ON THE DETERMINATION OF THE BEST MODELS IN MIXTURE EXPERIMENTS

Kadri Ulaş AKAY

University of Marmara, Departments of Mathematics, Goztepe Kampüsü, Kadiköy, 34722 ISTANBUL TURKEY, kadriulas@marmara.edu.tr

ABSTRACT

In this paper, an alternative approach was proposed for the determination of the models taken into account in the modeling of the mixture surface which is obtained on the experimental region. This approach depends on the examination of all possible subset regression models obtained for the mixture model. In addition, model control graphs are taken into account to determine the best models. In this situation, with the help of different subset regression models, a more comprehensive interpretation of the mixture system and the components can be obtained. Then, proposed approach has been investigated on flare data set which is widely known in literature.

Key Words: Mixture Model, All possible subset selection, Variable selection, Regression models, AMS-Mathematical Subject Classification Number: Primary 62K99, Secondary 62J07

KARMA DENEMELERDE EN İYİ MODELLERİN BELİRLENMESİ ÜZERİNE

ÖZET

Bu çalışmada, deneysel bölge üzerinde elde edilen karma yüzeyin modellenmesi için ele alınan modellerin belirlenmesinde alternatif bir yaklaşım önerilmiştir. Bu yaklaşım, bir karma model için elde edilen tüm olası alt küme regresyon modellerinin incelenmesine dayanmaktadır. Ayrıca en iyi modellerin belirlenmesi için model kontrol grafikleri göz önüne alınmıştır. Bu durumda, elde edilen farklı alt küme regresyon modelleri yardımıyla karma sistem ve bileşenler hakkında kapsamlı bir yorum elde edilebilir. Önerilen yaklaşım, literatürde çok bilinen flare veri kümesi üzerinde incelenmiştir.

Anahtar Kelimeler: Karma Model, Tüm olası alt küme seçimi, Değişken seçimi, Regresyon modelleri

1. Introduction

In mixture experiments, the measured response is assumed to depend only on the proportions of ingredients present in the mixture and not on the amount of mixture. For example, the response might be the tensile strength of stainless steel which is a mixture of iron, nickel, copper and chromium, or, it might be octane rating of a blend of gasolines. The purpose of mixture experiments is to build an appropriate model relating the response(s) to mixture components. The resulting models can be used to understand how the responses depend on the mixture components.

In a q-components mixture in which x_i represents the proportion of the ith components present in mixture,

$$0 \le x_i \le 1,\ i = 1,2,\dots,q,\ \sum_{i=1}^{q} x_i = 1 \qquad (1)$$

The composition space of the q components takes the form of a regular (q-1)-dimensional simplex. Physical, theoretical, or economic considerations often impose additional constraints on individual components,

$$0 \le L_i \le x_i \le U_i \le 1,\ i = 1,2,\dots,q \qquad (2)$$

where L_i and U_i denote lower and upper bounds, respectively. In general, restriction (2) reduce the constraint region given by (1) to an irregular (q-1)-dimensional hyperpolyhedron.

It is assumed that the response or property of interest, denoted by η, is to be expressed in terms of a suitable function f of the mixture variables x_i,

$$\eta = f\left(x_1, x_2, \dots, x_q\right) \qquad (3)$$

A typical model may thus be written,

$$y_i = \eta_i + \varepsilon_i \tag{4}$$

where ε_i is assumed that $\varepsilon_i \square \text{NID}(0,\sigma^2)$. The function form of the response $E(y) = f(x_1, x_2, ..., x_q)$ is usually not known. Often first- or second-degree polynomial approximation model can be used. Mixture model forms most commonly used in fitting data are the canonical polynomials introduced by Scheffé [8] in the form,

$$E(y) = \eta = \sum_{i=1}^{q} \beta_i x_i + \sum_{i=1}^{q} \sum_{i<j}^{q} \beta_{ij} x_i x_j \tag{5}$$

For modeling well-behaved systems, generally the Scheffé polynomials are adequate. For some situations, however, there are better modeling forms than Scheffé polynomials which could be used. For example, as an alternative to Scheffé mixture models, models including inverse term are used in order to model an extreme change in the response behavior of one or more components, which are close to boundary of the simplex region [4]. Following, quadratic model including an inverse term has been proposed by Draper and St. John,

$$E(y) = \sum_{i=1}^{q} \beta_i x_i + \sum_{i=1}^{q} \sum_{i<j}^{q} \beta_{ij} x_i x_j + \sum_{i=1}^{q} \gamma_i x_i^{-1} \tag{6}$$

Scheffé polynomial models fails to satisfy the modeling of additive effect of one component and at the same time accommodate the curvilinear blending effects of the remaining components. To model these effects jointly, Becker has developed a set of mixture models which are homogeneous of degree one [1]. They provide alternatives to the Scheffé polynomials. Becker's three second order models are of the form,

$$\begin{aligned} H1: \eta &= \sum_{i=1}^{q} \beta_i x_i + \sum_{i=1}^{q} \sum_{i<j}^{q} \beta_{ij} \min(x_i, x_j) \\ H2: \eta &= \sum_{i=1}^{q} \beta_i x_i + \sum_{i=1}^{q} \sum_{i<j}^{q} \beta_{ij} \frac{x_i x_j}{x_i + x_j} \\ H3: \eta &= \sum_{i=1}^{q} \beta_i x_i + \sum_{i=1}^{q} \sum_{i<j}^{q} \beta_{ij} (x_i x_j)^{1/2} \end{aligned} \tag{7}$$

In the H2 model, $x_i x_j / (x_i + x_j) = 0$ whenever $(x_i + x_j) = 0$.

As usual, we can represent the Scheffé canonical polynomial models, mixture models with inverse terms and Becker Homogenous models in matrix form by

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{8}$$

where $\mathbf{Y}$ is $n \times 1$ vector of observations on the response variable, $\mathbf{X}$ is $n \times p(\geq q)$ matrix, where p is number of terms in the model, $\boldsymbol{\beta}$ is the $p \times 1$ vector of parameters to be estimated and $\boldsymbol{\varepsilon}$ is $n \times 1$ vector of errors. It was assumed that the errors have the property

$$E(\boldsymbol{\varepsilon}) = 0,\ E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2 \mathbf{I}_n \tag{9}$$

where $\mathbf{I}_n$ is identity matrix and σ^2 is the error variance. Hence $E(\mathbf{Y}) = \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ where $\boldsymbol{\mu}$ is column vector of all expected responses. The least squares estimator for $\boldsymbol{\beta}$ is $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$ and variance-covariance matrix of $\mathbf{b}$ is $\text{var}(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1} \sigma^2$. A comprehensive reference on the design and analysis of mixture data is given by Cornell [2, 3].

All of the work on mixture models has been based on response surface concepts. A model is fitted to data by an experimental design. The response surface contours are examined to determine the region of the factor space where best values of the response can be obtained. The purpose of this paper is to present some methods which enable one to obtain a better understanding of a mixture system and the role of the different components. In the following sections, these methods are described.

2. Determination and Comparison of Mixture Models

In mixture experiments, reduction of the model is as much important as determination of the model because it is not a very good approach to add all the terms of the chosen model to itself. In a situation like this, the model may include meaningless interaction terms. It may also be hard to make comments on the mixture system as the parameter values may be affected. The sequential model fitting methods proposed by Draper and St. John for mixture experiments can be useful [4]. But, if there are many terms, it can require too much labor. There are various methods for choosing a regression model such as forward selection, backward elimination and stepwise regression when there are many candidate model terms. In addition, Cornell mentioned that the stepwise regression model can be investigated for various models in mixture experiments [2]. The objective is to obtain a model form that not only contains an adequate amount of information about the mixture system under investigation but whose form also makes sense. However these methods result in only one model and alternative models, with an equivalent or even better fit, are easily overlooked. A more preferable method than these methods is to fit all possible regression models, and to evaluate these according to some criterion. In this way a number of best regression models can be selected. In this case, alternative subset regression models, which can be used to model the mixture system on the simplex region, can be obtained. However the fitting of all possible regression models is very computer intensive. In order to find the best subset regression model "RESEARCH procedure" on GENSTAT was used [5]. While using this procedure, linear mixture terms $\left(x_1, x_2, \ldots, x_q\right)$ were kept in the model and all possible combinations for the rest of the terms were added to the linear mixture terms. From the models obtained, the models with terms $p-value < 0.05$ according to F statistics have been taken into account. However, in order to examine which of the models are adequate, model control graphs should be obtained. For the models whose model control graphs are adequate, a decision can be made by looking at R_A^2 and MSE values of the models. The proposed approach will be examined in the following part over the flare data set.

3. Flare Experiment

McLean and Anderson presented an example to illustrate their extreme-vertices design [6]. A flare is manufactured by mixing magnesium$\left(x_1\right)$, sodium nitrate$\left(x_2\right)$, strontium nitrate$\left(x_3\right)$, and binder$\left(x_4\right)$ under the following constraints,

$$0.40 \le x_1 \le 0.60 \qquad 0.10 \le x_3 \le 0.47$$
$$0.10 \le x_2 \le 0.47 \qquad 0.03 \le x_4 \le 0.08$$

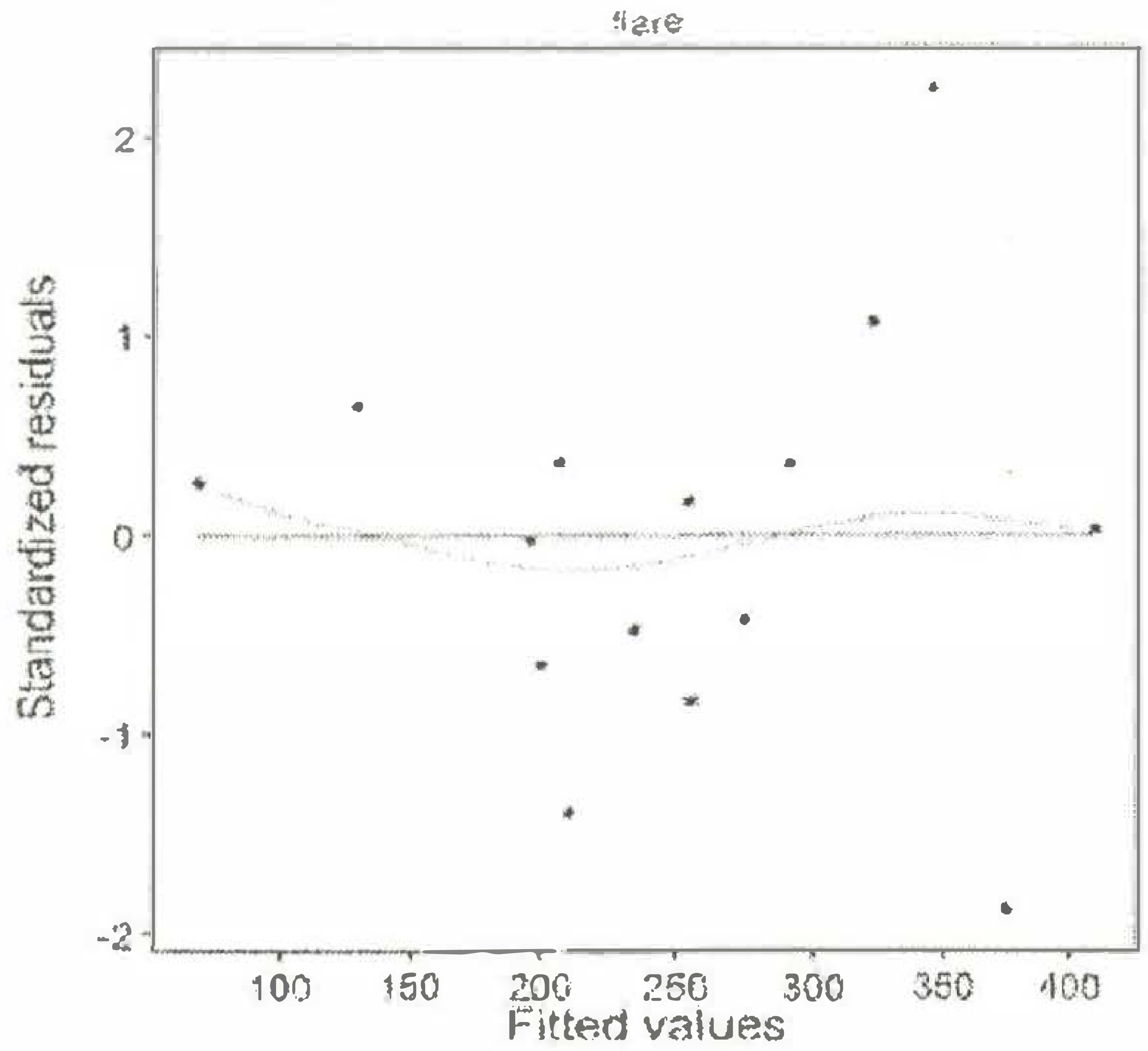
The component proportions for design points as well as the measured illumination values are given in Table 1.

Table 1. Components Proportions and Illumination Response Values for Flare Experiment

Blend No	*Component Proportions*				*Illumination (1000 candles)*
	x_1	x_2	x_3	x_4	
1	0.40	0.10	0.47	0.03	75
2	0.40	0.10	0.42	0.08	180
3	0.60	0.10	0.27	0.03	195
4	0.60	0.10	0.22	0.08	300
5	0.40	0.47	0.10	0.03	145
6	0.40	0.42	0.10	0.08	230
7	0.60	0.27	0.10	0.03	220
8	0.60	0.22	0.10	0.08	350
9	0.50	0.1000	0.3450	0.055	220
10	0.50	0.3450	0.1000	0.055	260
11	0.40	0.2725	0.2725	0.055	190
12	0.60	0.1725	0.1725	0.055	310
13	0.50	0.2350	0.2350	0.030	260
14	0.50	0.2100	0.2100	0.080	410
15	0.50	0.2225	0.2225	0.055	425

Snee and, Draper and St. John made a comparison of the mixture models for the flare data set [9, 4]. In addition, Draper and St. John used the backward elimination regression procedure [4]. On the other hand, Piepel and Cornell gave a summary of the models proposed for the flare data set till now [7]. When the control graphs of these models are investigated, it can be seen that they are not adequate and also they have meaningless interaction and inverse term. In this study, subset regression model for actual components will be given by using Scheffé, Homogenous H2 and Models including inverse term.

Subset regression models obtained from the modeling study done by using actual component for Scheffé, H2 and the models including inverse term are given in Tables 2-4 respectively (see Appendix). The values given in parenthesis in Tables show the standard errors of the predicted parameters. In addition, the terms shown with the symbol X are meaningless.



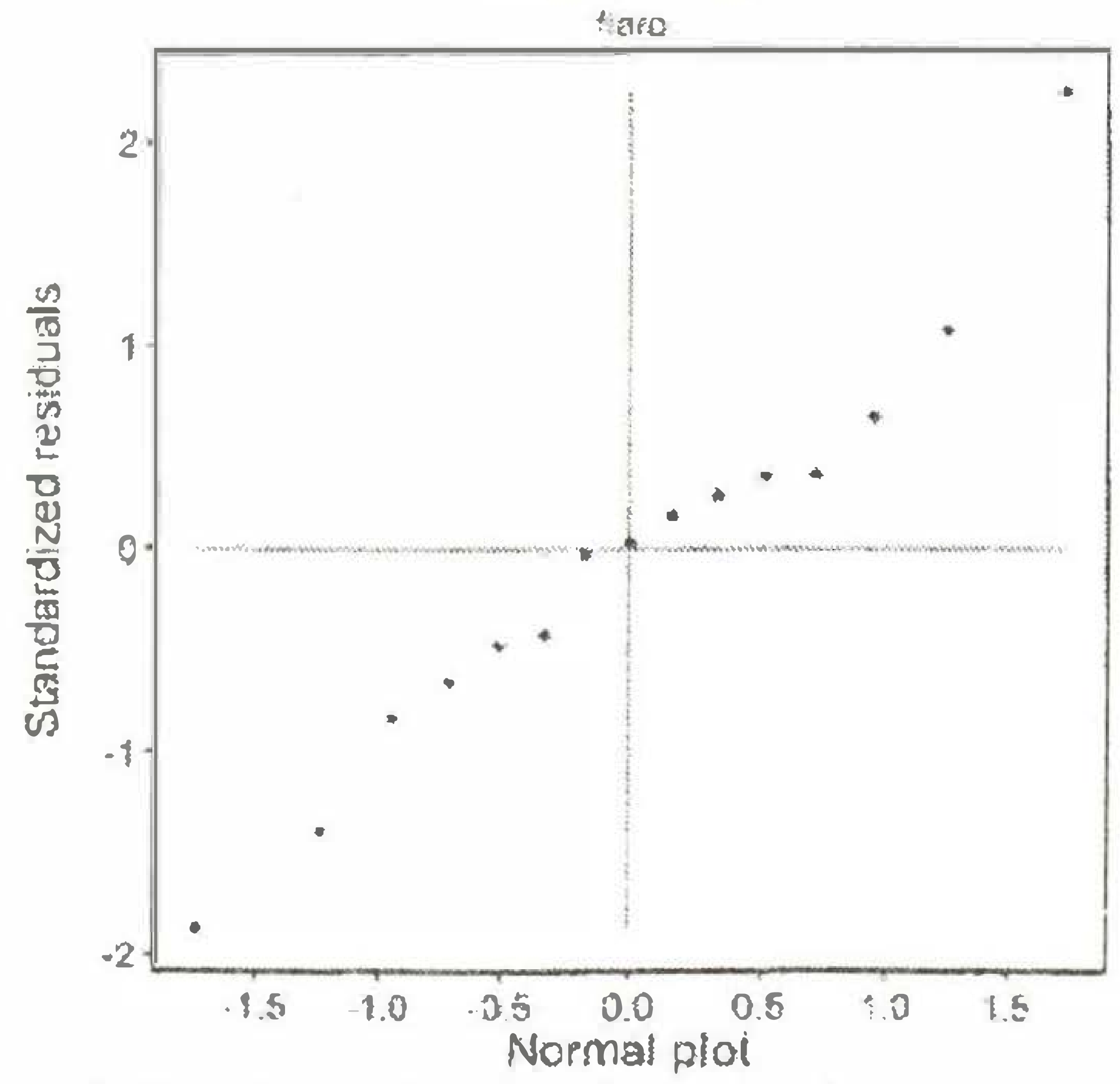


Figure 1. Model control graphs of model including inverse term

When the model control graphs for subset regression models are investigated, it can be seen that the models including inverse term are better than the other models. This is because the control graphs for Scheffé and H2 subset regression models show that these models are not adequate. In Table 4, only the control graphs of models including inverse term 2, 3 and 7 show that the models are adequate. If R_A^2 and MSE values are taken in to account, model 7 can be chosen by the researcher. The control graphs of model 7 are given in Figure 1.

The mixture surface for $x_4 = 0.03$ and $x_4 = 0.08$ on the experimental region for the model including inverse term is shown respectively in Figure 2.

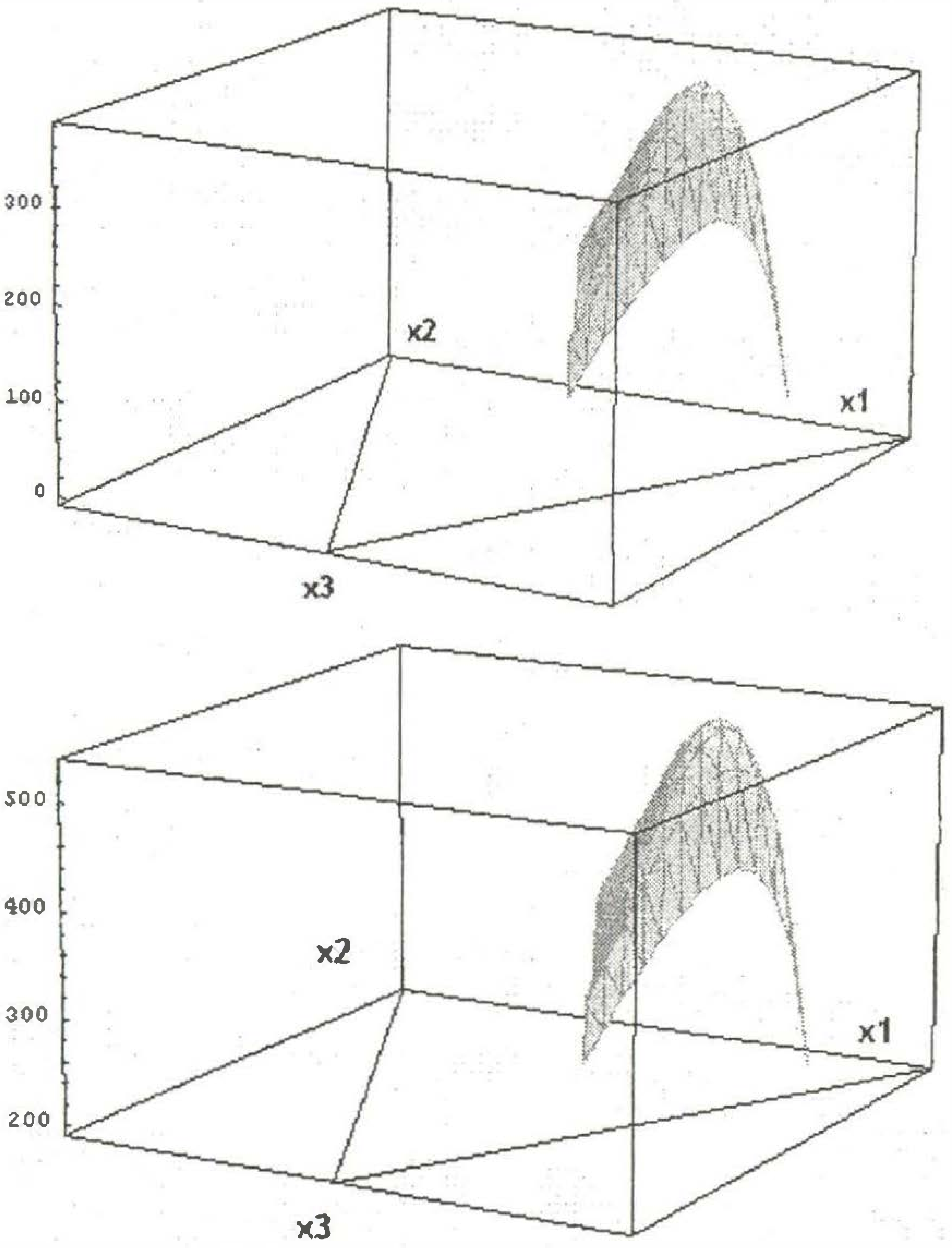


Figure 2. Mixture surfaces obtained for model including inverse terms

4. Conclusion

In this paper, subset regression models with different terms of alternative mixture models on the experimental region were obtained. A comprehensive research can be done about different subset regression models together with mixture system. The researcher can choose among this subset regression models whose model control graphs were adequate. In this study, our aim is not to make a comparison between mixture models but it is to obtain subset regression models which can be used in the modeling of the mixture system. Therefore, in this study R_A^2 and MSE values were taken into account for the determination of the best model.

Many researchers make a comparison of the models according to the numbers of terms the models include. Therefore, if the model includes few terms, this may make it easier to understand the model. However, as the number of the reasonable interaction terms of the model increase, it becomes easier to make a comment on the mixture system and to measure the effects of the component. Regression model including different numbers of term which can be used to model the mixture system can be chosen if the model control graphs are adequate.

As a result, the models obtained in Tables 2-4 differ from the regression models obtained with stepwise regression operations. On the other hand, meaningful regression terms can not always be obtained by using stepwise-type regression operations. The model control graphs of the models may not show if the models are adequate as well. For this reason, with the choice of all possible subset regression for mixture experiments better results can be obtained.

References

[1]. Becker, N. G., 1968. Models for the response of a mixture. Journal of the Royal Statistical Society, B, 30: 349-358

[2]. Cornell, J. A., 2000. Developing Mixture Models, Are we done?. Journal of Statistical Computation and Simulation, 66: 127-144

[3]. Cornell, J. A., 2002. Experiments with mixtures, 3 rd. ed. Wiley-Interscience

[4]. Draper, N. R. and R. C. St. John, 1977. A mixtures model with inverse terms. Technometrics, 19: 37-46

[5]. GENSTAT, Release 7.1, 2003. The Guide to Genstat Release 7.1: Part 2 Statistics

[6]. McLean, R. A. and V. L. Anderson, 1966. Extreme vertices design of mixture experiments. Technometrics, 8: 447-454

[7]. Piepel, G. F. and J. A. Cornell, 1994. Mixture Experiment Approaches: Examples, Discussion, and Recommendations. Journal of Quality Technology, 26: 177-196

[8]. Scheffé, H., 1958. Experiments with mixtures. Journal of the Royal Statistical Society, B, 20: 344-360

[9]. Snee, R. D., 1973. Techniques for the analysis of mixture data. Technometrics, 15: 517-528

Appendix

Table 2. The parameter predictions of subset regression models obtained by using Scheffé model

Scheffé	x_1	x_2	x_3	x_4	x_1x_2	x_1x_3	x_1x_4	x_2x_3	x_2x_4	x_3x_4	R_A^2	*MSE*
Best subset with 1 terms	469.897 (110.2)	–535.7 (236.8)	–716.5 (236.8)	2214.896 (736.1)	X	X	X	4345.936 (1820.3)	X	X	59.3	3720
Best subset with 2 terms	–1326.6 (683.6)	–2281 (974.9)	–2363 (974.9)	3983.158 (1029.4)	8121.991 (3299.6)	7899.748 (3299.6)	X	X	X	X	58.9	3752

(X is indicate meaningless terms)

Table 3. Parameter predictions of subset regression models obtained by using Becker H2 model

Becker (H2)	x_1	x_2	x_3	x_4	$\frac{x_1x_2}{x_1+x_2}$	$\frac{x_1x_3}{x_1+x_3}$	$\frac{x_1x_4}{x_1+x_4}$	$\frac{x_2x_3}{x_2+x_3}$	$\frac{x_2x_4}{x_2+x_4}$	$\frac{x_3x_4}{x_3+x_4}$	R_A^2	*MSE*
Best subset with 1 terms	287.692 (103.5)	–404.1 (162.6)	–584.9 (162.6)	2043.134 (666.3)	X	X	X	2442.910 (806.2)	X	X	66.7	3045
Best subset with 2 terms	–362.73 (210.5)	–1510.9 (480.9)	–1601.2 (480.9)	2110.746 (585.5)	3634.675 (1147.6)	3422.316 (1147.6)	X	X	X	X	74.2	2357

(X is indicate meaningless terms)

Table 4. Parameter predictions of subset regression models obtained by using models including inverse term

Models with Inverse Terms	*No*	x_1	x_2	x_3	x_4	x_1x_2	x_1x_3	x_1x_4	x_2x_3	x_2x_4	x_3x_4	$(x_1)^{-1}$	$(x_2)^{-1}$	$(x_3)^{-1}$	$(x_4)^{-1}$	R_A^2	*MSE*
Best subset with 1 terms	1	370.511 (92.9)	4680.802 (1444.0)	4499.992 (1444.0)	6890.397 (1574.7)	X	X	X	X	X	X	−1145.9 (347.829)	X	X	X	69.3	2801
	2	704.834 (147.3)	−427.168 (178.4)	308.715 (231.8)	2469.022 (700.1)	X	X	X	X	X	X	X	−35.62 (12.748)	X	X	64.1	3279
	3	682.848 (155.2)	449.420 (244.2)	−581.456 (187.9)	2445.641 (737.4)	X	X	X	X	X	X	X	X	−33.03 (13.427)	X	60.2	3638
Best subset with 2 terms	4	584.885 (122.4)	3294.105 (1357.8)	3739.807 (1258.3)	5972.952 (1383.1)	X	X	X	X	X	X	−871.19 (316.225)	−24.35 (10.71)	X	X	78.4	1977
	5	356.270 (183.2)	−187.926 (173.0)	−983.005 (548.6)	3038.738 (610.4)	X	4100.325 (1636.7)	X	X	X	X	X	−46.69 (11.22)	X	X	76.5	2147
	6	309.047 (190.5)	−935.823 (570.5)	−324.892 (179.9)	3056.605 (634.7)	4397.194 (1701.8)	X	X	X	X	X	X	X	−44.90 (11.666)	X	74.6	2321
Best subset with 3 terms	7	1340.480 (248.4)	1185.443 (549.8)	1107.848 (549.8)	3153.739 (562.4)	X	X	X	−8874.5 (3803.4)	X	X	X	−65.37 (19.0)	−61.36 (19.0)	X	81.1	1724

(X is indicate meaningless terms)