\bullet)$. The variance σ^2 can be estimated via

$$s_E^2 = \frac{S\left(\hat{\beta}_E\right)}{n-m} = \frac{\left(Y - X\hat{\beta}_E\right)^T\left(Y - X\hat{\beta}_E\right)}{n-m} \tag{7}$$

where $m = p + q$ is the total number of parameters in the EM.

Since the model is correctly structured, OLS estimation provides the Best Linear Unbiased Estimates (BLUE) for the model parameters, in the sense that the OLS parameter estimates have the smallest variance among all unbiased estimators (Beck and Arnold, 1977). Furthermore, any linear combination of the parameter estimates of the form $a^T\ \hat{\beta}_E$ also has the smallest variance among all unbiased estimators of $a^T\ \beta$, where a is a column vector of length m.

There are many ways in which a model that is used to represent a process can be misspecified (Hocking, 1976; Rao, 1971), including:

1. failure to include some or all of the explanatory variables (undermodelling);
2. inclusion of "extraneous" variables in the model (overfitting).

Even in the linear case, the true process may be complex, encompassing features such as a heteroskedastic structure for the stochastic components, or an error-in-variables structure for the explanatory variables (Beck and Arnold, 1977). The purpose of regression diagnostics is to evaluate model adequacy and reveal inadequacies.

In this paper, we focus on the consequences of undermodelling. In undermodelling, the analyst believes, or decides to use, a model of the form

$$Y = X_1\beta_1 + e \tag{8}$$

where $e = X_2\beta_2 + \varepsilon$ is the stochastic component combined with any model mismatch. We will refer to (8) as the SM.

For the SM, only the parameters associated with the explanatory variables in X_1 are estimated by minimizing the objective function

$$S(\beta_1) = \left(Y - X_1\beta_1\right)^T\left(Y - X_1\beta_1\right) \tag{9}$$

with respect to β_1, resulting in the OLS estimates $\hat{\beta}_{1S}$ as

$$\hat{\beta}_{1S} = \left(X_1^T X_1\right)^{-1} X_1^T Y \tag{10}$$

where the subscript "*S*" indicates the use of a SM. It is readily verified that (Draper and Smith, 1998)

$$\begin{aligned} E\left(\hat{\beta}_{1S}\right) &= \beta_1 + A_1\beta_2 \\ Cov\left(\hat{\beta}_{1S}\right) &= \sigma^2\left(X_1^T X_1\right)^{-1} \end{aligned} \tag{11}$$

where $A_1 = (X_1^T X_1)^{-1} X_1^T X_2$ is the projection of X_1 on X_2. These parameter estimates are generally biased, in that $E(\hat{\beta}_{1S}) \neq \beta_1$, unless $A_1\beta_2 = 0$, which only occurs when $\beta_2 = 0$, or when X_1 and X_2 are orthogonal. The modeller may use the residuals from the SM to estimate the noise variance σ^2,

$$s_S^2 = \frac{S\left(\hat{\beta}_{1S}\right)}{n-p} = \frac{\left(Y - X_1\hat{\beta}_{1S}\right)^T\left(Y - X_1\hat{\beta}_{1S}\right)}{n-p} \tag{12}$$

ignoring the influence of any model misspecification. The value of s_S^2 is often used to construct confidence intervals for parameter estimates and model predictions and to conduct model adequacy tests. However, since s_S^2 is a biased estimator of σ^2 (Hocking, 1976), statistical tests that use s_S^2 can be misleading.

When misspecified models are analyzed (and the modeller believes that the model is misspecified), it is common for the quality of parameter estimates (and model predictions) to be assessed using mean-squared-error (MSE) or mean-squared-error-matrix (MSEM) (Toutenburg and Trenkler, 1990; Price, 1982; Gunst and Mason, 1977; Lowerre, 1974), which account for both bias and variance. The MSEM and MSE for a parameter estimate $\hat{\beta}$ are defined as

$$\begin{aligned} MSEM\left(\hat{\beta}\right) &= E\left(\hat{\beta} - \beta\right)E\left(\hat{\beta} - \beta\right)^T \\ &= Cov\left(\hat{\beta}\right) + \Delta\Delta^T \end{aligned} \tag{13}$$

$$MSE\left(\hat{\beta}\right) = E\left(\hat{\beta}-\beta\right)^T E\left(\hat{\beta}-\beta\right) = Tr\left(MSEM\left(\hat{\beta}\right)\right) \quad (14)$$

where $\Delta = E(\hat{\beta}) - \beta$ is the bias, and $Tr(\bullet)$ denotes the trace of the quantity $(\bullet)$.

Miller (1990), Hocking (1976), Rao (1971), Toro-Vizcarrondo and Wallace (1968), Wallace (1964), Kabe (1963), Freund et al. (1961), Goldberg (1961) and Goldberg and Jochems (1961) provide important results concerned with the MSE of parameter estimates and model predictions when a SM structure is selected.

Quantitative Statements Regarding Misspecification

The following results are known for misspecified models in the linear case (Hocking, 1976; Rao, 1971):

1. The omission of a variable from the EM (the truth), introduces bias and decreases the variance in all of the parameter estimates (and model predictions) obtained using the SM;
2. The MSEs of all parameter estimates (and model predictions) are decreased when a single variable in the EM is deleted whose true parameter value is smaller in magnitude than the theoretical standard deviation of its least-squares estimate;
3. The estimate of the noise variance (Equation (12)) obtained from the SM residuals is upwardly biased;
4. The inclusion of an irrelevant variable in the model increases the variance and MSEs of all of the parameter estimates (and model predictions).

Quantitative Statements Regarding Misspecification

To enable a comparison of the properties of the estimates from an EM and a SM, it is helpful to partition the expected values of the parameter estimates and their covariance matrix from the EM as follows:

$$\begin{pmatrix}\hat{\beta}_{1E}\\ \hat{\beta}_{2E}\end{pmatrix} \sim \left(\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix}, \sigma^2\begin{pmatrix}\Gamma & \Psi\\ \Psi^T & \Omega\end{pmatrix}\right) \quad (15)$$

where

$$\begin{pmatrix}\Gamma & \Psi\\ \Psi^T & \Omega\end{pmatrix} = \begin{pmatrix}X_1^TX_1 & X_1^TX_2\\ X_2^TX_1 & X_2^TX_2\end{pmatrix}^{-1} \quad (16)$$

The elements of the composite covariance matrix can be obtained from sub-matrices using standard results from linear algebra (Beck and Arnold, 1977)

$$\begin{aligned}\Gamma &= \left(X_1^T(I_n-P_2)X_1\right)^{-1} = \left(X_1^TX_1\right)^{-1} + A_1\left(X_2^T(I_n-P_1)X_2\right)^{-1}A_1^T\\ \Omega &= \left(X_2^T(I_n-P_1)X_2\right)^{-1} = \left(X_2^TX_2\right)^{-1} + A_2\left(X_1^T(I_n-P_2)X_1\right)^{-1}A_2^T\\ \Psi &= -A_1\Omega\end{aligned} \quad (17)$$

where $P_1 = X_1\left(X_1^TX_1\right)^{-1}X_1^T$, $P_2 = X_2\left(X_2^TX_2\right)^{-1}X_2^T$, $A_2 = \left(X_2^TX_2\right)^{-1}X_2^TX_1$. A_2 is the projection of X_2 on X_1.

Based on the above theoretical results, comparisons between the SM and EM can be made, both for parameter estimates and for model predictions (or other linear combinations of the parameters).

Comparison of Parameter Estimates

Table 1 summarizes the expected values, covariance matrices and MSEM of parameter estimates obtained from using the SM and the EM. The expected noise variance estimates that would be obtained are given in the final row of the table.

Mean-Based Comparison of Parameter Estimates

The parameter estimates from the correctly specified EM are unbiased. The parameter estimates from the SM are biased except when $A_1 = 0$, which requires that X_2 be orthogonal to X_1, a situation not likely to be encountered in practice.

Variance-Based Comparison of Parameter Estimates

From Table 1, the difference between covariance matrices for $\hat{\beta}_{1E}$ and $\hat{\beta}_{1S}$ is

$$Cov\left(\hat{\beta}_{1E}\right) - Cov\left(\hat{\beta}_{1S}\right) = \sigma^2A_1\left(X_2^T(I_n-P_1)X_2\right)^{-1}A_1^T = \sigma^2A_1\Omega A_1^T \quad (18)$$

Since $Cov\left(\hat{\beta}_{2E}\right) = \sigma^2\Omega$, therefore Ω is positive definite, and the above difference will be positive semi-definite (Zhang, 1999). Let β_{1i} be the i^{th} element in β_1 ($i = 1,2,\ldots,p$). Using properties of positive semi-definite matrices (Zhang, 1999), the following inequalities are satisfied:

Individual Parameters

$$Var\left(\hat{\beta}_{1iS}\right) \le Var\left(\hat{\beta}_{1iE}\right) \quad (19a)$$

Total Variance

$$Tr\left(Cov\left(\hat{\beta}_{1S}\right)\right) \le Tr\left(Cov\left(\hat{\beta}_{1E}\right)\right) \quad (19b)$$

Generalized Variance (determinant)

$$\left|Cov\left(\hat{\beta}_{1S}\right)\right| \le \left|Cov\left(\hat{\beta}_{1E}\right)\right| \quad (19c)$$

Note that, the parameter estimates from the SM are biased, so the Gauss-Markov Theorem does not apply (Beck and Arnold, 1977).

In the case of overfitting, the SM would be correctly specified and the EM would contain redundant parameters, since the true value of β_2 is a zero vector. Both the SM and the EM would lead to unbiased parameter estimates because both the SM and the EM are correctly structured (Rao, 1971). However, the inclusion of the extraneous parameters results in less precision in the parameter estimates and in the predictions made using these parameters. In this scenario, the true covariance matrix of the parameters would be given by the entry in Table 1 under the column "Simplified Model (SM)," and the covariance matrix of the parameters for overfitting would be given by the entry in Table 1 under the column "Extended Model (EM)."

MSEM-Based Comparison of Parameter Estimates

From Table 1, the MSEM difference for $\hat{\beta}_{1E}$ and $\hat{\beta}_{1S}$ is

$$MSEM\left(\hat{\beta}_{1E}\right) - MSEM\left(\hat{\beta}_{1S}\right) = A_1\left(\sigma^2\left(X_2^T(I_n-P_1)X_2\right)^{-1} - \beta_2\beta_2^T\right)A_1^T \quad (20)$$

This difference is positive semi-definite if

$$\sigma^2\left(X_2^T(I_n-P_1)X_2\right)^{-1} - \beta_2\beta_2^T \ge 0 \quad (21)$$

A necessary and sufficient condition of Inequality (21) is that (Wang and Chow, 1994)

$$\beta_2^TX_2^T(I_n-P_1)X_2\beta_2 \le \sigma^2 \quad (22)$$

This inequality holds when the SM gives better (in sense of smaller MSEM) parameter estimates than the EM.

Inequality (22) has several appealing interpretations. First, we note that the last entry in Table 1 can be re-arranged as

$$\frac{\beta_2^T X_2^T (I_n - P_1) X_2 \beta_2}{\sigma^2 (n-p)} = \frac{E\left(s_S^2\right) - \sigma^2}{\sigma^2} \tag{23}$$

The right-hand side (RHS) of (23) is the fractional increase in the expected noise variance prediction that arises when the SM is used. If (22) holds, then

$$\frac{E\left(s_S^2\right) - \sigma^2}{\sigma^2} \le \frac{1}{n-p} \tag{24}$$

Inequality (24) provides an upper bound on the bias of the noise variance estimate obtained from the SM when Inequality (22) is satisfied (SM parameter estimates are better than the EM estimates).

Rearranging Inequality (22), we obtain the expression for a critical ratio R_C

$$R_C = \frac{\beta_2^T X_2^T (I_n - P_1) X_2 \beta_2}{q\sigma^2} \tag{25}$$

where q is the number of parameters contained in β_2. Inequality (22) is then equivalent to

$$R_C = \frac{\beta_2^T X_2^T (I_n - P_1) X_2 \beta_2}{q\sigma^2} \le \frac{1}{q} \tag{26}$$

Examination of (26) reveals that R_c becomes smaller (implying the SM tends to be better than the EM) in the following situations:

1. when there are high noise levels, i.e., σ^2 is large;
2. when the true absolute values of parameters in β_2 are small;
3. when there are high correlations among input variable settings so that the trace of $X_2^T(I_n - P_1)X_2$ is small;
4. when there are a limited number of experiments or a limited range of input conditions so that the trace of $X_2^T(I_n - P_1)X_2$ is small;

Proof of statements 1 and 3 above can be found in Abdullaev and Geidarov (1985).

Comparison of Model Predictions

Model developers and users often care more about model predictions or other linear combinations of parameters than the parameter estimates themselves. Imagine that the model parameters have been estimated using a matrix of input variable settings, $X \in \mathbb{R}^{n\times m}$, and then model predictions are made using other input settings $Z \in \mathbb{R}^{w\times m}$ where w is the total number of predictions to be made. Z can be partitioned in the same way as X into $Z_1 \in \mathbb{R}^{w\times p}$ and $Z_2 \in \mathbb{R}^{w\times q}$ corresponding to the partitioned parameters in (15). Two types of model predictions can be made

$$1.\text{SM predictions: } \hat{Y}_S = Z_1\hat{\beta}_{1S} \tag{27a}$$

$$2.\text{EM predictions: } \hat{Y}_E = Z\hat{\beta}_E = Z_1\hat{\beta}_{1E} + Z_2\hat{\beta}_{2E} \tag{27b}$$

Expressions for expected values, covariance matrices and the MSEM for these predictions are shown for two cases: (1) $Z = X$ (Table 2); and (2) $Z \neq X$ (Table 3). There is no extrapolation or interpolation in the first case, because predictions are made at the same conditions under which the data were collected. The second case corresponds to a new set of conditions for which predictions are desired.

Table 1. Comparison of parameter estimates and variance estimates from EM and SM

	Extended Model (EM)	Simplified Model (SM)
$E(\hat{\beta}_1)$	β_1	$\beta_1 + A_1\beta_2$
$E(\hat{\beta}_2)$	β_2	
$Cov(\hat{\beta}_1)$	$\sigma^2(X_1^TX_1)^{-1} + \sigma^2A_1(X_2^T(I_n-P_1)X_2)^{-1}A_1^T$	$\sigma^2(X_1^TX_1)^{-1}$
$Cov(\hat{\beta}_2)$	$\sigma^2(X_2^T(I_n-P_1)X_2)^{-1}$	
$MSEM(\hat{\beta}_1)$	$\sigma^2(X_1^TX_1)^{-1} + \sigma^2A_1(X_2^T(I_n-P_1)X_2)^{-1}A_1^T$	$\sigma^2(X_1^TX_1)^{-1}+A_1\beta_2\beta_2^TA_1^T$
$E(s^2)$	σ^2	$\sigma^2 + \frac{\beta_2^T X_2^T (I_n - P_1) X_2 \beta_2}{n-p}$

Table 2. Model predictions when $Z = X$

	SM Prediction $\hat{Y}_S$	EM Prediction $\hat{Y}_E$
$E(\hat{Y})$	$X_1\beta_1 + X_1A_1\beta_2$	$\beta_1 + A_1\beta_2$
$Cov(\hat{Y})$	σ^2P_1	$\sigma^2P_1 + \sigma^2(I_n-P_1)X_2\Omega_2X_2^T(I_n-P_1)$
$MSEM(\hat{Y})$	$\sigma^2P_1 + (I_n-P_1)X_2\beta_2\beta_2^TX_2^T(I_n-P_1)$	$\sigma^2P_1 + \sigma^2(I_n-P_1)X_2\Omega_2X_2^T(I_n-P_1)$

Table 3. Model predictions when $Z \neq X$

	SM Prediction $\hat{Y}_S$	EM Prediction $\hat{Y}_E$
$E(\hat{Y})$	$Z_1\beta_1 + Z_1A_1\beta_2$	$Z_1\beta_1 + Z_2\beta_2$
$Cov(\hat{Y})$	$\sigma^2Z_1(X_1^TX_1)^{-1}Z_1^T$	$\sigma^2Z_1(X_1^TX_1)^{-1}Z_1^T + \sigma^2(Z_1A_1-Z_2)\Omega(Z_1A_1-Z_2)^T$
$MSEM(\hat{Y})$	$\sigma^2Z_1(X_1^TX_1)^{-1}Z_1^T + (Z_1A_1-Z_2)\beta_2\beta_2^T(Z_1A_1-Z_2)^T$	$\sigma^2Z_1(X_1^TX_1)^{-1}Z_1^T + \sigma^2(Z_1A_1-Z_2)\Omega(Z_1A_1-Z_2)^T$

Mean-Based Comparison of Model Predictions
As seen in the first row of Tables 2 and 3, only the correctly structured EM provides unbiased model predictions.

Variance-Based Comparison of Model Predictions
From Table 3, in general, the difference between covariance matrices for $\hat{Y}_S$ and $\hat{Y}_E$ is

$$Cov\left(\hat{Y}_E\right) - Cov\left(\hat{Y}_S\right) = \sigma^2 \left(Z_1 A_1 - Z_2\right) \Omega \left(Z_1 A_1 - Z_2\right)^T \tag{28}$$

Since Ω is positive definite, the above difference is positive semi-definite, which means predictions from the SM cannot be more variable than those from the EM.

MSEM-Based Comparison of Model Predictions
The elements of the MSEM contain the covariances for the model predictions, plus the squared bias. From Table 3, the difference between the MSEM for $\hat{Y}_S$ and $\hat{Y}_E$ is

$$MSEM\left(\hat{Y}_E\right) - MSEM\left(\hat{Y}_S\right) = \left(Z_1 A_1 - Z_2\right)\left(\sigma^2 \left(X_2^T \left(I_n - P_1\right) X_2\right)^{-1} - \beta_2 \beta_2^T\right)\left(Z_1 A_1 - Z_2\right)^T \tag{29}$$

Following the same argument as for (21), (22) and (26), if $R_C \leq 1/q$, then $MSEM(\hat{Y}_S) \leq MSEM(\hat{Y}_E)$. This means that when few data points are available or the data are noisy or when there are high correlations between X_1 and X_2, the SM can be expected to provide better predictions than the properly specified EM.

A special case is that, when $Z = X$, the MSEM difference becomes

$$MSEM\left(\hat{Y}_E\right) - MSEM\left(\hat{Y}_S\right) = \left(I_n - P_1\right) X_2 \left(\sigma^2 \left(X_2^T \left(I_n - P_1\right) X_2\right)^{-1} - \beta_2 \beta_2^T\right) X_2^T \left(I_n - P_1\right) \tag{30}$$

and the MSE difference is

$$\begin{aligned} MSE\left(\hat{Y}_E\right) - MSE\left(\hat{Y}_S\right) &= Tr\left(\left(I_n - P_1\right) X_2 \left(\sigma^2 \left(X_2^T \left(I_n - P_1\right) X_2\right)^{-1} - \beta_2 \beta_2^T\right) X_2^T \left(I_n - P_1\right)\right) \\ &= \sigma^2 q - \beta_2^T X_2^T \left(I_n - P_1\right) X_2 \beta_2 \end{aligned} \tag{31}$$

In this special case, the SM is better if and only if

$$R_C = \frac{\beta_2^T X_2^T \left(I_n - P_1\right) X_2 \beta_2}{\sigma^2 q} \leq 1 \tag{32}$$

Note that Inequality (32) is a necessary and sufficient condition for $MSE(\hat{Y}_E) \geq MSE(\hat{Y}_S)$ and is less restrictive than (26). However, it only holds when model predictions are made using exactly the same input variable settings as those used in parameter estimation. If parameter estimation or extrapolation is the main purpose, expression (26) is more appropriate (Hocking, 1976). When there is only one variable in X_2 ($q = 1$), (26) and (32) become the same.

In summary, the literature on misspecified linear models provides information about the conditions under which a modeller can expect to get improved parameter estimates and model predictions when a SM is used. Unfortunately, the conditions in Inequalities (26) and (32) are based on the true values of the parameters in β_2 and on the true noise variance σ^2. In practical applications, the modeller does not know these true values, but can obtain estimated values, $\hat{\beta}_{2E}$ and s_E^2, from the data, assuming the correctly structured EM is available.

STRATEGY FOR ASSESSING UNCERTAINTY ABOUT WHICH MODEL IS BETTER

If estimated parameter values and noise variances from the EM are available, then an estimate of R_C is

$$\hat{R}_C = \frac{\hat{\beta}_{2E}^T X_2^T \left(I_n - P_1\right) X_2 \hat{\beta}_{2E}}{q s_E^2} \tag{33}$$

If we also assume that the stochastic component ε in model (4) is Normally distributed, then based on partial F tests (Montgomery et al., 2001), $\hat{R}_C$ follows a non-central F distribution $F_{q,n-m}(\delta)$ with q and $n - m$ degrees of freedom. The non-centrality parameter δ is

$$\delta = \frac{\beta_2^T X_2^T \left(I_n - P_1\right) X_2 \beta_2}{\sigma^2} \tag{34}$$

which equals qR_c. When $\beta_2 = 0$, $\hat{R}_C$ follows the more widely known *central F* distribution $F_{q,\,n-m}$, which is often used to test the null hypothesis that $\beta_2 = 0$ (Montgomery et al., 2001). Note that $\hat{\beta}_{2E}$ and s_E^2 are obtained by fitting parameters in the extended model, so the analyst must have access to an extended model that she believes to be well structured to compute $\hat{R}_C$ in Equation (33).

The construction of confidence intervals and hypothesis tests for many standard statistics that follow standard distributions (such as the Normal distribution, t distribution, central χ^2 distribution and central F distribution) is well established in introductory statistics textbooks. However, the construction of confidence intervals and hypothesis tests for R_C is complicated by the fact that $\hat{R}_C$ follows a non-central $F_{q,n-m}(\delta)$ distribution with unknown non-centrality parameter $\delta = qR_C$. As a result, an iterative algorithm is required to find appropriate confidence intervals. The following steps (Steiger, 2004) can be used to calculate the range $[\delta_L, \delta_U]$, the exact two-sided $100(1-\alpha)\%$ confidence interval for δ.

1. Calculate the cumulative probability p_C corresponding to $\hat{R}_C$ using the *central F* distribution with q and $n - m$ degrees of freedom. If p_C is less than $\alpha/2$, then δ_L and δ_U are both zero. The reason for this conclusion is that $\delta \geq 0$ by definition, and, if we did a one-sided hypothesis test at the $100(\alpha/2)\%$ significance level, we could reject the null hypothesis that δ is zero or larger than zero. If p_C is less than $(1 - \alpha/2)$, $\delta_L = 0$ and δ_U is calculated using Step 3. Otherwise, calculate δ_L and δ_U using Steps 2 and 3.
2. To calculate the lower limit, δ_L, iterate on the non-centrality parameter, so that the $(1 - \alpha/2)$ cumulative probability point (the critical value) of a non-central F distribution with q and $n - m$ degrees of freedom equals $\hat{R}_C$. This value is unique.
3. To calculate the upper limit, δ_U, iterate on the non-centrality parameter, so that the $\alpha/2$ cumulative probability point (the critical value) of a non-central F distribution with q and $n - m$ degrees of freedom equals $\hat{R}_C$. This value is unique.

Since $\delta = qR_C$, the two-sided $100(1-\alpha)\%$ confidence interval for R_c is $[\delta_L/q, \delta_U/q]$. This confidence interval for R_C contains all values of the null hypothesis that would not be rejected at the $100(1-\alpha)\%$ confidence level when testing the alternative hypothesis that $R_C \neq k$. Two values of k are of interest; $k = 1/q$ is used to test whether we are confident that Inequality (26) is satisfied and $k = 1$ is used to test Inequality (32). If $k \leq \delta_L/q$, we can be $100(1-\alpha/2)\%$ certain that $R_C > k$. and that the EM is better than the SM. If $k \geq \delta_U/q$, we can be $100(1-\alpha/2)\%$ certain that $R_C < k$ and that the SM is better than the EM. In the special case where only one parameter in the EM was not included in the SM ($q = 1$), Inequalities (26) and (32) become the same, and $k = 1$. The confidence intervals are readily computed using the cumulative non-central F distribution function in MATLAB® or other statistical software packages.

CONFIDENCE INTERVALS FOR PARAMETER ESTIMATES AND MODEL PREDICTIONS FROM THE SM

In situations in which the modeller has decided to use the misspecified SM, it is desirable to obtain appropriate confidence intervals for the parameters and model predictions. Such confidence intervals rely on good estimates of variance-covariance matrices. In the literature, three methods have been proposed for estimating variance-covariance matrices for parameter estimates: (1) the conventional method, which assumes the SM is truly structured (Montgomery et al., 2001); (2) the sandwich estimator (Waldorp et al., 2006; Waldorp et al., 2005; Seber and Wild, 2003; White, 1981); and (3) nonparametric bootstrapping (Waldorp et al., 2006; Good, 2005; Martinez and Martinez, 2002; Montgomery et al., 2001; Efron and Tibshirani, 1993). The last two methods have been recommended for use under model misspecification (Waldorp et al., 2006).

The conventional variance-covariance matrix of $\hat{\beta}_{1S}$ can be estimated by

$$\hat{\Sigma}_{CONV} = s^2 \left(X_1^T X_1 \right)^{-1} \tag{35}$$

where s^2 is an estimate of noise variance. s^2 could be determined from: (1) replicate runs (pooled variances) if replicates are available; (2) the EM (s_E^2), if the correctly structured extended model is available; (3) the SM (s_S^2); and (4) nonparametric bootstrapping (s_B^2). The conventional method for estimating the variance-covariance matrix for the parameter estimates requires the assumption that the SM is correctly structured ($\beta_2 = 0$). The estimated variance-covariance matrix, and the conventional confidence bounds that are determined from its diagonal elements, rely on the goodness of the noise variance estimate and the goodness of the assumption that $\beta_2 = 0$.

White (1981) proposed that the sandwich estimator, which is robust to model misspecification, should be used to estimate the variance-covariance matrix for parameter estimates when a misspecified model is used. The sandwich estimator is derived directly from the data without the requirement for a truly specified EM or a separate noise variance estimate from replicate experiments. The variance-covariance matrix of parameter estimates is defined as

$$\Sigma_{SANW} = \left(X_1^T X_1 \right)^{-1} \left(\sum_{i=1}^{n} \tilde{x}_i^T \tilde{x}_i \hat{e}_i^2 \right) \left(X_1^T X_1 \right)^{-1} \tag{36}$$

where $\tilde{x}_i$ is the i^{th} row of X_1 and $\hat{e}_i$ is the i^{th} residual from the SM ($i = 1, 2, \ldots, n$).

The sandwich estimator is a consistent estimator for the true variance-covariance matrix of parameter estimates even when the model is misspecified. The sandwich estimator has been investigated for models that are non-linear in the parameters (Donaldson and Schnabel, 1987), and has been used to obtain good estimate of the Cramér-Rao bound for the use of the Wald test in situations when the model is misspecified (Waldorp et al., 2005). Additionally, it is shown that sandwich estimator is robust against an incorrect assumption on the noise covariance (Waldorp et al., 2006; Waldorp et al., 2005).

Nonparametric bootstrapping is a computationally intensive procedure commonly used in situations when no analytical methods are available for determining confidence intervals and when the sample is representative of the population (Martinez and Martinez, 2002; Montgomery et al., 2001). The bootstrapping algorithm proceeds as follows:

1. Resample the original data (X_1, Y) B times with replacement, and for each resampled data set, estimate β_1 and the noise variance σ^2.
2. The final noise variance estimate is obtained as the average of the B individual noise variance estimates (Good, 2005),

$$s_B^2 = \frac{1}{B} \sum_{i=1}^{B} s_*^{2,i} \tag{37}$$

where $s_*^{2,i}$ is the i^{th} noise variance estimate ($i = 1, 2, \ldots, B$), and is calculated as

$$s_*^{2,i} = \frac{\left(Y^i - X_1^i \hat{\beta}_{1*}^i \right)^T \left(Y^i - X_1^i \hat{\beta}_{1*}^i \right)}{n - p} \tag{38}$$

(X_1^i, Y^i) is the i^{th} resampled pair, and $\hat{\beta}_{1*}^i$ is the i^{th} estimate of β_1 based on (X_1^i, Y^i).
3. The variance-covariance matrix for the parameter estimates is obtained directly from the parameter estimates as

$$\Sigma_{BOOT} = \frac{1}{B-1} \sum_{i=1}^{B} \left(\hat{\beta}_{1*}^i - \bar{\beta}_{1*} \right) \left(\hat{\beta}_{1*}^i - \bar{\beta}_{1*} \right)^T \tag{39}$$

where $\bar{\beta}_{1*}^i$ is the average of B estimates of β_1 (Waldorp et al., 2006).

Bootstrapping can be used to estimate the bias and to directly construct the confidence intervals for parameter estimates and model predictions. A detailed description of different algorithms and the limitations of each algorithm can be found in Efron and Tibshirani (1993). MATLAB® codes are available from Martinez and Martinez (2002). Chernick (1999) also described some examples from the literature when the bootstrapping method should not be used. Waldorp et al. (2006) showed that the nonparametric bootstrapping give better results than the conventional methods, especially when the noise is correlated.

ILLUSTRATIVE EXAMPLE

Theoretical Results

To illustrate the theoretical analysis provided in the third section, we consider a very simple example with three parameters in the SM ($p = 3$) and one additional parameter ($q = 1$) in the EM. Let the design matrix for the SM be $X_1 = (X_{11}\, X_{12}\, X_{13})$, an orthogonal matrix (obtained from a designed experiment) containing entries of $\pm r$. X_{1i} is the i^{th} column in X_1 ($i = 1,2,3$). We are interested in knowing whether or not the additional parameter β_2 in the EM should be estimated. We assume that the vector of experimental settings for the input corresponding to β_2 is correlated with the first column of X_1, which makes it difficult to obtain an independent estimate of β_2. In our example, we can adjust the amount of correlation between X_{11} and X_2 using an adjustable design factor λ, where $0 \le \lambda \le 1$. We let $X_2 = \lambda X_{11} + \sqrt{1-\lambda^2} W$ where W is a $(n \times 1)$ vector with entries of $\pm r$, which is orthogonal to all columns in X_1. Note that X_2 is correlated with the first column of X_1, but not with the other columns. We will consider different experimental designs, corresponding to different values of λ. When $\lambda = 0$, X_{11} and X_2 are uncorrelated, and when $\lambda = 1$, $X_{11} = X_2$.

For instance, when $n = 16$ data points are used and the input range for all independent variables is $r = 1$, the input settings are

$$X_1^T = \begin{pmatrix} X_{11}^T \\ X_{12}^T \\ X_{13}^T \end{pmatrix} = \begin{pmatrix} 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \end{pmatrix}^T \tag{40}$$

$$X_2 = \lambda X_{11} + \sqrt{1-\lambda^2} W$$

where W^T = (-1 -1 -1 -1 -1 -1 -1 -1 1 1 1 1 1 1 1 1)^T.

The EM is described by

$$\begin{aligned} Y &= X_1\beta_1 + X_2\beta_2 + \varepsilon \\ &= X_{11}\beta_{11} + X_{12}\beta_{12} + X_{13}\beta_{13} + X_2\beta_2 + \varepsilon \end{aligned} \tag{41}$$

where β_{1i} denotes the i^{th} parameter in β_1, and ε is independently and identically distributed with mean 0 and variance σ^2 following a Normal distribution. We selected this example because it can readily be used to study the influence of various parameters (e.g. correlation in the experimental design, input range and number of data points) on whether the SM or the EM will give better predictions. The covariance matrix of the parameter estimates obtained using the EM is

$$Cov\left(\hat{\beta}_E\right) = Cov\begin{pmatrix}\hat{\beta}_{11E}\\ \hat{\beta}_{12E}\\ \hat{\beta}_{13E}\\ \hat{\beta}_{2E}\end{pmatrix} = \frac{\sigma^2}{nr^2}\begin{pmatrix}\frac{1}{1-\lambda^2} & 0 & 0 & \frac{-\lambda}{1-\lambda^2}\\ 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ \frac{-\lambda}{1-\lambda^2} & 0 & 0 & \frac{1}{1-\lambda^2}\end{pmatrix} \tag{42}$$

As $\lambda \to 1$, $\sigma^2/(nr^2(1-\lambda^2))$, which is the variance of $\hat{\beta}_{11E}$ and of $\hat{\beta}_{2E}$, increases dramatically, and the correlation between $\hat{\beta}_{11E}$ and $\hat{\beta}_{2E}$ approaches –1. However, since X_2 is orthogonal to both X_{12} and X_{13}, the variances of $\hat{\beta}_{12E}$ and $\hat{\beta}_{13E}$ are not affected by λ.

The SM is

$$\begin{aligned} Y &= X_1\beta_1 + e \\ &= X_{11}\beta_{11} + X_{12}\beta_{12} + X_{13}\beta_{13} + e \end{aligned} \tag{43}$$

The expected values of the parameter estimates (from (11)) obtained using the SM are

$$E\left(\hat{\beta}_{1S}\right) = E\begin{pmatrix}\hat{\beta}_{11S}\\ \hat{\beta}_{12S}\\ \hat{\beta}_{13S}\end{pmatrix} = \begin{pmatrix}\beta_{11} + \lambda\beta_2\\ \beta_{12}\\ \beta_{13}\end{pmatrix} \tag{44}$$

It can be seen that $\hat{\beta}_{11S}$ will be biased unless $\lambda = 0$ or $\beta_2 = 0$. $\hat{\beta}_{12S}$ and $\hat{\beta}_{13S}$ are unbiased. The covariance matrix of $\hat{\beta}_{1S}$ is

$$Cov\left(\hat{\beta}_{1S}\right) = Cov\begin{pmatrix}\hat{\beta}_{11S}\\ \hat{\beta}_{12S}\\ \hat{\beta}_{13S}\end{pmatrix} = \frac{\sigma^2}{nr^2}\begin{pmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix} \tag{45}$$

Comparing with (42), it is seen that

$$Var\left(\hat{\beta}_{11S}\right) \le Var\left(\hat{\beta}_{11E}\right) \tag{46}$$

These two variances are equal only when there is no correlation between X_{11} and X_2 ($\lambda = 0$).

Based on the results summarized in Tables 1 and 2, the MSE of parameter estimates and model predictions from the SM and the EM are

$$\begin{aligned} MSE\left(\hat{\beta}_{1S}\right) &= \frac{\sigma^2 p}{nr^2} + \beta_2^2\lambda^2 \\ MSE\left(\hat{\beta}_{1E}\right) &= \frac{\sigma^2 p}{nr^2} + \frac{\lambda^2\sigma^2}{nr^2\left(1-\lambda^2\right)} \\ MSE\left(\hat{Y}_S\right) &= p\sigma^2 + nr^2\left(1-\lambda^2\right)\beta_2^2 \\ MSE\left(\hat{Y}_E\right) &= p\sigma^2 + \sigma^2 \end{aligned} \tag{47}$$

From the MSE expressions given in (47),

$$R_C = \frac{nr^2\left(1-\lambda^2\right)\beta_2^2}{\sigma^2} \tag{48}$$

The combinations of (σ^2, n, r, λ, β_2) that satisfy Inequalities (26) and (32) are very clear. R_C is small when: (1) σ^2 is large (high noise levels); (2) n is small (few data points); (3) r is small (small range of input conditions); (4) $\lambda \to 1$ (strong correlation among input variables in the SM with the remaining variables in the EM); and (5) β_2^2 is small (small values for the excluded parameters). This example will be used in Monte Carlo simulations described below.

Monte Carlo Simulations

In this section, Monte Carlo simulations are performed to illustrate the theoretical analysis in the previous sub-section. Note that, since there is only one additional parameter in the EM (q = 1), Inequalities (26) and (32) are the same.

We consider a set of experiments with n = 16 data points, and the input range r = 1, as described in (40) of earlier. Let the true parameter values be

$$\beta^T = \begin{pmatrix}\beta_{11} & \beta_{12} & \beta_{13} & \beta_2\end{pmatrix}^T = \begin{pmatrix}1 & -1 & 1 & -1\end{pmatrix}^T \tag{49}$$

and the noise variance σ^2 = 1. For this example, R_C is

$$R_C = \frac{nr^2\left(1-\lambda^2\right)\beta_2^2}{\sigma^2} = 16(1-\lambda^2) \tag{50}$$

Based on the above expression, the range of λ values that ensures that $R_C \le 1$ is $0.968 \le \lambda \le 1$. If λ is in this interval, the SM is better than the EM in the sense of MSE for parameter estimates and model predictions. To numerically demonstrate several of the concepts, let $\lambda = 0.99$, which gives R_C = 0.3184.

Comparison of Parameter Estimates and Model Predictions from the SM and EM

A total of 1000 simulated data sets were generated using different random noise sequences. Figure 1 shows a boxplot comparison of parameter estimates from the SM and the EM.

$\hat{\beta}_{11S}$ has much smaller variance than $\hat{\beta}_{11E}$ due to the strong correlation between X_2 and X_{11}, but $\hat{\beta}_{11S}$ is biased. Point estimates of parameters β_{12} and β_{13} from both models are the same, because X_2 is orthogonal to X_{12} and X_{13} .

Figure 2 shows the model predictions made at the 1st, 5th, 11th and 16th observation point (model predictions made at other observation points have similar behaviour). Model predictions from the EM have larger variances than those from the SM, which are biased. As expected, predictions from the SM are better, on the average, than those from the EM.

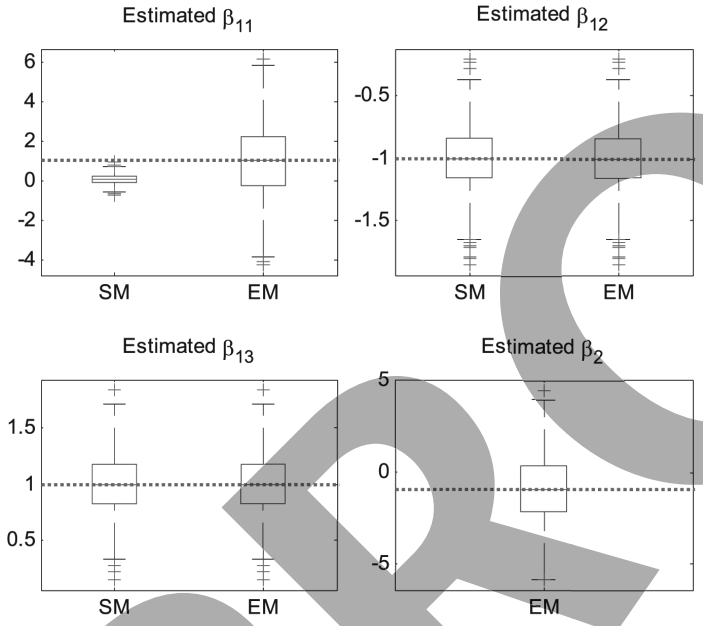
Deciding Whether to Use the SM or the EM

To illustrate the construction of confidence intervals for R_C using available data, we will consider two cases (i.e., $\hat{R}_C$ = 0.6585 and $\hat{R}_C$ = 1.5821 could be obtained from different simulated data sets) using the model described in the previous sub-section. Recall that the true value is R_C = 0.3184 The first value, $\hat{R}_C$ = 0.6585 corresponds to the *median* valùe from the probability density function for $\hat{R}_C$ (which is distributed as $F_{1,12}(\delta)$ with $\delta = qR_C$ = 0.3184). The second value, $\hat{R}_C$ = 1.5821, corresponds to the *mean* of the same distribution. The large difference between median and mean indicates that the distribution is right skewed.

Using the algorithm described in the "Strategy for Assessing Uncertainty About Which Model is Better" section, the two-sided

100(1 – α)% confidence interval for R_C is constructed for different values of α. When $\hat{R}_C = 0.6585$, the upper and lower limits are plotted in Figure 3. As α increases, the confidence limits narrow and converge to $\hat{R}_C = 0.6585$. The limits are very wide for typical values of α that are commonly recommended for confidence intervals (i.e., $\alpha \leq 0.1$). Although $R_C = 0.3184 < 1$, which indicates that the SM is better than the EM, the results in Figure 3 show that, for a reasonable value of α (near 0.10), the two-sided confidence interval for R_C is [0, 6]. Since 1.0 is within this range, we are unable to distinguish whether the SM gives better predictions than the EM (this is a Type II error).

The confidence limits obtained from the mean value of $\hat{R}_C = 1.5821$ are shown in Figure 4. For α near 0.10, the two-sided confidence interval is [0, 8.5], which is broader than in Figure 3.



Figur 1. Boxplot comparison of parameter estimates from the SM and the EM ····· shows the true parameter values used in the simulations.

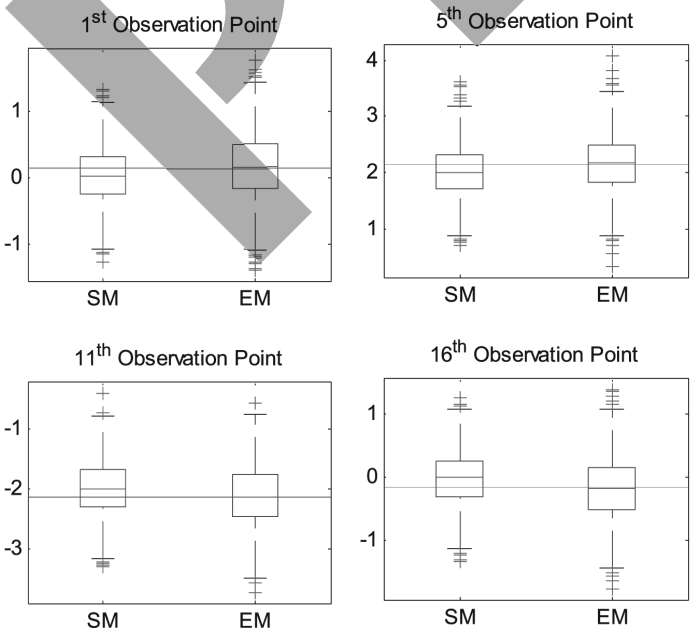


Figure 2. Boxplot comparison of model predictions from the SM and the EM ····· shows the noise-free response at given observation point.

The confidence intervals in Figures 3 and 4 are exact (Steiger, 2004). The high probability of mistakenly accepting the null hypothesis (concluding that the SM is not significantly better than the EM) arises from the limited data (few data points, limited input range and correlated design). In situations where there are more data points or there are more parameters in the EM ($q > 1$), the confidence intervals for R_C become narrower (higher degrees of freedom in the non-central F distribution). Narrower confidence intervals lead to better discrimination about which model is better. We are currently investigating the power of statistical tests to determine whether the EM or the SM is preferred using larger and more realistic models and data sets of interest to chemical engineers.

Statistical Inference based on the SM

In this section, we compare various estimates for the variance of parameter estimates from the SM and the additive noise ε.

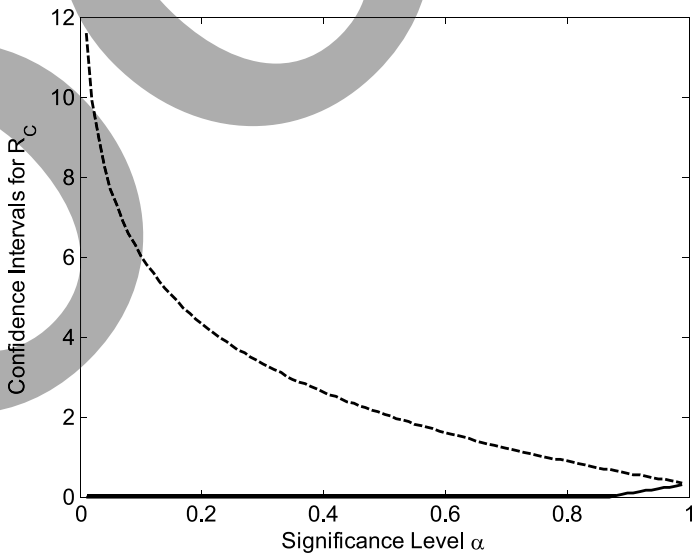


Figure 3. Two-sided confidence intervals for R_C for different values of the significance level α when $\hat{R}_C = 0.6585$. — — — — Upper confidence bound, ———— Lower confidence bound

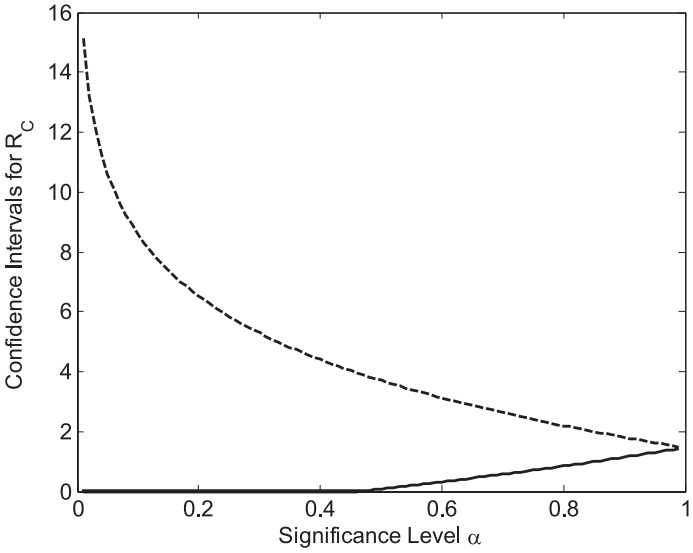


Figure 4. Two-sided confidence intervals for R_C for different values of the significance level α when $\hat{R}_C = 1.5821$. — — — — Upper confidence bound, ———— lower confidence bound

Two cases are considered: (1) $\lambda = 0.90$, which corresponds to $R_C = 3.04 > 1$ so that the EM is better; and (2) $\lambda = 0.99$ which corresponds to $R_C = 0.3184 < 1$, so that the SM is better (on average).

As there are no replicate experiments available in this example, we only consider using s_S^2 (from the SM residuals), s_E^2 (from the EM residuals) and s_B^2 (from nonparametric bootstrapping) to estimate the noise variance σ^2. To evaluate these three methods, a total of 1000 data sets were generated using different random noise sequences. For each simulated data set, $B = 200$ bootstraps were performed. In all simulations, we assumed that the noise is independently and identically distributed following a standard Normal distribution. The estimated noise variances from each situation ($\lambda = 0.90$ and $\lambda = 0.99$) are compared using boxplots in Figure 5. The dashed line is the true value of the noise variance used in the simulation ($\sigma^2 = 1$). It is seen that, in the case when the EM is better, s_S^2 is biased upward, and s_E^2 and s_B^2 provide good estimate of σ^2. However, in the case when the SM is better, s_S^2 and s_E^2 are good estimates of σ^2. Based on these simulations, it seems that, when the correctly structured EM is available, it should be used to estimate the noise variance because it provides an unbiased estimate. However, this issue requires further investigation because bias is not the entire problem. Perhaps other variance estimators will provide estimates with lower mean-squared-error.

The estimated variance of $\hat{\beta}_{11S}$ from the SM could be obtained by three methods: (1) conventional methods (Equation (35) with s_S^2 (SM) or s_E^2 (EM)); (2) sandwich estimator (Equation (36)); and (3) nonparametric bootstrapping (Equation (39)). The results from each situation (for both $\lambda = 0.90$ and $\lambda = 0.99$) are compared in Figure 6. Since $\hat{\beta}_{12S}$ and $\hat{\beta}_{13S}$ are the same as those obtained using the EM (Figure 1), we do not consider them in this section.

It is seen that, for our example, all the methods give similar results. Several nonparametric bootstrapping algorithms can be directly used to construct the confidence intervals for parameter estimates. For this particular example, there is no noticeable difference between the results obtained using the algorithms described by Efron and Tibshirani (1993) and by Martinez and Martinez (2002).

CONCLUSIONS AND RECOMMENDATIONS

In this paper, a number of important issues related to the use of simplified or misspecified models have been reviewed. Much of the statistical literature has focused on model validation and model use for models that are assumed to be statistically valid. There are however many instances when it is not possible to construct a model that is deemed statistically acceptable. There are other instances where the use of a truly structured extended model is undesirable due to the inherent complexity of the model. Additionally, there are instances where simplified or misspecified models can give predictions that are superior, in the mean-squared-error sense, to those from the extended model.

For models that are linear in the parameters, it is possible to study the implications of using simplified or extended models. This study was undertaken using both theoretical analysis and Monte Carlo simulations. The simplified model gives better parameter estimates and model predictions (on average) than the extended model if the inequalities (involving the critical ratio R_C) described in (26) and (32) are satisfied. In these situations, the simplified model is superior, even though the extended model is correctly structured and the simplified model is misspecified. It was demonstrated that these inequalities are satisfied when there are high noise levels, strong correlations among input variables, small number of experiments, a small range of independent variable settings, or small true values for parameters that are excluded from the simplified model. All of these situations are unfavourable for obtaining precise parameter estimates. When modellers are faced with uninformative and noisy data from poorly designed experiments, they should not try to estimate too many model parameters. Rather, they should confine themselves to fitting only a few key parameters that appear in the most important parts of their models. Unfortunately,

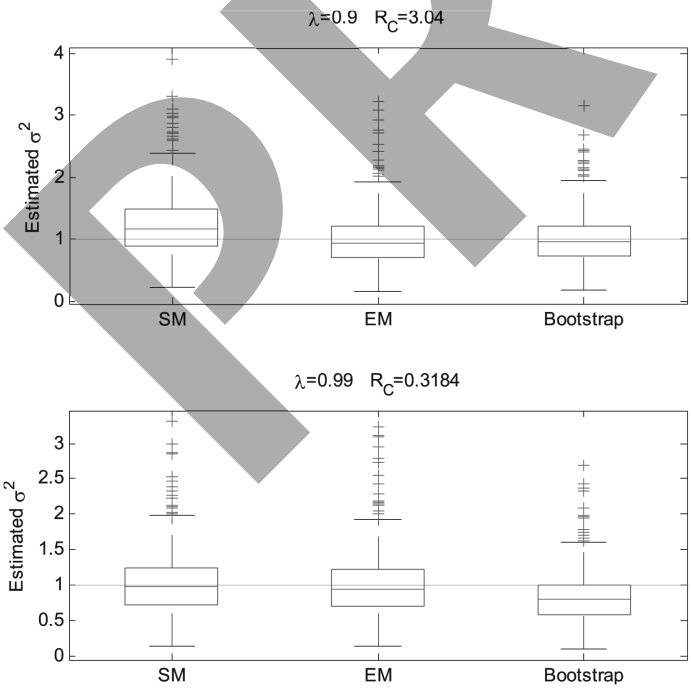


Figure 5. Comparison of estimated noise variances from the SM (s_S^2), the EM (s_E^2), and nonparametric bootstrapping (s_B^2)

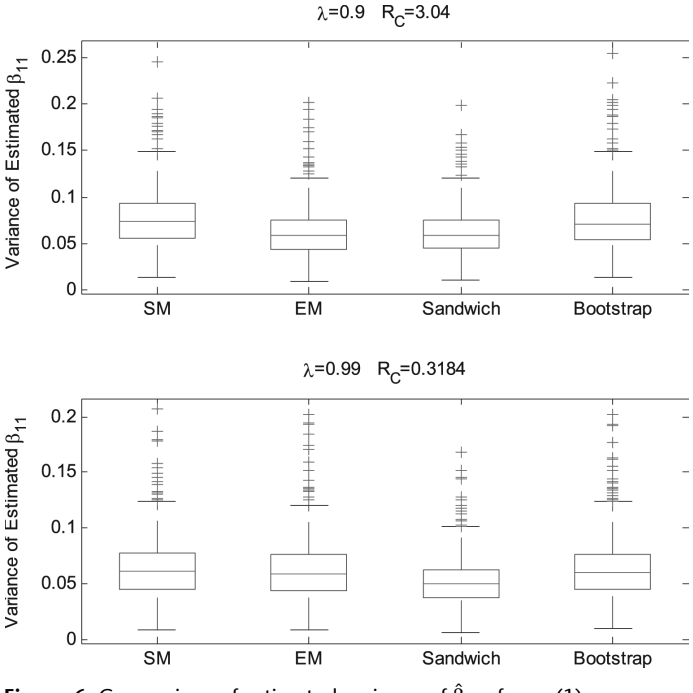


Figure 6. Comparison of estimated variance of $\hat{\beta}_{11S}$ from: (1) Conventional methods (based on the SM (s_S^2) and the EM (s_E^2)); (2) Sandwich estimator; and (3) nonparametric bootstrapping

there are considerable challenges in deciding which model is preferred, using the limited data that are available for parameter estimation. We have demonstrated how confidence intervals can be constructed for the critical ratio R_C. These intervals are often quite wide, especially when the number of parameters excluded is small. The result is that the statistical tests, while exact in their construction, can have poor discrimination properties for alternative hypotheses (either that the simplified model is better or that the extended model is better) when the data are uninformative. However, when the data are informative, and the terms left out of the simplified model are important, the lower confidence bound for R_C becomes greater than 1, and firm conclusions can be drawn that the extended model is better.

Several methods were investigated for estimating the noise variance and variance-covariance matrix of the estimated parameters obtained from the simplified or misspecified model. Interest in this area has been revived by the availability of inexpensive computing for computationally intensive methods such as the nonparametric bootstrapping.

The focus in this paper was on models that are linear in the parameters, but the results are very important for phenomenologically based models that are non-linear in the parameters. In these instances, there are often competing models that can be used. The difference in complexity between a simplified model and an extended model can be substantial. The application of the results in this paper can be used, in the first instance, on the linearized representation of the non-linear model. In these instances, X is replaced by a parametric sensitivity matrix, whose i^{th} column is $\partial f(X, \mathrm{b})/\partial \mathrm{b}_i|_{\beta=\hat{\beta}}$, where $\hat{\beta}$ is either a least-squares estimate of β or an initial guess for the parameter values. In non-linear models, the parametric sensitivity matrix, which corresponds to $X^T X$ for a linear model, is often ill-conditioned (Kou et al., 2005a, b; Bates and Watts, 1988), so that conditions under which the simplified model gives superior predictions are often present. Our future work will involve deciding how to simplify non-linear models so that the best possible predictions and parameter estimates can be obtained using limited data.

NOMENCLATURE

a	vector of coefficients
e	stochastic component
f, g	functions relating explanatory variables, parameters to response variable
k	number of explanatory variables in the true model
m	total number of parameters
n	number of observations
p	number of parameters in first part
p_C	cumulative probability of $\hat{R}_C$ based on central F distribution
q	number of parameters in second part
r	input range
s^2	sample variance
w	total number of predictions
x, z	single observation of explanatory variable
$\tilde{x}_i$	i^{th} row of X_1
y	single observation of response variable
A	auxiliary regression matrix
B	number of bootstraps
I	identity matrix
P	projection matrix
R_C	critical value
S	sum square residuals
W	vector of length n with entries of $\pm r$
X	matrix of regression variables
Y	response variables
Z	matrix of prediction variables

Greek Symbols

α	significance level
β, θ	unknown parameters
δ	non-centrality parameter
ε	stochastic component
λ	correlation factor
μ	expectation
σ^2	unknown noise variance
Γ	variance-covariance matrix of parameter estimates in the first part
Δ	bias
Σ	variance-covariance
Ψ	covariance matrix between the first part and the second part
Ω	variance-covariance matrix of parameter estimates in the second part

Superscripts

-1	inverse
T	transpose
$\hat{}$	estimated value
$\bar{}$	mean value

Subscripts

1	first partitioned part
2	second partitioned part
i	index
$BOOT$	results from nonparametric bootstrapping
$CONV$	results from conventional methods
E	extended model
L	lower confidence limit
S	simplified model
$SANW$	results from sandwich estimator
U	upper confidence limit
$*$	results from nonparametric bootstrapping

Abbreviations

min	minimization
Cov	covariance matrix
E	mathematical expectation
EM	correctly structured extended model
MSE	mean–squared-error
MSEM	mean-squared-error-matrix
OLS	ordinary least-squares
RHS	right-hand side
SM	simplified/misspecified model
Tr	trace
Var	variance

Others

$\mathbb{R}^n$	column vector of length n taking real values
$\mathbb{R}^{n\times m}$	(n × m) matrix taking real values

REFERENCES

Abdullaev, F. M. and E. K. Geidarov, "A Recursive 2-step Method of Least-Squares," Automat. Rem. Contr. **46**(1), 66–72 (1985).

Aerts, M. and G. Claeskens, "Bootstrap Tests for Misspecified Models, with Application to Clustered Binary Data," Comput. Stat. Data An. **36**(3), 383–401 (2001).
Bagajewicz, M. J. and E. Cabrera, "Data Reconciliation in Gas Pipeline Systems," Ind. Eng. Chem. Res. **42**(22), 5596–5606 (2003).
Bates, D. M. and D. G. Watts, "Nonlinear Regression Analysis and Its Applications," John Wiley & Sons, NY (1988).
Beck, J. V. and K. J. Arnold, "Parameter Estimation in Engineering and Science," John Wiley & Sons, NY (1977).
Bera, A. K., "Hypothesis Testing in the 20th Century with a Special Reference to Testing with Misspecified Models," in "Statistics for the 21st Century: Methodologies for Applications of the Future," C. R. Rao, G. J. Szekely, Marcel Dekkar, NY (2000), pp. 33–92.
Brendel, M., D. Bonvin and W. Marquardt, "Incremental Identification of Kinetic Models for Homogeneous Reaction Systems," Chem. Eng. Sci. **61**, 5404–5420 (2006).
Brooks, R. J. and A. M. Tobias, "Choosing the Best Model: Level of Detail, Complexity, and Model Performance," Math. Comput. Model. **24**(4), 1–14 (1996).
Chang, S., T. D. Waite and A. G. Fane, "A Simplified Model for Trace Organics Removal by Continuous Flow PAC Adsorption/Submerged Membrane Processes," J. Membrane Sci. **253**(1–2), 81–87 (2005).
Chernick, M. R., "Bootstrap Methods: A Practitioner's Guide," John Wiley & Sons, NY (1999).
Davison, A. C. and D. V. Hinkley, "Bootstrap Methods and Their Applications," Cambridge University Press, U.S.A. (1997).
Donaldson, J. R. and R. B. Schnabel, "Computational Experience with Confidence Regions and Confidence Intervals for Nonlinear Least Squares," Technometrics **29**(1), 67–82 (1987).
Draper, N. R. and H. Smith, "Applied Regression Analysis," 3rd ed, John Wiley & Sons, NY (1998).
Efron, B. and R. J. Tibshirani, "An Introduction to the Bootstrap," Chapman and Hall, London (1993).
Freund, R. J., C. W. Cluniesross and R. W. Vail, "Residual Analysis," J. Am. Stat. Assoc. **56**(293), 98–104 (1961).
Fushiki, T., "Bootstrap Prediction and Bayesian Prediction under Misspecified Models," Bernoulli **11**(4), 747–758 (2005).
Golbert, J. and D. R. Lewin, "Model-Based Control of Fuel Cells: (1) Regulatory Control," J. Power Sources **135**(1–2), 135–151 (2004).
Golden, R. M., "Making Correct Statistical Inferences Using a Wrong Probability Model," J. Math. Psychol. **39**, 3–20 (1995).
Goldberg, A., "Stepwise Least-Squares—Residual Analysis and Specification Error," J. Am. Stat. Assoc. **56**(296), 998–1000 (1961).
Goldberg, A. and D. B. Jochems, "Note on Stepwise Least-Squares," J. Am. Stat. Assoc. **56**(293), 105–110 (1961).
Good, P. I., "Resampling Methods: A Practical Guide to Data Analysis," 3rd ed, Birkhäuser, US (2005)
Gunst, R. F. and R. L. Mason, "Biased Estimation in Regression—Evaluation using Mean Squared Error," J. Am. Stat. Assoc. **72**(359), 616–628 (1977).
Hocking, R. R., "Analysis and Selection of Variables in Linear Regression," Biometrics **32**(1), 1–49 (1976).
Innis, G. and E. Rexstad, "Simulation Model Simplification Techniques," Simulation **41**(1), 7–15 (1983).
Kabe, D. G., "Stepwise Multivariate Linear-Regression," J. Am. Stat. Assoc. **58**(303), 770–773 (1963).
Kou, B., K. B. McAuley, C. C. Hsu and D. W. Bacon, "Mathematical Model and Parameter Estimation for Gas-Phase Ethylene/Hexene Copolymerization with Metallocene Catalyst," Macromol. Mater. Eng. **290**(6), 537–557 (2005a).
Kou, B., K. B. McAuley, C. C. Hsu, D. W. Bacon and K. Z. Yao, "Mathematical Model and Parameter Estimation for Gas-Phase Ethylene Homopolymerization with Supported Metallocene Catalyst," Ind. Eng. Chem. Res. **44**(8), 2428–2442 (2005b).
Lowerre, J. M., "Mean-Square Error of Parameter Estimates for Some Biased Estimators," Technometrics **16**(3), 461–464 (1974).
Lv, P., J. Chang, T. Wang, C. Wu and N. Tsubaki, "A Kinetic Study on Biomass Fast Catalytic Pyrolysis," Energ. Fuel. **18**(6), 1865–1869 (2004).
Maria, G., "A Review of Algorithms and Trends in Kinetic Model Identification for Chemical and Biochemical Systems," Chem. Biochem. Eng. Q. **18**(3), 195–222 (2004).
Martinez W. L. and A. R. Martinez, "Computational Statistics Handbook with MATLAB," Champman & Hall/CRC, U.S.A. (2002).
Mchaweh, A., A. Alsaygh, K. Nasrifar and M. Moshfeghian, "A Simplified Method for Calculating Saturated Liquid Densities," Fluid Phase Equilib. **224**(2), 157–167 (2004).
Miller, A. J., "Subset Selection in Regression," Chapman and Hall, London (1990).
Montgomery, D. C., E. A. Peck and G. G. Vining, "Introduction to Linear Regression Analysis," 3rd ed, John Wiley & Sons, NY (2001).
Montgomery, D. C. and G. C. Runger, "Applied Statistics and Probability for Engineers," 3rd ed, John Wiley & Sons, NY (2003).
O'Brien, S. M., L. L. Kupper and D. B. Dunson, "Performance of Tests of Association in Misspecified Generalized Linear Models," J. Stat. Plan. and Infer. **136**, 3090–3100 (2006).
Perregaard, J., "Model Simplification and Reduction for Simulation and Optimization of Chemical Processes," Comput. Chem. Eng. **17**(5–6), 465–483 (1993).
Price, J. M., "Comparisons among Regression—Estimators under the Generalized Mean-Square Error Criterion," Commun. Stat.-Theor. M. **11**(17), 1965–1984 (1982).
Rao, P., "Some Notes on Misspecification in Multiple Regressions," Am. Stat. **25**(5), 37–39 (1971).
Rao C. R. and Y. Wu, "On Model Selection," in "Model Selection," P. Lahiri, Institute of Mathematical Statistics, Beachwood, OH (2001), pp. 1–64.
Rexstad, E. and G. S. Innis, "Model Simplification—3 Applications," Ecol. Model. **27**(1–2), 1–13 (1985).
Romdhane, M. and C. Tizaoui, "The Kinetic Modelling of a Steam Distillation Unit for the Extraction of Aniseed (Pimpinelia Anisum) Essential Oil," J. Chem. Technol. Biot. **80**(7), 759–766 (2005).
Rosenberg, S. H. and P. S. Levy, "Characterization on Misspecification in General Linear Regression Model," Biometrics **28**(4), 1129–1133 (1972).
Seber, G. A. F. and C. J. Wild, "Nonlinear Regression," John Wiley & Sons, NJ (2003).
Steiger, J. H., "Beyond the F test: Effect Size Confidence Intervals and Tests fo Close Fit in the Analysis of Variance and Contrast Analysis," Psychol. Methods 9 (2), 164–182 (2004).

Sun, C. and J. Hahn, "Parameter Reduction for Stable Dynamical Systems based on Hankel Singular Values and Sensitivity Analysis," Chem. Eng. Sci. **61**, 5393–5403 (2006)
Toro-Vizcarrondo, C. and T. D. Wallace, "A Test of the Mean Square Error Criterion for Restrictions in Linear Regression," J. Am. Stat. Assoc. **63**(322), 558–572 (1968).
Toutenburg, H. and G. Trenkler, "Mean-Square Error Matrix Comparisons of Optimal and Classical Predictors and Estimators in Linear-Regression," Comput. Stat. Data An. **10**(3), 297–305 (1990).
Velilla S., "On the Bootstrap in Misspecified Regression Models," Comput. Stat. Data An. **36**(2), 227–242 (2001).
Waldorp, L. J., R. P. P. P. Grasman and H. M. Huizenga, "Goodness-of-fit and Confidence Intervals of Approximate Models," J. Math. Psychol. **50**, 203–213 (2006).
Waldorp, L. J., H. M. Huizenga and R. P. P. P. Grasman, "The Wald Test and Cramér-Rao Bound for Misspecified Models in Electromagnetic Source Analysis," IEEE T. Signal Proces **53**(9), 3427–3435 (2005).
Wallace, T. D., "Efficiencies for Stepwise Regressions," J. Am. Stat. Assoc. **59**(308), 1179–1182 (1964).
Wang, S. G. and S. C. Chow, "Advanced Linear Models: Theory and Applications," Marcel Dekker, NY (1994).
White, H., "Maximum Likelihood Estimation of Misspecified Models," Econometrica **50**(1), 1–26 (1982).
White, H., "Consequences and Detection of Misspecified Nonlinear Regression Models," J. Am. Stat. Assoc. **76**(374), 419–433 (1981).
Yoshida, H., Y. Takahashi and M. Terashima, "A Simplified Reaction Model for Production of Oil, Amino Acids, and Organic Acids from Fish Meat by Hydrolysis under Sub-Critical and Supercritical Conditions," J. Chem. Eng. Japan **36**(4), 441–448 (2003).
Zhang, F., "Matrix Theory: Basic Results and Techniques," Springer-Verlag, NY (1999).

Manuscript received February 3, 2007; revised manuscript received May 1, 2007; accepted for publication May 3, 2007.