Response Surfaces

Gary W. Oehlert

School of Statistics
University of Minnesota

December 2, 2013

. . .
I'm very well acquainted, too, with matters mathematical,
I understand equations, both the simple and quadratical,
About binomial theorem I'm teeming with a lot o' news,
With many cheerful facts about the square of the hypotenuse.
. . .

The Pirates of Penzance by Gilbert and Sullivan

Response surface methods are designed for

- Continuously variable factors
- A goal of optimization
- And perhaps a lesser goal of description

Widely used in industry and engineering.

Often used to get that last few percent of output after other methods have identified important factors.

Example. We want to know how arm length, rope length, and mass of counterweight affect the throw distance of a trebuchet.

The three factors can be continuously varied, and the function f gives the average throw distance for the parameter settings.

Gotcha example. A Harvey Wallbanger is a cocktail made from vodka, orange juice, and Galliano. We want to know how the amounts to add to get the best taste.

The “gotcha” is that only the proportions matter in this case, not the absolute amounts. For this we need a mixture design rather than a response surface design, even though there are similarities.

Models

> All models are wrong; some models are useful.
>
> George Box

We don't expect any of the models we use in response surface methods to be correct.

We do expect some of the models we use to be good enough to help us meet our goals.

Moral: don't get exercised over how ridiculously simplistic the models are.

Our response is a function of continuously variable factors $x_1, \ldots, x_q$ plus error, for example:

$$y_{ij} = f(x_{1i}, x_{2i}, \ldots, x_{qi}) + \epsilon_{ij}$$

We want to learn about f, for example:

- What shape does f have?
- Where is f optimized?
- Where is a constrained optimum?
- What direction do I move in from where I am now to improve response?

First order model:

$$f(x_{1i}, x_{2i}, \ldots, x_{qi}) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_q x_{qi}$$

Yes folks, a first order model is just an ordinary multiple regression.

Second order models:

$$\begin{aligned} f &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{qi} + \\ & \quad \beta_{11} x_{1i}^2 + \beta_{22} x_{2i}^2 + \cdots + \beta_{qq} x_{qi}^2 + \\ & \quad \beta_{12} x_{1i} x_{2i} + \beta_{13} x_{1i} x_{3i} + \cdots + \beta_{(q-1)q} x_{(q-1)i} x_{qi} \end{aligned}$$

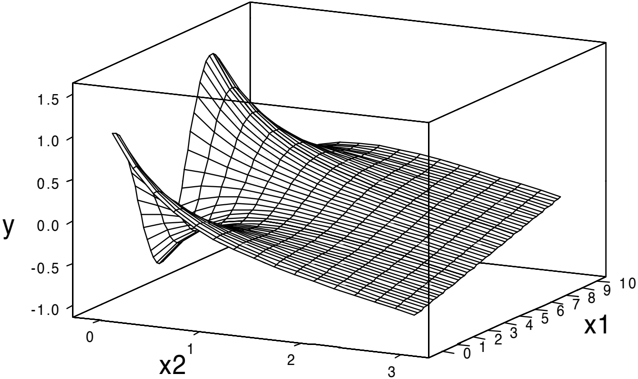
Linear terms, plus pure quadratic terms, plus cross product terms.

Yes, the models really are that simple.

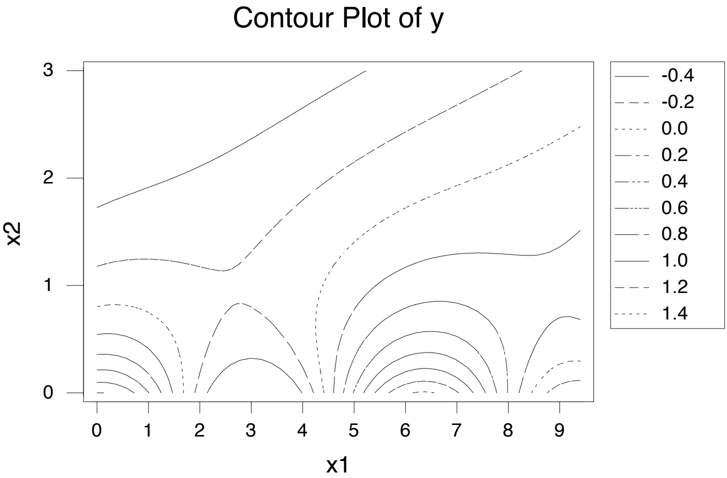
The remarkable thing is that they are usually good enough as long as:

1. The function is reasonably smooth (e.g., no jumps).
2. We are looking sufficiently locally.

Curious function:



Another view of the curious function:



For large values of X2, a first order model is adequate.

For small values of X2, we need a second order model. But, even this will only be approximately correct over a narrow range of X1.

More ways to write the first order model.

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_q x_{qi} + \epsilon_{ij} \\ &= \beta_0 + \sum_{k=1}^{q} \beta_k x_{ki} + \epsilon_{ij} \\ &= \beta_0 + \mathbf{x}_i' \boldsymbol{\beta} + \epsilon_{ij} \quad , \end{aligned}$$

where $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{qi})'$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_q)'$.

First order models describe tilted planes. The contours (level sets) are lines.

$\boldsymbol{\beta}$ is the direction of steepest ascent.

If you want to increase the function as quickly as possible, move in the direction of $\boldsymbol{\beta}$. That is, go from $\mathbf{x}$ to $\mathbf{x} + \lambda\boldsymbol{\beta}$ for some multiplier λ.

More ways to write the second order model.

$$
\begin{aligned}
y_{ij} &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_q x_{qi} + \\
&\quad \beta_{11} x_{1i}^2 + \beta_{22} x_{2i}^2 + \cdots + \beta_{qq} x_{qi}^2 + \\
&\quad \beta_{12} x_{1i} x_{2i} + \beta_{13} x_{1i} x_{3i} + \cdots + \beta_{1q} x_{1i} x_{qi} + \\
&\quad \beta_{23} x_{2i} x_{3i} + \beta_{24} x_{2i} x_{4i} + \cdots + \beta_{2q} x_{2i} x_{qi} + \\
&\quad \cdots + \beta_{(q-1)q} x_{(q-1)i} x_{qi} + \epsilon_{ij} \\
&= \beta_0 + \sum_{k=1}^{q} \beta_k x_{ki} + \sum_{k=1}^{q} \beta_{kk} x_{ki}^2 + \sum_{k=1}^{q-1} \sum_{l=k+1}^{q} \beta_{kl} x_{ki} x_{li} + \epsilon_{ij} \\
&= \beta_0 + \mathbf{x}_i' \boldsymbol{\beta} + \mathbf{x}_i' \mathcal{B} \mathbf{x}_i + \epsilon_{ij} \quad ,
\end{aligned}
$$

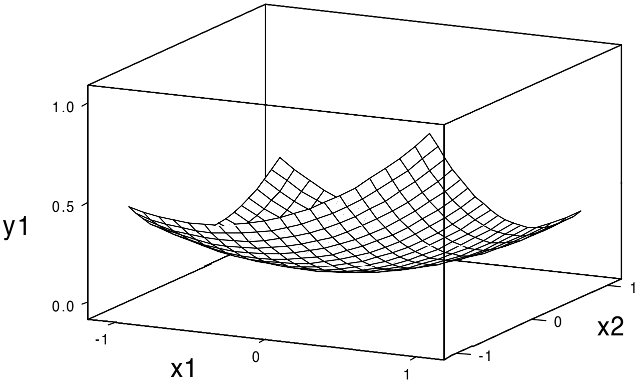
where once again $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{qi})'$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_q)'$, and $\mathcal{B}$ is a $q \times q$ matrix with $\mathcal{B}_{kk} = \beta_{kk}$ and $\mathcal{B}_{kl} = \mathcal{B}_{lk} = \beta_{kl}/2$.

The diagonal of $\mathcal{B}$ has the pure quadratic coefficients, and the off diagonal entries are the cross product coefficients.

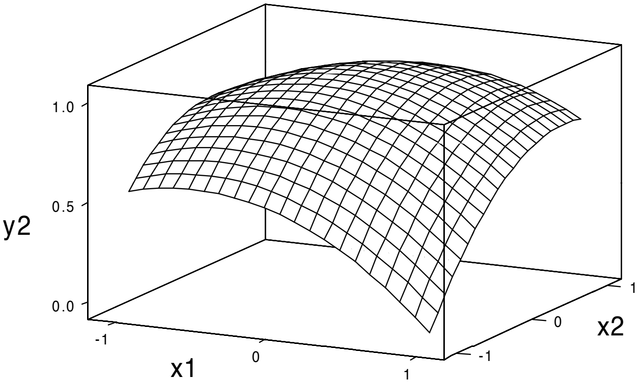
Note that the model only includes the kl cross product for $k < l$; the matrix form with $\mathcal{B}$ includes both kl and lk, so the coefficients are halved to take this into account.

Second order models can describe many shapes.

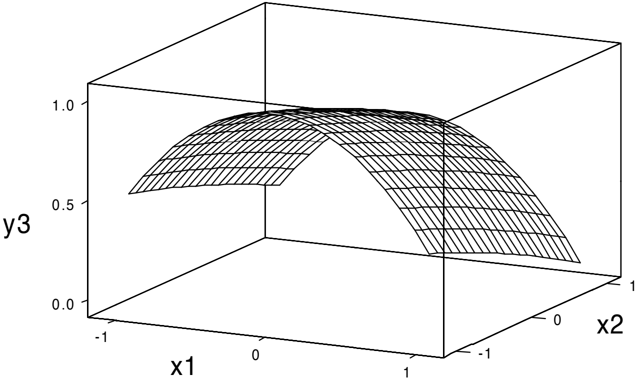
A minimum.



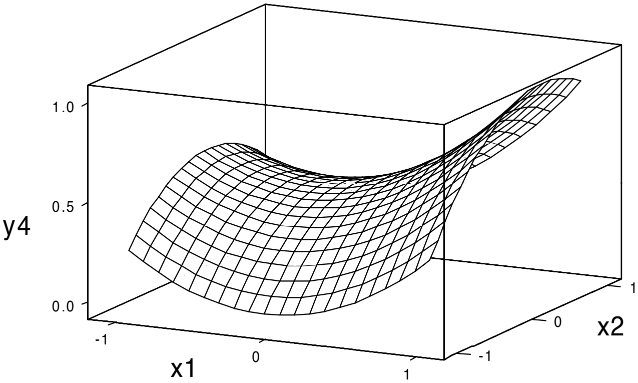
A maximum.



A ridge.



A saddle.



These shapes are determined by $\mathcal{B}$.

We will determine a point $\mathbf{x}_0$ and a matrix H and make canonical variables

$$\mathbf{v} = H'(\mathbf{x} - \mathbf{x}_0)$$

In the canonical variables, the function is

$$f_v(\mathbf{v}) = f_v(0) + \sum_{k=1}^{q} \lambda_k v_k^2 \quad ,$$

where $f_v(0) = f(\mathbf{x}_0)$

If all of the λ_is are positive, then the shape has a minimum.

If all of the λ_is are negative, then the shape has a maximum.

If all of the λ_is are one sign, but one or more of them is close to zero, then you get a ridge (or a trench).

If some λ_i are positive and some are negative, then you get a saddle.

Algebra technicalities:

The matrix H contains the eigenvectors of $\mathcal{B}$ as columns.

The λ_is are the eigenvalues of $\mathcal{B}$.

$\mathbf{x}_0 = -\frac{1}{2}\mathcal{B}^{-1}\boldsymbol{\beta}$

$f(\mathbf{x}_0) = \beta_0 - \frac{1}{4}\boldsymbol{\beta}'\mathcal{B}^{-1}\boldsymbol{\beta}$

The closer you get to a ridge system (some zero λ_is) the more poorly $\mathbf{x}_0$ is defined from a numerical perspective (it's like dividing by 0).

Designs to fit Models

Our basic approach is to fit the first order model until it is no longer tenable; then we fit the second order model.

Thus our designs must give us data to:

- Fit the model.
- Estimate error.
- Estimate lack of fit.

We would prefer also to have efficiency.

Coded variables.

Coded variables rescale the xs so that 0 is in the center of the design and $+1$ and –1 are “standard” steps up and down from the center.

Models are more easily expressed in coded variables.

Coded variables are less collinear, simplifying estimation of terms.

To fit a first order model, use a 2^q factorial with “corners” at $+1$ and –1 (in coded variables).

Then add multiple center points at the origin (all 0s). (How many?)

Any resolution III or greater fraction can be used instead of the full factorial.

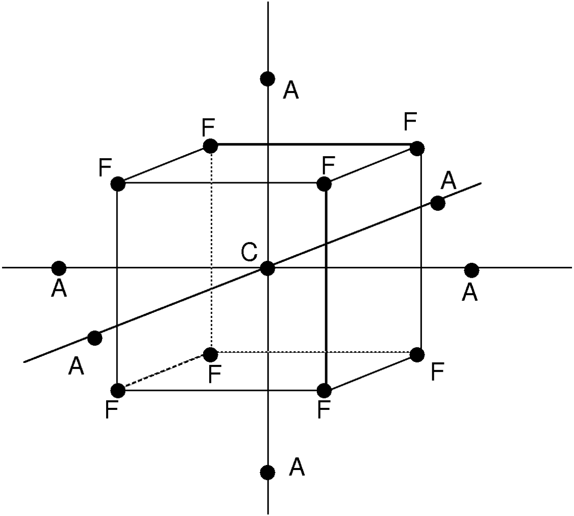
SS between center points is pure error. All other residual SS is lack of fit.

To fit a second order model, use a central composite design.

1. Begin with a 2^q factorial with “corners” at $+1$ and –1 (in coded variables). A resolution V fraction can be used instead.
2. Add center points at 0.
3. Add axial points at $(\pm\alpha, 0, 0, \ldots, 0)$, $(0, \pm\alpha, 0, 0, \ldots, 0)$, . . . $(0, 0, \ldots, 0, \pm\alpha, 0)$, $(0, 0, \ldots, 0, \pm\alpha)$.

Need to select α and number of center points.

See Table 19.1 of text for α and n_0 in the case that the design is run in incomplete blocks. This gives blocks orthogonal to treatment effects.


A
F
F
F
F
A
C
A
A
A
F
F
F
F
A

When not blocked, there are heuristics, but no rules.[1]

Rotatable designs choose $\alpha = 2^{(q-k)/4}$ (k could be 0). These designs do not favor one direction over another.

Face centered designs use $\alpha = 1$. The are more practical in some situations.

Putting the non-center points on a sphere $\alpha = \sqrt{q}$ works well when the sphere is the zone of interest.

Setting n_0 so that prediction variance is the same at 0 and at distance 1 is called uniform precision; this is not very compelling.

[1] "The code is more of what you'd call guidelines than actual rules."

Overall plan

If we know that we are working in a area of curvature, begin with a second order design.

If we might be off on the edge somewhere, start with a first order design.

Check lack of fit in first order design.

1. If there is lack of fit, augment first order design to second order design (add axial points and more center points according to blocking rules).
2. If no lack of fit, move off in direction of steepest ascent (assuming bigger is better) as long as we keep ascending.
3. Once we go over the top, put down another first order design and return to step 1.

Once we are in a second order design setting, do canonical analysis (see below) to find optimum.

If optimum is outside region of experimentation, move in that direction and try another second order design.

In all cases, we need to keep checking on which terms and factors are important, on lack of fit, and on the usual assumptions.

Analysis

First order analysis is just linear regression.

Pure error is error from replication. We can get this by fitting a model in R that “joins” the predicting variables together.

Anova comparing the regular model to this joined model is a lack of fit test.

We need to worry about the usual assumptions.

Second order analysis is just polynomial regression.

The canonical analysis works on the matrix of coefficients. We have special R functions to do that.

Note: if we have some noise variables in the model, the canonical analysis will be highly suspect. Test all coefficients for a variable before deleting it from model.

Constraints

Often we must optimize under constraints.