Journal of the KSQC
Vol. 8, No. 1 (1980)

Empirical process optimization through response surface experiments and model building *

SUNG H. PARK **

Abstract

In many industrial processes, there are more than two responses (i.e., yield, percent impurity, etc.) of interest, and it is desirable to determine the optimal levels of the factors (i.e., temperature, pressure, etc.) that influence the responses. Suppose the response relationships are assumed to be approximated by second-order polynomial regression models. The problems considered in this paper is, first, to propose how to select polynomial terms to fit the multivariate regression surfaces for a given set of data, and, second, to propose how to analyze the data to obtain an optimal operating condition for the factors. The proposed techniques were applied for empirical process optimization in a tire company in Korea. This case is presented as an illustration.

1. INTRODUCTION

In empirical investigations of the relationship between a response variable (y) and several independent variables ($x_1, x_2, ..., x_k$) to determine optimum operating conditions for process control, the response relationship is often assumed to be approximated by a second-order polynomial regression model,

$$y = \beta_o + \sum_{i=1}^{k} \beta_i x_i + \sum_{i \le j}^{k} \beta_{ij} x_i x_j + e$$

where the values of the x_i's are in some region of interest, R, over which the polynomial approximation is to be used, and e is a random error whose distribution is usually normal with mean zero and variance σ^2. However, a question is raised in using the polynomial model. Does he have to include every term in the polynomial to fit the response surface?

One of the problems considered in this paper is that of selecting subsets of polynomial terms and model-building from a given polynomial model so as to achieve "improved" response surfaces in estimation of the response. Such improvement in fitting the response surfaces would be very helpful to determine optimum operating conditions of independent variables for quality control, and to explore the functional relationship with better precision.

The author has proposed a criterion in PARK

* 本 研究는 79年度 文教部 研究費에 의하여 遂行된 것임.

** Department of Computer Science & Statistics
Seoul National University

[8] to select subsets of polynomial terms to fit the response surface, when the precision of response over a region of interest is of main concern. This selection technique is applied in this paper for determination of a proper model. For other selection techniques, see DRAPER & SMITH [3], MALLOWS [6], HELMS [5] and PARK [9, 10].

In empirical response surface experiments engineers frequently need to analyze multi-response data, and it is desirable to find the optimal levels of the independent factors that influence the responses. Several authors (BOX & DRAPER [1], HELLER & STAATS [4], MYERS & CARTER [7], BOX ET AL. [2]) have considered a variety of estimation problems associated with multi-response data. However, the special case discussed in this paper has not been treated explicitly. Suppose we have q responses on k independent factors, $x' = (x_1, x_2, ..., x_k)$, and the following relationships are assumed to be true in a region of interest,

$R = \{(x_1, x_2, ..., x_k) : a_i \leq x_i \leq b_i, i = 1, 2, ..., k\}$,

$$y_m(\underline{x}) = f_m(\underline{x}) + e_m \qquad \text{(Eq. 1)}$$

where

$$f_m(\underline{x}) = \beta_{mo} + \sum_{i=1}^{k} \beta_{mi} x_i + \sum_{i \leq j}^{k} \beta_{mij} x_i x_j.$$

Without loss of generality, suppose $y_1(\underline{x})$ is the primary response (most important response) and $y_2(\underline{x})$, $y_3(\underline{x})$, ..., $y_q(\underline{x})$ are the secondary responses (the responses which should satisfy some levels of specification). What the engineers and scientists usually want to know is to find the optimal levels of $\underline{x}$ such that

$$\underset{\underline{x}}{\text{Maximize}} \quad \hat{y}_1(\underline{x}) \qquad \text{(Eq. 2)}$$

$$\text{subject to} \quad \hat{y}_m(\underline{x}) \geq C_m \ (\text{or} \leq C_m),\ m = 2, 3, ..., q \qquad \underline{x} \in R,$$

where $\hat{y}_m(\underline{x})$ is a proper estimate of $f_m(\underline{x})$ for a given set of data on $(y_{1i}, y_{2i}, ..., y_{qi}; x_{1i}, x_{2i}, ..., x_{ki})$, i = 1, 2, ..., n. It is suggested that one select the necessary polynomial terms from (Eq. 1) by the technique proposed by PARK [8] and fit the regression models to obtain $\hat{y}_m(\underline{x})$ by the ordinary least squares method.

2. SELECTION CRITERION

Suppose there are $n \leq t$ observations on a t-vector of input polynomial terms, $\underline{x}' = (1, x_1, x_2, ..., x_1^2, x_2^2, x_1 x_2 ...)$. The response variable is frequently expressed in vector notation as

$$y = \underline{x}'\underline{\beta} + e \qquad \text{(Eq. 3)}$$

where β is the t-vector of unknown regression coefficients and it is assumed that the residuals, e, are identically and independently distributed with mean zero and unknown variance, σ^2. Let r denote the number of terms which are deleted from (Eq. 3), and p = t - r denote the number of terms which are retained in the final equation.

Let (Eq. 3) be written in partitioned vector form as

$$y = \underline{x}_p' \underline{\beta}_p + \underline{x}_r' \underline{\beta}_r + e$$

where $\underline{x}_p$ contains the retained terms and $\underline{x}_r$ contains the deleted terms. Let $\hat{\underline{\beta}}$ with components $\hat{\underline{\beta}}_p$ and $\hat{\underline{\beta}}_r$ denote the least squares estimator of $\underline{\beta}$ and let $\tilde{\underline{\beta}}_p$ denote the subset least squares estimator of $\underline{\beta}_p$ when the polynomial terms in x_r are deleted from the model.
That is,

$$\hat{\underline{\beta}} = (X'X)^{-1} X'Y \text{ and } \tilde{\underline{\beta}}_p = (X_p'X_p)^{-1} X_p'Y$$

where X and X_p are the matrices of values taken by the polynomial terms in $\underline{x}$ and $\underline{x}_p$, respectively, at each of the design points, and Y is the n-vector of observed responses. If we use the full model then the estimated value of the response at a particular input $\underline{x}$ is $\hat{y}(\underline{x}) = \underline{x}'\hat{\underline{\beta}}$. On the other hand, if the subset model with x_I deleted is used,

the estimated response is $\tilde{y}_p(\underline{x}_p) = \underline{x}_p' \tilde{\underline{\beta}}_p$.

The proposed criterion is to select the p polynomial terms which maximize the quantity,

$$Q = \int_R [MSE(y) - MSE(\tilde{y}_p)] \, dW(\underline{x}).$$

which is integrated over the region of interest, R. In this $W(\underline{x})$ is a weighting function that can be treated as a probability distribution function on R. The $W(\underline{x})$ allows for differential importance of the estimator of $f(\underline{x})$ at different points in the region and can be specialized to a discrete set of points if desired.

PARK [8] shows that, after replacement of the parameters σ^2 and $\underline{\beta}_r$ by their estimates from the current data using the full model, the quantity to be miximized is

$$\hat{Q} = \hat{\sigma}^2 \{ T_r [(X'X)^{-1} M] - T_r [X_p' X_p)^{-1} M_{pp}] \} - \hat{\underline{\beta}}_r' [A' M_{pp} A - 2A' M_{pr} + M_{rr}] \hat{\underline{\beta}}_r, \quad \text{(Eq. 4)}$$

where

$$M = \int_R \underline{x}\underline{x}' \, dW(\underline{x}),$$

$$M_{ij} = \int_R \underline{x}_i \underline{x}_j \, dW(\underline{x})$$

$$A = (X_p' X_p)^{-1} X_p' X_r,$$

and T_r denotes trace. The first term in $\hat{Q}$ is the integrated difference of $Var(\hat{y}) - Var(\tilde{y}_p)$, which is always non-negative, and the last term in Q is the integrated squared bias of $\tilde{y}_p$. Therefore, in essence, the criterion is to look for a subset of polynomial terms whose gain in precision is not offset by the squared bias over the whole region of interests, R.

3. OPTIMIZATION PROCEDURE

Suppose we select the proper polynomial terms from the method proposed in the previous section, and fit q equations by the least squares method, and obtain q estimated responses, $\hat{y}_m(\underline{x})$, $m = 1, 2, \ldots, q$, where $\underline{x}' = (x_1, x_2, \ldots, x_k)$ in this section. We want to find $\underline{x} \in R$ which maximizes $\hat{y}_1(\underline{x})$ subject to

$$\hat{y}_m(\underline{x}) \leq C_m \ , \ m = 2, 3, \ldots, q$$

as described in (Eq. 2).

Let $R^* = \{ (\underline{x}) : y_m(\underline{x}) \leq C_m \ , \ m = 2, 3, \ldots, q, \text{ and } \underline{x} \in R \}$. This R^* is the region for $\underline{x}$ such that every point $\underline{x} \in R^*$ satisfies the secondary response constraints. Suppose $R_m^* = \{ (\underline{x}) : \hat{y}_m(\underline{x}) \leq c_m, \ \underline{x} \in R \}$. Then, obviously, R^* is the intersection of R_m^*, $m = 2, 3, \ldots, q$. That is,

$$R^* = R_2^* \cap R_3^* \cap \cdots \cap R_q^*. \quad \text{(Eq. 5)}$$

Usually, it is difficult to obtain R_m^* for each $\hat{y}_m(\underline{x})$. To obtain R_m^* easily, a computer program was written to sketch the contours of $\hat{y}_m(\underline{x})$, so that R_m^* can be identified. The computer program was written in FORTRAN IV, and can be obtained from the author by request.

Suppose $\underline{x}^*$ is the optimal point of $\underline{x}$ which belongs to R^*, and maximizes $\hat{y}_1(\underline{x})$. The point $\underline{x}^*$ and its vicinity can be obtained by plotting the contours of $\hat{y}_1(\underline{x})$ and observing the values of $\hat{y}_1(\underline{x})$ on R^*. This whole procedure can be best explained by the following industrial experiments, which the author helped for the experimental design and the analysis of the data.

4. EXAMPLE

The data used in this example were provided by a tire manufacturing company in Korea. The company conducted an experiment in 1979 to find a better combination of raw materials for tire-making to improve the reliability of an industrial tire, named 1100-20. The company achieved its foal to find a better combination through the experimental design and the analysis of data which will be the experimental design and the analysis of data which will be described in the rematinder of this paper.

The laboratory scientists of the company found that the 300% modulus (y_1) would be the primary response, and the amount of heat (y_2) would be the secondary response. Also they found that two raw materials of chemicals, R101 and U100 (called x_1 and x_2 respectively), are the factors of interest which might affect the two responses. They decided to choose the 3^2 factorial disign with 3 replicates for y_1 and 2 repliate for y_2 in each level combination, and the levels of each factor are $(\alpha-3, \alpha, \alpha+3)=(-1, 0, +1)$ for x_1 and $(\beta-1, \beta, \beta+1) = (-1, 0, +1)$ for x_2, where $(x_1, x_2) = (\alpha, \beta)$ is the present level of combination for the two chemicals.

What they want to find out is the best combination of x_1 and x_2 in the range of (-1, +1) that will maximizes y_1 subject to the constraint that y_2 is not greater than a certain level of response. Table 1 and Table 2 show the data for this 3^2 factorial experiments.

	$x_1=-1$ $(\alpha-3)$	$x_1=0$ (α)	$x_1=+1$ $(\alpha+3)$
$x_2=-1$ $(\beta-1)$	90 85 93	114 100 100	108 124 112
$x_2=0$ (β)	133 119 108	97 114 97	109 102 122
$x_2=+1$ $(\beta+1)$	108 103 109	100 105 107	125 112 121

Table 1 : 300% modulus (y_1) data

	$x_1=-1$ $(\alpha-3)$	$x_1=0$ (α)	$x_1=+1$ $(\alpha+3)$
$x_2=-1$ $(\beta-1)$	17.2 14.4	20.3 21.1	23.8 22.8
$x_2=0$ (β)	17.5 18.5	21.4 19.7	21.1 22.7
$x_2=+1$ $(\beta+1)$	17.7 17.5	20.3 19.4	22.1 22.1

Table 2 : Amount of heat (y_2 : °C) data

The second-degree polynomial regression model was fitted for y_1 and y_2, respectively, and the selection rule proposed in Section 2 was applied. The result is that all polynomial terms (x_1, x_2, $x_1^{2\prime}$, x_2^2, x_1x_2) for y_1 are retained in the model, but for y_2, all terms except x_2 are retained in the model. Therefore, the resulted least squares equations obtained from the computer program are:

$$\hat{y}_1(\underline{x}) = 106.96 + 4.83x_1 + 3.56x_2 + 6.39x_1^2 - 4.78x_2^2 - 3.17x_1x_2,$$

$$\hat{y}_2(\underline{x}) = 20.51 + 2.65x_1 - 0.58x_1^2 - 0.21x_2^2 - 0.78x_1x_2.$$

Also the computer program gives Fig. 1 and Fig. 2 for the contours of $\hat{y}_1(\underline{x})$ and $\hat{y}_2(\underline{x})$, respectively, in the range of (-1, +1) for x_1 and x_2.

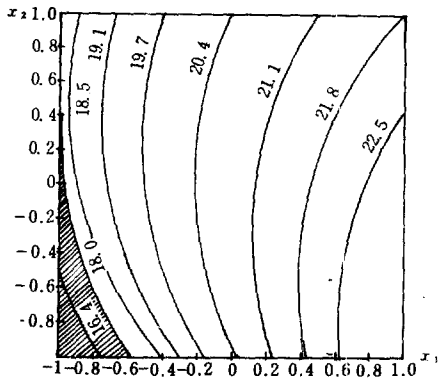

Fig. 1: Contours for $\hat{y}_2(\underline{x})$.

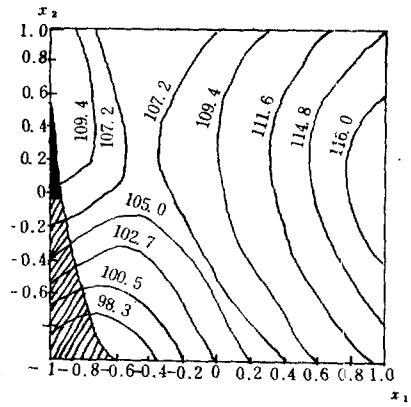

Fig. 2: Contours for $\hat{y}_1(\underline{x})$.

The company wants to find out the values of x_1 and x_2 in the region of interest, $R = \{(x_1, x_2) : -1 \leq x_i \leq 1, i = 1, 2\}$, such that

$$\underset{\underline{x}}{\text{Maximize}} \quad \hat{y}_1(\underline{x})$$

subject to $\hat{y}_2(\underline{x}) \leq 18.0\,^{\circ}C$.

The satisfactory region R^* for the secondary response is sketched in Fig. 1, and in this region of (x_1, x_2) the secondary response is not greater than 18 degrees of temperature. Next, this region R^* is sketched in Fig. 2 to identify the optimal point $x^* \in R^*$ that maximizes $\hat{y}_1(\underline{x})$. It turns out that the shaded dark part in Fig. 2 is the optimal region and the top part of the region is the point, x^*. Therefore, the present levels of x_1 and x_2 should move to the vicinity of (-1, 0.5) for (x_1, x_2) to improve the 300% modulus while the secondary response is satisfactory.

The company found that the new optimal levels, $(x_1, x_2) = (-1, 0.5) = (\alpha-3, \beta+1/2)$, are really bettern than the previously used optimal levels $(x_1, x_2) = (0, 0) = (\alpha, \beta)$, and the company decided to take this new combination for the two chemicals. Also the company found that the reliability of the tire is increased by 10% by the adoption of this new combination.

5. ACKNOWLEDGEMENT

This work was pertially supported by the research fund of the Ministry of Education, Korea Government, in 1979. The author is grateful for the graduate students in the department of computer science and statistics, Seoul National University, who helped me in writing the computer program used in this paper.

REFERENCES

[1] Box, G.E.P. and Draper, N.R.; "The Bayesian Estimation of Common Parameters from Several Responses"; *Biometrika*, Vol. 52, 1965, pp. 355-365.

[2] Box, G.E.P., Hunter, W.G., MacGregor, J.E. and Erjavec, J.; "Some Problems Associated with the Analysis of Multiresponse Data"; *Technometrics*, Vol. 15, 1973, pp. 33-49.

[3] Draper, N. and Smith, H.; *Applied Regression Analysis;* John Wiley & Sons; New York; 1966.

[4] Heller, N.B. and Staats, G.E.; "Response Surface Techniques for Dual Response Systems"; *Technometrics*, Vol. 15, 1973, pp. 113-123.

[5] Helms, R.; "The Average Estimated Variance Criterion for the Selection of Variables Problem in General Linear Models"; *Technometrics*, Vol. 16, 1974, pp. 261-274.

[6] Mallows, C.; "Some Comments on C_p"; *Technometrics*, Vol. 15, 1973, pp. 661-675.

[7] Myers, R.H. and Carter, W.H.; "Response Surface Techniques for Dual Response Systems"; *Technometrics*, Vol. 15, 1973, pp. 301-318.

[8] Park, S.H.; "Selection of Polynomial Terms for Response Surface Experiments"; *Biometrics*, Vol. 33, 1977, pp. 225-229.

[9] Park, S.H.; "On Screening of Variables for Response Surface Experiments with Mixtures"; *The Journal of the Korean Statistical Society*, Vol. 6, 1977, pp. 103-116.

[10] Park, S.H.; "Selecting Contrasts among Parameters in Scheffe's Mixture Models: Screening Components and Model Reduction"; *Technometrics*, Vol. 20, 1978, pp. 273-279.