Analysis of Variance and Design of Experiments-II

MODULE VIII

LECTURE - 33

RESPONSE SURFACE DESIGNS

Dr. Shalabh
Department of Mathematics & Statistics
Indian Institute of Technology Kanpur

An important objective of the design of experiment is the comparison of treatments whose nature can be either qualitative or quantitative.

The objective in both the cases is to detect structure of some form among the treatment effects. The methods of regression analysis can be used in case the treatments are quantitative in nature.

If the treatments are represented by the level of one treatment factor then the dependence of treatment effects on treatments can be represented by a response curve.

If the treatments are level combinations of two or more treatment factors, then a response surface can be used. Such curves can be used to make judgments about treatment structure and to know the relationship between treatments and responses, or between input and output variables.

Such knowledge of relationship is important if one wants to know the treatment combination which gives the optimal. The optimum can be defined in terms of highest or lowest response depending upon the situation.

The exact relationship is never known to the experimenter but an attempt is made to approximate it. This can be achieved by using the methods of experimental design and regression analysis. Such methods are referred to as a **response surface methodology (RSM).**

We consider a simple example to illustrate the application of RSM.

The relationship $y = x_1^2 + x_2^2$ can be pictured as a two-dimensional surface in a three-dimensional space giving the dependence of y on x_1 and x_2.

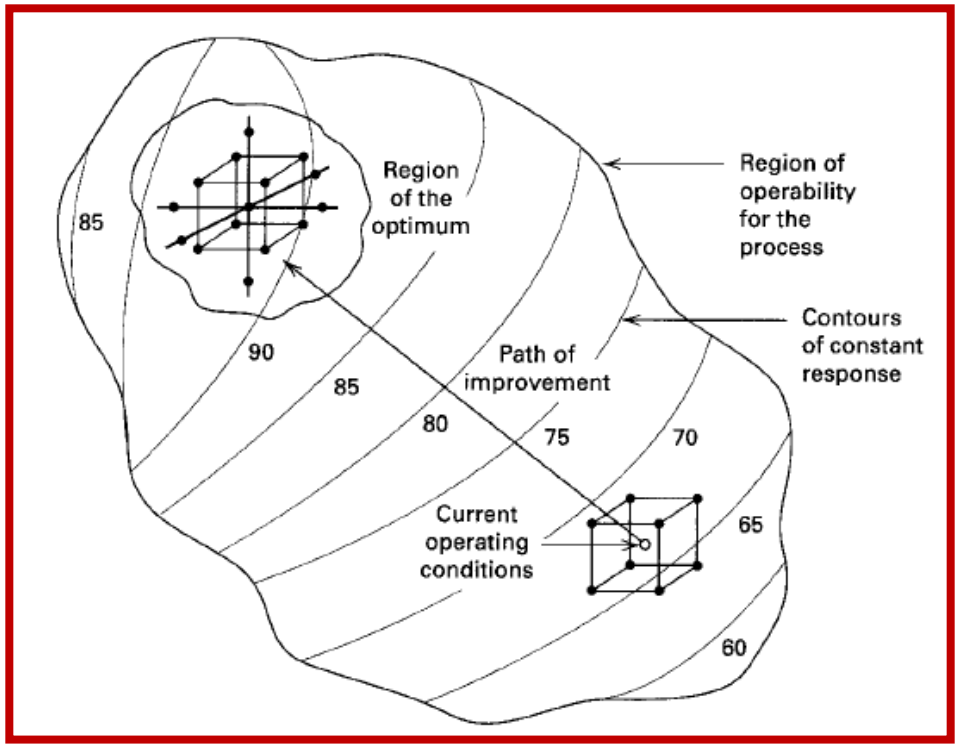
If we change the units for the input variables to $x_1^* = 2x_1$ and $x_2^* = 3x_2$, then the relationship becomes

$$y = \frac{1}{4}x_1^{*2} + \frac{1}{9}x_2^{*2}.$$

Note that y is constant on the curves $x_1^2 + x_2^2 =$ constant, which is a circle. So on circles in the $(x_1, x_2)-$ plane, y is constant on the curves $\frac{1}{9}x_1^{*2} + \frac{1}{25}x_2^{*2} =$ constant, which is now an ellipse in the $(x_1^*, x_2^*)-$ plane.

Obviously, these two surfaces are quite different from each other and this illustrate that the choice of surface depends on the choice of units of plotting also. This type of consideration are always kept in mind while doing with RSM.

RSM is a sequential procedure. Often when the experimenter is at a point on the response surface which is far away from the point of optimum, then there is a little curvature present in the system. In such a situation, the first order RSM will be appropriate. This is presented in the following figure:



The sequential nature of response surface methodology

This figure is reproduced with permission of John Wiley & Sons, Inc. from Montgamery, Douglas, C., 2001, 'Design and analysis of experiments', 5th edition. The author gratefully acknowledges the support.

The objective of experimenter is to lead along a path of improvement toward the general vicinity of the **optimum** in an efficient way. Once the experimenter is close to that region, a more elaborate model, e.g., such as second order model, may be employed. Then the analysis can be performed to locate the optimum.

Formulation of the problem

Suppose we have *k* quantitative factors F_1, F_2, …, F_k which affect the particular response. Each factor has continuous levels within a certain interval; e.g., F_i has levels X_i with $X_{iL} \le X_i \le X_{iU} (i = 1, 2, ..., k)$. A hypercube $\{X_{iL} \le X_i \le X_{iU}; (i = 1, 2, ..., k)\}$ can be defined as the **operational region (OR).** In the OR every level combination $(X_1, X_2, .., X_k)$ represents a feasible operating condition. Assuming that each such setting can be controlled (essentially without error) by the experimenter, a response is considered as a function of $(X_1, X_2, .., X_k)$ and is associated as

$$\eta = \phi\left(X_1, X_2, ..., X_k : \theta_1, \theta_2, ..., \theta_q\right) = \phi(X; \theta)$$

where $\theta_1, \theta_2, ..., \theta_q$ are parameters, $X = (X_1, X_2, ..., X_k)'$ and $\theta = (\theta_1, \theta_2, ..., \theta_q)'$. Now the true response $\eta = \eta(X_1, X_2, ..., X_k)$, and the form of the functional relationship ϕ at any given point in OR are unknown. Only observed responses $y = y(X)$ are available and attempt is made to approximate $\phi(X, \theta)$ by a polynomial function $f(X, \beta)$ in *X*. Then consider a model of the form

$$y(X) = f(X, \beta) + \varepsilon(X),$$

in place of $\eta = \phi(X, \theta)$ where $\beta = (\beta_1, \beta_2, ..., \beta_m)'$ are unknown parameters and $\varepsilon(X)$ represents random error.

Ideally, the experimenter wants to have $y(X)$ available for a sufficiently fine grid in OR so that approximate ϕ or a realization of ϕ can be adequately approximated. It is difficult in real experiments to do so. Instead of that, the experimenter has only a relatively small number of points (these are sometimes referred to as **runs** or **experiments**) and they are usually confined to a region called as **experimental region (ER)** or region of interest. Obviously, such an ER is contained in OR.

The basic ideas behind this method are the following:

- Based on the limited available knowledge about the process under study, the experimenter chooses an ER.
- Assuming that the response surface for ER is sufficiently smooth so that it can be approximated by a lower polynomial, say, first or second degree polynomial.
- Then an appropriate treatment and error control design can be chosen to estimate the coefficients of the polynomial.
- From this, the response can be predicted for any point in ER. If one of these points attains the optimal response then one may have reached an optimum which may only be an optimum either locally or globally.
- If the fitted response surface indicates that the optimum may only be outside ER then the experimenter can choose a new ER and repeat the whole process until the (predicted) optimum can be located more precisely.`

This procedure leads to two sources of error:

i. There can be experimental and sampling error in estimating the function $f(X;\beta)$ and

ii. some bias may be introduced is approximating $\phi(X;\theta)$ due to the inadequacy of $f(X;\beta)$.

The objective of response surface designs is to minimize these errors. The basic requirements for such designs are as follows:

1. The design should allow $f(X,\beta)$ to be estimated with reasonable precision in ER under the assumption that a polynomial $f(X;\beta)$ of degree approximates $\phi(X;\theta)$ sufficiently well.
2. A provision in design should be there to check whether the chosen $f(X;\beta)$ provides a satisfactory fit to the response surface or whether a different polynomial is more appropriate.
3. The design should not contain a large number of experimental points.
4. The design should be available for blocking of the experimental points.
5. One should be able to modify the design in case the polynomial of degree to which the polynomial is fitted is not found to be adequate and a polynomial of next higher degree needs to be fitted.

Now we discuss the basic tools and designs of RSM and point out the connection to treatment and error-control designs.

First-order models and designs

First-order regression model

The response surface function ϕ is approximated by a first-order polynomial within a small region based on the k input variables $X_1, X_2, ..., X_k$ as follows

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \varepsilon.$$

Here β_i is the regression coefficient associated with X_i and measures the change in the response y due to a change in the input variable X_i. This kind of information is available from the main effect from a factorial experiment where each factor has two levels. A good choice of a response surface design for such a situation is a 2^k factorial, or a fraction of it.

Suppose 2^k factorial is considered as a choice of response surface design then there are 2^k experimental points $(X_1, X_2, ..., X_k)_j$ say, with $j = 1, 2, .., 2^k$ With each level combination being replicated r times in a CRD, there are $N = r2^k$ experimental runs. The low and high level of the i^{th} factor are denoted by X_{i0} and X_{i1}, respectively. If the experimenter decides to use the coded levels X_i in place of X_{i0} and X_{i1}, then the coded levels are expressed as

$$X_i = \frac{X_i - \bar{X}}{\frac{1}{2}(X_{i1} - X_{i0})}$$

With such a transformation, the low level becomes $x_{i0} = -1$ and high level becomes $x_{i1} = 1.$

The model $y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \varepsilon$ is rewritten as as

$$y(x_1, x_2, ..., x_k)_\ell = \beta_0^* + \sum_{i=1}^{k} \beta_i^* x_i + \varepsilon(x_1, x_2, ..., x_k)_\ell,$$

where $x_i = \pm 1,$ or in matrix notation as

$$y = (L, \quad D)\beta^* + \varepsilon,$$

where *y* is the *N* x 1 vector of observation, *L* is an *N* x 1 vector of unity elements, *D* is the *N* x *k* design-model matrix consisting of elements -1's and 1's, $\beta^* = (\beta_0^*, \beta_1^*, ..., \beta_k^*)'$ and ε is the *N* x 1 vectors of errors.

More specifically, let

$$D = (d_1, d_2, ..., d_k)$$

where each $d_i, i = 1, 2, ..., k$ is a *N* x 1 column vector, we know that each d_i has $r2^{k-1}$ elements equal to -1 and $r2^{k-1}$ elements equal to 1. Thus $L'd_i = 0$ for every *i*. The $d_i's$ are orthogonal to each other.