Response Surface Methodology[1]

CASOS Technical Report

Kathleen M. Carley, Natalia Y. Kamneva, Jeff Reminga

October 2004

CMU-ISRI-04-136

Carnegie Mellon University

School of Computer Science

ISRI - Institute for Software Research International

CASOS - Center for Computational Analysis of Social and Organizational Systems

[1] This work was supported in part by NASA # NAG-2-1569, Office of Naval Research Grant N00014-02-1-0973, "Dynamic Network Analysis: Estimating Their Size, Shape and Potential Weaknesses", Office of Naval Research, N00014-97-1-0037, "Constraint Based Team Transformation and Flexibility Analysis" under "Adaptive Architectures", the DOD and the National Science Foundation under MKIDS. Additional support was provided by the center for Computational Analysis of Social and Organizational Systems (CASOS) (http://www.casos.cs.cmu.edu) and the Institute for Software Research International at Carnegie Mellon University. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the National Science Foundation, or the U.S. government.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
OCT 2004		**00-10-2004 to 00-10-2004**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Response Surface Methodology	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Carnegie Mellon University,School of Computer Science,Pittsburgh,PA,15213**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**31**	

Keywords: Response Surface Methodology (RSM), regression analysis, linear regression model, regressors, variable selection, model building, full model, multicollinearity, ridge regression, unit length scaling**,** condition number, optimization, Simulated Annealing, global optimum

Abstract

There is a problem faced by experimenters in many technical fields, where, in general, the response variable of interest is y and there is a set of predictor variables $x_1, x_2, ..., x_k$. For example, in Dynamic Network Analysis (DNA) Response Surface Methodology (RSM) might be useful for sensitivity analysis of various DNA measures for different kinds of random graphs and errors.

In Social Network Problems usually the underlying mechanism is not fully understood, and the experimenter must approximate the unknown function g with appropriate empirical model

$y = f(x_1, x_2, ..., x_k) + \varepsilon$, where the term ε represents the error in the system.

Usually the function f is a first-order or second-order polynomial. This empirical model is called a response surface model.

Identifying and fitting from experimental data an appropriate response surface model requires some use of statistical experimental design fundamentals, regression modeling techniques, and optimization methods. All three of these topics are usually combined into Response Surface Methodology (RSM).

Also the experimenter may encounter situations where the full model may not be appropriate. Then variable selection or model-building techniques may be used to identify the best subset of regressors to include in a regression model. In our approach we use the simulated annealing method of optimization for searching the best subset of regressors. In some response surface experiments, there can be one or more near-linear dependences among regressor variables in the model. Regression model builders refer to this as multicollinearity among the regressors. Multicollinearity can have serious effects on the estimates of the model parameters and on the general applicability of the final model.

The RSM is also extremely useful as an automated tool for model calibration and validation especially for modern computational multi-agent large-scale social-networks systems that are becoming heavily used in modeling and simulation of complex social networks.

The RSM can be integrated in many large-scale simulation systems such as BioWar, ORA and is currently integrating in Vista, Construct, and DyNet.

This report describes the theoretical approach for solving of these problems and the implementation of chosen methods.

Table of Contents

List of Tables

1. Introduction

1.1 Response Surface Methodology

Response Surface Methodology (RSM) is a collection of statistical and mathematical techniques useful for developing, improving, and optimizing processes [1].

The most extensive applications of RSM are in the particular situations where several input variables potentially influence some performance measure or quality characteristic of the process. Thus performance measure or quality characteristic is called the **response**. The input variables are sometimes called **independent variables,** and they are subject to the control of the scientist or engineer. The field of response surface methodology consists of the experimental strategy for exploring the space of the process or independent variables, empirical statistical modeling to develop an appropriate approximating relationship between the yield and the process variables, and optimization methods for finding the values of the process variables that produce desirable values of the response. In this report we will concentrate on the second strategy: statistical modeling to develop an appropriate approximating model between the response *y* and independent variables $\xi_1, \xi_2, ..., \xi_k$.

In general, the relationship is

$$y = f(\xi_1, \xi_2, ..., \xi_k) + \varepsilon ; \tag{1.1}$$

where the form of the true response function *f* is unknown and perhaps very complicated, and ε is a term that represents other sources of variability not accounted for in *f*. Usually ε includes effects such as measurement error on the response, background noise, the effect of other variables, and so on. Usually ε is treated as a statistical error, often assuming it to have a normal distribution with mean zero and variance σ^2 . Then

$$E(y) = \eta = E\,[f(\xi_1, \xi_2, ..., \xi_k)] + E(\varepsilon) = f(\xi_1, \xi_2, ..., \xi_k); \tag{1.2}$$

The variables $\xi_1, \xi_2, ..., \xi_k$ in Equation (1.2) are usually called the **natural variables,** because they are expressed in the natural units of measurement, such as degrees Celsius, pounds per square inch, etc. In much RSM work it is convenient to transform the natural variables to **coded variables** $x_1, x_2, ..., x_k$, which are usually defined to be dimensionless with mean zero and the same standard deviation. In terms of the coded variables, the response function (1.2) will be written as

$$\eta = f(x_1, x_2, ..., x_k); \tag{1.3}$$

Because the form of the true response function *f* is unknown, we must approximate it. In fact, successful use of RSM is critically dependent upon the experimenter's ability to develop a suitable approximation for *f*. Usually, a low-order polynomial in some relatively small region of

the independent variable space is appropriate. In many cases, either a **first-order** or a **second-order** model is used.

The first-order model is likely to be appropriate when the experimenter is interested in approximating the true response surface over a relatively small region of the independent variable space in a location where there is little curvature in *f*.

For the case of two independent variables, the first-order model in terms of the coded variables is

$$\eta = \beta_o + \beta_1 x_1 + \beta_2 x_2 ; \tag{1.4}$$

The form of the first-order model in Equation (1.4) is sometimes called a **main effects model**, because it includes only the main effects of the two variables x_1 and x_2. If there is an **interaction** between these variables, it can be added to the model easily as follows:

$$\eta = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 ; \tag{1.5}$$

This is the first-order model with interaction. Adding the interaction term introduces curvature into the response function.

Often the curvature in the true response surface is strong enough that the first-order model (even with the interaction term included) is inadequate. A second-order model will likely be required in these situations. For the case of two variables, the second-order model is

$$\eta = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 ; \tag{1.6}$$

This model would likely be useful as an approximation to the true response surface in a relatively small region.

The second-order model is widely used in response surface methodology for several reasons:

1. The second-order model is very flexible. It can take on a wide variety of functional forms, so it will often work well as an approximation to the true response surface.

2. It is easy to estimate the parameters (the β's) in the second-order model. The method of least squares can be used for this purpose.

3. There is considerable practical experience indicating that second-order models work well in solving real response surface problems.

In general, the first-order model is

$$\eta = \beta_o + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \tag{1.7}$$

and the second-order model is

$$\eta = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + \sum_{j=1}^{k} \beta_{jj} x_j^2 + \sum_{i<} \sum_{j=2}^{k} \beta_{ij} x_i x_j \qquad (1.8)$$

In some infrequent situations, approximating polynomials of order greater than two are used. The general motivation for a polynomial approximation for the true response function *f* is based on the Taylor series expansion around the point $x_{10}, x_{20}, ..., x_{k0}$.

Finally, let's note that there is a close connection between RSM and **linear regression analysis.** For example, consider the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$$

The β's are a set of unknown parameters. To estimate the values of these parameters, we must collect data on the system we are studying. Because, in general, polynomial models are linear functions of the unknown β's, we refer to the technique as linear regression analysis.

1.2 Response Surface Methodology and Robust Design

RSM is an important branch of experimental design. RSM is a critical technology in developing new processes and optimizing their performance. The objectives of quality improvement, including reduction of variability and improved process and product performance, can often be accomplished directly using RSM.

It is well known that variation in key performance characteristics can result in poor process and product quality. During the 1980s [2, 3] considerable attention was given to process quality, and methodology was developed for using experimental design, specifically for the following:

1. For designing or developing products and processes so that they are robust to component variation.

2. For minimizing variability in the output response of a product or a process around a target value.

3. For designing products and processes so that they are robust to environment conditions.

By **robust** means that the product or process performs consistently on target and is relatively insensitive to factors that are difficult to control.

Professor Genichi Taguchi [2, 3] used the term **robust parameter design** (RPD) to describe his approach to this important problem. Essentially, robust parameter design methodology prefers to reduce process or product variation by choosing levels of controllable factors (or parameters) that make the system insensitive (or robust) to changes in a set of uncontrollable factors that represent most of the sources of variability. Taguchi referred to these uncontrollable factors as **noise factors**. RSM assumes that these noise factors are uncontrollable in the field, but can be controlled during process development for purposes of a designed experiment.

Considerable attention has been focused on the methodology advocated by Taguchi, and a number of flaws in his approach have been discovered. However, the framework of response

surface methodology allows easily incorporate many useful concepts in his philosophy [1]. There are also two other full-length books on the subject of RSM [4, 5].

In our technical report we are concentrated mostly on building and optimizing the empirical models and practically do not consider the problems of experimental design.

1.3 The Sequential Nature of the Response Surface Methodology

Most applications of RSM are **sequential** in nature.

Phase 0: At first some ideas are generated concerning which factors or variables are likely to be important in response surface study. It is usually called a **screening experiment**. The objective of factor screening is to reduce the list of candidate variables to a relatively few so that subsequent experiments will be more efficient and require fewer runs or tests. The purpose of this phase is the identification of the important independent variables.

Phase 1: The experimenter's objective is to determine if the current settings of the independent variables result in a value of the response that is near the optimum. If the current settings or levels of the independent variables are not consistent with optimum performance, then the experimenter must determine a set of adjustments to the process variables that will move the process toward the optimum. This phase of RSM makes considerable use of the first-order model and an optimization technique called the **method of steepest ascent (descent)**.

Phase 2: Phase 2 begins when the process is near the optimum. At this point the experimenter usually wants a model that will accurately approximate the true response function within a relatively small region around the optimum. Because the true response surface usually exhibits curvature near the optimum, a second-order model (or perhaps some higher-order polynomial) should be used. Once an appropriate approximating model has been obtained, this model may be analyzed to determine the optimum conditions for the process.

This sequential experimental process is usually performed within some region of the independent variable space called the **operability region or experimentation region or region of interest.**

2. Building Empirical Models

2.1 Linear Regression Model

In the practical application of RSM it is necessary to develop an approximating model for the true response surface. The underlying true response surface is typically driven by some unknown physical mechanism. The approximating model is based on observed data from the process or system and is an empirical model. Multiple regression is a collection of statistical techniques useful for building the types of empirical models required in RSM.

The first-order **multiple linear regression model** with two independent variables is

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \qquad (2.1)$$

The independent variables are often called **predictor variables** or **regressors**. The term "linear" is used because Equation (2.1) is a linear function of the unknown parameters $\beta_o, \beta_1, and \beta_2$.

In general, the response variable *y* may be related to *k* regressor variables. The model

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon \quad (2.2)$$

is called a multiple linear regression model with *k* regressor variables. The parameters β_j, *j=0,1,…k,* are called the **regression coefficients**.

Models that are more complex in appearance than Equation (2.2) may often still be analyzed by multiple linear regression techniques. For example, considering adding an **interaction term** to the first-order model in two variables

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon \quad (2.3)$$

As another example, consider the **second-order** response surface model in two variables

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon \quad (2.4)$$

In general, any regression model that is linear in the parameters (the β-values) is a linear regression model, regardless of the shape of the response surface that it generates.

2.2 Estimation of the Parameters in Linear Regression Models

The method of least squares is typically used to estimate the regression coefficients in a multiple linear regression model. Suppose that $n > k$ observations on the response variable are available, say $y_1, y_2, ..., y_n$. Along with each observed response y_i, we will have an observation on each regressor variable, let x_{ij} denote the *i*th observation or level of variable x_j (see Table 2.1).

The model in terms of the observations may be written in matrix notation as

$$\mathbf{y = X\beta + \varepsilon} \quad (2.5)$$

where

y is an *n x 1* vector of the observations,

X is an *n x p* matrix of the levels of the independent variables,

β is a *p x 1* vector of the regression coefficients, and

ε is an *n x 1* vector of random errors.

Table 1: Data for Multiple Linear Regression

y	x_1	x_2	...	x_k
y_1	x_{11}	x_{12}	...	x_{1k}
y_2	x_{21}	x_{22}	...	x_{2k}
.	.	.	...	.
.	.	.	...	.
.	.	.	...	.
y_n	x_{n1}	x_{n2}	...	x_{nk}

We wish to find the vector of least squares estimators, **b**, that minimizes

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \quad (2.6)$$

After some simplifications, the least squares estimator of $\boldsymbol{\beta}$ is

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad (2.7)$$

It is easy to see that **X'X** is a *p x p* symmetric matrix and **X'y** is a *p x 1* column vector. The matrix **X'X** has the special structure. The diagonal elements of **X'X** are the sums of squares of the elements in the columns of **X**, and the off-diagonal elements are the sums of cross-products of the elements in the columns of **X.** Furthermore, the elements of **X'y** are the sums of cross-products of the columns of **X** and the observations $\{y_i\}$.

The fitted regression model is

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} \quad (2.8)$$

In scalar notation, the fitted model is

$$\hat{y}_i = b_0 + \sum_{j=1}^{k} b_j x_{ij}, \qquad i = 1,2,\ldots,n$$

The difference between the observation y_i and the fitted value $\hat{y}_i$ is a **residual,** $e_i = y_i - \hat{y}_i$.

The *n x 1* vector of residuals is denoted by

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} \quad (2.9)$$

2.3 Model Adequacy Checking

It is always necessary to

1. Examine the fitted model to ensure that it provides an adequate approximation to the true system;
2. Verify that none of the least squares regression assumptions are violated. Now we consider several techniques for checking model adequacy.

2.3.1 Properties of the Least Squares Estimators

The method of least squares produces an **unbiased estimator** of the parameter $\boldsymbol{\beta}$ in the multiple linear regression model. The important parameter is the sum of squares of the residuals

$$\mathrm{SS}_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = \mathbf{e'e} \tag{2.10}$$

Because **X'Xb = X'y**, we can derive a computational formula for SS_E:

$$\mathrm{SS}_E = \mathbf{y'y - b'X'y} \tag{2.11}$$

Equation (2.11) is called the **error** or **residual sum of squares**.

It can be shown that an unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{SS_E}{n-p} \tag{2.12}$$

where

n is a number of observations and

p is a number of regression coefficients.

The total sum of squares is

$$SS_T = y'y - \frac{(\sum_{i=1}^{n} y_i)^2}{n} = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n} \tag{2.13}$$

Then the coefficient of multiple determination R^2 is defined as

$$R^2 = 1 - \frac{SS_E}{SS_T} \tag{2.14}$$

R^2 is a measure of the amount of reduction in the variability of *y* obtained by using the regressor variables $x_1, x_2, ..., x_k$ in the model. From inspection of the analysis of variance identity equation (Equation (2.14)) we can see that $0 \le R^2 \le 1$. However, a large value of R^2 does not necessarily imply that the regression model is good one. Adding a variable to the model will always increase R^2, regardless of whether the additional variable is statistically significant or not. Thus it is possible for models that have large values of R^2 to yield poor predictions of new observations or estimates of the mean response.

Because R^2 always increases as we add terms to the model, some regression model builders prefer to use an adjusted R^2 statistic defined as

$$R^2_{adj} = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)} = 1 - \frac{n-1}{n-p}(1-R^2) \qquad (2.15)$$

In general, the adjusted R^2 statistic will not always increase as variables are added to the model. In fact, if unnecessary terms are added, the value of R^2_{adj} will often decrease. When R^2 and R^2_{adj} differ dramatically, there is a good chance that nonsignificant terms have been included in the model.

We are frequently interested in testing hypotheses on the individual regression coefficients. Such tests would be useful in determining the value of each of the regressor variables in the regression model. For example, the model might be more effective with the inclusion of additional variables, or perhaps with the deletion of one or more of the variables already in the model.

Adding a variable to the regression model always causes the sum of squares for regression to increase and the error sum of squares to decrease. We must decide whether the increase in the regression sum of squares is sufficient to warrant using the additional variable in the model. Furthermore, adding an unimportant variable to the model can actually increase the mean square error, thereby decreasing the usefulness of the model.

2.3.2 Residual Analysis

The residuals from the least squares fit, defined by $e_i = y_i - \hat{y}_i$, $i = 1, 2, \ldots, n$, play an important role in judging model adequacy. Many response surface analysts prefer to work with **scaled residuals**, in contrast to the ordinary least squares residuals. These scaled residuals often convey more information than do the ordinary residuals.

The standardizing process scales the residuals by dividing them by their average standard deviation. In some data sets, residuals may have standard deviations that differ greatly. There is some other way of scaling that takes this into account. Let's consider this.

The vector of fitted values $\hat{y}_i$ corresponding to the observed values y_i is

$$\mathbf{\hat{y} = Xb = X(X'X)^{-1}X'y = Hy} \qquad (2.16)$$

The *n x n* matrix $\mathbf{H=X(X'X)^{-1}X'}$ is usually called the **hat** matrix because it maps the vector of observed values into a vector of fitted values. The hat matrix and its properties play a central role in regression analysis.
Since $\mathbf{e = y - \hat{y}}$ there are several other useful ways to express the vector of residuals

$$\mathbf{e = y - Xb = y - Hy = (I - H)y} \tag{2.17}$$

The **prediction error sum of squares (PRESS)** proposed in [6, 7], provides a useful residual scaling

$$PRESS = \sum_{i=1}^{n} \left(\frac{e_i}{1-h_{ii}}\right)^2 \tag{2.18}$$

From Equation (2.18) it is easy to see that the PRESS residual is just the ordinary residual weighted according to the diagonal elements of the hat matrix h_{ii}. Generally, a large difference between the ordinary residual and the PRESS residual will indicate a point where the model fits the data well, but a model built without that point predicts poorly.

3. Variable Selection and Model Building in Regression

In response surface analysis it is customary to fit the **full model** corresponding to the situation at hand. It means that in steepest ascent we usually fit the full first-order model, and in the analysis of the second-order model we usually fit the full quadratic.

Nevertheless, an experimenter may encounter situations where the full model may not be appropriate; that is, a model based on a subset of the regressors in the full model may be superior. **Variable selection** or **model-building** techniques usually is used to identify the best subset of regressors to include in a regression model [8,9]. Now we give a brief presentation of regression model-building and variable selection methods, introduce our method of variable selection and illustrate their application to a response surface problem. We assume that there are *K* candidate regressors denoted $x_1, x_2, ..., x_k$ and a single response variable *y*. All models will have an intercept term β_0, so that the full model has *K + 1* parameters.

It is shown in [8,9] that there is a strong motivation for correctly specifying the regression model: Leaving out important regressors introduces bias into the parameter estimates, while including unimportant variables weakens the prediction or estimation capability of the model.

3.1 Procedures for Variable Selection

Now we will consider several of the more widely used methods for selecting the appropriate subset of variables for a regression model. We will also introduce our approach based on the optimization procedure used for selecting the best model from the whole set of models and

finally we will discuss and illustrate several of the criteria that are typically used to decide which subset of the candidate regressors leads to the best model.

3.1.1 All Possible Regression

This procedure requires that the analyst fit all the regression equations involving one-candidate regressors, two-candidate regressors, and so on. These equations are evaluated according to some suitable criterion, and the best regression model selected. If we assume that the intercept term β_0 is included in all equations, then there are *K* candidate regressors and there are 2^K total equations to be estimated and examined. For example, if *K = 4*, then there are $2^4 = 16$ possible equations, whereas if *K = 10*, then there are $2^{10} = 1024$. Clearly the number of equations to be examined increases rapidly as the number of candidate regressors increases.

Usually the analysts restrict the candidate variables for the model to those in the full quadratic polynomial and require that all models obey the principal of **hierarchy.** A model is said to be hierarchical if the presence of higher-order terms (such as interaction and second-order terms) requires the inclusion of all lower-order terms contained within those of higher order. For example, this would require the inclusion of both main effects if a two-factor interaction term was in the model. Many regression model builders believe that hierarchy is a reasonable model-building practice when fitting polynomials.

3.1.2 Stepwise Regression Methods

Because evaluating all possible regressions can be burdensome computationally, various methods have been developed for evaluating only a small number of subset regression models by either adding or deleting regressors one at a time. These methods are generally referred to as **stepwise-type procedures**. They can be classified into three broad categories: (1) forward selection, (2) backward elimination, and (3) stepwise regression, which is a popular combination of procedures (1) and (2).

Forward Selection

This procedure begins with the assumption that there are no regressors in the model other than the intercept. An effort is made to find an optimal subset by inserting regressors into the model one at a time. The first regressor selected for entry into the equation is the one that has the largest simple correlation with the response variable *y*. Suppose that this regressor is x_1. This also the regressor that will produce the largest value of the *F*-statistic for testing significance of regression. This regressor is entered if the *F*-statistic exceeds a preselected *F*-value, say F_{in} (or *F*-to-enter). The second regressor chosen for entry is the one that now has the largest correlation with *y* after adjusting for the effect of the first regressor entered (x_1) on *y*. We refer to these correlations as partial correlations. They are the simple correlations between the residuals from the regression $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$ and the residuals from the regressions of the other candidate regressors on x_1, say $\hat{x}_j = \hat{a}_{0j} + \hat{a}_{1j} x_1$, *j = 2, 3, …, K.*

In general, at each step the regressor having the highest partial correlation with *y* (or equivalently the largest partial *F*-statistic given the other regressors already in the model) is added to the model if its partial *F*-statistic exceeds the preselected entry level F_{in}. The procedure terminates either when the partial *F*-statistic at a particular step does not exceed F_{in} or when the last candidate regressor is added to the model.

Backward Elimination

Forward selection begins with no regressors in the model and attempts to insert variables until a suitable model is obtained. Backward elimination attempts to find a good model by working in the opposite direction. That is, we begin with a model that includes all *K* candidate regressors. Then the partial *F*-statistic (or a *t*-statistic, which is equivalent) is computed for each regressor as if it were the last variable to enter the model. The smallest of these partial *F*-statistics is compared with a preselected value, F_{out} (or *F*-to-remove); and if the smallest partial *F*-value is less than F_{out}, that regressor is removed from the model. Now a regression model with $K - 1$ regressors is fitted, the partial *F*-statistic for this new model calculated, and the procedure repeated. The backward elimination algorithm terminates when the smallest partial *F*-value is not less than the preselected cutoff value F_{out}.

Backward elimination is often a very good variable selection procedure. It is particularly favored by analysts who like to see the effect of including all the candidate regressors, just so that nothing obvious will be missed.

Stepwise Regression

The two procedures described above suggest a number of possible combinations. One of the most popular is the stepwise regression algorithm. This is a modification of forward selection in which at each step all regressors entered into the model previously are reassessed via their partial *F*-or *t*-statistics. A regressor added at an earlier step may now be redundant because of the relationship between it and regressors now in the equation. If the partial *F*-statistic for a variable is less than F_{out}, that variable is dropped from the model.

Stepwise regression requires two cutoff values, F_{in} and F_{out}. Some analysts prefer to choose $F_{in} = F_{out}$, although this is not necessary. Sometimes we choose $F_{in} > F_{out}$, making it more difficult to add a regressor than to delete one.

3. 2 General Comments on Stepwise-Type Procedures

The stepwise regression algorithms described above have been criticized on various grounds, the most common being that none of the procedures generally guarantees that the best subset regression model of any size will be identified. Furthermore, because all the stepwise-type procedures terminate with one final equation, inexperienced analysts may conclude that they

have found a model that is in some sense optimal. Part of the problem is that it is likely that there is not one best subset model, but several equally good ones.

The analyst should also keep in mind that the order in which the regressors enter or leave the model does not necessarily imply an order of importance to the variables. It is not unusual to find that a regressor inserted into the model early in the procedure becomes negligible at a subsequent step. For example, suppose that forward selection chooses x_4 (say) as the first regressor to enter. However, when x_2 (say) is added at a subsequent step, x_4 is no longer required because of high positive correlation between x_2 and x_4. This is a general problem with the forward selection procedure. Once a regressor has been added, it cannot be removes at a later step.

Note that forward selection, backward elimination, and stepwise regression do not necessarily lead to the same choice of final model. The correlation between the regressors affects the order of entry and removal. Some users have recommended that all the procedures be applied in the hopes of either seeing some agreement or learning something about the structure of the data that might be overlooked by using only one selection procedure. Furthermore, there is not necessarily any agreement between any of the stepwise-type procedures and all possible regressions.

For these reasons, stepwise-type variable selection procedures should be used with caution. Some analysts prefer the stepwise regression algorithm followed by backward elimination. The backward elimination algorithm is often less adversely affected by the correlative structure of the regressors than is forward selection. They also found it helpful to run the problem several times with different choices of F_{in} and F_{out}. This will often allow the analyst to generate several different models for evaluation.

3.3 Our Approach: Using Optimization Procedure for Variable Selection

All previous approaches are burdensome computationally and none of the procedures generally guarantees that the best subset regression model of any size will be identified. This fact determined our decision to use the optimization procedure for variable selection in a model. We use **simulated annealing** (SA) method to search for the optimal model.

The rough idea of simulated annealing is that it first picks a random move. If the move improves the objective function, then the algorithm accepts the move. Otherwise, the algorithm makes the move with some probability less than one:

$$p = exp\ [-(E2-E1) / kT] \qquad (3.1)$$

The probability decreases exponentially with the "badness" of the move - the amount

(E2 – E1) by which the evaluation is worsened.

A second parameter T is also used to determine the probability. At higher values of T, "bad" moves are more likely to be allowed. As T gets closer to zero, they become more and more unlikely, until the algorithm behaves more or less like a local search. The schedule input determines the value of T as a function of how many cycles already have been completed [10].

Simulated Annealing was first used extensively to solve VLSI layout problems in the early 1980s [11]. Since that, it has been used in Operations Research to successfully solve a large number of optimization problems such as the Traveling Salesman problem and various scheduling problems [12]. We also successfully used simulated annealing method in our work for improving organizational design [13]. In fact, simulated annealing can be used as a global optimizer for difficult functions. Due to the similar approach with random steps in the search process, we came up with the idea to use SA for the selection of variables for search of the optimal model in realistically complex modeling situations.

An initial *K* variable subset of a full set of *P* (*P* might be as large as 300 – 500) variables is randomly selected and passed on to a Simulated Annealing algorithm. The algorithm then selects a random subset in the neighborhood of the current state (neighborhood of a subset S being defined as the family of all *K*-variable subsets which differ from S by a single regressor), and decides whether to replace the current subset according to the Simulated Annealing rule, i.e., either (i) always, if the alternative subset's value of the criterion is higher; or (ii) with probability defined by Equation (2.19) if the alternative subset's value of the optimization criterion is lower than that of the current solution, where the parameter *T* (temperature) decreases throughout the iterations of the algorithm. We suggest that for each cardinality *K*, the stopping criterion for the algorithm is the number of iterations which is controlled by the user. Also controlled by the user are the initial temperature and the rate of geometric cooling of the temperature.

The user has also the option to specify his initial model to compute regression analysis and to compare the resulting statistics with the optimized results.

3.4 Variable Selection: Results

We used the Example A.1.1 from [1] to compare our results of variable selection procedure with results presented in [1]. Table 2.1 presents the results of running a rotatable central composite design on a process used to make a polymer additive for motor oil. The response variable of interest is the average molecular weight (M_n), and the two process variables are reaction time in minutes and catalyst addition rate. The table shows the design in terms of both the natural variables and the usual coded variables. In this case full quadratic model was selected because the *F*-statistic for the quadratic terms (over the contribution of the linear terms) was large and because the linear model displayed strong lack of fit. There is also some indication here that a subset model might be more appropriate than the full quadratic.

The authors of [1] used the all-possible-regressions procedure to identify a model. They restricted the candidate variables for the model to those in the full quadratic polynomial and required that all models obey the principal of hierarchy. As we already said, a model is hierarchical if the presence of higher-order terms (such as interaction and second-order terms) requires the inclusion of all lower-order terms contained with those of higher order.

Table 2: Factors and Response for Example A.1.1

Time ξ_1	Catalyst ξ_2	Time x_1	Catalyst x_2	$y = M_n$
30	4	-5	-1	2320
40	4	5	-1	2925
30	6	-5	1	2340
40	6	5	1	2000
27.93	5	-7.07	0	3180
42.07	5	7.07	0	2925
35	3.586	0	-1.414	1930
35	6.414	0	1.414	1860
35	5	0	0	2980
35	5	0	0	3075
35	5	0	0	2790
35	5	0	0	2850
35	5	0	0	2910

Table 2.3 presents the all-possible regressions results. The values of the residual sum of squares, the residual mean square, R^2, R^2_{adj}, and PRESS are given for each model. Table 2.3 also shows the value of the C_p **statistic** for each subset model. The C_p statistic is a measure of the total mean squared error for the *p*-term regression model

$$C_p = \frac{SS_E(p)}{\hat{\sigma}^2} - n + 2p \qquad (3.2)$$

where $SS_E(p)$ is the error sum of squares for the p-term model and

$\hat{\sigma}^2 = MS_E$(full model)

If the *p*-term model has negligible bias, then it can be shown that

$E(C_p \mid zero\ bias) = p$

Therefore, the values of C_p for each regression model under consideration should be evaluated relative to *p*. The regression equations that have substantial bias will have values of

C_p that are greater than *p*. Then we can choose as the "best" regression equation either a model with minimum C_p or a model with a slightly larger C_p that does not contain as much bias (i.e., $C_p \cong p$) as the minimum.

Table 3: All Possible Regressions Results for Example A.1.1

Terms in Model	SS Residual	MS Residual	R^2	R^2_{adj}	PRESS	Cp
x_1	**2,607,485.0**	**237,044.1**	**0.0004**	**-0.0904**	**3,728,849.2**	**86.26**
x_2	**2,482,610.9**	**225,691.9**	**0.0483**	**-0.0382**	**4,240,155.7**	**81.70**
x_1, x_2	**2,481,469.0**	**248,146.9**	**0.0487**	**-0.1415**	**4,963,701.9**	**83.66**
x_1, x_2, x_1x_2	**2,258,212.8**	**250,912.5**	**0.1343**	**-0.1542**	**5,379,306.7**	**77.50**
x_1, x_2, x_1^2	**2,386,620.7**	**265,180.1**	**0.0851**	**-0.2199**	**5,736,756.9**	**82.19**
x_1, x_2, x_1x_2, x_1^2	**2,163,364.5**	**270,420.6**	**0.1707**	**-0.2440**	**6,580,579.4**	**76.04**
x_1, x_2, x_2^2	**429,842.7**	**47,760.3**	**0.8352**	**0.7803**	**1,123,679.6**	**10.70**
x_1, x_2, x_1x_2, x_2^2	**206,586.5**	**25,823.3**	**0.9208**	**0.8812**	**853,249.0**	**4.55**
$x_1, x_2, x_1x_2, x_1^2, x_2^2$	**191,604.2**	**27,372.0**	**0.9265**	**0.8741**	**1,087,751.3**	**6.00**
x_1, x_1^2	**2,512,636.7**	**251,263.7**	**0.0368**	**-0.1558**	**4,399,694.1**	**84.80**
x_2, x_2^2	**430,984.6**	**43,098.5**	**0.8343**	**0.8017**	**864,737.5**	**8.75**
$x_2, x_1x_2, x_1^2, x_2^2$	**192,750.0**	**24,093.3**	**0.9261**	**0.8892**	**734,700.0**	**4.04**
x_2, x_1, x_2, x_2^2	**184,650.0**	**23,080.9**	**0.9292**	**0.8938**	**210,987.0**	**2.59**

The Table 2.3 contains 13 models. The first 11 models were considered in [1]. All of them follow the hierarchical rule. The two last models were obtained as a result of the optimization process that we applied to variable selection. The selection of the appropriate subset model is usually based on the summary statistics given in Table 2.3. Note that among the first 12 models considered in [1], the subset model containing the terms x_1, x_2, x_1x_2, x_2^2 has the smallest residual mean square, the largest adjusted R^2, the smallest value of PRESS, and $C_p = C_5 = 4.55$, which is just less than $p = 5$, so this equation contains little bias. Nevertheless, the two last models that we found have definitely much better statistics. If regression model builder had followed hierarchical principal, he/she never would have found them. As you can see the full model is not

the best model in this particular case and the simulated annealing method allows us to find better models that were not originally considered by model builders. Since all considered statistics usually change accordingly we had chosen residual sum of squares (2.11) as an optimization criterion. The user can also choose any other statistic as a criterion and the output of all statistics is shown in Table 2.3. All statistics for all models were also double checked by MATLAB and completely coincide with results given by the optimization process.

4. A Simulation Framework for Response Surface Methodology

4.1 Response Surface Methodology as an Automated Tool for Model Validation

Modern computational large-scale social-networks simulation systems are becoming heavily used due to their efficiency and flexibility in modeling and simulation of complex social networks. Since these models are increasing in complexity and size they require more significant time and efforts for their validation, optimization, improvement and understanding of their behavior.

One example of such multi-agent social-network model named BioWar is presented in [14]. BioWar is a spatial city-scale multi-agent social-network model capable of simulating the effects of biological weapon attacks against the background of naturally-occurring diseases on a demographically realistic population. Response Surface Methodology (RSM) might be extremely useful as an automated tool that can be used to calibrate such multi-agent models as BioWar, and facilitate validation and understanding, thereby increasing model fidelity and reliability and giving the user some feedback for analysis and insight.

Every simulation model has its own specific that should be taken into consideration when we try to specify independent and dependent variables. For example, within ORA the RSM tool can use one group of measures like a person's height, age, weight as independent variables and another like the amount of money earned as the dependent variable. We can also consider some of DNA measures like Resource Congruence or Task Exclusivity as independent variables while some other DNA measure like Network Closeness or Betweenness Centralization can serve as dependent variable. If the RSM is operating on the node level measures then independent variables can be any node level measure. There is a particular interest also in consideration of some means or standard deviations of the graph level measures as dependent or independent variables.

The RSM is also useful for searching the input combination that maximizes the output of a real system or its simulation such as BioWar or ORA. When validating or optimizing a stochastic simulation model, one tries to estimate the model parameters that optimize specific stochastic output of the simulation model. A simulation framework is especially intended for simulation models where the calculation of the corresponding stochastic objective function is very expensive or time-consuming. RSM is frequently used for the optimization of stochastic simulation models [15]. This methodology is based on approximation of the stochastic objective function by a low order polynomial on a small sub region of the domain. The coefficients of the polynomial are estimated by ordinary least squares method applied to a number of observations of the stochastic objective function. To this end, the objective function is evaluated in an arrangement of points referred to as an experimental design [1, 15]. Based on the fitted

polynomial, the local best point is derived, which is used as a current estimator of the optimum and as a center point of a new region of experimentation, where again the stochastic objective function is approximated by a low order polynomial. In non-automated optimization, RSM is an interactive process in which the user gradually gains understanding of the nature of the stochastic objective function. In an automated RSM algorithm, however, human intervention during the optimization process is excluded. A good automated RSM algorithm should therefore include some degree of self-correction mechanisms [16].

4.2 Steps of Response Surface Methodology in Automated Validation Process

Usually automated RSM validation process includes the following steps:

1. Approximate the simulation response function in the current region of interest by a first-order model. The first-order model is described by Equation (1.7). Estimators of the regression coefficients (the β's) are determined by using ordinary least squares. To this end, the objective function is evaluated in the points of an experimental design, which is a specific arrangement of points in the current region of interest. Although there are many designs to choose from, usually a fractional two-level factorial design [14] is used, often augmented by the center point of the current region of experimentation [1]. The advantages of this design are that it is orthogonal, what means that the variance of the predicted response is minimal, gives unbiased estimators of the regression coefficients and can quite easily be augmented to derive a second-order design.

2. Test the first-order model for adequacy. Before using the first-order model to move into a direction of improved response, it should be tested if the estimated first-order model adequately describes the behavior of response in the current region of experimentation. It is necessary to remember that the total number of observations should be always larger than the number of regression coefficients. Moreover, multiple observations are needed in the center point of the region of experimentation. Estimation of adequacy is usually performed using the analysis of variance (ANOVA) table. It allows decide when to accept the first-order model. The decisions include choosing the significance levels for the test involved.

3. Perform a line search in the steepest descent direction. If the first-order model is accepted, then this model is used for determining the direction where improvement of the simulation response is expected. The steepest descent direction is given by $(-b_1,...,-b_k)$. A line search is performed from the center point of the current region of experimentation in this direction to find a point of improved response. This point is taken as the estimator of the optimum of the simulation response function in the *n*th iteration, and is used as the center point of the region of the experimentation in the *(n + 1)*th iteration. The line search is stopped when no further improvement is observed.

4. Solve the inadequacy of the first-order model. If the first-order model is not accepted, then either there is some evidence of pure curvature or interaction between the factors in the current region of experimentation. Usually, this is solved by approximating the simulation response by a second-order polynomial. However, the optimization algorithm becomes less efficient especially if it occurs very early during the

optimization process. There is alternative solution that recommends reduce the size of the region of experimentation by decreasing the step sizes. In this way this region can possibly become small enough to ensure that a first-order approximation is an adequate local representation of the simulation response function. Another solution is to increase the simulation size used to evaluate a design point or to increase the number of replicated observations done in the design points. This may ensure that a significant direction of steepest descent is indeed found. At the start of the algorithm it should be decided which actions will be taken when the first-order model is rejected. For example, depending on the p-value found for the lack-of-fit test, one could decide to apply a second-order approximation or to decrease the size of the region of experimentation.

5. Approximate the objective function in the current region of experimentation by a second-order model. The coded second-order model is given by the Equation (1.8). The regression coefficients of the second-order model are again determined by using ordinary least squares method applied to observations performed in an experimental design. The most popular class of second-order designs is the central composite design (CCD) [1]. This design can be easily constructed by augmenting the fractional factorial design that is used for estimating the first-order model.

6. Testing the second-order model for adequacy. Similar to the first-order model, it should be tested if the estimated second-order model adequately describes the behavior of the response in the current region of experimentation before using the model.

7. Solve the inadequacy of the second-order model. If the second-order model is found not to be adequate, then one should reduce the size of the region of experimentation or increase the simulation size used in evaluating a design point. In RSM it is not customary to fit a higher than second-order polynomial.

8. Perform canonical analysis. If the second-order model is found to be adequate, then canonical analysis is performed to determine the location and the nature of the stationary point of the second-order model. The estimated second-order approximation can be written as follows:

$$\hat{\mathbf{Y}} = b_0 + \mathbf{x'b} + \mathbf{x'Bx} \qquad (4.1)$$

where b_0, **b**, and **B** are the estimates of the intercept, linear, and second-order coefficients, respectively.

The stationary point $\mathbf{x}_s$ of the second-order polynomial is determined by

$$\mathbf{x}_s = -\frac{1}{2}\mathbf{B}^{-1}\mathbf{b} \qquad (4.2)$$

If all eigenvalues of **B** are positive(negative), then the quadratic surface has a minimum(maximum) at the stationary point $\mathbf{x}_s$. If the eigenvalues have mixed signs, then the stationary point $\mathbf{x}_s$ is a saddle point.

9. Perform ridge analysis. It is not advisable to extrapolate the second-order polynomial beyond the current region of experimentation, because the fitted model is not reliable outside the experimental region [1]. Therefore, if the stationary point is a minimum that lies outside the current region of experimentation, it is not accepted as the center of the next region of experimentation. In the case if the stationary point is a maximum or a saddle point, then the stationary point is rejected as well. In these cases, ridge analysis is performed, which means a search for a new stationary point $\mathbf{x}_s$ on a given radius R such that the second order model has a minimum at this stationary point [1]. Using Lagrange analysis, the stationary point is given by

$$(\mathbf{B} - \mu\mathbf{I})\mathbf{x} = (-\tfrac{1}{2})\,\mathbf{b} \qquad (4.3)$$

As a result, for a fixed μ, a solution **x** of Equation (4.3) is a stationary point on $R = (\mathbf{x'x})^{1/2}$. In the working regions of μ, namely $\mu > \lambda_k$ or $\mu < \lambda_1$, where λ_1 is the smallest eigenvalue of **B** and λ_k is the smallest eigenvalue of **B,** R is a monotonic function of μ. As a result, a computer algorithm for ridge analysis involves the substitution of $\mu > \lambda_k$ (for a design maximum response) and increases μ until radii near the design perimeter are encountered. Future increases in μ results in coordinates that are closer to the design center. The same applies for $\mu < \lambda_1$ (for desired minimum response), with decreasing values of μ being required.

10. Accept the stationary point. The stationary point will be used as the center point of the next experimental region. The analyst should decide whether the first-order or a second-order model is used to approximate the simulation response surface in this region. This decision can be based on the results of the canonical analysis. For example, if a minimum was found, it could be useful to explore a region around this minimum with a new second-order approximation. Opposite, if a maximum or a saddle point was found, the optimum could still be located far away from the current experimental region. In this case, approximating this region with a first-order model and consequently performing a line search would be preferable. Allowing this we return to the first phase of our algorithm. It is considered to be a powerful self-correction mechanism.

11. An enhanced algorithm is introduced in [17]. In this algorithm the authors use the gradient of the second-order model in the center point of the current region and the results of the canonical analysis to determine the direction of steepest descent. Next, the perform a line search using this direction, resulting in a new center of an experimental region. In this region they already approximate the simulation response surface by a first-order model.

12. Stopping criterion. Usually it is recommended ending the optimization process if the estimated optimal simulation response value does not improve sufficiently anymore or if the experimental region becomes too small. In the case of restricted budget we can stop if a fixed number of evaluations has been performed. Next, a confidence interval on the response at the estimator for the optimum can be determined.

Very often the natural sequential deployment of RSM allows the user to make intelligent choices of variable ranges. What is often even more important is for the response surface analysis to reveal important information about the nature of the simulation model and the roles of the variables. The computation of a stationary point, a canonical analysis, or a ridge analysis may lead to important information about the simulated process, and in the long run it might be very valuable. Using of RSM as an automated validation tool also helps to make the validation process of the simulation model less time and resource consuming and less prone to bias. Using the RSM equivalent rather then the multi-agent simulation model on some stages of simulation process also eliminates the wait time for generating results under a variety of conditions and to address a number of policy issues. This approach is fairly common in electrical engineering when the detailed simulation model is originally used for design and then the RSM analog used for its validation and then on a daily basis as a fast approximation of the simulation model. All these considerations forced us to decide that the RSM validation mechanism will be a valuable tool that needs be integrated with our existing simulation tools such as the DyNet, BioWar, Construct, Vista, and ORA.

5. Multicollinearity and Biased Estimation in Regression

5.1 Definition of Multicollinearity

In some response surface experiments, there can be one or more near-linear dependences among the regressor variables in the model. Regression model builders refer to this as **multicollinearity** among the regressors. Multicollinearity can have serious effects on the estimates of the model parameters and on the general applicability of the final model. In this chapter, we give a brief introduction to the multicollinearity problem along with biased estimation, one of the parameter estimation techniques useful in dealing with multicollinearity.

The effects of multicollinearity may be easily demonstrated. Consider a regression model with two regressor variables x_1 and x_2, and suppose that x_1 and x_2 have been standardized by subtracting the average of that variable from each observation and dividing by the square root of the corrected sum of squares. This **unit length scaling**, as it is called, results in the matrix **X'X** having the form of a correlation matrix; that is, the main diagonals are 1 and the off-diagonals are the simple correlation between regressor x_i and regressor x_j. The model is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, \qquad i = 1,2,...,n \tag{4.1}$$

Now if the response is also centered, then the estimate of β_0 is zero. The $(X'X)^{-1}$ matrix for this model is

$$C = (X'X)^{-1} = \begin{bmatrix} 1/(1-r_{12}^2) & -r_{12}/(1-r_{12}^2) \\ -r_{12}/(1-r_{12}^2) & 1/(1-r_{12}^2) \end{bmatrix} \tag{4.2}$$

where r_{12} is the simple correlation between x_1 and x_2.

Now, if multicollinearity is present, x_1 and x_2 are highly correlated, and $|r_{12}| \to 1$. In such a situation, the variances and covariances of the regression coefficients become very large. The large variances for b_j imply that the regression coefficients are very poorly estimated. Note the effect of multicollinearity is to introduce a near-linear dependence in the columns of **X.** As $r_{12} \to 1$ or -1, this linearity becomes exact.

Similar problems occur when multicollinearity is present and there are more than two regressor variables. In general, the diagonal elements of the matrix $C = (X'X)^{-1}$ can be written as

$$C_{jj} = \frac{1}{(1 - R_j^2)}, \qquad j = 1, 2, \ldots, k \tag{4.3}$$

where R_j^2 is the coefficient of multiple determination resulting from regressing x_j on the other $k - 1$ regressor variables. Clearly, the stronger the linear dependence of x_j on the remaining regressor variables (and hence the stronger the multicollinearity), the larger the value of R_j^2 will be. It is usually said that the variance of b_j is inflated by the quantity $(1 - R_j^2)^{-1}$. Consequently, (4.3) is usually called the variance inflation factor for b_j. Note that these factors are the main diagonal elements of the inverse of the correlation matrix. They are an important measure of the extent to which multicollinearity is present.

Although the estimates of the regression coefficients are very imprecise when multicollinearity is present, the fitted model may still be useful. For example, suppose we wish to predict new observations. If these predictions are required in the region of the *x*-space where the multicollinearity is in effect, then often satisfactory results will be obtained, because while individual b_j may be poorly estimated, the function $\sum_{j=1}^{k} \beta_j x_{ij}$ may be estimated quite well. On the other hand, if the prediction of new observations requires extrapolation, then generally we would expect to obtain poor results. Successful extrapolation usually requires good estimates of the individual model parameters.

Multicollinearity arises for several reasons. It will occur when the analyst collects the data such that a constraint of the form $\sum_{j=1}^{k} a_j x_j = 0$ holds among the columns of **X** (the a_j are constants, not all zero). For example, if four regressor variables are the components of a mixture, then such a constraint will always exist because the sum of the component proportions is always constant. Usually, however, these constraints do not hold exactly, and the analyst does not know that they exist.

5.2 Detection of Multicollinearity

There are several ways to detect the presence of multicollinearity. We will briefly discuss some of the more important of these.

1. The variance inflation factors, defined in (4.3), are very useful measures of multicollinearity. The larger the variance inflation factor, the more severe the multicollinearity. Some authors have suggested that if any variance inflation factors exceed 10, then multicollinearity is a problem. Other authors consider this value too liberal and suggest that the variance inflation factors should not exceed 4 or 5.

2. The determinant of **X'X** in correlation form may also be used as a measure of multicollinearity. The value of this determinant can range between 0 and 1. When the value of the determinant is 1, the columns of **X** are orthogonal (i.e., there is no intercorrelation between the regressor variables), and when the value is 0, there is an exact linear dependence among the columns of **X**. The smaller the value of the determinant, the greater the degree of multicollinearity.

3. The eigenvalues, or characteristic roots, of **X'X** in correlation form provide a measure of multicollinearity. The eigenvalues of **X'X** are the roots of the equation

$$|\mathbf{X'X} - \boldsymbol{\lambda}\mathbf{I}| = 0 \qquad (4.4)$$

One or more eigenvalues near zero implies that multicollinearity is present. If $\lambda_{\max}$ and $\lambda_{\min}$ denote the largest and smallest eigenvalues of **X'X,** then the **condition number** $k = \lambda_{\max} / \lambda_{\min}$ is less than 100, there is little problem with multicollinearity.

4. Sometimes inspection of the individual elements of the correlation matrix can be helpful in detecting multicollinearity. If an element $|r_{ij}|$ is close to one, then x_i and x_j may be strongly multicollinear. However, when more than two regressor variables are involved in a multicollinear fashion, the individual r_{ij} are not necessarily large. Thus, this method will not always enable us to detect the presence of multicollinearity.

5. If the *F*-test for significance of regression is significant, but test on the individual regression coefficients are not significant, then multicollinearity may be present.

Since method 4 does not always allow us to detect multicollinearity and method 5 is more complicated in implementation, we implemented 3 first methods in our code. Methods 1 and 3 require the user specified parameters that might be subjective. Our experience shows that combination of first 3 methods works very well in detection of multicollinearity.

5.3 Multicollinearity Remedial Measures

Several remedial measures have been proposed for resolving the problem of multicollinearity. Augmenting the data with new observations specifically designed to break up

the approximate linear dependences that currently exist is often suggested. However, sometimes this is impossible for economic reasons, or because of the physical constraints that relate the x_j.

Another possibility is to delete certain terms from the model. This suffers from the disadvantage of discarding the information contained in the deleted terms.

Because multicollinearity primarily affects the stability of the regression coefficients, it would seem that estimating these parameters by some method that is less sensitive to multicollinearity than ordinary least squares would be helpful. **Ridge regression** is one of several methods that have been suggested for this. In ridge regression, the regression coefficients estimates are obtained by solving

$$\mathbf{b}^*(\theta) = \mathbf{(X'X + \theta I)}^{-1}\mathbf{X'y} \qquad (4.5)$$

where $\theta \geq 0$ is a constant. Generally, values of θ in the interval $0 \leq \theta \leq 1$ are appropriate. The ridge estimator $\mathbf{b}^*(\theta)$ is not an unbiased estimator of **β**, as is the ordinary least squares estimator **b,** but the mean square error of $\mathbf{b}^*(\theta)$ will be smaller than that of **b.** Thus ridge regression seeks to find a set of regression coefficients that is more **stable** in the sense of having a small mean square error. Because multicollinearity usually results in ordinary least squares estimators that may have extremely large variances, ridge regression is suitable for situations where the multicollinearity problem exists.

To obtain the ridge regression estimator from Equation (4.5), the user must specify a value for the constant θ. Generally, there is an optimum θ for any problem. In general, the variance of $\mathbf{b}^*(\theta)$ is a decreasing function of θ, while the squared bias $[\mathbf{b} - \mathbf{b}^*(\theta)]^2$ is an increasing function of θ. Choosing the value of θ involves trading off these two properties of $\mathbf{b}^*(\theta)$**.** A good discussion of the practical aspects of ridge regression may be found in [18].

Multicollinearity is usually not a big problem in well-designed and well-executed response surface experiment. However, a poorly designed or poorly executed response surface experiment can have substantial problems with multicollinearity. For example, mixture experiments may often have substantial multicollinearity. A mixture problem is a special type of response surface problem when the response variables of interest in the problem are a function of the proportions of the different ingredients used in its formulation. While we traditionally think of mixture problems in the product design or formulation environment, they also occur in many other settings. In addition to use of ridge regression as a model-building method for mixture problems, they also require special experimental design techniques [1].

6. Limitations and Future Extensions

The main limitation of this work is that we operated with fairly small models. We plan to test our approach on large models with 100 and more variables. We also plan to include models with different terms besides polynomial like logarithmical, exponential, etc.

The RSM can be easily integrated in many large-scale simulation systems such as BioWar, ORA and is currently integrating with Vista, Construct, and DyNet. Some research has been done to provide the integration of the RSM with BioWar.

We also started the implementation of the interface for the response surface analyzer that will allow the user to input his/her own parameters and see the results of analyzer in more convenient format. The current version of the analyzer already allows the user to work in two different modes: to run the response surface analyzer or run the linear regression on his/her own model. In the future the interface will allow the user make a comparison of the two models: user specified model and built by the response surface analyzer.

7. System Requirements

The response surface analyzer is written in C++ and currently runs on Windows XP using an Intel processor. The interface of the response surface analyzer will be implemented in Java.

References

[1] Myers Raymond H. & D.C. Montgomery, 2002. “Response Surface Methodology: process and product optimization using designed experiment,.” A Wiley-Interscience Publication.

[2] Taguchi, G., 1986. “Introduction to Quality Engineering,” Asian Productivity Organization, UNIPUB, White Plains, NY.

[3] Taguchi, G., 1987. “System of Experimental Design: Engineering Methods to Optimize Quality and Minimize Cost,” UNIPUB/Kraus International, White Plains, NY.

[4] Box, G. E. P. and N.R. Draper, 1987. “Empirical Model-Building and Response Surfaces,” Jon Wiley & Sons, New York.

[5] Khuri, A.I. and J.A. Cornell, 1996. “Response Surfaces,” 2nd edition, Marcel Dekker. New York.

[6] Allen, D.M., 1971, “Mean Square Error of Prediction as a Criterion for Selecting Variables,” Technometrics, 13, 469-475.

[7] Allen, D.M., 1974, “The Relationship Between Variable Selection and Data Augmentation and a Method for Prediction,” Technometrics, 16, 125-127.

[8] Montgomery, D.C., E.A. Peck, and & G.G. Vining, 2001, “Introduction to Linear Regression Analysis,” 3rd edition, John Willey & Sons, New York.

[9] Myers R.H., 1990, “Classical and Modern Regression with Applications,” 2nd edition, Duxbury Press, Boston.

[10] Russel, S. J. & P. Norvig, 1995, “Artificial Intelligence: A Modern Approach”, Prentice Hall.

[11] Kirkpatrick, S. and C.D. Gelatt and M.P. Vecchi, 1983,”Optimization by Simulated Annealing,” Science, vol. 220, Number 4598, pp. 671-680.

[12] Cerny V., 1985, “A Thermodynamical Approach to the Traveling Salesman Problem: An Efficient Simulation Algorithm”, Journal of Optimization Theory and Applications, Number 45, pages 41-51.

[13] Carley Kathleen M. & Natalia Y. Kamneva, 2004, “A Network Optimization Approach for Improving Organizational Design”, Carnegie Mellon University, School of Computer Science, Institute for Software Research International, Technical Report CMU-ISRI-04-102.

[14] Carley Kathleen M., D. Fridsma, E. Casman, N. Altman, J. Chang, B. Kaminsky, D. Nave, and A. Yahia, 2003, “BioWar: Scalable Multi-Agent Social and Epidemiological Simulation of Bioterrorism Events”, *2003 NAACSOS conference proceedings,* Pittsburgh, PA

[15] Kleijnen, J. P. C., 1998. “Experimental Design for Sensitivity Analysis, Optimization, and Validation of Simulation Models.” In *Handbook of simulation: principles, methodology, advances, applications and practice*, ed. J. Banks, 173–223. New York: John Wiley & Sons.

[16] Neddermeijer, H.G., van Oortmarssen G.J., Piersma N., Dekker R., 2000. “A Framework for Response Surface Methodology for Simulation Optimization Models.” *Proceedings of the 2000 Winter Simulation Conference* (edited by J.A. Joines, R.R. Barton, K Kang, and P.A. Fishwick), pp. 129-136.

[17] Joshi., S., H.D. Sherali and J.D. Tew, 1998, “An Enhanced Response Surface Methodology (RSM) Algorithm Using Gradient Deflection and Second-Order Search Strategies”, Computers and Operations Research, 25 (7/8), pp. 531 - 541.

[18] Marquardt, D.M., and R.D. Snee, 1975, “Ridge Regression in Practice”, The American Statistician, 29, pp. 3 – 20.