THE RESPONSE SURFACE METHODOLOGY

Nuran Bradley
Department of Mathematical Sciences
Indiana University of South Bend
E-mail Address: nbradley@iusb.edu

Submitted to the faculty of the
Indiana University South Bend
in partial fulfillment of requirements for the degree of

MASTER OF SCIENCE

in

APPLIED MATHEMATICS & COMPUTER SCIENCE

Advisor

Dr. Yi Cheng
Department of Mathematical Sciences

Committee:
Dr. Zhong Guan
Dr. Dana Vrajitoru

Accepted by the Graduate Faculty, Indiana University South Bend, in partial fulfillment of the requirements for the degree of Master of Science.

Master's Thesis Committee

_____

Chairperson, Yi Cheng, Ph.D.

_____

Zhong Guan, Ph.D.

_____

Dana Vrajitoru, Ph.D.

_____

David R. Surma Ph.D.
Graduate Director
Applied Mathematics and Computer Science

Dedication

This thesis is dedicated to my husband Eric and daughters: Selin and Melody.

Hayatimin en güzel yanlari sizleri çok seviyorum!

Abstract

The experimentation plays an important role in Science, Engineering, and Industry. The experimentation is an application of treatments to experimental units, and then measurement of one or more responses. It is a part of scientific method. It requires observing and gathering information about how process and system works. In an experiment, some input *x*'s transform into an output that has one or more observable response variables *y*. Therefore, useful results and conclusions can be drawn by experiment. In order to obtain an objective conclusion an experimenter needs to plan and design the experiment, and analyze the results.

There are many types of experiments used in real-world situations and problems. When treatments are from a continuous range of values then the true relationship between *y* and *x*'s might not be known. The approximation of the response function $y = f(x_1, x_2, \ldots, x_q) + e$ is called *Response Surface Methodology*. This thesis puts emphasis on designing, modeling, and analyzing the *Response Surface Methodology*. The three types of *Response Surface Methodology*, the first-order, the second-order, and three-level fractional factorial, will be explained and analyzed in depth. The thesis will also provide examples of application of each model by numerically and graphically using computer software.

Acknowledges

First and foremost I would like to thank my advisor, Dr. Yi Cheng. Without her patience and encouragement, this thesis would not have been possible. I sincerely appreciate invaluable academic and personal support I received from her throughout this thesis.

I would also thank the rest of my thesis committee members: Dr. Zhong Guan and Dr. Dana Vrajitoru for their valuable feedbacks and suggestions helped me to improve the thesis in many ways.

I also respectfully acknowledge Dr. Morteza Shafii-Mousavi and Dr. Yu Song, Chair, Department of Mathematical Science at IUSB for giving me their valuable time to guide me throughout my higher education.

I thank you all for teaching me and giving me a higher mind!

Finally, my further gratitude goes to my family: my husband Eric, my beautiful girls: Selin and Melody, my mother Reyhan, my sister Beyhan and her son Baran. Thank you for your love, support, and patience. I am truly blessed to have you as my family.

TABLE OF CONTENTS

List of Figures

List of Figures

List of Tables

1. Introduction

As an important subject in the statistical design of experiments, the *Response Surface Methodology* (*RSM*) is a collection of mathematical and statistical techniques useful for the modeling and analysis of problems in which a response of interest is influenced by several variables and the objective is to optimize this response (Montgomery 2005). For example, the growth of a plant is affected by a certain amount of water x_1 and sunshine x_2. The plant can grow under any combination of treatment x_1 and x_2. Therefore, water and sunshine can vary continuously. When treatments are from a continuous range of values, then a Response Surface Methodology is useful for developing, improving, and optimizing the response variable. In this case, the plant growth y is the response variable, and it is a function of water and sunshine. It can be expressed as

$$y = f(x_1, x_2) + e$$

The variables x_1 and x_2 are independent variables where the response y depends on them. The dependent variable y is a function of x_1, x_2, and the experimental error term, denoted as e. The error term e represents any measurement error on the response, as well as other type of variations not counted in f. It is a statistical error that is assumed to distribute normally with zero mean and variance s^2. In most *RSM* problems, the true response function f is unknown. In order to develop a proper approximation for f, the experimenter usually starts with a low-order polynomial in some small region. If the response can be defined by a linear function of independent variables, then the approximating function is a **first-order model**. A first-order model with 2 independent variables can be expressed as

$$y = b_0 + b_1 x_1 + b_2 x_2 + e$$

If there is a curvature in the response surface, then a higher degree polynomial should be used. The approximating function with 2 variables is called a **second-order model**:

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_{11} x_{11}^2 + b_{22} x_{22}^2 + b_{12} x_1 x_2 + e$$

In general all *RSM* problems use either one or the mixture of the both of these models. In each model, the levels of each factor are independent of the levels of other factors. In order to get the most efficient result in the approximation of polynomials the proper experimental design must be used to collect data. Once the data are collected, the *Method of Least Square* is used to estimate the parameters in the polynomials. The response surface analysis is performed by using the fitted surface. The **response surface designs** are types of designs for fitting response surface. Therefore, the objective of studying *RSM* can be accomplish by

(1) understanding the topography of the response surface (local maximum, local minimum, ridge lines), and

(2) finding the region where the optimal response occurs. The goal is to move rapidly and efficiently along a path to get to a maximum or a minimum response so that the response is optimized.

2. Literature Reviews

The *RSM* is important in designing, formulating, developing, and analyzing new scientific studying and products. It is also efficient in the improvement of existing studies and products. The most common applications of *RSM* are in Industrial, Biological and Clinical Science, Social Science, Food Science, and Physical and

Engineering Sciences. Since *RSM* has an extensive application in the real-world, it is also important to know how and where *Response Surface Methodology* started in the history. According to Hill and Hunter, *RSM* method was introduced by G.E.P. Box and K.B. Wilson in 1951 (Wikipedia 2006). Box and Wilson suggested to use a first-degree polynomial model to approximate the response variable. They acknowledged that this model is only an approximation, not accurate, but such a model is easy to estimate and apply, even when little is known about the process (Wikipedia 2006). Moreover, Mead and Pike stated origin of RSM starts 1930s with use of *Response Curves* (Myers, Khuri, and Carter 1989).

According to research conducted (Myers, Khuri, and Carter 1989), the *orthogonal design* was motivated by Box and Wilson (1951) in the case of the first-order model. For the second-order models, many subject-matter scientists and engineers have a working knowledge of the *central composite designs* (CCDs) and *three-level designs* by Box and Behnken (1960). Also, the same research states that another important contribution came from Hartley (1959), who made an effort to create a more economical or *small composite design*. There exist many papers in the literatures about the response surface models. In contrast, 3-level fractional design has limited works. Thus, 3-level fractional design is an open research subject. *Fractional Factorial Experiment Design for Factor at 3-Levels* (Connor and Zelen 1959) is a helpful resource conducting this kind of design. Many three-level fractional factorial designs and more importantly their alias tables can be found in their study.

According to (Myers, Khuri, and Carter 1989), the important development of optimal design theory in the field of experimental design emerged following Word World

II. Elfving (1952, 1955, 1959), Chernoff (1053), Kiefer (1958, 1959, 1960, 1962), and Kiefer and Wolfowitz were some of the various authors who published their work on optimality.

One of the important facts is whether the system contains a maximum or a minimum or a saddle point, which has a wide interest in industry. Therefore, *RSM* is being increasingly used in the industry. Also, in recent years more emphasis has been placed by the chemical and processing field for finding regions where there is an improvement in response instead of finding the optimum response (Myers, Khuri, and Carter 1989). In result, application and development of *RSM* will continue to be used in many areas in the future.

3. Response Surface Methods and Designs

Response Surface Methods are designs and models for working with continuous treatments when finding the optima or describing the response is the goal (Oehlert 2000). The first goal for Response Surface Method is to find the optimum response. When there is more than one response then it is important to find the compromise optimum that does not optimize only one response (Oehlert 2000). When there are constraints on the design data, then the experimental design has to meet requirements of the constraints. The second goal is to understand how the response changes in a given direction by adjusting the design variables. In general, the response surface can be visualized graphically. The graph is helpful to see the shape of a response surface; hills, valleys, and ridge lines. Hence, the function $f(x_1, x_2)$ can be plotted versus the levels of x_1 and x_2 as shown as Figure 3.1 .

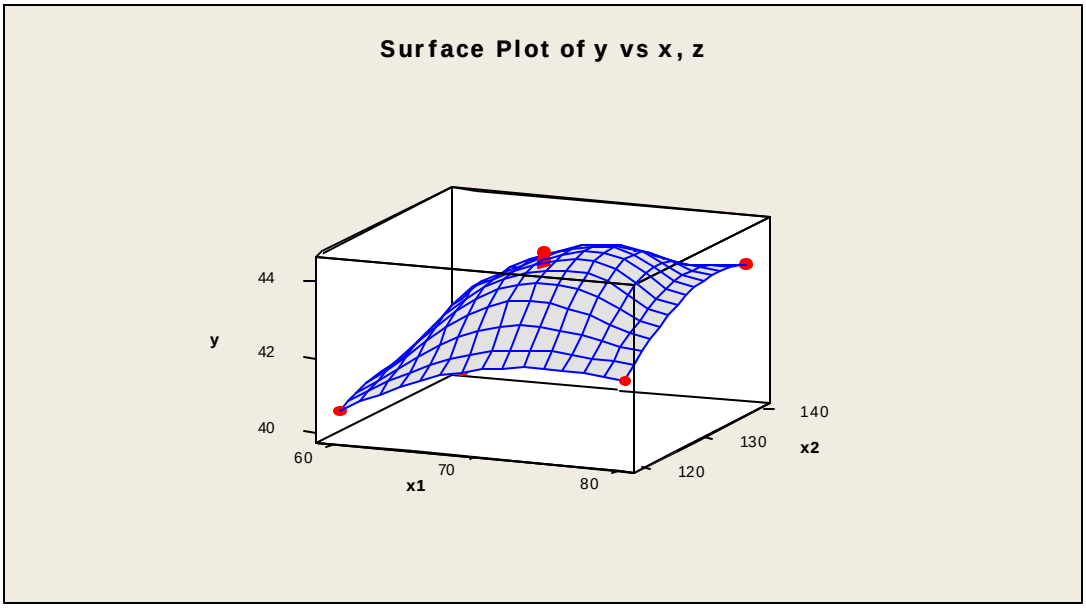


Figure 3.1 Response surface plot

$$y = f(x_1, x_2) + e$$

In this graph, each value of x_1 and x_2 generates a y-value. This three-dimensional graph shows the response surface from the side and it is called a **response surface plot**. Sometimes, it is less complicated to view the response surface in two-dimensional graphs. The contour plots can show contour lines of x_1 and x_2 pairs that have the same response value y. An example of contour plot is as shown in Figure 3-2.

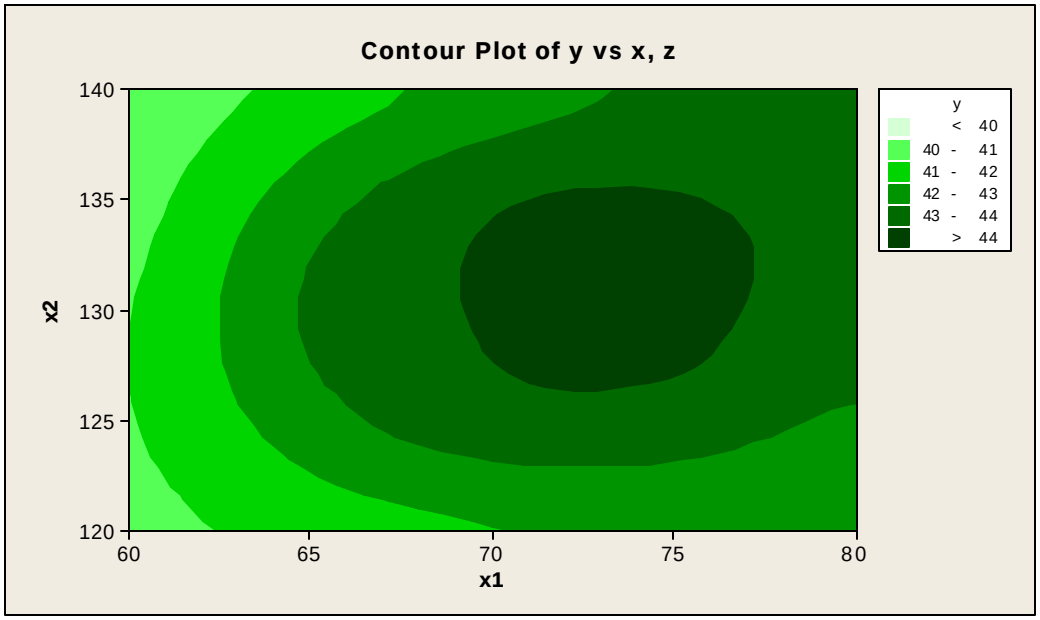


Figure 3-2 Contour plot

In order to understand the surface of a response, graphs are helpful tools. But, when there are more than two independent variables, graphs are difficult or almost impossible to use to illustrate the response surface, since it is beyond 3-dimension. For this reason, response surface models are essential for analyzing the unknown function *f*.

4. First-Order Model

4.1 Analysis of a First-Order Response Surface

The relationship between the response variable *y* and independent variables is usually unknown. In general, the low-order polynomial model is used to describe the response surface *f*. A polynomial model is usually a sufficient approximation in a small region of the response surface. Therefore, depending on the approximation of unknown function *f*, either first-order or second-order models are employed.

Furthermore, the approximated function *f* is a first-order model when the response is a linear function of independent variables. A first-order model with *N* experimental runs carrying out on *q* design variables and a single response *y* can be expressed as follows:

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots\ldots\ldots + b_q x_{iq} + e_i \qquad (i = 1, 2, \ldots\ldots, N)$$

The response *y* is a function of the design variables $x_1, x_{2,\ldots,} x_q$, denoted as *f*, plus the experimental error. A first-order model is a *multiple-regression* model and the b_j's are regression coefficients. I will explain multiple-regression in Section 4.1.1.

4.1.1 Multiple Regression Model

The relationship between a set of independent variables and the response *y* is determined by a mathematical model called *regression model*. When there are more than two independent variables the regression model is called *multiple-regression model*. In general, a *multiple-regression model* with *q* independent variable takes the form of

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots\ldots\ldots + b_q x_{iq} + e_i \qquad (i = 1, 2, \ldots\ldots, N)$$

$$= b_0 + \sum_{j=1}^{q} b_j x_{ij} + e_i \qquad (j = 1, 2, \ldots\ldots\ldots, q)$$

where $n > q$. The parameter $ß_j$ measures the expected change in response *y* per unit increase in x_i when the other independent variables are held constant. The i^{th} observation and j^{th} level of independent variable is denoted by x_{ij}. The data structure for the *multiple-regression model* is shown in Table 4.1.

Table 4.1 Data for Multiple-Regression Model

y	x_1	x_2	…	x_q
y_1	x_{11}	x_{12}	…	x_{1q}
y_2	x_{21}	x_{22}	…	x_{2q}
.	.	.	…	.
.	.	.	…	.
.	.	.	…	.
y_n	x_{n1}	x_{n2}	…	x_{nq}

The *multiple-regression model* can be written in a matrix form

$$\mathbf{y} = \mathbf{X}\boldsymbol{ß} + \boldsymbol{e}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{(nX1)} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} \cdots & x_{1q} \\ 1 & x_{21} & x_{22} \cdots & x_{2q} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \cdots & x_{nq} \end{bmatrix}_{(nXk)} \quad \boldsymbol{ß} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_q \end{bmatrix}_{(kX1)} \quad \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}_{(nX1)}$$

$\boldsymbol{y}$ is an (n X 1) vector of observations, $\boldsymbol{X}$ is an (n X k) matrix of levels of independent variables, $\boldsymbol{ß}$ is a (k X 1) vector of regression coefficients, and $\boldsymbol{e}$ is an (n X 1) vector of random errors (Montgomery 2005).

If X is a (k X k) matrix, then the linear system $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{ß} + \boldsymbol{e}$ has a unique *least squares* solution given by $\hat{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$. The estimated regression equation is

$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{b}$, it can also represent as $\hat{y}_i = \hat{b}_0 + \sum_{j=1}^{q} \hat{b}_j x_{ij} \quad i = 1,2,....,n.$

4.2 Designs for Fitting the First-Order Model

First-order model is used to describe the flat surfaces that may or may not be tilted. This model is not suitable for analyzing maximum, minimum, and ridge lines.

The first-order model approximation of the function f is reasonable when f is not too curved in that region and the region is not too big. First-order model is assumed to be an adequate approximation of true surface in a small region of the x's (Montgomery 2005). At this point my motivation is to illustrate a first-order model. The authors Dean and Voss give a data set for fractionation experiment that is conducted by M. Sosada (1993) in their case study sets. The reason I wanted to study this real-life experiment is, it allows me to work on two different response variables. This Case Study also allows me to demonstrate when first-order model is adequate to the given data versus when it is not. With this respect, it is essential to illustrate a first-order design.

…… Case Study - 1 ……

M. Sosada (1993) studied the effects of extraction time (t), solvent volume (V), ethanol concentration (C), and temperature (T) on the yield and phosphatidylcholine enrichment (PCE) of deoiled rapeseed lecithin when fractionated with ethanol.

Initially, a single-replicate 2^4 experiment was conducted, augmented by three center points. The design also included the sample variance of these three observations $s_c^2 = 1.120$ of PCE and $s_c^2 = 0.090$ of *Yield*. The results for the 16 factorial points are shown as the first 16 runs in Table 4.2 (Dean 1999).

Table 4.2 Process Data for fitting the First-Order Model

Natural Variables				Coded Variables				Responses	
t	V	C	T	*A*	*B*	*C*	*D*	*Yield*	*PCE*
15	10	98	25	1	1	1	1	27.6	43.8
5	5	98	25	-1	-1	1	1	16.6	27.2
15	5	92	25	1	-1	-1	1	15.4	23.6
5	10	92	25	-1	1	-1	1	17.4	26.2
15	5	98	15	1	-1	1	-1	17	27.8
5	10	98	15	-1	1	1	-1	19	30.2
15	10	92	15	1	1	-1	-1	17.4	25.2
5	5	92	15	-1	-1	-1	-1	12.6	18.8
15	5	98	25	1	-1	1	1	18.6	28.8
5	10	98	25	-1	1	1	1	22.4	36.8
15	10	92	25	1	1	-1	1	21.4	33.4
5	5	92	25	-1	-1	-1	1	14	21.0
15	10	98	15	1	1	1	-1	24	38.0
5	5	98	15	-1	-1	1	-1	15.6	23.6
15	5	92	15	1	-1	-1	-1	13	20.2
5	10	92	15	-1	1	-1	-1	14.4	22.6

In order to simplify the calculation, it is appropriate to use *coded variables* for describing independent variables in the (-1, 1) interval. The independent variables are rescaled therefore 0 is in the middle of the center of the design, and ±1 are the distance from the center with direction. The variables *t*, *V*, *C* and *T* are usually called *natural variables*, because they are expressed in the natural units of measurement. Therefore, if *t*, *V*, *C* and *T* denote the natural variables reaction time, volume, concentration, and temperature respectively then the transformation of these natural variables to coded variables is

$$A = \frac{t - 10}{5} \qquad B = \frac{V - 7.5}{2.5} \qquad C = \frac{Con - 95}{3} \qquad D = \frac{T - 20}{5}$$

The complete calculation of the coded variables is shown in Table 4.2. I illustrated the geometric view of the response variables *PCE* and *Yield* in Figure 4.1 and 4.2 respectively.

Figure 4.1 The Geometric View of the Response Variable *PCE*

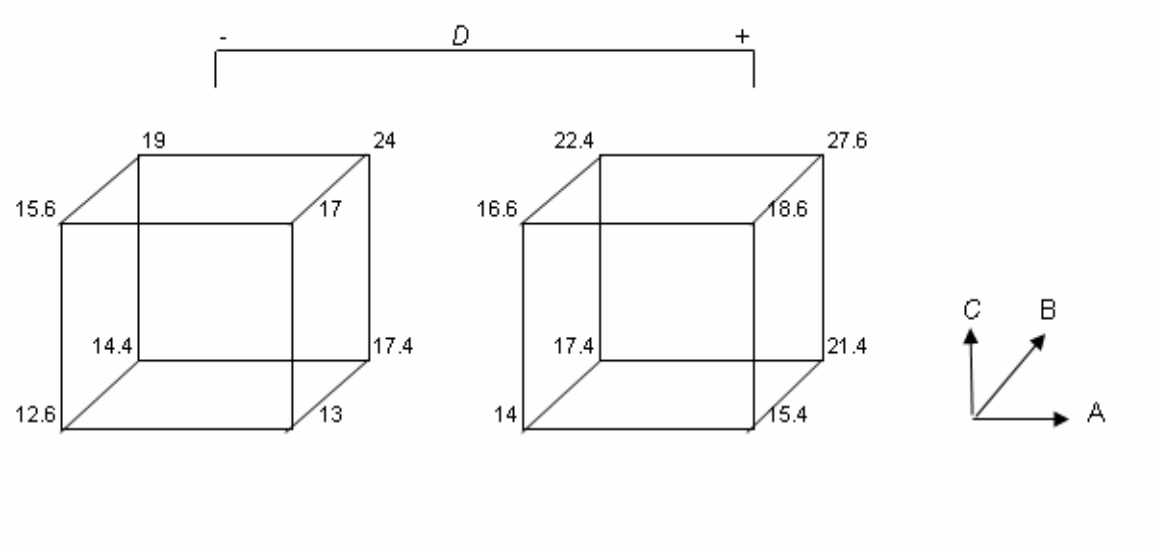
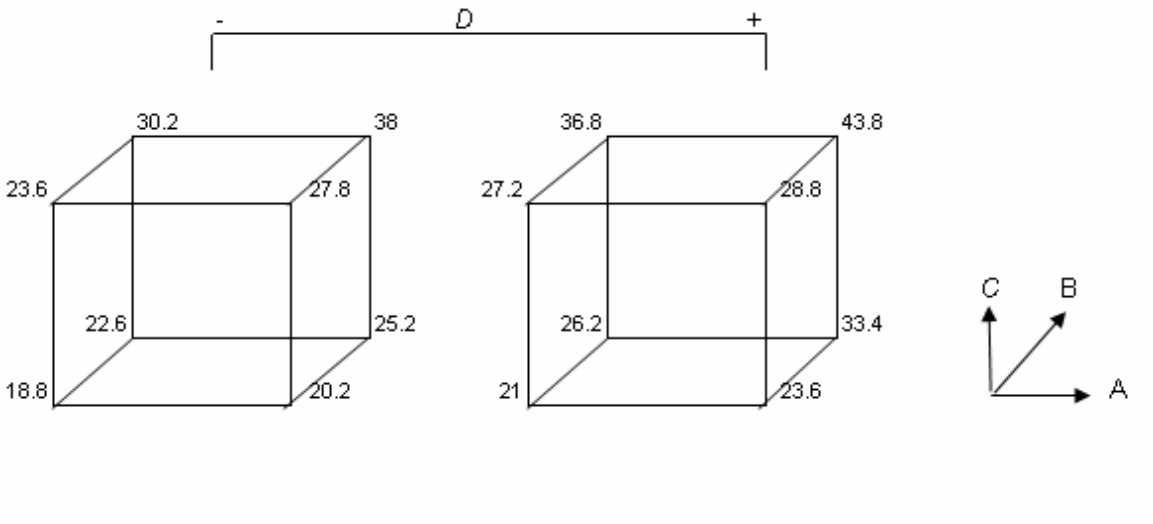


Figure 4.2 The Geometric View of the Response Variable *Yield*



4.2.1 Orthogonal First-Order Design

The experimenter needs to design a model to be efficient. For that reason, I have to take estimation of variances into consideration. The *orthogonal first-order designs* minimize the variance of the regression coefficients $\hat{b}_j$. A first-order design is

orthogonal if the off-diagonal elements of the (**X´X**) matrix are all zero (Montgomery 2005). Consequently, the cross-products of the columns of the ***X*** matrix sum to zero, the inverse matrix of (**X´X**) can be obtained easily, and all of the regression coefficients are uncorrelated. When the columns of the ***X*** matrix are mutually orthogonal then the levels of the corresponding variables are linearly independent. I demonstrated the matrix calculation for Case Study-1 using excel. The results are shown as follows:

Figure 4.3 Multiple Linear Model

X =

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	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

X'X =

16	0	0	0	0
0	16	0	0	0
0	0	16	0	0
0	0	0	16	0
0	0	0	0	16

$(X'X)^{-1}$

0.0625	0	0	0	0
0	0.0625	0	0	0
0	0	0.0625	0	0
0	0	0	0.0625	0
0	0	0	0	0.0625

X' =

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
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

The regression coefficients can be obtained by using the formula $\hat{b} = (X'X)^{-1}X'y$.

X'y =

447.2
34.4
65.2
65.2
34.4

43.8
27.2
23.6
26.2
27.8
30.2
25.2
18.8

$$\hat{b} = \begin{vmatrix} 27.950 \\ 2.150 \\ 4.075 \\ 4.075 \\ 2.150 \end{vmatrix} \qquad \mathbf{y_{PCE}} = \begin{vmatrix} 28.8 \\ 36.8 \\ 33.4 \\ 21.0 \\ 38.0 \\ 23.6 \\ 20.2 \\ 22.6 \end{vmatrix}$$

The fitted regression model for *PCE* is

$$\hat{y}_{PCE} = 27.950 + 2.150\ A + 4.075\ B + 4.075\ C + 2.150\ D \qquad \textbf{(4.1)}$$

The similar matrix form can be used to calculate the regression model for *Yield*.

$$\mathbf{X'y} = \begin{vmatrix} 286.000 \\ 22.800 \\ 40.400 \\ 35.600 \\ 20.800 \end{vmatrix} \qquad \mathbf{y_{yield}} = \begin{vmatrix} 27.60 \\ 16.60 \\ 15.40 \\ 17.40 \\ 17.00 \\ 19.00 \\ 17.40 \\ 12.60 \\ 18.60 \\ 22.40 \\ 21.40 \\ 14.00 \\ 24.00 \\ 15.60 \\ 13.00 \\ 14.40 \end{vmatrix}$$

$$\hat{b} = \begin{vmatrix} 17.900 \\ 1.400 \\ 2.550 \\ 2.200 \\ 1.275 \end{vmatrix}$$

The fitted regression model for *Yield* is

$$\hat{y}_{Yield} = 17.90 + 1.40\ A + 2.55\ B + 2.20\ C + 1.275\ D \qquad \textbf{(4.2)}$$

4.3 Model Adequacy Checking

In this section, I am going to analyze the model adequacy. It is important to examine the fitted model if the model provides an adequate approximation of the true response surface. I will use normality, analysis of variance, regression analysis, and lack of fit test to examine both of the models. I used Minitab to conduct the regression analysis and the variance of analysis of *PCE* and *Yield*. The results are shown respectively in Figure 4.4 and 4.5.

Figure 4.4 Analysis of Variance of Purified Lecithin - PCE

Regression Analysis: PCE versus A, B, C, D

```
The regression equation is
PCE = 28.0 + 2.15 A + 4.08 B + 4.08 C + 2.15 D


Predictor     Coef  SE Coef      T      P
Constant   27.9500   0.5666  49.33  0.000
A           2.1500   0.5666   3.79  0.003
B           4.0750   0.5666   7.19  0.000
C           4.0750   0.5666   7.19  0.000
D           2.1500   0.5666   3.79  0.003


S = 2.26635   R-Sq = 92.3%   R-Sq(adj) = 89.5%


Analysis of Variance

Source          DF      SS      MS      F      P
Regression       4  679.30  169.83  33.06  0.000
Residual Error  11   56.50    5.14
Total           15  735.80
```

Figure 4.5 Analysis of Variance of Purified Lecithin – Yield

Regression Analysis: Yield versus A, B, C, D

```
The regression equation is
Yield = 17.9 + 1.40 A + 2.55 B + 2.20 C + 1.28 D


Predictor     Coef  SE Coef      T      P
Constant   17.9000   0.3434  52.13  0.000
```

```
A           1.4000   0.3434   4.08  0.002
B           2.5500   0.3434   7.43  0.000
C           2.2000   0.3434   6.41  0.000
D           1.2750   0.3434   3.71  0.003


S = 1.37345   R-Sq = 92.0%   R-Sq(adj) = 89.1%


Analysis of Variance

Source          DF       SS      MS      F      P
Regression       4  238.850  59.713  31.65  0.000
Residual Error  11   20.750   1.886
Total           15  259.600
```

Even though, the Figures 4.4 and 4.5 can be produced using a variety of computer software, it is imperative for me to show how to calculate and analyze them. The table of analysis of variance for significance of the regression is given as follows:

Table 4.3 Analysis of Variance for Significance of Regression

Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	SS_R	q	MS_R	MS_R/MS_E
Error or Residuals	SS_E	$N - q - 1$	MS_E	
Total	SS_T	$N - 1$		

N is observations
q is the number of independent variable

The *error sum of squares* SS_E is a measurement of the amount of variation explained by the regression, the smaller the SS_E, the better the regression model. The following is called the *decomposition of the total variation*.

$$SS_E = SS_T - SS_R$$

$$SS_T = y'y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} \qquad SS_R = \hat{b}'X'y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

$$SS_E = y'y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} - \left[\hat{b}'X'y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right]$$

$$SS_E = y'y - \hat{b}'X'y$$

I demonstrated the process of the decomposition of variance for the response variables *PCE* and *Yield*. The process of the decomposition of variance for *PCE* is shown as follows:

$y'y = 13235.04$

$\hat{b}'X'y = 13178.54$ and

$\left(\sum_{i=1}^{n} y_i\right)^2 / n = 12499.24$

$\boldsymbol{SS_R}$ = 13178.54 - 12499.24 = 679.30
$\boldsymbol{SS_E}$ = 13235.04 - 13178.54 = 56.50
$\boldsymbol{SS_T}$ = 13235.04- 12499.24 = 735.80

$\boldsymbol{MS_R} = SS_R / q = 679.30 / 4 = 169.825$

$\boldsymbol{MS_E} = SS_E / N - q - 1 = 56.50 / (16-4-1) = 5.136$

Therefore, the statistic $\boldsymbol{F}$ is $\frac{SS_R / q}{SS_E /(N-q-1)} = \frac{MS_R}{MS_E} = 33.063.$

The process of the decomposition of variance for *Yield* is shown as follows:

$y'y = 5386.16$

$\hat{b}'X'y = 5365.41$ and

$\left(\sum_{i=1}^{n} y_i\right)^2 / n = 5126.56$

$\boldsymbol{SS_R}$ = 5365.41 - 5126.56 = 238.85
$\boldsymbol{SS_E}$ = 5386.16 – 5365.41 = 20.75
$\boldsymbol{SS_T}$ = 5386.16 – 5126.56 = 259.60

$\boldsymbol{MS_R} = SS_R / q = 238.85 / 4 = 59.712$

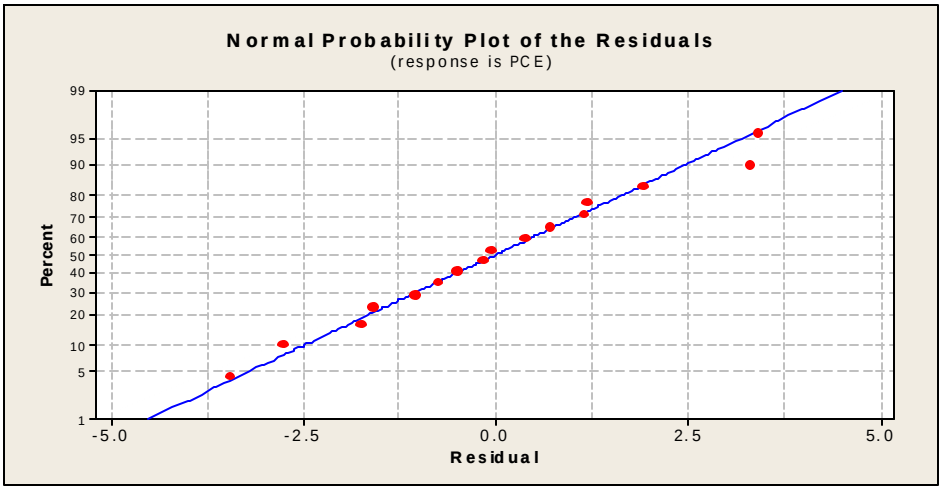
$\boldsymbol{MS_E} = SS_E / N - q - 1 = 20.75 / (16-4-1) = 1.886$

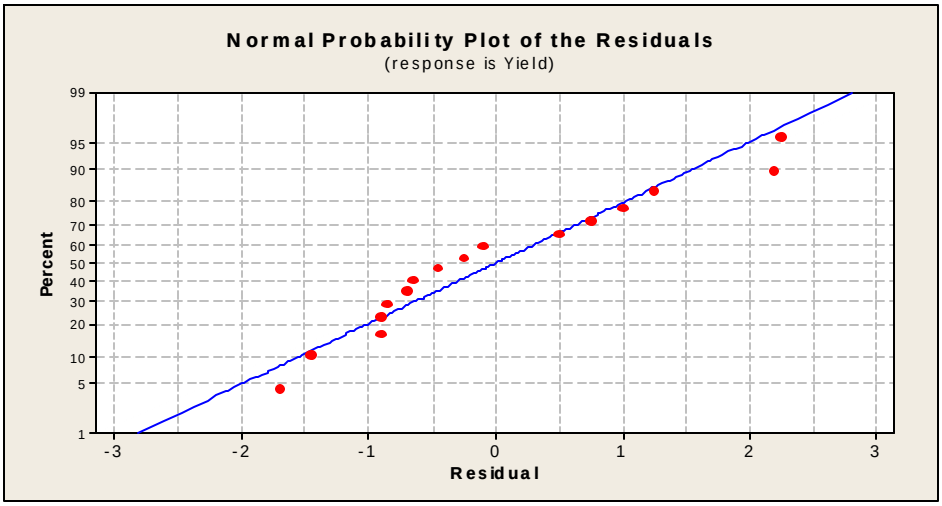
Therefore, the observed statistic $\boldsymbol{F}$ is $\frac{SS_R / q}{SS_E /(N-q-1)} = \frac{MS_R}{MS_E} = 31.655$. I will apply these statistics to the significance test in the next section.

4.3.1 The Test for Significance of Regression

A good estimated regression model shall explain the variation of the dependent variable in the sample. There are certain tests of hypotheses about the model parameters that can help the experimenter in measuring the effectiveness of the model. The first of all, these tests require for the error term e_i's to be normally and independently distributed with mean zero and variance s^2. To check this assumption, I graphed the normal probability of residuals for Case Study-1 as shown in Figure 4.6.

Figure 4.6 Normal Probability Plot of the Residuals





If the residuals plot approximately along a straight line, then the normality assumption is satisfied. In this study, the residuals can be judged as normally distributed; therefore normality assumptions for both of the responses are satisfied. The error term is the difference between the observed value y_i and the corresponding fitted value $\hat{y}_i$, that is, $e_i = y_i - \hat{y}_i$. As a result of this assumption, observations y_i are also normally and independently distributed. Therefore, the test for the significance of the regression can be applied to determine if the relationship between the dependent variable y and independent variables $x_1, x_{2,\ldots}, x_q$, exists. The proper hypotheses are

$H_0 : ß_1 = ß_2 = \ldots = ß_q = 0$ vs

$H_0 : ß_j \; ? \; 0$ for at least one j.

The statistic F is compared to the critical $F_{a,q,N-q-1}$, if observed F-value is greater than the critical F, then H_0 will be rejected. Equivalently, H_0 is rejected when P-value for the statistic F is less than significant level a. As a result, the hypothesis for the statistical analysis of response variable *RSE* can be written as:

$H_0 : ß_1 = ß_2 = ß_3 = ß_4 = 0$ vs

$H_1 : ß_j \neq 0$ for at least one j.

At the significant level $a = 0.05$, the critical value $F_{.05,4,11} = 3.36$ is < the observed $F = 33.063$. Also, P-value from Figure 4.4 for the statistic F is less than a. There is a significant statistical evidence to reject the null hypothesis. It implies that at least one of the independent variables – time (A), volume (B), concentration (C), and temperature (D) - contributes significantly to the model.

I used the same method to test for the significance of the regression model for the response variable *Yield*. Using a % 5 level of significance, the critical value $F_{.05,4,11} = 3.36$ is < the observed $F = 31.655$. Again, there is a linear relationship between the independent variables – time (A), volume (B), concentration (C), and temperature (D) - and the response variable *Yield* of purified lecithin.

How well the estimated model fits the data can be measured by the value of R^2. The R^2 lies in the interval [0,1]. When R^2 is closer to the 1, the better the estimation of regression equation fits the sample data. In general, the R^2 measures percentage of the variation of y around $\bar{y}$ that is explained by the regression equation. However, adding a variable to the model always increased R^2, regardless of whether or not that variable statistically significant. Thus, some experimenter rather using *adjusted-* $\overline{R}^2$. When variables are added to the model, the *adjusted-* $\overline{R}^2$ will not necessarily increase. In actual fact, if unnecessary variables are added, the value of *adjusted -* $\overline{R}^2$ will often decrease. For instance, consider the regression models in Case Study-1. I calculated the R^2 and the *adjusted-* $\overline{R}^2$ for both of the models. I showed these results earlier in Figure 4.4 and 4.5.

RSE

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T} = \frac{679.30}{735.80} = .923$$

$$\overline{R}^2 = 1 - \frac{SS_E/(n-q-1)}{SS_T/(n-1)} = \frac{56.50/(16-4-1)}{735.80/(15)} = .895$$

Yield

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T} = \frac{238.85}{259.60} = .92$$

$$\overline{R}^2 = 1 - \frac{SS_E/(n-q-1)}{SS_T/(n-1)} = \frac{20.75/(16-4-1)}{259.60/(15)} = .89$$

Both of R^2 and $\overline{R}^2$ are statistically significant for the response variables *RSE* and *Yield*. It suggests that the estimated regression equations for the Case Study-1 fit the data well. At this point, there is no sufficient reason to reject the initial regression Equations 4.1 and 4.2 for *PCE* and *Yield* of purified lecithin respectively.

4.3.2 The Test for Individual Regression Coefficients

In order to determine whether given variables should be included or discluded from the model, I need to test hypotheses for the individual regression coefficients. The simple analysis starts with a main effects plot. A main effects plot is a plot of the means of the response variable for each level of a factor. It allows an experimenter to obtain a general idea of which main effects may be important. The main effect is calculated by subtracting the overall mean for the factor from the mean for each level. The Figure 4.7 and 4.8 show the locations of the main effects for *PCE* and *Yield* respectively.

Figure 4.7 Main Effects Plot of Purified Lecithin Phosphatidylcholine Enrichment

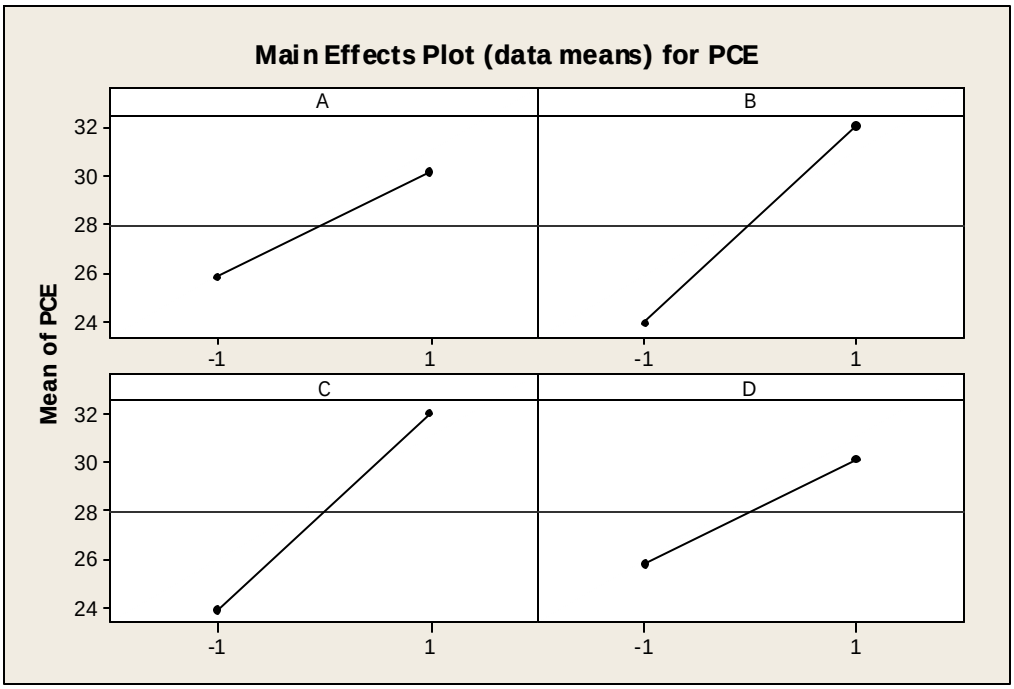
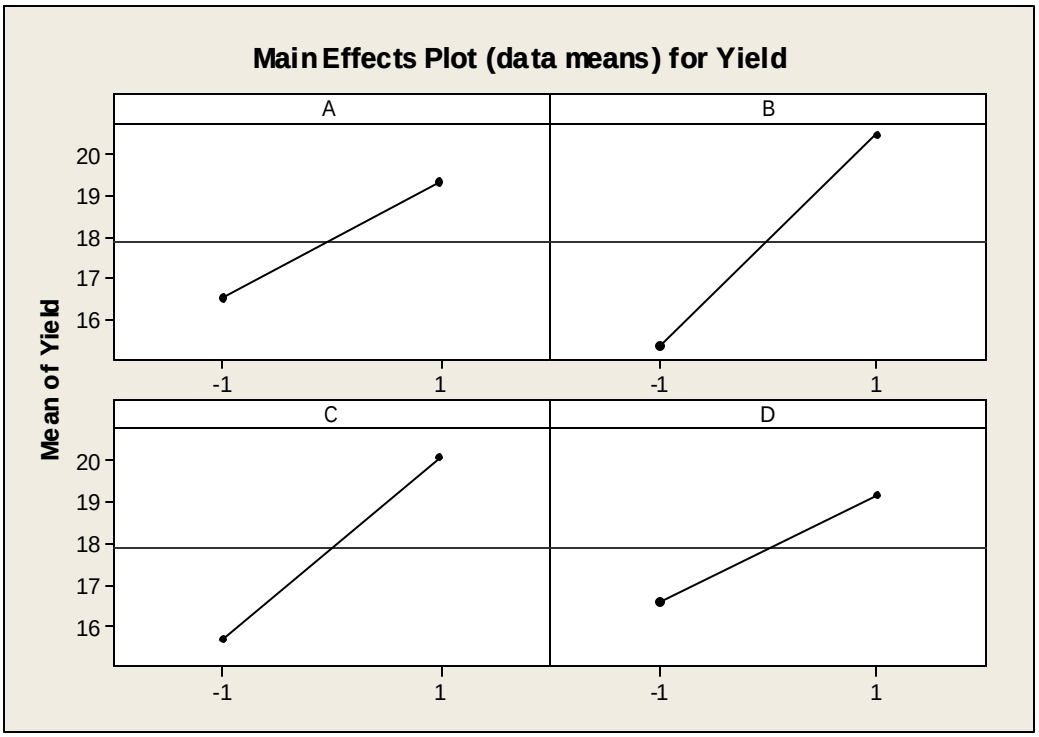


Figure 4.8 Main Effects Plot of Purified Lecithin Yield



My analysis indicates that the factors *A*, *B*, *C*, and *D* increase when they move from the low level to the high level of purified lecithin Phosphatidylcholine Enrichment (*PCE*) and *Yield*. Each level of the factors affects the response differently. Each factor at their high level results in higher mean responses comparing to that at the low level. Alternatively, the factors *B* and *C* appear to have a greater effect on the responses, with a steeply slope. If the slope is close to zero, the magnitude of the main effect would be small. The main effect plots are helpful in visualizing which factors affect the response the most, but in order to determine the significance of the factors, I have to conduct an appropriate statistical test, a *t*-test, to identify the significance of the main factors.

In general, an *F*-test is used to test for more than one coefficient or, joint hypotheses. When the hypotheses test is particular to one coefficient at a time, then *t*-test is more common. To examine the significant contribution of the independent variables to the phosphatidylcholine enrichment (*PCE*), I did the following calculations for the following hypotheses:

$H_0: \beta_{time} = 0$ $\quad$ $H_1: \beta_{time} \neq 0$

$H_0: \beta_{volume} = 0$ $\quad$ $H_1: \beta_{volume} \neq 0$

$H_0: \beta_{conc.} = 0$ $\quad$ $H_1: \beta_{conc.} \neq 0$

$H_0: \beta_{temp} = 0$ $\quad$ $H_1: \beta_{temp} \neq 0$

The test for this hypothesis is called *t*-statistic, expressed as

$t_0 = \dfrac{\hat{b}_j}{\sqrt{\hat{s}^2 W_{jj}}}$ where W_{jj} is the diagonal elements of $(X'X)^{-1}$ corresponding to $\hat{b}_j$. The denominator $\sqrt{\hat{s}^2 W_{jj}}$ is called the *standard error* of the regression coefficient $\hat{b}_j$,

because $se(\hat{b}_j) = \sqrt{\hat{s}^2 W_{jj}}$. The values of $se(\hat{b}_j)$ are also found in Figure 4.2. Recall

from earlier calculation that $\hat{b}_j = \begin{bmatrix} 27.950 \\ 2.150 \\ 4.075 \\ 4.075 \\ 2.150 \end{bmatrix}$, $(X'X)^{-1} = \begin{bmatrix} .0625 & 0 & 0 & 0 & 0 \\ 0 & .0625 & 0 & 0 & 0 \\ 0 & 0 & 0.625 & 0 & 0 \\ 0 & 0 & 0 & .0625 & 0 \\ 0 & 0 & 0 & 0 & .0625 \end{bmatrix}$,

and $\hat{s}^2 = 5.14$. Consequently, t-statistics are computed below:

$$t_A = \frac{\hat{b}_1}{\sqrt{\hat{s}^2 W_{11}}} = \frac{2.150}{\sqrt{5.14 * .0625}} = 3.778 , \quad t_B = \frac{\hat{b}_2}{\sqrt{\hat{s}^2 W_{22}}} = \frac{4.075}{\sqrt{5.14 * .0625}} = 7.19 ,$$

$t_C = 7.19$, and $t_D = 3.778$.

These *t*-statistic values are compared with the *critical t*-values. The null hypothesis H_0: $ß_j = 0$ is rejected if the observed $|t_0|$ > critical value $t_{\frac{a}{2}, N-q-1}$. The level of significance is at 5 percent, that is, $a = .05$. Noting that

$|t_A| = |t_D| = 3.778 > \ t_{.025,11} = 2.201$ and

$|t_B| = |t_C| = 7.19 \ > \ t_{.025,11} = 2.201,$

the null hypotheses $H_0: ß_{time} = 0$, $H_0: ß_{volume} = 0$, $H_0: ß_{conc} = 0$, and $H_0: ß_{temp} = 0$ are rejected. I concluded that the independent variables: time (*A*), volume (*B*), concentration (*C*), and temperature (*D*), all contribute significantly to the response variable phosphatidylcholine enrichment (*PCE*).

Furthermore, I used a similar test for the hypothesis on the individual regression coefficients for the yield of purified lecithin. Using the coefficients $\hat{b}_j = \begin{bmatrix} 17.900 \\ 1.400 \\ 2.550 \\ 2.200 \\ 1.275 \end{bmatrix}$ and $\hat{s}^2 = 1.886$, the *t*-statistics were computed as follows:

$$t_A = \frac{\hat{b}_1}{\sqrt{\hat{s}^2 W_{11}}} = \frac{1.40}{\sqrt{1.886 * .0625}} = 4.077, \quad t_B = \frac{\hat{b}_2}{\sqrt{\hat{s}^2 W_{22}}} = \frac{2.550}{\sqrt{1.886 * .0625}} = 7.427,$$

$$t_C = \frac{\hat{b}_1}{\sqrt{\hat{s}^2 W_{11}}} = \frac{2.20}{\sqrt{1.886 * .0625}} = 6.4078, \quad t_D = \frac{\hat{b}_2}{\sqrt{\hat{s}^2 W_{22}}} = \frac{1.275}{\sqrt{1.886 * .0625}} = 3.713.$$

Note that

$|t_A| = 4.08 > \quad t_{.025,11} = 2.201,$

$|t_B| = 7.43 > \quad t_{.025,11} = 2.201,$

$|t_C| = 6.41 > \quad t_{.025,11} = 2.201,$ and

$|t_D| = 3.71 > \quad t_{.025,11} = 2.201.$

All *t*-statistics are larger than the critical *t*-value. I concluded that the independent variables, the time (*A*), the volume (*B*), the concentration (*C*), and the temperature (*D*), all contribute significantly to the model.

4.3.3 Center Points in a 2^q Design

In addition to the orthogonal design, the standard first-order design is a 2^q factorial with a center point. These designs consist of factorial points n_f and the center points n_c. The center points are observations collected at the center points $x_i = 0$ (i = 1,2,…, q). The replicated points at the center points can be used to calculate the pure

error. Also, the contrast between the mean of the center points and the mean of the factorial points provides a test for the *lack of fit* in a 2^q design. The lack of fit of a first-order model occurs when the model does not adequately represent the mean response as a function of the factor level (Angela 1999). The analysis of variance of generic lack of fit test for the first order model is given in Table 4.4.

Table 4.4 Analysis of Variance of Lack of Fit

Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Residuals	SS_E	$N - q - 1$	MS_E	
Lack of fit	SS_{LOF}	$n_d - q - 1$	MS_{LOF}	MS_{LOF}/MS_{PE}
Pure Error	SS_{PE}	$N - n_d$	MS_{PE}	

N observations
q number of independent variables
n_d *distinct design points*

The residual sum of squares SS_E can be partitioned into two components,

$$SS_E = SS_{PE} + SS_{LOF}$$

where SS_{PE} is the sum of squares due to the pure error and SS_{LOF} is sum of squares due to the lack of fit. The replicates at the center can be used to calculate the mean squares for the pure error, where $\bar{y}_c$ is the average of the n_c runs at the center point

$$MS_{PE} = \frac{SS_{PE}}{n_c - 1} = center\sum_{i=1}^{n_c} (y_i - \bar{y}_c)^2 / n_c - 1 .$$

The mean squares for lack of fit is

$$MS_{LOF} = \frac{SS_{LOF}}{n_d - q - 1}$$

then the ratio

$$F_0 = \frac{MS_{LOF}}{MS_{PE}}$$

is used to test the null hypothesis of the lack of fit.

Recall the Case Study-1, the authors Dean and Voss stated in their case study sets that design included n_c = 3 center points observation of each response variables *PCE* and *Yield*. Since, these additional observations are not included in the data set; I can use the given sample variance values of each three observations to test for lack of fit.

For testing the lack of fit for the response variable *Yield* of purified lecithin, the following computations are carried out:

$N = n_f + n_c = 16 + 3 = 19$,

n_d = 16 + (1 center point) = 17,

$df_{LOF} = n_d - q - 1 = 12$,

$df_{PE} = N - n_d = 2$, and

$s_c^2 = .09$.

Since the factorial points included no replication,

$s_c^2 = \sum_{i=1}^{n_c} (y_i - \bar{y}_c)^2 / n_c - 1$ will imply that $MS_{PE} = s_c^2 = .09$.

Therefore, $SS_{PE} = MS_{PE} * (N - n_d) = .18$,

$SS_{E(19\ runs)} = SS_{E(16\ runs)} + s_c^2 * (n_c - 1) = 20.750 + .18 = 20.93$,

$SS_E = SS_{PE} + SS_{LOF}$ implies $SS_{LOF} = 20.750$, and

$MS_{LOF} = \frac{SS_{LOF}}{n_d - q - 1} = 1.729$.

The test statistic for the lack of fit $F_0 = \frac{MS_{LOF}}{MS_{PE}} = 19.213$ is compared to the critical $F_{a,n_d-q-1,N-n_d}$ value.

The comparison shows that $F_0 = 19.213 \tilde{} F_{.05,12,2} = 19.41$. Since the observed statistic F_0 value is slightly less than the critical F-value, I cannot conclude the significance of regression model by this test at significance level $a = 0.05$. The analysis of variance for *Yield* is given in Table 4.5. Therefore, I will conduct a more appropriate model, such as a second-order model, and I will study in Section 5.2. However, the analysis of the response variable *Yield* will still be continued in a single replicate of the 2^q design in Section 4.4.

Table 4.5 Analysis of Variance of Lack of Fit for *Yield*

Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Residuals	*20.93*	*14*	*1.495*	
Lack of fit	*20.75*	*12*	*1.729*	*19.213*
Pure Error	*.18*	*2*	*.09*	

I carried out same type of calculation on the data set for *PCE* using the sample variance $s_c^2 = 1.120$. My result of the analysis of variance is given in Table 4.6.

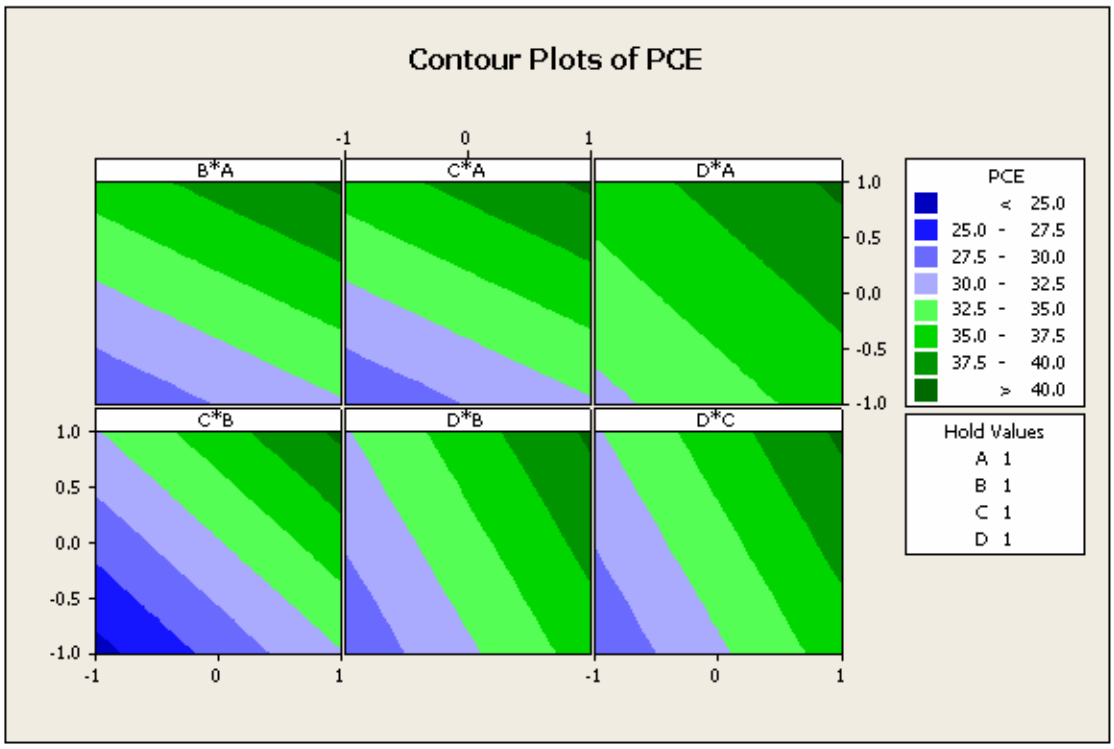
Table 4.6 Analysis of Variance of Lack of Fit for *RCE*

Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Residuals	58.74	*14*	*4.196*	
Lack of fit	*56.50*	*12*	*4.708*	*4.204*
Pure Error	*2.24*	*2*	*1.12*	

The test statistic $F_0 = 4.204$ is smaller than the critical $F_{.05,12,2} = 19.41$ value. There is no significant evidence of lack of fit at $a = 0.05$. Therefore, I can conclude that the true response surface is explained by the linear model.

I can also use contour plot to visualize the response surface. These plots show how the response variable relates to the two factors at a time. Since there are four factors; each time two factors will be hold at a constant level when plotting the other two factors. The response surface changes when the holding levels are changed. Therefore, it is important to select the holding levels for the other factors. In general, the optimum levels for factorial model with no curvature will be at one of the corners. The analysis of the main effects plot in Figure 4.7 indicates that the best optimum setting includes Time, Volume, Concentration, and Temperature, all at their high levels. These settings can be used as the hold values for each factors when it was not included in the plot (Minitab). The Figure 4.9 shows the contour plots of *PCE*.

Figure 4.9 the Contour Plots of *PCE*



Since the response surface is a plane, the contour plots are parallel straight lines. The analysis of the contour plots is as follows:

B*A: This plot indicates that how variables, Volume and Time, are related to the *PCE* of deoiled rapeseed lecithin while the other factors, Concentration and Temperature, are held constant at high level 1. The response is at its highest (greater than 40) at the darkest region of the graph (upper right corner).

C * A: This plot indicates that how variables, Concentration and Time, are related to the *PCE* of deoiled rapeseed lecithin while the other factors, Volume and Temperature, are held constant at high level 1. The response is at its highest (greater than 40) at the darkest region of the graph (upper right corner).

D * A: This plot indicates that how variables, Temperature and Time, are related to the *PCE* of deoiled rapeseed lecithin while the other factors, Volume and

Concentration, are held constant at high level 1. The response is at its highest (greater than 40) at the darkest region of the graph (upper right corner).

C * B: This plot indicates that how variables, Concentration and Volume, are related to the *PCE* of deoiled rapeseed lecithin while the other factors, Time and Temperature, are held constant at high level 1. The response is at its highest (greater than 40) at the darkest region of the graph (upper right corner).

D * B: This plot indicates that how variables, Temperature and Volume, are related to the *PCE* of deoiled rapeseed lecithin while the other factors, Time and Concentration, are held constant at high level 1. The response is at its highest (greater than 40) at the darkest region of the graph (upper right corner).

D * C: This plot indicates that how variables, Temperature and Concentration, are related to the *PCE* of deoiled rapeseed lecithin while the other factors, Time and Volume, are held constant at high level 1. The response is at its highest (greater than 40) at the darkest region of the graph (upper right corner).

In order to maximize the phosphatidylcholine enrichment (*PCE*) of deoiled rapeseed lecithin when fractionated with ethanol, I can choose high level settings for Extraction Time, Solvent Volume, Ethanol Concentration, and Temperature. The final estimated regression model using the coded variables is expressed as follows:

$$\hat{y}_{PCE} = 27.950 + 2.150\ A + 4.075\ B + 4.075\ C + 2.150\ D$$

I found the maximum predicted response is $\hat{y}_{PCE} = 40.40$, achieved when all four factors are at their high level (1).

4.4 A Single Replicate of the 2^q Design

In general, the 2^q design can be large; therefore availability of resources allows an experimenter to run a single replicate of a design. The Case Study-1 is a single replicate 2^4 design. An earlier analysis concluded that there may be a possibility that the regression model for the *Yield* is not sufficiently explained by the main effects. Therefore, I need to study the impact of the interactions. Recall that the data did not include the 3 center points, thus I can continue to analyze the data by using a single replicate of the 2^4 design. The design matrix of main effects and their interactions are shown in Table 4.7.

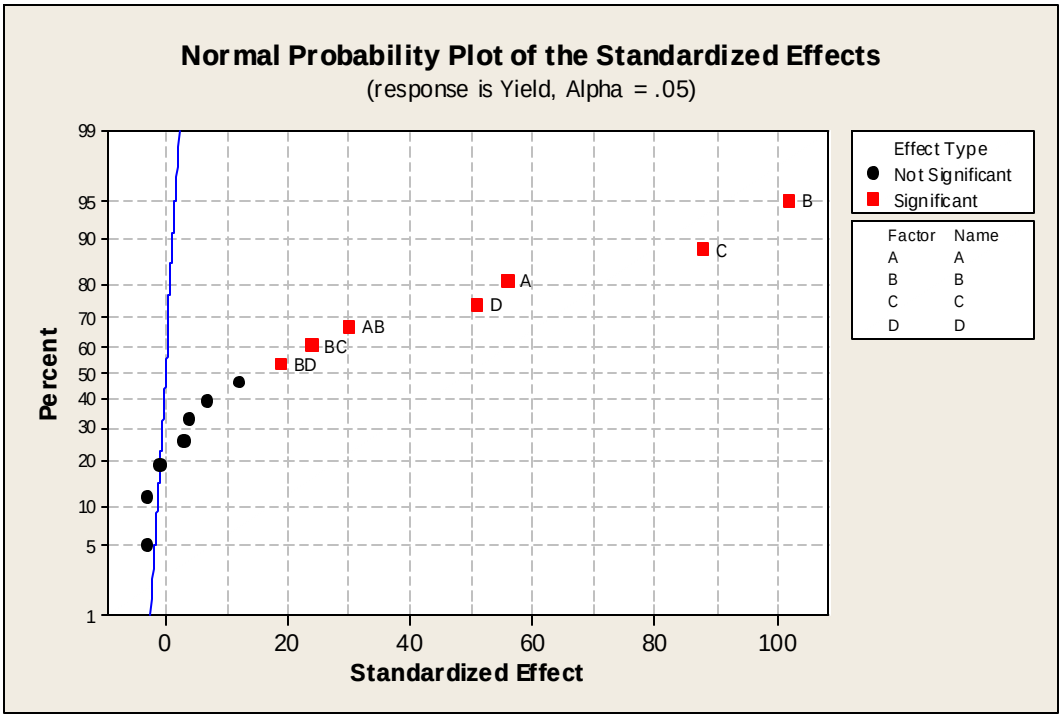
Table 4.7 The Main Effects and Interactions for 2^4 Design

A	*B*	*C*	*D*	*AB*	*AC*	*AD*	*BC*	*BD*	*CD*	*ABC*	*ABD*	*ACD*	*BCD*	*ABCD*	*Yield*
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	27.6
-1	-1	1	1	1	-1	-1	-1	-1	1	1	1	-1	-1	1	16.6
1	-1	-1	1	-1	-1	1	1	-1	-1	1	-1	-1	1	1	15.4
-1	1	-1	1	-1	1	-1	-1	1	-1	1	-1	1	-1	1	17.4
1	-1	1	-1	-1	1	-1	-1	1	-1	-1	1	-1	1	1	17
-1	1	1	-1	-1	-1	1	1	-1	-1	-1	1	1	-1	1	19
1	1	-1	-1	1	-1	-1	-1	-1	1	-1	-1	1	1	1	17.4
-1	-1	-1	-1	1	1	1	1	1	1	-1	-1	-1	-1	1	12.6
1	-1	1	1	-1	1	1	-1	-1	1	-1	-1	1	-1	-1	18.6
-1	1	1	1	-1	-1	-1	1	1	1	-1	-1	-1	1	-1	22.4
1	1	-1	1	1	-1	1	-1	1	-1	-1	1	-1	-1	-1	21.4
-1	-1	-1	1	1	1	-1	1	-1	-1	-1	1	1	1	-1	14
1	1	1	-1	1	1	-1	1	-1	-1	1	-1	-1	-1	-1	24
-1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	1	-1	15.6
1	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1	-1	-1	13
-1	1	-1	-1	-1	1	1	-1	-1	1	1	1	-1	1	-1	14.4

The problem of running an analysis of this saturated model is that I cannot get an estimate of error. Here is the reason. There are 15 degrees of freedom in such experiment, with 4 degrees of freedom for main effects, 6 degrees of freedom for 2-factor interactions, 4 degrees of freedom for 3-factor interactions, and 1 degree of freedom for

4-factor interaction. Consequently, there are no degrees of freedom left to estimate the error variance. Therefore, one way to analyze the unreplicated factorial design is to examine the normality of the estimated effects. The experimenter can use a normal effects plot to determine the statistical significance of both main and interaction effects. The effects that are not significant will fall along a line, on the other hand, the significant effects will stray farther from the line. The Figure 4.10 illustrates the normal plot of these effects.

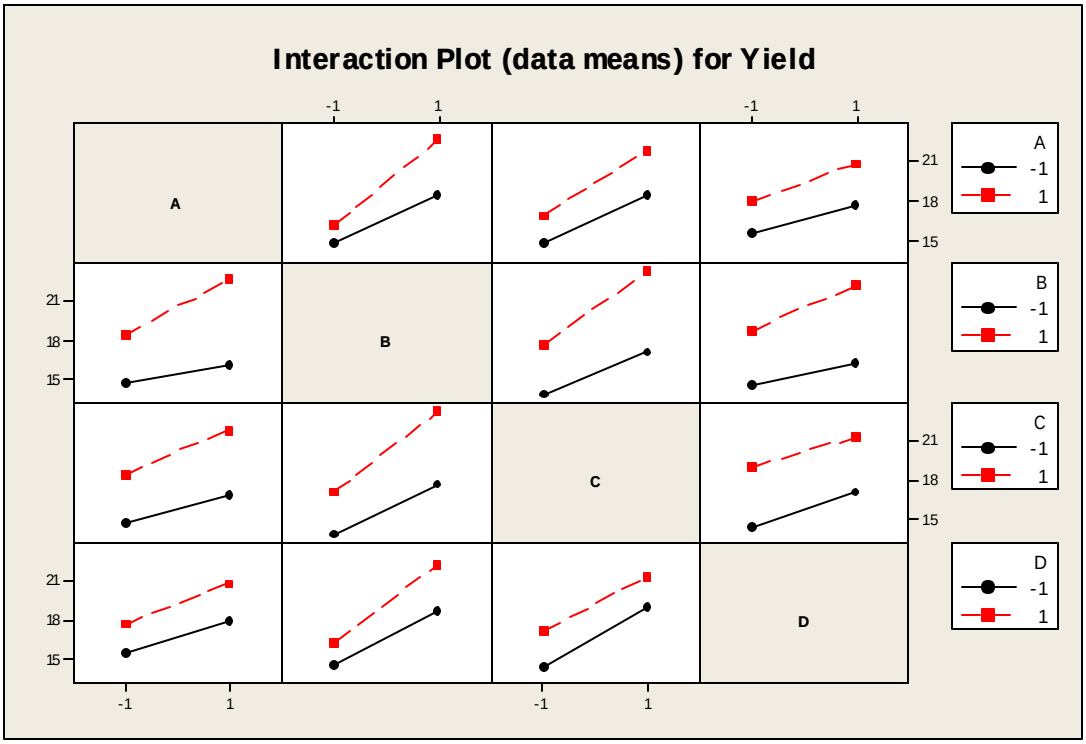
The Figure 4.10 The Normal Plot of Effects



My analysis is concluded that the main effects *A*, *B*, *C*, and *D* and the interactions *AB*, *BC*, and *BD* are significant. Since they lie on right hand side of the line, their contribution has a positive effect on the model. The rest of the effects lie along the line are negligible. The factor Solvent Volume (*B*) appears to have a largest effect because it

lies furthest from the line. The lower term interaction plot as shown in Figure 4.11 can also be a helpful resource in visualizing interactions.

Figure 4.11 The Plot of Low-Order Interactions



This interaction plot confirms the significance of *AB*, *BC*, and *BD* interactions as stated earlier. Interaction occurs when one factor does not produce the same effect on the response at different levels of another factor. Therefore, if the lines of two factors are parallel, there is no interaction. On the contrary, when the lines are far from being parallel, the two factors are interacting. In each case of *AB*, *BC*, and *BD* interactions, the response yield increases when the line moves from the low level (-1) to high level (1). For example, the factor *A* effect is small when the factor *B* is at the low level and large

when the factor *B* is at the high level. It appears that the best result is obtained when each of the factors: *A*, *B*, *C*, and *D* is at their high level.

Another strategy is to analyze the significance of data using the *sparsity of effects principle*. This principle assumes that most systems are dominated by some main effects and low-order interactions, and most high-order interactions are negligible (Myers 1995). I assumed that the highest interaction component *ABCD* is negligible and its mean square can be used to obtain an error term. Table 4.8 gives the analysis of the factorial design in this respect.

Figure 4.12 Analysis of Variance of the 2^4 Factorial Design for Purified Lecithin – Yield

Regression Analysis: Yield versus A, B, ...

```
The regression equation is
Yield = 17.9 + 1.40 A + 2.55 B + 2.20 C + 1.28 D + 0.750 AB + 0.300 AC
        + 0.175 AD + 0.600 BC + 0.475 BD - 0.0750 CD + 0.100 ABC - 0.0250 ABD
        - 0.0750 ACD + 0.0750 BCD
```

Predictor	Coef	SE Coef	T	P
Constant	17.9000	0.0250	716.00	0.001
A	1.40000	0.02500	56.00	**0.011**
B	2.55000	0.02500	102.00	**0.006**
C	2.20000	0.02500	88.00	**0.007**
D	1.27500	0.02500	51.00	**0.012**
AB	0.75000	0.02500	30.00	**0.021**
AC	0.30000	0.02500	12.00	0.053
AD	0.17500	0.02500	7.00	0.090
BC	0.60000	0.02500	24.00	**0.027**
BD	0.47500	0.02500	19.00	**0.033**
CD	-0.07500	0.02500	-3.00	0.205
ABC	0.10000	0.02500	4.00	0.156
ABD	-0.02500	0.02500	-1.00	0.500
ACD	-0.07500	0.02500	-3.00	0.205
BCD	0.07500	0.02500	3.00	0.205

```
S = 0.1   R-Sq = 100.0%   R-Sq(adj) = 99.9%
```

Analysis of Variance for Yield (coded units)

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Main Effects	4	238.850	238.850	59.7125	5971.25	0.010
2-Way Interactions	6	20.390	20.390	3.3983	339.83	0.041
3-Way Interactions	4	0.350	0.350	0.0875	8.75	0.248

```
Residual Error       1    0.010    0.010   0.0100
Total               15  259.600
```

As it appeared on the normal plot of effects, 2-factor interactions are significant at the level of % 5 significance. Meanwhile, the 3-factor interactions do not appear to contribute significantly to the model. The *t*-tests reveal that the main effects of *A*, *B*, *C*, and *D* and the interactions *AB*, *BC*, and *BD* are significant. My result confirms previous graphical analysis of normal plot effects and interactions plot. In order to get the best response surface result for yield of purified lecithin, I can consider the main effects at their high level. The following estimated equation is my final model for the single replicated 2^4 factorial design for the response variable *Yield*.

$$\hat{y}_{Yield} = 17.90 + 1.40\,A + 2.55\,B + 2.2\,C + 1.275\,D + 0.75\,AB + 0.60\,BC + 0.475\,BD$$

Therefore, predicted *Yield* of deoiled rapeseed lecithin when fractionated by ethanol is

$$\hat{y}_{Yield} = 17.90 + 1.40\,(1) + 2.55\,(1) + 2.2\,(1) + 1.275\,(1) + 0.75\,(1)(1) + 0.60\,(1)(1) + 0.475\,(1)(1)$$

$$\hat{y}_{Yield} = 27.15$$

where all four factors; Extraction Time, Solvent Volume, Ethanol Concentration, and Temperature are at high level (+1) .

4.5 Conclusion of the first-order model

A first-order model uses low-order polynomial terms to describe some part of the response surface. This model is appropriate for describing a flat surface with or without tilted surfaces. Usually a first-order model fits the data by least squares. Once the estimated equation is obtained, an experimenter can examine the normal plot, the main effects, the contour plot, and ANOVA statistics (F-test, t-test, R^2, the adjusted R^2, and lack of fit) to determine adequacy of the fitted model. Lack of fit of the first-order model happens when the response surface is not a plane. If there is a significant lack of fit of the first-order model, then a more highly structured model, such as second-order model, may be studied in order to locate the optimum.

5. Second-Order Model

5.1 Analysis of a Second-Order Response Surface

When there is a curvature in the response surface the first-order model is insufficient. A second-order model is useful in approximating a portion of the true response surface with parabolic curvature. The second-order model includes all the terms in the first-order model, plus all quadratic terms like $b_{11}x_{1i}^2$ and all cross product terms like $b_{13}x_{1i}x_{3j}$. It is usually expressed as

$$\begin{aligned} y &= b_0 + \sum_{j=1}^{q} b_j x_j + \sum_{i=1}^{q} b_{jj} x_j^2 + \sum\sum_{i<j} b_{ij} x_i x_j + e \\ &= b_0 + x'_i\, b + x'_i\, b x_i + e_{ij}, \end{aligned}$$

where $x_i = (x_{1i}, x_{2i}, ..., x_{iq})'$, $b = (b_1, b_2, ..., b_q)'$.

The second-order model is flexible, because it can take a variety of functional forms and approximates the response surface locally. Therefore, this model is usually a good estimation of the true response surface. Also, as I described in Section 4.1.1, the method of least squares can be applied to estimate the coefficients b_j in a second-order model.

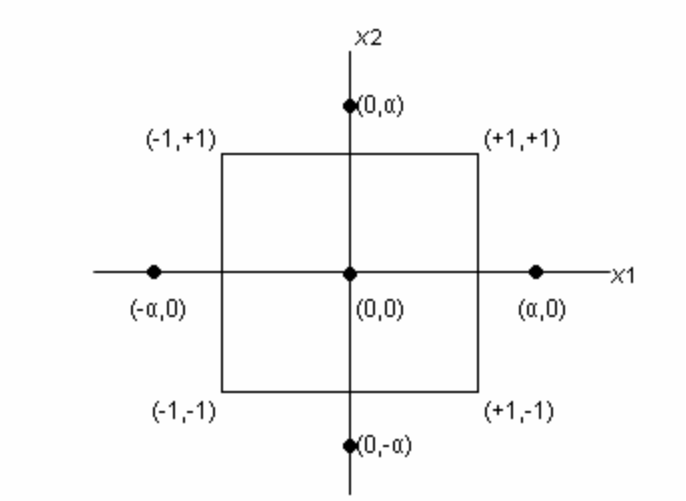
5.2 Designs for Fitting Second-Order Model

There are many designs available for fitting a second-order model. The most popular one is the *central composite design* (CCD). This design was introduced by Box and Wilson. It consists of factorial points (from a 2^q design and 2^{q-k} *fractional factorial design*), central points, and axial points. The following is the representation of $2q$ axial points:

x_1	x_2	…	x_q
-a	0	…	0
a	0	…	0
0	-a	…	0
0	a	…	0
.	.	…	.
.	.	…	.
.	.	…	.
0	0	…	-a
0	0	…	a

CCD was often developed through a *sequential experimentation*. When a first-order model shows an evidence of *lack of fit*, axial points can be added to the quadratic terms with more center points to develop CCD. The number of center points n_c at the origin and the distance *a* of the axial runs from the design center are two parameters in the CCD design. The center runs contain information about the curvature of the surface, if the curvature is significant, the additional axial points allow for the experimenter to obtain an efficient estimation of the quadratic terms. The Figure 5.1 illustrates the graphical view of a central composite design for $q = 2$ factors.

Figure 5.1 Central Composite Design for $q = 2$



There are couples of ways of choosing a and n_c. First, CCD can run in incomplete blocks. A block is a set of relatively homogeneous experimental conditions so that an experimenter divides the observations into groups that are run in each block. An incomplete block design may be conducted when all treatment combinations cannot be run in each block. In order to protect the shape of the response surface, the block effects need to be orthogonal to treatment effects. This can be done by choosing the correct a and n_c in factorial and axial blocks.

Also, a and n_c can be chosen so that the CCD is not blocked. If the precision of the estimated response surface at some point x depends only on the distance from x to the origin, not on the direction, then the design is said to be *rotatable* (Oehlert 2000). When the rotatable design is rotated about the center, the variance of $\hat{y}$ will remain same. Since the reason for using response surface analysis is to located unknown optimization, it makes sense to use a rotatable design that provides equal precision of estimation of the surface in all directions. The choice of a will make the CCD design rotatable by using either $a = 2^{q/4}$ for the full factorial or $a = 2^{(q-k)/4}$ for a fractional factorial.

5.2.1 Orthogonal Central Composite Design

Occasionally, a central composite design may contain only one observation at each of the n_f factorial points and $2q$ axial points, and with n_c observations at the center. This design is known as Khuri and Cornell orthogonal if

$$\left(n_f + 2a^2\right)^2 = n_f n$$

where n is the total number of observations and $n = n_f + 2q + n_c$ (Dean 1999). Orthogonal central composite design with only one observation is achieved by appropriate choice of a and n_c. Consequently, the value of a would be

$$a = \left(\frac{\sqrt{n_f n} - n_f}{2} \right)^{1/2}$$

Recall from the Case Study-1, the test statistic for lack of fit indicated that the initial model for fitting the response variable *Yield* was not adequate. Therefore, Sosada chose to augment the 16 factorial points of the first-order design into a 25-run central composite design (Dean 1999). The data set are as follows:

Table 5.1 Data for Yield of Deoiled Rapeseed Lecithin when Fractionated with Ethonal

A	*B*	*C*	*D*	*Yield*
1	1	1	1	27.6
-1	-1	1	1	16.6
1	-1	-1	1	15.4
-1	1	-1	1	17.4
1	-1	1	-1	17
-1	1	1	-1	19
1	1	-1	-1	17.4
-1	-1	-1	-1	12.6
1	-1	1	1	18.6
-1	1	1	1	22.4
1	1	-1	1	21.4
-1	-1	-1	1	14
1	1	1	-1	24
-1	-1	1	-1	15.6
1	-1	-1	-1	13
-1	1	-1	-1	14
0	0	0	0	22.6
1.414	0	0	0	23.4
-1.414	0	0	0	20.6
0	1.414	0	0	22.6
0	-1.414	0	0	13.4
0	0	1.414	0	20.6
0	0	-1.414	0	15.6
0	0	0	1.414	21
0	0	0	-1.414	17.6

This design contains 25 numbers of observations, and 8 axial points with 1 center point. In order to determine if this design orthogonal central composite design, I applied the test by Khuri and Cornell.

n = 25
n_c = 1
n_f = 16
$2q$ = 8
a = 1.414

$$a = \left(\frac{\sqrt{n_f n} - n_f}{2} \right)^{1/2} = \left(\frac{\sqrt{16 * 25} - 16}{2} \right)^{1/2} = 1.414$$

since, $\left(n_f + 2a^2\right)^2 = n_f n \Rightarrow \left(16 + 2 * 1.414^2\right)^2 = 16 * 25 \quad 400 = 400$

this design is orthogonal.

The analysis of a second-order model is usually done by computer software. The analysis of variance for fitting the data to the second-order and contour plots will help characterize the response surface. In this section, my goal is to fit the second-order model using central composite design. I will investigate the adequacy of the second-order model for *Yield* of deoiled rapeseed lecithin when fractionated with ethanol. The ANOVA and regression analysis for the response variable *Yield* are shown in Figure 5.2.

Figure 5.2 Analysis of Purified Lecithin Yield

Central Composite Design

```
Factors:       4     Replicates:     1
Base runs:    25     Total runs:    25
Base blocks:   1     Total blocks:   1

Two-level factorial: Full factorial

Cube points:             16
Center points in cube:    1
Axial points:             8
Center points in axial:   0

Alpha: 1.414
```

Response Surface Regression: Yield versus A, B, C, D

The analysis was done using coded units.

Estimated Regression Coefficients for Yield

Term	Coef	SE Coef	T	P
Constant	21.4480	0.4313	49.732	0.000
A	1.3180	0.1607	8.200	**0.000**
B	2.6905	0.1607	16.740	**0.000**
C	2.1136	0.1607	13.150	**0.000**
D	1.2604	0.1607	7.842	**0.000**
A*A	0.4200	0.2541	1.653	0.129
B*B	-1.5800	0.2541	-6.217	**0.000**
C*C	-1.5300	0.2541	-6.021	**0.000**
D*D	-0.9300	0.2541	-3.660	**0.004**
A*B	0.7500	0.1797	4.174	**0.002**
A*C	0.3000	0.1797	1.669	0.126
A*D	0.1750	0.1797	0.974	0.353
B*C	0.6000	0.1797	3.339	**0.008**
B*D	0.4750	0.1797	2.643	**0.025**
C*D	-0.0750	0.1797	-0.417	0.685

S = 0.7188   R-Sq = 98.6%   R-Sq(adj) = 96.7%

Analysis of Variance for Yield

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Regression	14	368.056	368.056	26.2897	50.88	0.000
Linear	4	300.637	300.637	75.1593	145.47	0.000
Square	4	47.029	47.029	11.7572	22.76	0.000
Interaction	6	20.390	20.390	3.3983	6.58	0.005
Residual Error	10	5.167	5.167	0.5167		
Total	24	373.222				

The regression equation is

```
Yield = 21.4 + 1.32 A + 2.69 B + 2.11 C + 1.26 D + 0.420 A^2 - 1.58 B^2
        - 1.53 C^2 - 0.930 D^2 + 0.750 AB + 0.300 AC + 0.175 AD + 0.600 BC
        + 0.475 BD - 0.075 CD
```

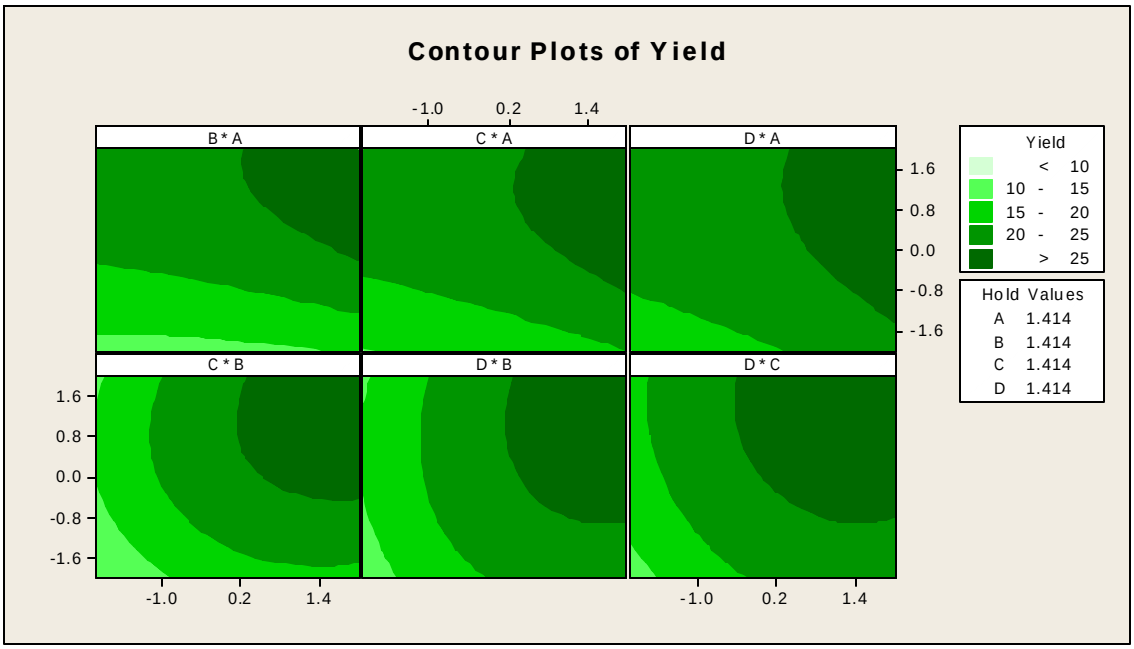
The Minitab computes the linear, quadratic, and interactions terms in the model. My analysis of variance indicates that there are significant interactions between the factors. The small *p*-values for linear and square terms also point out that their contribution is significant to the model. Since, there are no replicated center points; the software cannot obtain a lack-of-fit. But, small *p*-values for the interactions and the squared terms suggest there is curvature in the response surface.

Moreover, the main effects can be referred to as significant at an individual .05 significant level. The quadratic terms, B^2, C^2, and D^2 and interaction terms *AB*, *BC*, and *BD*, significantly contribute to the response model at a = 0.05. As a result, my final model for the response variable *Yield* is concluded as follows:

$$\hat{y}_{Yield} = 21.448 + 1.32A + 2.69B + 2.11C + 1.26D - 1.58B^2 - 1.53C^2 - 0.93D^2 \\ 0.75AB + 0.60BC + 0.48BD$$

Since the response surface is explained by the second-order model, it is necessary to analyze the optimum setting. The graphical visualization is very helpful in understanding the second-order response surface. Specifically, contour plots can help characterize the shape of the surface and locate the optimum response approximately. I graphed the contour plot of purified lecithin *Yield* as is shown in Figure 5.3.

Figure 5.3 Contour Plot of Purified Lecithin *Yield*



Since the response surface is not a plane, it is more complicated to determine the optimum value. But, it appears to be each of the main factors is related to the response variable *Yield* at their high level. At this point, I need a more efficient procedure to find the optimum conditions for the model.

5.3 Analyzing the Stationary Point

The second-order models illustrate quadratic surfaces such as minimum, maximum, ridge, and saddle. If there exits an optimum then this point is a *stationary point*. The stationary point is the combination of design variables where the surface is at either a maximum or a minimum in all directions. If the stationary point is a maximum in some direction and minimum in another direction, then the stationary point is a *saddle point*. When the surface is curved in one direction but is fairly constant in another

direction, then this type of surface is called *ridge* system (Oehlert 2000). The stationary point can be found by using matrix algebra. The fitted second-order model in matrix form is follows:

$$\hat{y} = \hat{b}_0 + x'\boldsymbol{b} + x'\boldsymbol{B}x$$

The derivative of $\hat{y}$ with respect to the elements of the vector x is

$$\frac{\partial \hat{y}}{\partial x} = \boldsymbol{b} + 2\boldsymbol{B}x = 0$$

Therefore, the solution to stationary point is

$$x_s = -\frac{1}{2}\boldsymbol{B}^{-1}\boldsymbol{b}$$

where $\boldsymbol{b} = \begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \\ \vdots \\ \hat{b}_q \end{bmatrix}$ and $\boldsymbol{B} = \begin{bmatrix} \hat{b}_{11}, & \hat{b}_{12}/2, & \cdots, & \hat{b}_{1q}/2 \\ & \hat{b}_{22}, & \cdots, & \hat{b}_{2q}/2 \\ & & \ddots & \\ sym. & & & \hat{b}_{qq}/2 \end{bmatrix}$

$\boldsymbol{b}$ is a (q X 1) vector of the first-order regression coefficients and $\boldsymbol{B}$ is a (q X q) symmetric matrix whose main diagonal elements are the quadratic coefficients ($\hat{b}_{ii}$) and whose off-diagonal elements are one-half the mixed quadratic coefficients (b_{ij} $i \neq j$) (Montgomery 2005). In result, the estimated response value at the stationary point can be calculated as

$$\hat{y}_s = \hat{b}_0 + \frac{1}{2}x'_s\,\boldsymbol{b}$$

Therefore, I used excel to find the location of the stationary point for *Yield*. The calculations are as follows:

B =

0.42	0.375	0.15	0.088
0.375	-1.580	0.3	0.2
0.15	0.3	-1.53	-0.04
0.088	0.2	-0.04	-0.93

b =

1.318
2.691
2.114
1.260

$\mathbf{B}^{-1}$ =

0.42	0.375	0.15	0.0875
0.375	-1.58	0.3	0.2375
0.15	0.3	-1.53	-0.0375
0.0875	0.2375	-0.0375	-0.93

The stationary point using the equation $x_s = -\frac{1}{2}\boldsymbol{B}^{-1}\boldsymbol{b}$ is

$\mathbf{x_s}$ =

-0.995
1.412
1.138
0.249

Please see Attachment – I, II for detailed matrix computations.

At this instant, I can find the stationary point in terms of the natural variables: time, volume, concentration, and temperature.

$$-.995 = \frac{\boldsymbol{t} - 10}{5} \qquad 1.412 = \frac{\boldsymbol{V} - 7.5}{2.5} \qquad 1.138 = \frac{\boldsymbol{Con} - 95}{3} \qquad .249 = \frac{\boldsymbol{T} - 20}{5}$$

These calculations result in $\boldsymbol{t}$ = 5.025 ˜ 5 minutes of reaction time, $\boldsymbol{V}$ = 11.029 ˜ 11 liter solvent volume, ***Con*** = 98.414 ˜ 98 percent of ethanol concentration, and $\boldsymbol{T}$ = 21.243 ˜ 21 ^{0}C temperature. Using the equation $\hat{y}_s = \hat{b}_0 + \frac{1}{2}x'_s\,\boldsymbol{b}$, I can find that estimated maximum response *Yield* of deoiled rapeseed lecithin at the stationary point is

$$\hat{y}_{yield} = 24.05$$

Thus, I can conclude that this level of main factors setting will result in best optimum solution for the *Purified Lecithin Yield* of deoiled rapeseed lecithin when fractionated with ethanol.

5.4 Conclusion of the Second-Order Model

When the first-order model shows a significant lack of fit, then an experimenter can use a second-order model to describe the response surface. There are many designs available to conduct a second-order design. The central composite design is one of the most popular ones. An experimenter can start with 2^q factorial point, and then add center and axial points to get central composite design. Adding the axial points will allow quadratic terms to be included into the model. Second-order model describes quadratic surfaces, and this kind of surface can take many shapes. Therefore, response surface can represent maximum, minimum, ridge or saddle point. Contour plot is a helpful visualization of the surface when the factors are no more than three. When there are more than three design variables, it is almost impossible to visualize the surface. For that reason, in order to locate the optimum value, one can find the stationary point. Once the stationary point is located, either an experimenter can draw a conclusion about the result or continue in further studying of the surface.

6. Three-level Fractional Factorial Design

6.1 The 3-level Factorial Design

In addition to the second-order model, when the curvature in the response surface is concerned, an experimenter can design a model using a three-level factorial design. The factorial designs are widely used in experiments, when an experimenter needs to evaluate the joint effects of several controllable factors on the response. The 3^q factorial design is a factorial arrangement with q factors, each at three levels. The levels of factor refer to as low, intermediate, and high, represented by the digit 0 (low), 1 (intermediate), and 2 (high). For instance, in a 3^3 design, 021 indicates the treatment combination corresponding to factor *A* at the low level, *B* at the high level, and *C* at the intermediate level. When the measurements on the response variable contain all possible combinations of the levels of the factors, this type of experimental design is called *a complete factorial experiment*,

In general, the 3^q design require many runs, therefore it is unlikely that all 3^q runs can be carried out under homogeneous conditions. As a result, the confounding in blocks is unavoidable. A complete factorial experiment can be placed in the blocks of unit, where units in the same block are homogeneous. This type of the design technique is called *confounding*. The complete blocks include every treatment in every block; on the contrary, the incomplete blocks do not include all the treatments or treatment combinations in each block. The incomplete blocks are less efficient than complete blocks due to the lose of some information (usually the higher order interactions). Meanwhile, confounded factorials will tolerate more efficient result in main effects and low-order interactions.

The 3^q design can be confounded in 3^s blocks, each with 3^{q-s} units, where $q > s$. For instance, suppose that the $q = 3$ and $s = 2$. The 3^3 factorial design is confounded in $3^3 = 9$ incomplete blocks, each with $3^{3-2} = 3^1$ units. First, it is necessary to define a contrast by choosing a factorial effect to confound with blocks. The general defining contrast is

$$L = a_1 x_1 + a_2 x_2 + \ldots\ldots + a_q x_q,$$

where a_i represents the exponents on the i^{th} factor in the effect to be confounded and x_i is the level of the i^{th} factor in a particular treatment combination (Montgomery 2005). Thus, x_i takes the values of 0 (low level), 1 (intermediate level), or 2 (high level), where a_i is 0, 1, or 2. At this point, before I study more complex 3-level factorial design, I would like to construct a small example. For example, if I let AB^2 and AC to be the two components of interaction chosen to construct the design. The two defining contrasts for assigning runs to blocks are

$$L_1 = x_A + 2x_B = u \ \textbf{(mod 3)} \qquad u = 0, 1, 2$$
$$L_2 = x_A + x_C = h \ \textbf{(mod 3)} \qquad h = 0, 1, 2$$

The L equations can take only the values of 0, 1, or 2 because of L (mod 3). As a result, the treatment combinations in the 3^3 design assigned to the blocks based on the values of u and h, denoted as u/h block. For example, the 121 treatment combination has an u value of

$$u = 1(1) + 2(2) + 0(1) = 5 \text{ (mod 3)} = 2$$

and has an h value of

$$h = 1(1) + 2(0) + 1(1) = 2$$

Therefore, the treatment combination 121 will be assigned to the 2/2 block. I calculated the rest of the values of treatment combinations and their assigned block *u*/*h* using Excel. The results are shown in Figure 6.1.

Figure 6.1 Confounding a 3^3 Design in 9 blocks

Treatments

A	*B*	*C*	*u*	*h*
0	0	0	0	0
0	0	1	0	1
0	0	2	0	2
0	1	0	2	0
0	1	1	2	1
0	1	2	2	2
0	2	0	1	0
0	2	1	1	1
0	2	2	1	2
1	0	0	1	1
1	0	1	1	2
1	0	2	1	0
1	1	0	0	1
1	1	1	0	2
1	1	2	0	0
1	2	0	2	1
1	2	1	2	2
1	2	2	2	0
2	0	0	2	2
2	0	1	2	0
2	0	2	2	1
2	1	0	1	2
2	1	1	1	0
2	1	2	1	1
2	2	0	0	2
2	2	1	0	0
2	2	2	0	1

0 / 0	**0 / 1**	**0 / 2**
0 0 0	0 0 1	0 0 2
1 1 2	1 1 0	1 1 1
2 2 1	2 2 2	2 2 0
1 / 0	**1 / 1**	**1 / 2**
0 2 0	0 2 1	0 2 2
1 0 2	1 0 0	1 0 1
2 1 1	2 1 2	2 1 0
2 / 0	**2 / 1**	**2 / 2**
0 1 0	0 1 1	0 1 2
1 2 2	1 2 0	1 2 1
2 0 1	2 0 2	2 0 0

The block where the treatment combinations satisfying $u = 0$ and $h = 0$ is called a *principal block*, that is 0/0 block. A principal block will always include the treatment combination 000…0 represented by ***I***. The principal block ***I*** act as an identity, that is, anything added by ***I*** is just itself. In this example, the principal block 0/0 contains the treatment combinations 000, 112, and 221. In general, the treatment combinations in the

principal block form a group with respect to addition modulus 3, and it is called a *group-theoretic* property. This implies that any element in the principal block may be generated by the addition of two other elements in the principal block modulus 3. The operation $\oplus$ is used to add the factor levels individually and reduce module 3. Referring to the Figure 6.1, one can see that $112 \oplus 221 = 000$, $112 \oplus 112 = 221$, and $221 \oplus 221 = 112$. Also, treatment combinations in any of the other blocks may be generated by adding one element in that block by each element in the principal block modulus 3. For instance, since 100 is one of the other blocks, elements of 1/1 block can be computed as

$$000 \oplus 100 = 100 \pmod 3$$

$$112 \oplus 100 = 212 \pmod 3$$

$$221 \oplus 100 = 021 \pmod 3.$$

Confounding a three-series design into nine blocks uses two components of interaction. Thus, eight degrees of freedom will be confounded with blocks. The four degrees of freedom confounded along with the components of interaction AB^2 and AC. Therefore, the additional four degrees of freedom are from the *generalized interactions* of the defining effects. These interactions can be written in a three series with exponents of 0, 1, or 2, with the first nonzero exponent always being a 1. If the first letter exponent is not 1, the entire expression is squared and the exponents are reduced modulus 3. As a result, if P_1 and P_2 are defining effects, then their generalized interactions are $P_1 * P_2$ and $P_1 * P_2^2$. The generalized interaction of AB^2 and AC are

$$P_1 P_2 = (AB^2)(AC) = (A^2 B^2 C)^2 \qquad \text{the leading exponent is 2, so square it}$$

$$= A^4 B^4 C^2$$

$$= ABC^2$$

$P_1P_2^2 = (AB^2)(AC)^2 = A^3B^2C^2$ reduce exponents modulo 3

$= B^2C^2 = (B^2C^2)^2$ the leading exponent is 2, so square it

$= BC$

More generally, when there are s independent defining contrasts, then $(3^s - 2s - 1)/2 = p$ other effects are automatically confounded due to their generalized interactions with original effects.

The one concern about 3^q design is that it can require a large number of runs even for moderate values of q. For instance, consider a 3^9 design with a single replicate would have 19,683 observations. If the design is confounded in $3^{9-6} = 27$ incomplete blocks, then each block will require 27 observations. Therefore, the *fractional factorial* design might be an alternative approach when dealing with a large number of factors.

6.2 The 3-level Fractional Factorial Design

A *fractional factorial* design is a revision of a factorial design without having to run the full factorial design. The fractional factorial design partitions full 3^q runs into blocks, but running only one of the blocks. This design allows an experimenter to get information on the main effects and the low-order interactions. A fractional factorial model can be conducted to study the response surface. I worked out the following Case Study using a 3-level fractional factorial design.

……. Case Study 2 ……

The proposed design and analysis strategy is illustrated with the data from a 27-run experiment (Taguchi 1987), which was from a study about the PVC insulation for electric

wire. The objective of the study is to understand the compounding method of plasticizer, stabilizer, and filler for avoiding embrittlement of PVC insulation, as well as to find the most suitable process conditions. All nine factors are continuous and their levels are chosen to be equally spaced. Among the factors, two are about plasticizer: DOA (denoted by *A*) and n-DOP (*B*); two about stabilizer: Tribase (*C*) and Dyphos (*D*); three about filler: Clay (*E*), Titanium white (*F*), and Carbon (*G*); the remaining two are about the process condition: the number of revolutions of screw (*H*) and the cylinder temperature (*J*). The measured response is the embrittlement temperature. The design matrix and data are given in Table 6.1.

Table 6.1. Design matrix and response data, PVC insulation data.

run	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*J*	response
1	0	0	0	0	0	0	0	0	0	5
2	0	0	0	0	1	1	1	1	1	2
3	0	0	0	0	2	2	2	2	2	8
4	0	1	1	1	0	0	0	2	2	-15
5	0	1	1	1	1	1	1	0	0	-6
6	0	1	1	1	2	2	2	1	1	-10
7	0	2	2	2	0	0	0	1	1	-28
8	0	2	2	2	1	1	1	2	2	-19
9	0	2	2	2	2	2	2	0	0	-23
10	1	0	1	2	0	1	2	0	1	-13
11	1	0	1	2	1	2	0	1	2	-17
12	1	0	1	2	2	0	1	2	0	-7
13	1	1	2	0	0	1	2	2	0	-23
14	1	1	2	0	1	2	0	0	1	-31
15	1	1	2	0	2	0	1	1	2	-23
16	1	2	0	1	0	1	2	1	2	-34
17	1	2	0	1	1	2	0	2	0	-37
18	1	2	0	1	2	0	1	0	1	-29
19	2	0	2	1	0	2	1	0	2	-27
20	2	0	2	1	1	0	2	1	0	-27
21	2	0	2	1	2	1	0	2	1	-30
22	2	1	0	2	0	2	1	2	1	-35
23	2	1	0	2	1	0	2	0	2	-35
24	2	1	0	2	2	1	0	1	0	-38
25	2	2	1	0	0	2	1	1	0	-39
26	2	2	1	0	1	0	2	2	1	-40
27	2	2	1	0	2	1	0	0	2	-41

There are nine factors: *A, B, C, D, E, F, H, G,* and *J,* each at level 0, 1, or 2. Since a 3^9 design is costly, the model was designed using fraction. This model is a 3^{9-6} fractional factorial design with 27 runs. Thus, a 3^{9-6} design is the 1/729 fraction of the 3^9 design, where the fraction contains 3^{9-6} runs. In general, a $\frac{1}{3^s} * 3^q = 3^{q-s}$ design is the $\frac{1}{3^s}$ fraction of the 3^q design for $q > s$, where the fraction contains 3^{q-s} runs. In order to construct a 3^{q-s} fractional factorial design, the treatment combinations are grouped into blocks. Firstly, *s* components of interactions should be selected, and then 3^q treatment combinations should be partitioned into 3^s blocks, each with 3^{q-s} units. Secondly, the generalized interactions of *s* effects should be identified. This experimental plan, as 3^q design, will allow homogenous grouping of the experimental material in blocks.

Alternatively, a 3^{q-s} fractional factorial design can be constructed by writing down the treatment combination of full 3^{q-s} factorial design and then introducing the additional *s* factors by equating them to the components of interactions. The group formed by the *s* defining words (*generator*) is called *defining contrast subgroup*. Let 1, 2,…,*q-s* denote the *q-s* independent columns of the 0, 1, 2's that generate the 3^{q-s} runs in the design. The remaining *s* columns, *q-s*+1,…, *q*, can be generated as the interactions of the first *q-s* columns (Wu 2000). For example, the procedure for constructing a one-ninth fraction of the 3^3 design is to write down the 3^2 full factorial design in the factors *A* and *B* as in Figure 6.2. If I let AB^2C^2 be the word or the generator of the design, then the factor *C* can be represent by the notation

$$C = AB^2.$$

More generally, if exponents of AB^2C^2 are $a_1 = 1, a_2 = a_3 = 2$, where $d_1 = (3 - a_3)a_1$ = (3 – 2) (1) (mod 3) = 1, $d_2 = (3 - a_3)a_2$ = (3 – 2) (2) = 2 (mod 3), the levels of x_3 can be equated by the following relationship

$$x_3 = d_1 x_1 + d_2 x_2$$

which follows

$$x_3 = 1x_1 + 2x_2 .$$

The Figure 6.2 illustrates the 3^{3-1}design with the defining relation AB^2C^2.

Figure 6.2 3^{3-1} fractional design

A	*B*	$C = AB^2$
0	0	0
0	1	2
0	2	1
1	0	1
1	1	0
1	2	2
2	0	2
2	1	1
2	2	0

A second approach can be used to construct a 3^{9-6} fractional design in PVC insulation for electric wire. Thus, the factors *A*, *B*, and *E* are used to define the treatment combinations of a full 3^3 design. The rest of the six columns found by equating $s = 6$ factors to components of interactions; A^2B^2C, AB^2D, A^2E^2F, AE^2G, BE^2H, and A^2BE^2J. The levels of the last six factors satisfy these following equations:

$$x_C = x_A + x_B$$

$$x_D = 2x_A + x_B$$

$$x_F = x_A + x_E$$

$$x_G = 2x_A + x_E$$

$$x_H = 2x_B + x_E$$

$x_J = x_A + 2x_B + x_E$

For this experiment, the factor *C* is assigned to the column that equals the addition of the columns for factor *A* and *B*. Since the column for *C* is used to estimate the main effect of *C* and also for the interaction effect between *A* and *B*, consequently the data from such design is not capable of distinguishing the estimate of *C* from the estimate of *AB* (Wu 2000). In fact, when *C* is estimated, it is actually *C*+*AB* that is estimated. In this case, the factor of main effect *C* is said to be *aliased* with the *AB* interaction. This aliasing relation is denoted below.

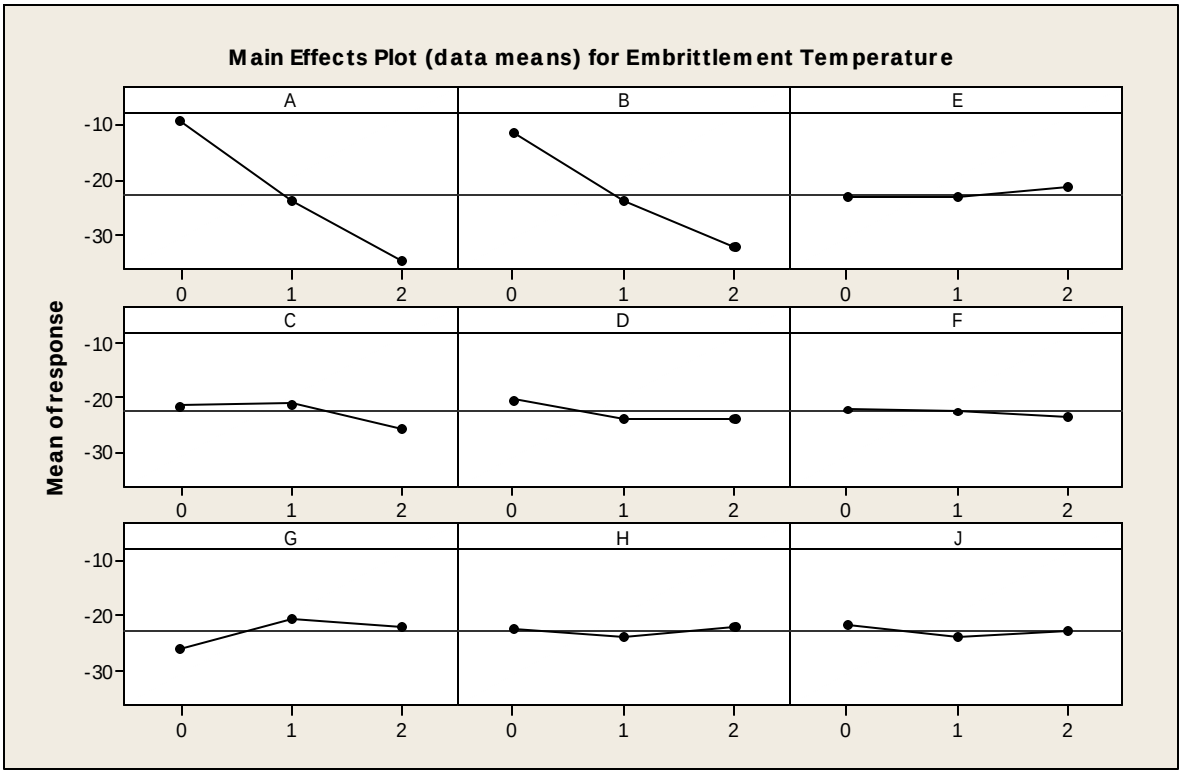
$C = AB$ or $\boldsymbol{I} = ABC^2$,

where $\boldsymbol{I}$ is the identity in group theory since $0 + z = z$ for any z. The other alias relations for effects are: $D = A^2B$, $F = AE$, $G = A^2E$, $H = B^2E$, and $J = AB^2E$. The one-729th fraction is defined by $\boldsymbol{I} = ABC^2 = A^2BD^2 = AEF^2 = A^2EG^2 = B^2EH^2 = AB^2EJ^2$. In order to obtained the rest of the defining contrast group I noted the complexity of the alias relationship in 3^{9-6} fractional factorial design. If a 3^{9-6} has 6 generators P_1, P_2, P_3, P_4, P_5, and P_6, then the each constant is aliased to $1 + 3^1 + 3^2 + 3^3 + 3^4 + 3^5 = 364$ splits, noting that $s - 1 = 5$; these splits aliased to $\boldsymbol{I}$ are of the form $P_1^{i_1} P_2^{i_2} P_3^{i_3} P_4^{i_4} P_5^{i_5} P_6^{i_6}$ where exponents are 0, 1, or 2, and the first nonzero exponent is 1. The rest of the splits are aliased to $3^s - 1 = 3^6 - 1 = 728$ other splits. Furthermore, the aliases of a split *W* are products of the form $W\left(P_1^{i_1} P_2^{i_2} P_3^{i_3} P_4^{i_4} P_5^{i_5} P_6^{i_6}\right)$, where the exponents i_j are consent to range over all $3^6 = 729$ combinations of 0, 1, or 2. Therefore, there are $1 + 3^1 + 3^2 + 3^3 + 3^4 = 121$ sets of aliases in addition to the aliases of $\boldsymbol{I}$ (Dehlert 2000), noting that $q - s - 1 = 4$. In result, each of the 121 sets of aliases has 729 names of interactions.

6.2.1 Analysis of Three-level Fractional Factorial Design

When the design deals with complex aliasing, it is very complicated to separate the large number of aliased effects and to interpret their significance. Therefore, as in Case Study - 2, this kind of method can be used for *screening design*s. They are used to estimate the main effects but not their interactions. For this reason, using Minitab, I can determine the significance of the main factors. The simple analysis starts with the main effects plot. A main effects plot is a plot of the means of the response variable for each level of a factor, which allows me to obtain a general idea of the possibly important main effects. I showed the locations of main effects for PVC insulation for electric wire in Figure 6.3.

Figure 6.3 Main Effects Plot of Strength Embrittlement Temperature



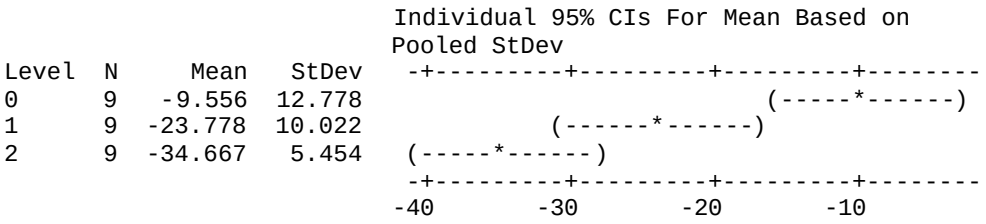
Analysis of the above main effect plots indicates that a main effect occurs when the mean response changes across the levels of a factor. Therefore, I can identify the strength of the effects of embrittlement temperature across factors by using the main effects plots as stated below.

- Factors ***A*** and ***B*** decrease when they move from the high level to the low level of embrittlement temperature.
- Factors ***D***, ***E***, ***F***, ***H***, and ***J*** remain practically the same when they move from the high level to the low level of embrittlement temperature.
- Factors ***C*** and ***G*** increase when they move from the low level to the intermediate level and then decrease from the intermediate level to the high level of embrittlement temperature.

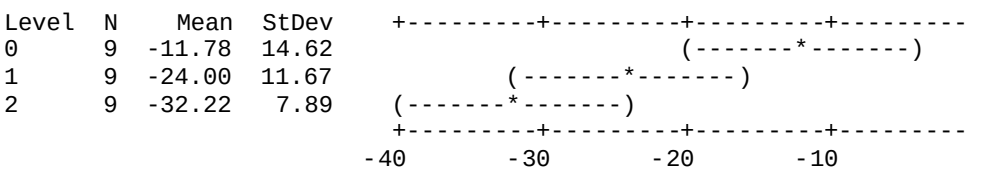
My analysis concluded that the levels of factors *D, E, F, H,* and *J* affect the response in a similar way. It seems no visible main effect is present, since the lines are almost parallel to the x-axis. On the other hand, the levels of factors *A, B, C,* and *G* appear to affect the response differently. The levels of *A* and *B* factors have larger difference in the vertical position of the plotted points, that is, steeply slopes. Consequently, the levels of these factors appear to have a greater affect on the response embrittlement temperature. I also showed the levels of each factor's mean values that are given in Table 6.2.

Table 6.2 Mean values of each levels of factors

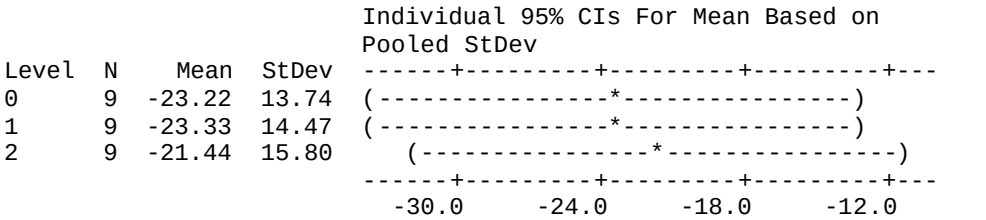
Mean values of factor A



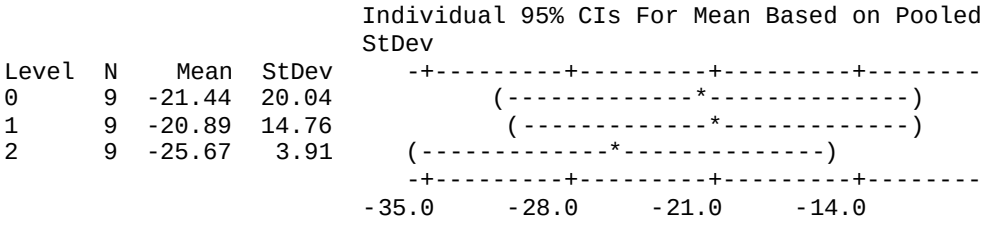
Mean values of factor B



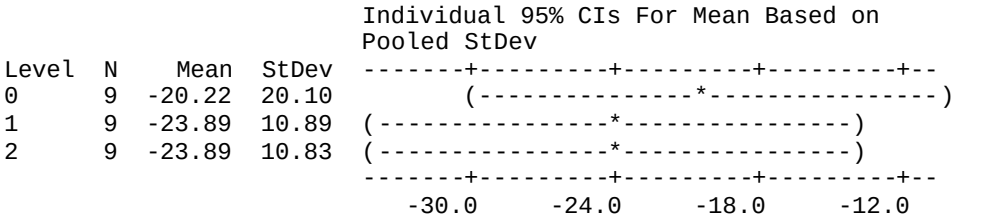
Mean values of factor E



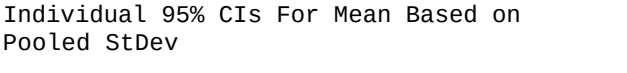
Mean values of factor C



Mean values of factor D



Mean values of factor F



```
Level  N    Mean  StDev  ------+---------+---------+---------+---
0      9  -22.11  14.17    (----------------*----------------)
1      9  -22.44  14.77    (----------------*----------------)
2      9  -23.44  15.17  (----------------*----------------)
                         ------+---------+---------+---------+---
                           -30.0     -24.0     -18.0     -12.0
```

Mean values of factor G

```
                         Individual 95% CIs For Mean Based on Pooled
                         StDev
Level  N    Mean  StDev     -+---------+---------+---------+--------
0      9  -25.78  14.57     (-------------*-------------)
1      9  -20.33  14.03               (-------------*-------------)
2      9  -21.89  14.95          (--------------*-------------)
                            -+---------+---------+---------+--------
                         -35.0     -28.0     -21.0     -14.0
```

Mean values of factor H

```
                         Individual 95% CIs For Mean Based on
                         Pooled StDev
Level  N    Mean  StDev  ------+---------+---------+---------+---
0      9  -22.22  14.80    (----------------*----------------)
1      9  -23.78  13.54  (----------------*----------------)
2      9  -22.00  15.68    (----------------*----------------)
                         ------+---------+---------+---------+---
                           -30.0     -24.0     -18.0     -12.0
```

Mean values of factor J

```
                         Individual 95% CIs For Mean Based on
                         Pooled StDev
Level  N    Mean  StDev  ------+---------+---------+---------+---
0      9  -21.67  15.84     (----------------*----------------)
1      9  -23.78  13.67  (----------------*----------------)
2      9  -22.56  14.49    (----------------*----------------)
                         ------+---------+---------+---------+---
                           -30.0     -24.0     -18.0     -12.0
```

Although a table of means and a plot of main effects provide useful information, in order to confirm the results I need to use a more formal analysis of the data. Therefore, once again I can use Minitab to obtain statistically significant analysis of the data. The analysis of variance of PVC insulation data is as follows:

Figure 6.4 Analysis of variance of PVC insulation

Response Surface Regression: response versus A, B, E, C, D, F, G, H, J

The analysis was done using coded units.

Estimated Regression Coefficients for response

Term	Coef	SE Coef	T	P
Constant	2.2222	2.8034	0.793	0.439
A	-12.5556	0.9017	-13.925	**0.000**
B	-10.2222	0.9017	-11.337	**0.000**
E	0.8889	0.9017	0.986	0.338
C	-2.1111	0.9017	-2.341	***0.032***
D	-1.8333	0.9017	-2.033	0.058
F	-0.6667	0.9017	-0.739	0.470
G	1.9444	0.9017	2.157	***0.046***
H	0.1111	0.9017	0.123	0.903
J	-0.4444	0.9017	-0.493	0.628

S = 3.825 R-Sq = 95.2% R-Sq(adj) = 92.7%

Analysis of Variance for response

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Regression	9	4953.22	4953.22	550.358	37.61	0.000
Linear	9	4953.22	4953.22	550.358	37.61	0.000
Residual Error	17	248.78	248.78	14.634		
Total	26	5202.00				

In order to determine which factors are significantly related to the response, I can use the least squares regression to analyze the variability of data. The Figure 6.4 provides a statistical summary of the main effects. The main effects *A, B, C,* and *G* are significant at the 0.05 α-level. Meanwhile the main effects *D, E, F, H,* and *J* do not appear to contribute to the response at the 0.05 α-level. This result confirms my earlier conclusion completed by using the main effect plots. The graphical analysis of the effects allows me to visually identify the important effects, while the statistical analysis confirms which factors are significantly related to the response.

My next step is to determine interactions effects that are significant. In order for an interaction to be significant, at least one of its parent factors should be significant (Wu 2000). This fundamental principle for factorial effects is called the *effect heredity*

principle. Since four of the parent factors are identified as significant, I cannot rule out any of the interactions between the main effects. Hence, the analysis can start with interaction plots as shown in Figure 6.5. This graph displays a full interactions plot matrix. Each pair of factors provides the summary below:

- ***A* and *B***: Both of the rows indicate that factors *A* and *B* interact.

  Row 1: The lines for the three levels of factors *A* decrease but at different rates while the level of factor *B* increases.

  Row 2: The lines for the three levels of factors *B* decrease but at different rates while the level of factor *B* increases.

- ***A* and *C***: Both of the rows indicate that factors *A* and *C* interact.

  Row 1: The level 0 decreases, the level 1 first increases then it decreases, and the level 2 first decreases and then increases for factor *A*, while the level of *C* increases.

  Row 2: The level 0 and the level 1 decrease at different rates, and the level 2 stays about the same as *C* increases, while the level of *A* increases.

- ***B* and *C***: Both of the rows indicate that factors *B* and *C* interact.

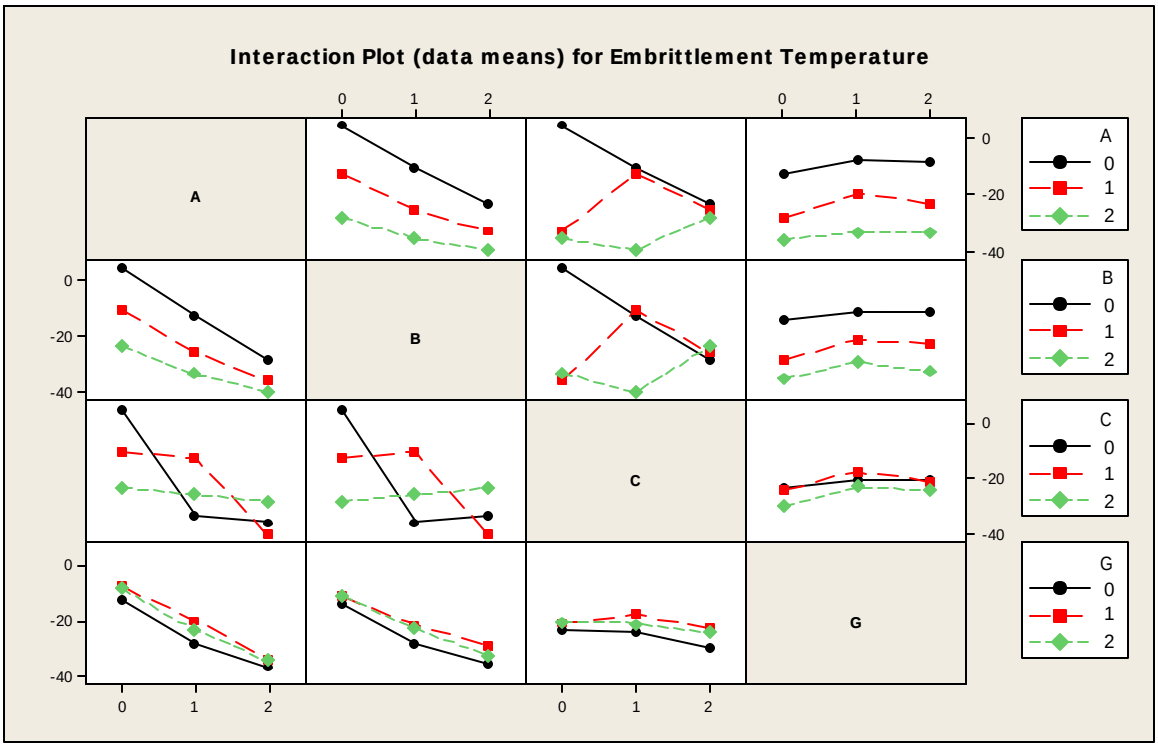
  The movement of the levels of *B* and *C* is almost identical to the movement of the levels of *A* and *C*.

- ***A* and *G* – *B* and *G* – *C* and *G***:

  All three lines for each of the factor interactions are approximately parallel.

  The factor *G* does not appear to be interacting with factors *A*, *B*, and *C*.

Figure 6.5 Interaction Plot of Embrittlement Temperature



In addition to estimating the eight degrees of freedom in the main effects *A*, *B*, *C*, and *G*, there are 26 – 8 = 18 degrees of freedom left for estimating two-factor interactions and the error term. Each two-factor interaction has 4 degrees of freedom. Consequently, there will not be enough degrees of freedom for estimating six of the two-factor interactions using ANOVA. An alternative method is to decompose each of the main effects into linear and quadratic components: linear versus linear, linear versus quadratic, quadratic versus linear, or quadratic versus quadratic. Note that if y_0, y_1, and y_2 correspond to the observations at level 0, 1, and 2, the linear effect is defined as

$$y_2 - y_1$$

and the quadratic effect as

$$(y_2 + y_0) - 2y_1.$$

Furthermore, the linear and quadratic effects are represented by two mutually orthogonal vectors:

$$A_l = \frac{1}{\sqrt{2}}(-1,0,1), \qquad A_q = \frac{1}{\sqrt{6}}(1,-2,-1)$$

$\sqrt{2}$ and $\sqrt{6}$ are the scaling constants and they will be dropped in the table for simplicity (Wu 2000). Therefore, I can apply the formulas above to the columns of *A*, *B*, *C*, and *G* in Table 6.1 to facilitate A_l, A_q, B_l, B_q, C_l, C_q and G_l, G_q. For instance, I can obtain the column of AB_{ll} by multiplying A_l and A_q. The rest of the two-way interactions are obtained similarly. The decomposed main effects are given in Table 6.3.

Table 6.3 The decomposition of Main Effects

A_l	A_q	B_l	B_q	C_l	C_q	G_l	G_q	$(AB)_{ll}$	$(AC)_{ll}$	$(AG)_{ll}$	BC_{ll}	$(BG)_{ll}$	$(CG)_{ll}$
-1	1	-1	1	-1	1	-1	1	1	1	-1	1	1	1
-1	1	-1	1	-1	1	0	-2	1	1	2	1	0	0
-1	1	-1	1	-1	1	1	1	1	1	-1	1	-1	-1
-1	1	0	-2	0	-2	-1	1	0	0	-1	0	0	0
-1	1	0	-2	0	-2	0	-2	0	0	2	0	0	0
-1	1	0	-2	0	-2	1	1	0	0	-1	0	0	0
-1	1	1	1	1	1	-1	1	-1	-1	-1	1	-1	-1
-1	1	1	1	1	1	0	-2	-1	-1	2	1	0	0
-1	1	1	1	1	1	1	1	-1	-1	-1	1	1	1
0	-2	-1	1	0	-2	1	1	0	0	0	0	-1	0
0	-2	-1	1	0	-2	-1	1	0	0	0	0	1	0
0	-2	-1	1	0	-2	0	-2	0	0	0	0	0	0
0	-2	0	-2	1	1	1	1	0	0	0	0	0	1
0	-2	0	-2	1	1	-1	1	0	0	0	0	0	-1
0	-2	0	-2	1	1	0	-2	0	0	0	0	0	0
0	-2	1	1	-1	1	1	1	0	0	0	-1	1	-1
0	-2	1	1	-1	1	-1	1	0	0	0	-1	-1	1
0	-2	1	1	-1	1	0	-2	0	0	0	-1	0	0
1	1	-1	1	1	1	0	-2	-1	1	-2	-1	0	0
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	0	-2	-1	1	0	-2	0	-1	-2	0	0	0
1	1	0	-2	-1	1	1	1	0	-1	1	0	0	-1
1	1	0	-2	-1	1	-1	1	0	-1	1	0	0	1
1	1	1	1	0	-2	0	-2	1	0	-2	0	0	0
1	1	1	1	0	-2	1	1	1	0	1	0	1	0
1	1	1	1	0	-2	-1	1	1	0	1	0	-1	0

As a result of my table, I can use regression analysis to identify the two-factor interactions that are significant.

Figure 6.6 Regression Analysis of factors Main Factors and Two-way Interactions

Regression Analysis: response versus Al, Aq, ...

```
* (BC)ll is highly correlated with other X variables
* (BC)ll has been removed from the equation.


The regression equation is
response = - 22.7 - 12.6 Al + 0.556 Aq - 10.2 Bl + 0.667 Bq - 0.278 Cl
           - 0.278 Cq + 1.94 Gl - 1.17 Gq + 3.67 (AB)ll - 0.00 (AC)ll
           + 0.222 (AG)ll - 0.083 (BG)ll + 0.583 (GC)ll
```

Predictor	Coef	SE Coef	T	P
Constant	-22.6667	0.5115	-44.31	0.000
Al	-12.5556	0.6265	-20.04	**0.000**
Aq	0.5556	0.3617	1.54	0.149
Bl	-10.2222	0.8859	-11.54	**0.000**
Bq	0.6667	0.4176	1.60	0.134
Cl	-0.2778	0.8859	-0.31	0.759
Cq	-0.2778	0.4176	-0.67	0.518
Gl	1.9444	0.6265	3.10	**0.008**
Gq	-1.1667	0.3617	-3.23	**0.007**
(AB)ll	3.667	1.253	2.93	**0.012**
(AC)ll	-0.000	1.253	-0.00	1.000
(AG)ll	0.2222	0.4430	0.50	0.624
(BG)ll	-0.0833	0.7673	-0.11	0.915
(GC)ll	0.5833	0.7673	0.76	0.461

```
S = 2.65784   R-Sq = 98.2%   R-Sq(adj) = 96.5%
```

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	13	5110.17	393.09	55.65	0.000
Residual Error	13	91.83	7.06		
Total	26	5202.00			

It appears that BC_{ll} component is not included in the analysis. Most computer programs will print out an error message indicating that they are unable to estimate the coefficients of the collinear variables. The collinearity occurs when the relative movements of one variable will be matched exactly by the relative movement of the other variable (Studenmund 2006). As I mentioned earlier via interactions plot in Figure 6.5,

the two-way interactions AC_{ll} and BC_{ll} are completely explained by each other's movements. As a result, I can exclude one of the redundant interactions from the model.

Also, the main effect *C* was found significant as shown in Figure 6.4, but when I reduced the model, the factor *C* became highly insignificant. Recall that *C* is aliased with *AB*, therefore when using Minitab to estimate *C*, it also estimates the interaction effect between *A* and *B*. The result of adding the two-way interaction of *AB* in the reduced model caused components of *C* to become insignificant. Consequently, since the main effects *A* and *B* and their interaction AB_{ll} are highly significant, it is appropriate to remove the components of *C* from the data. The components A_l, B_l, G_l, G_q, and AB_{ll} appear to be significant at the 0.05 α-level. At this point, I can analyze the response surface by fitting the data in a second-order model.

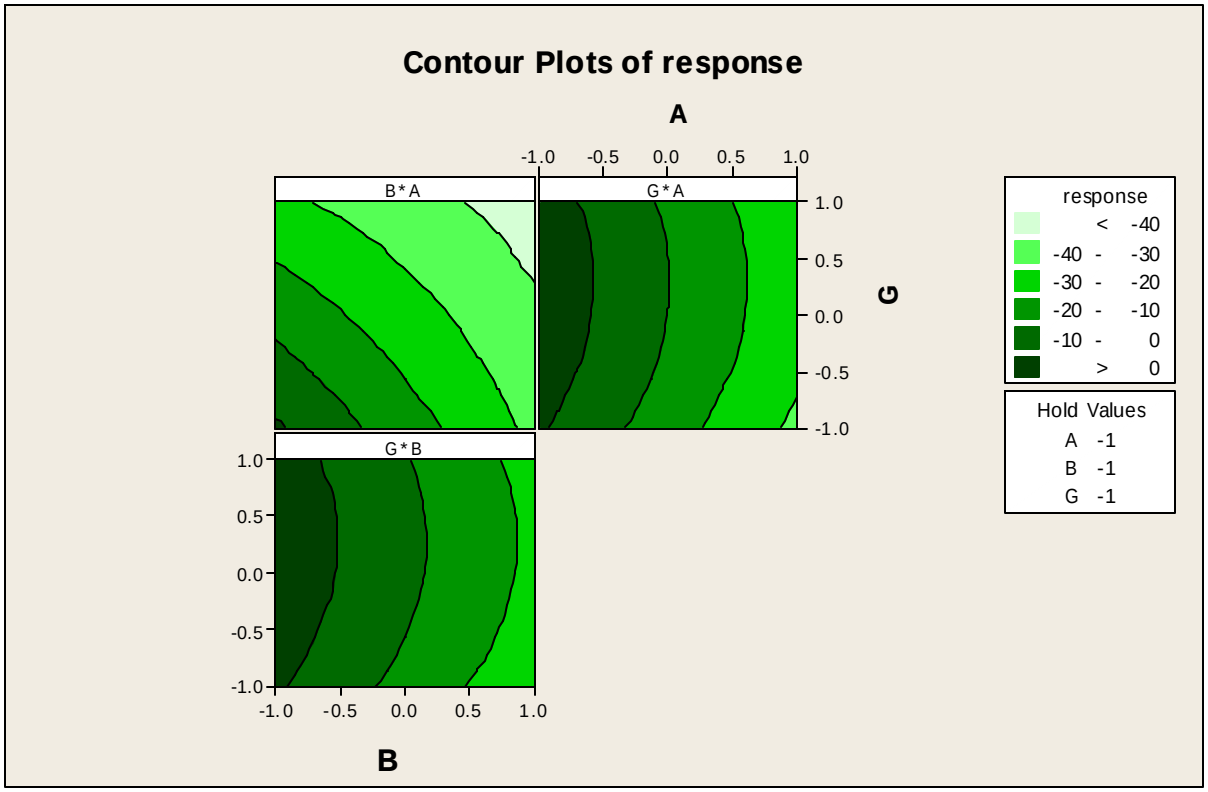
Figure 6.7 The Final Model

Response Surface Regression: response versus A, B, G

```
The analysis was done using coded units.

Estimated Regression Coefficients for response

Term         Coef  SE Coef        T      P
Constant  -20.333   0.8665  -23.465  0.000
A         -12.556   0.6127  -20.491  0.000
B         -10.222   0.6127  -16.683  0.000
G           1.944   0.6127    3.173  0.005
G*G        -3.500   1.0613   -3.298  0.003
A*B         4.083   0.7504    5.441  0.000

S = 2.600   R-Sq = 97.3%   R-Sq(adj) = 96.6%


Analysis of Variance for response

Source          DF   Seq SS   Adj SS   Adj MS       F      P
Regression       5  5060.08  5060.08  1012.02  149.75  0.000
  Linear         3  4786.50  4786.50  1595.50  236.09  0.000
  Square         1    73.50    73.50    73.50   10.88  0.003
  Interaction    1   200.08   200.08   200.08   29.61  0.000
Residual Error  21   141.92   141.92     6.76
Total           26  5202.00
```

```
Estimated Regression Coefficients for response using data in uncoded units

Term          Coef
Constant   1.08333
A         -16.6389
B         -14.3056
G          8.94444
G*G       -3.50000
A*B        4.08333
```

Figure 6.8 Contour plot of Embrittlement Temperature



My best regression model to describe the relationship between the embrittlement temperature and the factors is

$$y = -20.33 - 12.56x_A - 10.22x_B + 4.08x_Ax_B + 1.94x_G - 3.50x_G^2$$

A contour plot can show only two factors at a time, for that reason each factor is held at a constant level. In order to avoid the embrittlement temperature of PVC insulation of an electric wire, the analysis of response surface indicates that the plasticizers DOA (*A*) and n-DOP (*B*) should set to level (-1, 0) and carbon (*G*) should set to level (0, 1).

6.3 The Conclusion of the Three-Level Fractional Factorial Design

The factorial designs are widely used in experiments when the curvature in the response surface is concerned. All treatment factors have 3-levels in the three-level factorial design. This design requires many runs, as a result, the confounding in blocks can be used. Also, the fractional factorial design can be an alternative approach when the number of factors gets large.

The three-level fractional factorial design partitions the full 3^q runs into blocks, but it only runs one of the blocks. This design is more efficient, it allows collecting information on the main effects and on the low-order interactions. The one problem with three-level fractional factorial is that when number of factors is large, it becomes very complicated to separate the aliased effects and to interpret their significance. For this reason, when q is large, most of the time this kind of design is used for screening designs. After an appropriate design is conducted, the response surface analysis can be done by any statistical computer software and then statistical analyses can be applied to draw the appropriate conclusions.

Appendix –I

X =

	A	B	C	D	A^2	B^2	C^2	D^2	AB	AC	AD	BC	BD	CD	Yield
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	27.6
1	-1	-1	1	1	1	1	1	1	1	-1	-1	-1	-1	1	16.6
1	1	-1	-1	1	1	1	1	1	-1	-1	1	1	-1	-1	15.4
1	-1	1	-1	1	1	1	1	1	-1	1	-1	-1	1	-1	17.4
1	1	-1	1	-1	1	1	1	1	-1	1	-1	-1	1	-1	17
1	-1	1	1	-1	1	1	1	1	-1	-1	1	1	-1	-1	19
1	1	1	-1	-1	1	1	1	1	1	-1	-1	-1	-1	1	17.4
1	-1	-1	-1	-1	1	1	1	1	1	1	1	1	1	1	12.6
1	1	-1	1	1	1	1	1	1	-1	1	1	-1	-1	1	18.6
1	-1	1	1	1	1	1	1	1	-1	-1	-1	1	1	1	22.4
1	1	1	-1	1	1	1	1	1	1	-1	1	-1	1	-1	21.4
1	-1	-1	-1	1	1	1	1	1	1	1	-1	1	-1	-1	14
1	1	1	1	-1	1	1	1	1	1	1	-1	1	-1	-1	24
1	-1	-1	1	-1	1	1	1	1	1	-1	1	-1	1	-1	15.6
1	1	-1	-1	-1	1	1	1	1	-1	-1	-1	1	1	1	13
1	-1	1	-1	-1	1	1	1	1	-1	1	1	-1	-1	1	14.4
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	22.6
1	1.41	0	0	0	2	0	0	0	0	0	0	0	0	0	23.4
1	-1.4	0	0	0	2	0	0	0	0	0	0	0	0	0	20.6
1	0	1.4	0	0	0	2	0	0	0	0	0	0	0	0	22.6
1	0	-1.4	0	0	0	2	0	0	0	0	0	0	0	0	13.4
1	0	0	1.4	0	0	0	2	0	0	0	0	0	0	0	20.6
1	0	0	-1.4	0	0	0	2	0	0	0	0	0	0	0	15.6
1	0	0	0	1.4	0	0	0	2	0	0	0	0	0	0	21
1	0	0	0	-1.4	0	0	0	2	0	0	0	0	0	0	17.6

Design Matrix for analyzing the stationary point for response variable *Yield*.

X'X=

25	0	0	0	0	20	20	20	20	0	0	0	0	0	0
0	20	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	20	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	20	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	20	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	24	16	16	16	0	0	0	0	0	0
20	0	0	0	0	16	24	16	16	0	0	0	0	0	0
20	0	0	0	0	16	16	24	16	0	0	0	0	0	0
20	0	0	0	0	16	16	16	24	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	16	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	16	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	16	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	16	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	16	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	16

The matrix multiplication of X transposes and X

Appendix - II

$(X'X)^{-1}$

0.4	0	0	0	0	-0	-0	-0	-0	0	0	0	0	0	0
0	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0.1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0.1	0	0	0	0	0	0	0	0	0	0
-0	0	0	0	0	0.1	-0	-0	-0	0	0	0	0	0	0
-0	0	0	0	0	-0	0.1	-0	-0	0	0	0	0	0	0
-0	0	0	0	0	-0	-0	0.1	-0	0	0	0	0	0	0
-0	0	0	0	0	-0	-0	-0	0.1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0.1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0.0625	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0.063	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0.063	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0.063	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.06

Inverse of matrix of X'X.

X'y =

463.800
26.359
53.809
42.270
25.208
374.373
358.378
358.778
363.577
12.000
4.800
2.800
9.600
7.600
-1.200

$\hat{b} =$

21.448	b0
1.318	b1
2.691	b2
2.114	b3
1.260	b4
0.420	b11
-1.580	b22
-1.530	b33
-0.930	b44
0.750	b12
0.300	b13
0.175	b14
0.600	b23
0.475	b24
-0.075	b34

The result for $\hat{b}$'s using the matrix multiplication $\hat{b} = (X'X)^{-1} X'y$

Bibliography

Anderson, Mark J. and Whitcomb, Patrick J. 2004. Design solutions from concept through manufacture: Response surface methods for process optimization. *Desktop Engineering.* http://www.deskeng.com/ (accessed May 16, 2006).

Box, E.P. George, Hunter, J. Stuart, and Hunter, G. William. 2005. *Statistics for Experiments.* New Jersey: John Willey and Sons, Inc.

Dean, Angela and Voss, Daniel. 1999. *Design and Analysis of Experiments.* New York: Springer.

Minto, Charles. 2006. Response Surface Modeling of Drug Interactions. http://eurosiva.org/Archive/Vienna/abstracts (accessed May 16, 2006).

Montgomery, Douglas C. 2005. *Design and Analysis of Experiments: Response surface method and designs.* New Jersey: John Wiley and Sons, Inc.

Myers, Raymond H., Khuri, Andre I. and Carter, Walter H., Jr. 1989. Response surface methodology: *1966-1988*. *Technometrics* 31 (2): 137-153 http://www.jstor.org/ (accessed January 29, 2007).

Myers, Raymond H. and Montgomery, Douglas C. 1995. *Response Surface Methodology: process improvement with steepest ascent, the analysis of response Surfaces, experimental designs for fitting response surfaces,* 183-351. New York: John Wiley and Sons, Inc.

Oehlert, Gary W. 2000. *Design and analysis of experiments: Response surface design.* New York: W.H. Freeman and Company

Shi, Y. and Weimer P.J. 1992. Response surface analysis of the effects of pH and Dilution rate on ruminococcus flavefaciens FD-1 in cellulose-fed continues Culture. *American Society for Microbiology* 58 (8): 2583-2591.

Studenmund, A.H. 2006. *Using Econometrics*. New York: Pearson Education, Inc.

Verseput, Richard. 2000. Digging into DOE: Selecting the right central composite Design for response surface methodology applications. *Quality Digest*. http://www.qualitydigest.com/june01/html/doe.html (accessed June 7, 2006).

Wu, Chien Fu and Hamada, Michael. 2000. *Experiments: planning, analysis, and parameter design optimization*. New York: Wiley – Interscience.

W.S., Connor and Zelen, Marvin. 1959. *Fractional factorial experiment designs for factors at three-levels.* Washington: U.S. Gov.

2006. Response surface methodology. http://en.wikipedia.org/wiki/response_surface_methodology (accessed January 22, 2007).