Experimental Design

Unit 11

Werner Müller

Institute of Applied Statistics

June 4th 2025

16.1 Introduction to regression experiments: second-order polynomial models

We now consider designs for experiments in which the anticipated model is quadratic in the controlled variables.

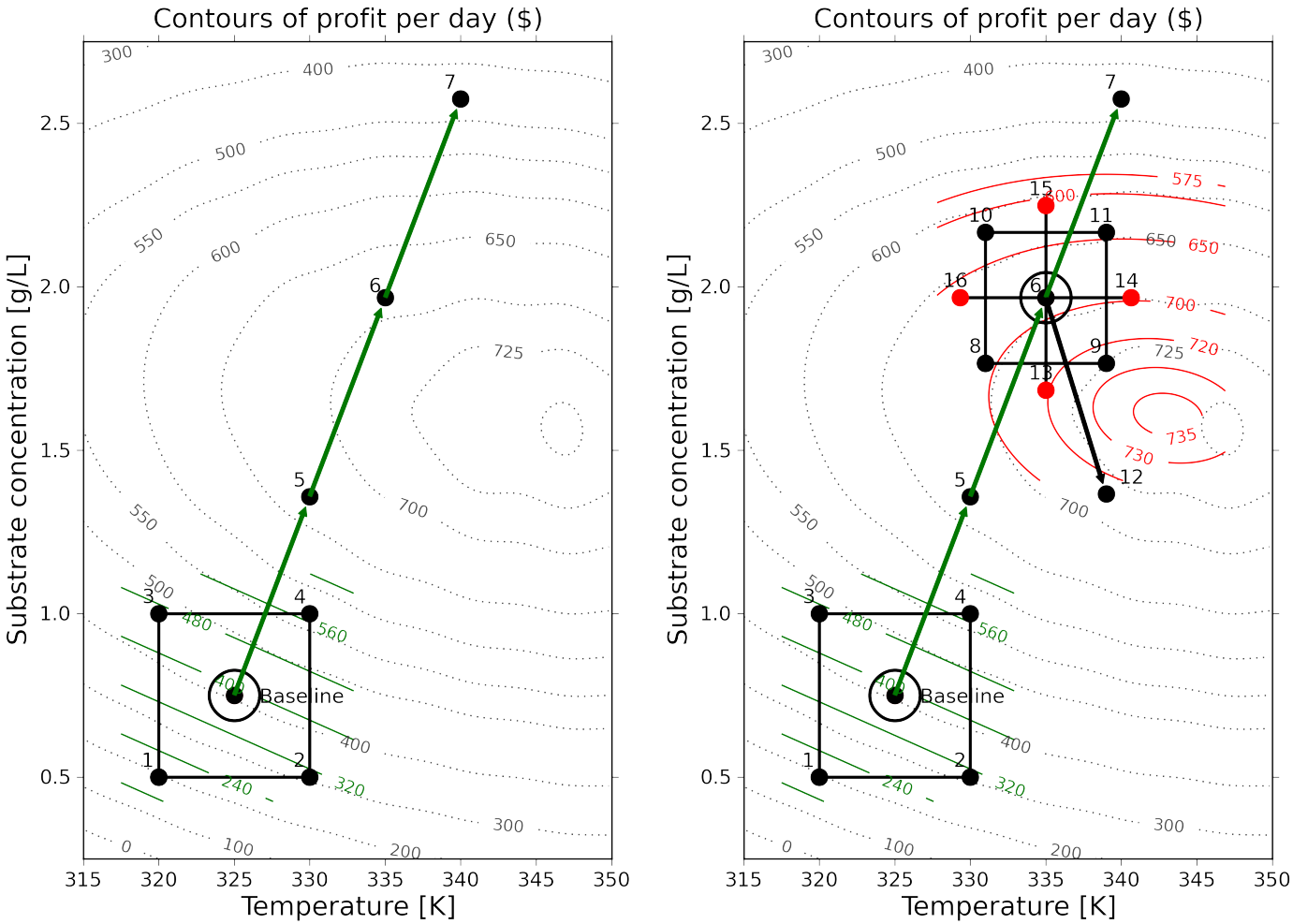
Response surface optimization strategies are often based on a series of experiments, focusing on

- first-order regression in the early iterations to establish paths of „steepest ascent“ to new experimental regions, and
- second-order regression in the final step(s) when the conditions that will optimize the process are within the region of the design.

Of course, regression is possible with even higher-order polynomial models but the number of parameters in polynomial models of order 3 or more is quite large, requiring experiments that are often impractically large.

Regression experiments designed for quadratic polynomials provide more information about the nature of the expected response-versus-controlled variable relationship than those designed for first-order polynomial models, at the cost of a larger number of required experimental runs.

A typical example



taken from `https://learnche.org/pid/design-analysis-experiments/index`

16.2 Quadratic polynomial models

We again define a d-dimensional vector $\mathbf{x} = (x_1, x_2, \ldots, x_d)^T$, a point in the d-dimensional experimental region R specifying a particular experimental treatment. A second-order, or quadratic, polynomial model for the expected response may be written as:

$$\mathsf{E}(y) = \alpha + \sum_{i=1}^{d} x_i \beta_i + \sum_{i=1}^{d} x_i^2 \beta_{ii} + \sum_{i=1}^{d-1} \sum_{j=i+1}^{d} x_i x_j \beta_{ij}$$

An equivalent matrix notation that is sometimes more convenient is:

$$\mathsf{E}(y) = \alpha + \mathbf{x}^T \boldsymbol{\beta}_1 + \mathbf{x}^T \mathbf{B}_2 \mathbf{x}$$

where $\boldsymbol{\beta}_1$ is a d-vector with $(\boldsymbol{\beta}_1)_i = \beta_i$ and
$\mathbf{B}_2$ is a $(d \times d)$-matrix with $(\mathbf{B}_2)_{ii} = \beta_{ii}$ and $(\mathbf{B}_2)_{ij} = (\mathbf{B}_2)_{ji} = \frac{\beta_{ij}}{2}$.

Quadratic polynomial models (cont.)

In many regression experiments, an important goal is to identify values of the independent variables that result in minimal or maximal expected response.

For known coefficients of the quadratic model, the *stationary point* or points at which derivatives of $\mathsf{E}(y)$ with respect to each x_i are zero can be identified as

$$\frac{\partial}{\partial \mathbf{x}} \mathsf{E}(y) = \boldsymbol{\beta}_1 + 2\mathbf{B}_2\mathbf{x}_0 = \mathbf{0}$$

When $\mathbf{B}_2$ is regular, the solution and can be written as:

$$\mathbf{x}_0 = -\frac{\mathbf{B}_2^{-1}\boldsymbol{\beta}_1}{2}$$

$\mathbf{x}_0$ maximizes $\mathsf{E}(y)$ if all eigenvalues of $\mathbf{B}_2$ are negative, and minimizes $\mathsf{E}(y)$ if all eigenvalues of $\mathbf{B}_2$ are positive.

Quadratic polynomial models (cont.)

The regression coefficients are unknown quantities and can only be estimated, the (estimated) coefficients from the fitted model can be substituted in the above expressions to yield an estimated stationary point $\mathbf{x}_0$ for an additional treatment.

$$\hat{\mathbf{x}}_0 = -\frac{\hat{\mathbf{B}}_2^{-1}\hat{\boldsymbol{\beta}}_1}{2}$$

But eigenvalues of $\hat{\mathbf{B}}_2$ might be mixed in sign, such that the estimated stationary point neither maximizes nor minimizes the fitted response model.

Even more attention should be directed to the possibility that $\hat{\mathbf{x}}_0$ might not be physically meaningful at all in the context of the problem. $\hat{\mathbf{x}}_0$ can be forced to lie in the experimental region by constrained optimization techniques.

Generally, estimated stationary points that lie outside the region in which an experiment is conducted should be regarded with suspicion, because extrapolations based on polynomial models are often unreliable even when the model is an adequate approximation to the truth in a limited region.

Quadratic polynomial models (cont.)

For an n-run experiment, we can write a matrix model for the full experiment as:

$$\mathbf{y} = \alpha\mathbf{1} + \mathbf{X}_2\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \alpha\mathbf{1} + \mathbf{X}_L\boldsymbol{\beta}_L + \mathbf{X}_{PQ}\boldsymbol{\beta}_{PQ} + \mathbf{X}_{MQ}\boldsymbol{\beta}_{MQ} + \boldsymbol{\varepsilon}$$

- each row of $\mathbf{X}_L$ is the transpose of the vector $\mathbf{x}$ associated with the corresponding run,
- $\boldsymbol{\beta}_L$ is the vector of „linear“ polynomial coefficients,
- $\mathbf{X}_{PQ}$, like $\mathbf{X}_L$, is a $(n \times d)$-matrix, $(\mathbf{X}_{PQ})_{ij}$ is the square of $(\mathbf{X}_L)_{ij}$,
- $\boldsymbol{\beta}_{PQ}$ is the set of „pure quadratic“ coefficients - those associated with squared controlled variables,
- $\mathbf{X}_{MQ}$ is a $(n \times \frac{d(d-1)}{2})$-matrix with the pairwise products of all distinct pairs of elements from the vector $\mathbf{x}$
- $\boldsymbol{\beta}_{MQ}$ contains the „mixed quadratic“ coefficients - those associated with products of two controlled variables.

16.3 Designs for second-order models

In order to support estimation of the coefficients in a quadratic polynomial model, an experimental design must include at least 3 distinct values for each controlled variable.

16.3.1 Complete three-level factorial designs

In the complete factorial design (CFD) each controlled variable takes on three coded levels within R.

Denoting the levels as $\{-f, 0, f\}$ for each variable, and adding $n_c - 1$ additional runs at the center point $\mathbf{x} = \mathbf{0}$, to provide replicate information from which σ^2 can be estimated, such a design contains a total of $n = 3^d + n_c - 1$ runs.

The design matrices for a three-level factorial plan in $d = 2$ controlled variables, with $n_c = 2$ center points, can be seen on the next slide.

Complete three-level factorial designs (cont.)

$$\mathbf{X}_L = \begin{pmatrix} -f & -f \\ -f & 0 \\ -f & f \\ 0 & -f \\ 0 & 0 \\ 0 & f \\ f & -f \\ f & 0 \\ f & f \\ 0 & 0 \end{pmatrix} \qquad \mathbf{X}_{PQ} = \begin{pmatrix} f^2 & f^2 \\ f^2 & 0 \\ f^2 & f^2 \\ 0 & f^2 \\ 0 & 0 \\ 0 & f^2 \\ f^2 & f^2 \\ f^2 & 0 \\ f^2 & f^2 \\ 0 & 0 \end{pmatrix} \qquad \mathbf{X}_{MQ} = \begin{pmatrix} f^2 \\ 0 \\ -f^2 \\ 0 \\ 0 \\ 0 \\ -f^2 \\ 0 \\ f^2 \\ 0 \end{pmatrix}$$

Complete three-level factorial designs are popular plans when d is relatively small. However, for larger values of d, the number of experimental runs required is impractically large in many settings.

16.3.2 Central composite designs

The most widely used experimental plan in situations where a quadratic polynomial model is anticipated is the central composite design (CCD).

It can be thought of as being comprised of three „subdesigns“ each located so that its „center of gravity“ corresponds to the center point $\mathbf{x} = \mathbf{0}$.

- an n_f-run orthogonal two-level design; either a full factorial plan in the d controlled variables, or a regular fractional factorial plan of resolution at least V,
- a collection of n_c runs taken at the center point, and
- an „axial“ subdesign of $2d$ treatments, each of which is defined by setting one of the controlled variables to a standard nonzero value a, either positive or negative, and all other controlled variables to zero.

The overall size of the design is then $n = n_f + 2d + n_c$.

Central composite designs (cont.)

The design matrices for a central composite design in $d = 2$ controlled variables, with $n_c = 2$ center points, can be written as

$$\mathbf{X}_L = \begin{pmatrix} -f & -f \\ -f & f \\ f & -f \\ f & f \\ -a & 0 \\ a & 0 \\ 0 & -a \\ 0 & a \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \quad \mathbf{X}_{PQ} = \begin{pmatrix} f^2 & f^2 \\ f^2 & f^2 \\ f^2 & f^2 \\ f^2 & f^2 \\ a^2 & 0 \\ a^2 & 0 \\ 0 & a^2 \\ 0 & a^2 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \quad \mathbf{X}_{MQ} = \begin{pmatrix} f^2 \\ -f^2 \\ -f^2 \\ f^2 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

16.3.3 Box-Behnken designs

Box-Behnken designs (BBD) are symmetric three-level designs, consisting of a combination of two-level factorial plans, each constructed using only a subset of the controlled variables.

For d controlled variables, one first selects a balanced incomplete block design (BIBD) for d treatments in b blocks of size $n_f < d$. This BIBD is just used in the construction process of the BBD.

Each treatment in the BIBD is associated with one of the regression variables.

For each block in the BIBD, a two-level factorial design or regular fractional factorial design of resolution at least V is selected for only those variables associated with BIBD treatments included in the block; all other controlled variables are zero in these runs.

n_c center point runs are added to form a Box-Behnken regression design of $n = b\, n_f + n_c$ runs.

Box-Behnken designs (cont.)

A quadratic regression design for $d = 3$ controlled variables can be formed using the BIBD for three treatments in three blocks of two units each:

1	2	1	3	2	3

Using a complete 2^2 factorial design in the pairs of factors associated with each block, a Box-Behnken design with $n_c = 3$ center point runs can then be constructed:

$$\mathbf{X}_L = \begin{pmatrix} -f & -f & 0 \\ -f & f & 0 \\ f & -f & 0 \\ f & f & 0 \\ -f & 0 & -f \\ -f & 0 & f \\ f & 0 & -f \\ f & 0 & f \\ 0 & -f & -f \\ 0 & -f & f \\ 0 & f & -f \\ 0 & f & f \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \mathbf{X}_{PQ} = \begin{pmatrix} f^2 & f^2 & 0 \\ f^2 & f^2 & 0 \\ f^2 & f^2 & 0 \\ f^2 & f^2 & 0 \\ f^2 & 0 & f^2 \\ f^2 & 0 & f^2 \\ f^2 & 0 & f^2 \\ f^2 & 0 & f^2 \\ 0 & f^2 & f^2 \\ 0 & f^2 & f^2 \\ 0 & f^2 & f^2 \\ 0 & f^2 & f^2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \mathbf{X}_{MQ} = \begin{pmatrix} f^2 & 0 & 0 \\ -f^2 & 0 & 0 \\ -f^2 & 0 & 0 \\ f^2 & 0 & 0 \\ 0 & f^2 & 0 \\ 0 & -f^2 & 0 \\ 0 & -f^2 & 0 \\ 0 & f^2 & 0 \\ 0 & 0 & f^2 \\ 0 & 0 & -f^2 \\ 0 & 0 & -f^2 \\ 0 & 0 & f^2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

16.3.4 Augmented pairs designs

The augmented pairs designs (APD) are constructed by combining sets of design points.

- The first set of points is a two-level fractional factorial plan of size n_f and of resolution at least III (or a full two-level factorial design for d = 2); the n_f points in this two-level design are specified by the vectors of controlled variables $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n_f})$.
- The second set of points contains one experimental run determined by each distinct pair of runs in the first set, specified as: $\mathbf{x}_{ij} = -\frac{\mathbf{x}_i + \mathbf{x}_j}{2}$
- The third set of points are n_c replicated runs at the design region center point.

The augmented pairs design contains a total of $n = n_f + \begin{pmatrix} n_f \\ 2 \end{pmatrix} + n_c$ runs.

Augmented pairs designs (cont.)

For $d = 3$, the design matrices for an augmented pairs design containing $n_c = 3$ center point runs are:

$$\mathbf{X}_L = \begin{pmatrix} f & f & f \\ f & -f & -f \\ -f & f & -f \\ -f & -f & f \\ -f & 0 & 0 \\ 0 & -f & 0 \\ 0 & 0 & -f \\ 0 & 0 & f \\ 0 & f & 0 \\ f & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \mathbf{X}_{PQ} = \begin{pmatrix} f^2 & f^2 & f^2 \\ f^2 & f^2 & f^2 \\ f^2 & f^2 & f^2 \\ f^2 & f^2 & f^2 \\ f^2 & 0 & 0 \\ 0 & f^2 & 0 \\ 0 & 0 & f^2 \\ 0 & 0 & f^2 \\ 0 & f^2 & 0 \\ f^2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \mathbf{X}_{MQ} = \begin{pmatrix} f^2 & f^2 & f^2 \\ -f^2 & -f^2 & f^2 \\ -f^2 & f^2 & -f^2 \\ f^2 & -f^2 & -f^2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

The first set of $n_f = 4$ points here is based on a 2^{3-1}_{III} fraction.

16.4 Design scaling and information

All designs in each of the four classes described above have two important „balance“ properties:

- The elements in each column of $\mathbf{X}_L$ have a zero sum.
- The elements in each column of $\mathbf{X}_{MQ}$ have a zero sum.

If we use a model containing only the „nuisance parameter“ α, the hat matrix is of form $\mathbf{H}_1 = \frac{1}{n}\mathbf{J}$.

As a result, $(\mathbf{I} - \mathbf{H}_1)\mathbf{X}_L = \mathbf{X}_L$ and $(\mathbf{I} - \mathbf{H}_1)\mathbf{X}_{MQ} = \mathbf{X}_{MQ}$.

The zero-sum property cannot hold for the columns of $\mathbf{X}_{PQ}$, however, because all nonzero elements of this matrix must be positive. But

- The elements in each column of $\mathbf{X}_{PQ}$ have the same average, say a_{PQ}.

So $(\mathbf{I} - \mathbf{H}_1)\mathbf{X}_{PQ} = \mathbf{X}_{PQ} - a_{PQ}\mathbf{J}_{n\times d}$.

Design scaling and information (cont.)

We can derive the form of a_{PQ} for each design:

design	a_{PQ}	design	a_{PQ}
CFD	$2 \cdot 3^{d-1} \cdot \frac{f^2}{n}$	BBD	$r \cdot 2^{n_f} \cdot \frac{f^2}{n}$
CCD	$2^d \cdot \frac{f^2}{n} + \frac{2a^2}{n}$	ADP	$(n_f + \frac{n_f}{2}(\frac{n_f}{2} - 1))\frac{f^2}{n}$

where r is the replication factor for the BIBD on which the BBD is based.

Because the columns of $\mathbf{X}_L$ are orthogonal to each other and to those in $\mathbf{X}_{MQ}$ the design information matrices of the 4 second order models are

$$\mathcal{I}_{\mathsf{CFD}} = \begin{pmatrix} 6 \cdot f^2 \mathbf{I}_d & 0 & 0 \\ 0 & f^4(\frac{12}{5}\mathbf{I}_d + \frac{2}{5}\mathbf{J}_d) & 0 \\ 0 & 0 & 4 \cdot f^4 \mathbf{I}_{\frac{d(d-1)}{2}} \end{pmatrix} \quad \mathcal{I}_{\mathsf{BBD}} = \begin{pmatrix} 8 \cdot f^2 \mathbf{I}_d & 0 & 0 \\ 0 & f^4(\frac{104}{15}\mathbf{I}_d + \frac{44}{15}\mathbf{J}_d) & 0 \\ 0 & 0 & 4 \cdot f^4 \mathbf{I}_{\frac{d(d-1)}{2}} \end{pmatrix}$$

$$\mathcal{I}_{\mathsf{CCD}} = \begin{pmatrix} 6 \cdot f^2 \mathbf{I}_d & 0 & 0 \\ 0 & f^4(\frac{12}{5}\mathbf{I}_d + \frac{2}{5}\mathbf{J}_d) & 0 \\ 0 & 0 & 4 \cdot f^4 \mathbf{I}_{\frac{d(d-1)}{2}} \end{pmatrix} \quad \mathcal{I}_{\mathsf{ADP}} = \begin{pmatrix} 6 \cdot f^2 \mathbf{I}_d & 0 & 0 \\ 0 & f^4(\frac{42}{13}\mathbf{I}_d + \frac{16}{13}\mathbf{J}_d) & 0 \\ 0 & 0 & 4 \cdot f^4 \mathbf{I}_{\frac{d(d-1)}{2}} \end{pmatrix}$$

In the computation of $\mathcal{I}_{\mathsf{CCD}}$ we set $a = f$.

Design scaling and information (cont.)

In the three-level plans above have coded controlled variable values restricted to $-f, 0$ and f, so the overall scale of $\mathcal{I}$ for these designs is governed by the value of f.

Consider two designs from any of the above classes that are identical except that design A is scaled by f_A and design B is scaled by f_B. If the information matrix for design A is $\mathcal{I}_A$, then the information matrix for design B is:

$$\mathcal{I}_B = \begin{pmatrix} \frac{f_B}{f_A}\mathbf{I}_d & 0 \\ 0 & \frac{f_B^2}{f_A^2}\mathbf{I}_{\frac{d(d+1)}{2}} \end{pmatrix} \cdot \mathcal{I}_A \cdot \begin{pmatrix} \frac{f_B}{f_A}\mathbf{I}_d & 0 \\ 0 & \frac{f_B^2}{f_A^2}\mathbf{I}_{\frac{d(d+1)}{2}} \end{pmatrix}$$

All functions of the information matrix that reflect statistical performance (noncentrality and estimation precision) are superior for the design that has larger „span“ as measured by f. For these designs, „f“ is usually simply coded as „1“.

16.5 Orthogonal blocking

May blocking be arranged such that efficiency is not lost? Equivalently, we can ask whether $\mathcal{I}$ is the same when block parameters are included in the model as when α is the only nuisance parameter, i.e. whether the design is blocked orthogonally.

Recall that for unblocked designs, the four design classes we have considered are each such that:

- The elements in each column of $\mathbf{X}_L$ have a zero sum.
- The elements in each column of $\mathbf{X}_{MQ}$ have a zero sum.
- The elements in each column of $\mathbf{X}_{PQ}$ have the same average a_{PQ}.

Hence for the blocked case, $(\mathbf{I} - \mathbf{H}_1)\mathbf{X}_L$ and $(\mathbf{I} - \mathbf{H}_1)\mathbf{X}_{MQ}$ will be as they are in the unblocked case if the elements of each column of $\mathbf{X}_L$ and $\mathbf{X}_{MQ}$ sum to zero within each block.

$(\mathbf{I} - \mathbf{H}_1)\mathbf{X}_{PQ}$ will be as it is in the unblocked case if the elements of each column of $\mathbf{X}_{PQ}$ have an average value of a_{PQ} within each block. The number of center points in each block can be adjusted to meet the conditions for orthogonal blocking.

16.6 Split-plot designs

The general structure for split-plot regression experiments is essentially the same regardless of the order of model used in analysis

The values of some controlled variables remain constant within blocks/plots, while the values of other controlled variables differ across units/split-plots within a block/plot.

Model monomials (x_i, x_i^2, and $x_i x_j$) that have constant values within each block are assessed in the whole-plot section of the ANOVA decomposition, relative to a MSE reflecting block-to-block variation.

Other model terms for which values change from unit to unit within blocks are assessed in the split-plot section of the ANOVA decomposition, relative to a MSE reflecting unit-to-unit (within-block) variation.

16.7 Bias due to omitted model terms

Quadratic regression models are useful and popular because they can approximate many (but not all) „curved“ functions reasonably well over a limited domain.

It is important to understand how much estimation bias might result from the omission of higher-order (i.e., greater than two) monomial terms from the model.

The phenomenon of coefficient estimate bias is really general, and can occur with any linear model when some important terms are omitted (or cannot be estimated based on the selected design).

Suppose the data were actually generated by a model of form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{W}\boldsymbol{\varphi} + \boldsymbol{\varepsilon}$$

and we fit a model of matrix form $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$ to the data.

Bias due to omitted model terms (cont.)

The design matrix $\mathbf{X}$ would include the collection of columns from $\mathbf{1}$, $\mathbf{X}_L$, $\mathbf{X}_{PQ}$, and $\mathbf{X}_{MQ}$, while $\mathbf{W}$ might contain columns associated with the omitted third-order monomials (e.g., $x_i^2 x_j$).

If $\mathbf{X}$ has full rank the unique least-squares estimate of $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. The „true“ expectation of $\hat{\boldsymbol{\theta}}$ then is

$$\mathsf{E}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{\varphi}$$

i.e. the bias of each estimated coefficient is determined by the value of the omitted coefficients $\boldsymbol{\varphi}$, and the form of $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}$, often called the *alias matrix*.

An experimental design for which the *alias matrix* contains elements of relatively small absolute value is often preferred, because this offers relatively more protection against coefficient bias if it turns out that $\boldsymbol{\varphi}$ is nonzero.

17.1 Introduction to optimal design

In many cases, the designs presented so far are optimal, in the sense that no other experimental plan in the same number of runs can provide more precise estimates or powerful tests against broad collections of alternative hypotheses for the parameters of interest, given the assumed model.

Optimal design provides a more direct connection between experimental design and statistical performance by framing design selection as an optimization problem, in which standard errors are minimized or noncentrality parameters are maximized.

17.2 Optimal design fundamentals

The framework for optimal design construction requires specification of three entities:

- *The experimental region R*: the finite or infinite set of values for $\mathbf{x}$, the scalar or vector of independent variables that defines a treatment. E.g.
  $R = \{1, 2, \ldots, t\}$ for an experiment in t unstructured treatments,
  $R = \{0, 1\}^f$ in two-level factorial experiments,
  $R = [-1, 1]^d$ for regression problems when the d independent variables have each been scaled to $[-1, 1]$.
- *The (here linear) model*:

$$M: \quad \mathbf{y} = \mathbf{t}_{\mathbf{x}}^T \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

  where $\mathbf{t}_{\mathbf{x}}$ is a function of the elements of $\mathbf{x}$.
  E.g. a main effects model for an experiment in f two-level factors yields $\mathbf{t}_{\mathbf{x}}^T = (\mathbf{1}, \mathbf{x}^T)$ of $f + 1$ elements
  a quadratic regression model in d independent variables leads to a vector $\mathbf{t}_{\mathbf{x}}$ of $1 + 2d + \begin{pmatrix} d \\ 2 \end{pmatrix}$ elements.

Optimal design fundamentals (cont.)

- *The criterion function*: For any specified design, $D = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, the concept of optimal design requires a function, $\varphi_M(D)$, that can be used as a measure of quality of the inference that can be expected from the resulting data.
  This might be the power of a test (maximize) or the standard deviation of an estimate (to be minimized).

  The criterion function also depends on the form of the model M, e.g. designs that are very good for factorial models containing only main effects can be very poor when interactions of higher order are included.

The general idea is to identify the design or designs, comprised of runs from the experimental region, that maximize or minimize the criterion function, e.g.:

$$D_{opt} = \arg\max_{D} \ \varphi_M(D) \quad \text{such that} \quad D = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \quad \text{and} \quad \mathbf{x}_i \in R$$

Optimal design fundamentals (cont.)

Optimal designs are especially valuable with:

- *Nonstandard experimental regions*: Some designs require special standard experimental regions. E.g. $R = \{0, 1\}^f$ for fractional factorial designs or $R = [-1, 1]^d$ for central composite designs (CCDs).

  Physical operating constraints may require $x_1 + 2x_2 \leq 2$, say; i.e., $R = \{\mathbf{x} : \ -1 \leq x_i \leq +1, x_1 + 2x_2 \leq 2\}$.

  Because optimal designs are region-specific, they can be constructed for any proposed experimental region.

- *Nonstandard models*: Standard designs such as balanced incomplete block designs (BIBDs), regular fractional factorials, and CCDs are widely used because they perform well in the wide variety of circumstances.

  Because optimal designs are model-specific, they can be constructed for any proposed linear model, not just the „standard“ forms.

Optimal design fundamentals (cont.)

- *Nonstandard experiment size*: For many of the classes of designs we have examined, there are restrictions on the value of n

  Because optimal designs are the solution to optimization problems for which n can be specified to be any desired integer value, attention need not be limited to the values that are required or convenient for any particular class of designs.

A major disadvantage of optimal design is that construction of a design requires solution of a mathematical optimization problem that is usually of high dimension, can often be approached only numerically, and for which „true“ optimal solutions sometimes cannot be practically verified.

17.3 Optimality criteria
17.3.1 A-optimality

„A“ optimality refers to designs for which the average variance of estimates of interest is minimized.

suppose we consider the basic linear model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

with full rank $\mathbf{X}$ and $\mathcal{I} = \mathbf{X}^T\mathbf{X}$. If the estimates of interest are the (entire) set of elements of $\hat{\boldsymbol{\theta}}$ we note that their individual variances can be written as:

$$\mathsf{Var}\left((\hat{\boldsymbol{\theta}})_i\right) = \sigma^2(\mathcal{I}^{-1})_{ii}$$

So an A-optimal design minimizes

$$\varphi_M(D) = \text{trace}(\mathcal{I}^{-1})$$

over possible designs.

17.3.2 D-optimality

„D“-optimality refers to designs for which the determinant of the covariance matrix of estimates of interest is minimized.

This determinant is monotonically related to the volume of the simultaneous confidence ellipsoid for the parameters. For the standard linear model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

with full rank $\mathbf{X}$ and $\mathcal{I} = \mathbf{X}^T\mathbf{X}$ such designs minimize

$$\varphi_M(D) = |\mathcal{I}^{-1}|$$

or equivalently, maximize

$$\varphi_M(D) = |\mathcal{I}|$$

over possible designs. The latter form is often preferred in practice since it avoids the unnecessary numerical step of inverting $\mathcal{I}$.

17.3.3 G-optimality, I-optimality

D- and A-optimality are criteria for *parameter estimation*.

When *model predictions* $\hat{\mathbf{y}}$ for unobserved values for $\mathbf{x}$ are of primary interest, G-optimality (for „global“) and I-optimality (for „integrated“) are the appropriate criteria.

G-optimal designs are those for which the largest (with respect to $\mathbf{x} \in R$) value of $\mathsf{Var}\,(\hat{\mathbf{y}}(\mathbf{x}))$ is minimized, and so can be implemented by minimizing

$$\varphi_M(D) = \max_{\mathbf{x} \in R} \mathbf{t}_\mathbf{x}^T \mathcal{I}^{-1} \mathbf{t}_\mathbf{x}$$

Rather than minimizing the largest predictive variance, I-optimality is defined so as to minimize the average (over R) response variance, and so can be implemented by minimizing

$$\varphi_M(D) = \int_R \mathbf{t}_\mathbf{x}^T \mathcal{I}^{-1} \mathbf{t}_\mathbf{x}\, w(\mathbf{x})\, d\mathbf{x}$$

for an appropriate weight function w (which may be omitted if all regions of equal volume in R should receive the same weight).

17.3.4.1 A factorial example

Suppose a two-level factorial design is required for fitting the model:

$$y = \mu + \alpha x_1 + \beta x_2 + \gamma x_3 + (\alpha\beta) x_1 x_2 + \varepsilon$$

where the three independent variables have been coded so that $R = \{-1, +1\}^3$.

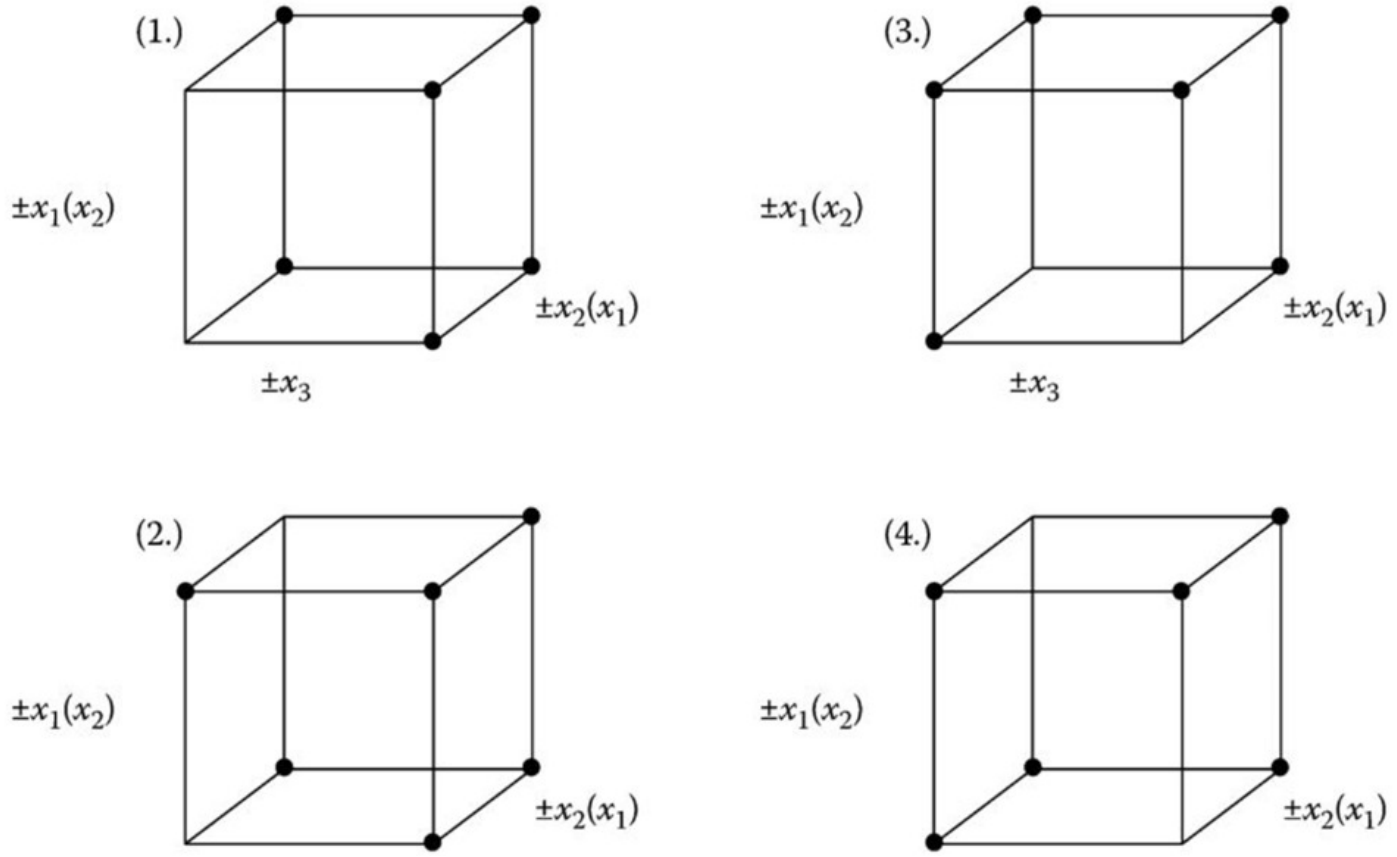
Further, suppose that each experimental run is relatively expensive or time-consuming, and only $n = 6$ unblocked experimental runs are possible.

The question arises how to select the treatments for the 6 runs (out of 8 different treatments) to get an A- or D-optimal design respectively.

In this case we may enumerate all possible designs: There are $\binom{8}{6} = 28$ designs with 6 different treatments and $5 \cdot \binom{8}{5} = 280$ designs with 5 different treatments. Designs with fewer than 5 treatments cannot be used to estimate the 5 model parameters ...

A factorial example (cont.)

For all possible designs we compute the design criterion and choose the designs which optimize the criterion.



Designs (1.) and (2.) are A-optimal, all 4 designs above are D-optimal.

17.3.4.2 A regression example

Consider the design of a small regression experiment in which a quadratic polynomial model in $d = 1$ independent variable is to be fit based on only $n = 3$ experimental runs.

We are interested in estimating the first- and second-order coefficients, β_1 and β_{11}. $R = [-1, +1]$ is the interval of allowable values for the independent variable x, so the number of possible designs is infinite.

However, here we may compute the D-optimal design analytically:

The three values of x to be used are c (center value), $c - r\Delta$ and $c + (1 - r)\Delta$, where Δ is the difference between the largest and smallest x.

The design matrix then is

$$\mathbf{X} = \begin{pmatrix} 1 & c - r\Delta & (c - r\Delta)^2 \\ 1 & c & c^2 \\ 1 & c + (1 - r)\Delta & (c + (1 - r)\Delta)^2 \end{pmatrix}$$

A regression example (cont.)

With $\mathbf{X}_1 = \mathbf{1}$ we get

$$\mathbf{X}_{2|1} = \frac{\Delta}{3} \begin{pmatrix} -r-1 & \Delta(r^2+2r-1) - 2c(r+1) \\ 2r-1 & -\Delta(2r^2-2r+1) + 2c(2r-1) \\ 2-r & \Delta(r^2-4r+2) - 2c(r-2) \end{pmatrix}$$

Finally the determinant of the (2×2) information matrix is:

$$|\mathbf{X}_{2|1}^T \mathbf{X}_{2|1}| = \frac{\Delta^2}{3} (r(1-r))^2$$

and $x = \{-1, 0, 1\}$ is the D-optimal 3-point design for this problem.

17.4 Algorithms

Most often it is not possible to analytically derive or verify optimal experimental designs, optimal designs almost always have to be constructed numerically - which is difficult because

- optimization is over many variables - often all of the elements of each $\mathbf{x}$ in D.
- The objective function is often optimized for several designs, i.e. there are often many optimal designs of a given size for a given experimental region, model and criterion.
- Many near-optimal designs that are dissimilar to optimal plans may also exist.

Designs are often constructed using algorithms that begin with an arbitrary or random starting design, and make a series of iterative changes with the purpose of improving the quality of the design at each step.

Algorithms (cont.)

The most widely used approach to the numerical construction of optimum experimental designs is undoubtedly through the use of various *point-exchange algorithms*.

For finite R, a simple point-exchange algorithm, is:

- Specify the N-point experimental region R, the model M, the experiment size n, and the criterion function φ_M to be used.
- Specify a „starting design“ D_0 and compute $\varphi_M(D_0)$.
- Construct the N designs, each of $n+1$ points, by adding one additional point from R to D_0. Computing φ_M for all these designs and identify the best design D_0^+.
- Construct the $n+1$ designs, each of n points, that consist of all but one of the points in D_0^+, compute φ_M for each design and identify the best design D_1.
- Continue the add-and-delete process as long as the design criterion is improved.

More elaborate point-exchange algorithms have also been developed.

Software and Packages

There is an enormous variety of DoE software, most notably

commerical solutions:

- JMP `https://www.jmp.com/de_at/home.html`
- Design-Expert `https://www.statease.com/software/design-expert/`
- Minitab `https://www.additive-net.de/de/software/produkte/minitab/minitab/doe`

and R-packages, see `https://cran.r-project.org/web/views/ExperimentalDesign.html`