Contents

12

Response surfaces

Very often we want to model $\mathbb{E}(Y \mid \boldsymbol{x})$ for continuosly distributed $\boldsymbol{x}$, not just binary variables as we could handle with 2^k factorials. Those can be used to study continuous variables by choosing two levels for them. However, once we know which variables are the most important and have perhaps a rough idea of the range in which to explore them we may then want to map out $\mathbb{E}(Y \mid \boldsymbol{x})$ more precisely for the subset of most important variables. We would like to estimate an arbitrary ***response surface*** $\mathbb{E}(Y \mid \boldsymbol{x})$ in those key variables. The literature on response surface models is mostly about estimating first order (linear) and second order (quadratic) polynomial models in $\boldsymbol{x}$, so most of the practical methods do not have the full generality that the term 'response surface' suggests.

If we are operating near an optimum value of $\mathbb{E}(Y \mid \boldsymbol{x})$ then a quadratic model might capture the most important aspects of $\mathbb{E}(Y)$. If that optimum is on the boundary of a constraint region then a local linear model might be very suitable. Local linear models are also very suitable in screening out the most important variables.

The material for this lecture is based largely on these texts Box and Draper (1987), Myers et al. (2016) and Wu and Hamada (2011) and these survey articles Myers et al. (1989) and Khuri and Mukhopadhyay (2010). Additional material on optimal design comes from Atkinson et al. (2007).

12.1 Center points

Designs at just two levels for each component of $\boldsymbol{x}$ let us fit the first order model

$$\mathbb{E}(Y \mid \boldsymbol{x}) \doteq \beta_0 + \sum_{j=1}^{p} \beta_j x_j. \tag{12.1}$$

We can also fit a "ruled surface" model like

$$\mathbb{E}(Y \mid \boldsymbol{x}) \doteq \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \sum_{1 \leqslant j < k \leqslant p} \beta_{jk} x_j x_k \tag{12.2}$$

though some of the cross terms might be subject to aliasing with higher order interactions, with each other (resolution IV) or with main effects (resolution III). Equation (12.2) leaves out term for $\beta_{jj} x_j^2$.

If x_{ij} takes only two levels, then the most we can do with it is fit a two parameter model such as a linear one. To fit a third parameter, such as curvature, we need a third level. For that we can take some center points. When we have been sampling $x_{ij} \in \{-1, 1\}$ we might then take some additional runs with $x_{ij} = 0$.

The simplest strategy is to add one or more center points with $\boldsymbol{x}_i = 0$. Put in a center point (maybe several). E.g. for $p = 2$ we could use

$$\begin{array}{c} \begin{matrix} x_1 & x_2 \end{matrix} \\ \begin{bmatrix} -1 & -1 \\ -1 & 1 \\ 1 & -1 \\ 1 & 1 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix} \end{array}.$$

From the repeated center points, we can estimate σ^2 or at least $\mathrm{var}(Y \mid \boldsymbol{x} = 0)$. We can also use that data to estimate this model

$$\beta_0 + \sum_{j=1}^{p} \beta_j x_j + \sum_{1 \leqslant j < k \leqslant p} \beta_{jk} x_j x_k + \gamma \sum_{j=1}^{p} x_j^2.$$

Notice that there is only one coefficient γ for all of the squared terms. This is $\gamma = \sum_{j=1}^{p} \beta_{jj}$ in our usual notation. The reason is that in a design with two levels plus a center point we have $x_{ij} = \pm x_{ij'}$ for all $i = 1, \ldots, n$ and all $1 \leqslant j < j' \leqslant d$. This then implies that $x_{ij}^2 = x_{ij'}^2$ and so all of the quadratic terms are perfectly confounded.

We might run this centerpoint design in a case where we expect little curvature but just want to be able to make a check on it. If there is convincing evidence that γ differs from zero by an important amount, then we know there

is curvature and a first order model is problematic. If $\hat{\gamma}$ is not significantly different from zero then this is consistent with there being no curvature but does not prove it. The true β_{jj} might sum to nearly zero. In other words we have a ***one way diagnostic***. When it provides evidence of curvature we can be confident that it is there, but when it does not provide such evidence we cannot be confident that there isn't any curvature. Readers might be familiar with one way diagnostics in Markov chain Monte Carlo methods: they can reliably detect slow mixing but ordinarily cannot establish good mixing.

Another use for centerpoints is that, as remarked above, we might want an estimate of $\mathrm{var}(Y \mid \boldsymbol{x} = 0)$. The variance of Y at the centerpoint might be a reasonable variance to use for planning even if the true variance depends on $\boldsymbol{x}$.

Yet another use for them is to serve as 'control runs' that this page from NIST `https://www.itl.nist.gov/div898/handbook/pri/section3/pri337.htm` describes as "To provide a measure of process stability and inherent variability". Their advice is ***not*** to include the centerpoints in the randomized experimental order. Instead they recommend spacing those points out evenly through the run. For instance with 16 points in a 2 level design, they might add a centerpoint at the beginning, middle and end of experimentation, bringing the total to $n = 19$ runs. The other 16 points would be placed in a random order in the remaining 16 experimental positions.

12.2 Three level factorials

There is a theory of 3^k factorial designs and 3^{k-p} fractional factorial designs that parallels the case for two level designs. Not surprisingly, the expense grows more quickly with k than we get for 2 level designs.

In a three level design we take $x_{ij} \in \{-1, 0, 1\}$ after rescaling. For variables that take widely different values these levels might be what we get after a logarithmic transformation. For instance we might use $x_{ij} \in \{-1, 0, 1\}$ to encode a quantity that originally took values 100, 200 and 400.

When $k = 1$ our experiment at 3 level has two degrees of freedom for treatments. These are usually expressed through two contrasts: a linear contrast $\bar{Y}_1 - \bar{Y}_{-1}$, and a quadratic contrast $2\bar{Y}_0 - \bar{Y}_1 - \bar{Y}_{-1}$. These are orthogonal contrasts.

With k effects A, B, C, $\cdots$ we find 2 degrees of freedom for A, 4 degrees of freedom for a two factor interaction like AB, 8 degrees of freedom for a three factor interaction like ABC, and so on. So things get expensive. The ANOVA table for a three level design can be partitioned as in Table 12.1.

The ANOVA table for a three level design has terms for the product of k quadratic effects such as $A_Q \times B_Q \times C_Q$. We might well use some of those interactions in an error term, just as we did for two factor designs to mitigate the high cost of three level experiments. We could also plot estimated effects in a QQ plot to identify important variables.

Source	df
A	2
A_L	1
A_Q	1
B	2
B_L	1
B_Q	1
AB	4
$A_L \times B_L$	1
$A_L \times B_Q$	1
$A_Q \times B_L$	1
$A_Q \times B_Q$	1

Table 12.1: An ANOVA table for a three level factorial.

x	y	x+y mod 3	x+2y mod 3
0	0	0	0
0	1	1	2
0	2	2	1
1	0	1	1
1	1	2	0
1	2	0	2
2	0	2	2
2	1	0	1
2	2	1	0

Table 12.2: This is a 3^{4-2} fractional factorial. It is also an orthogonal array in that every pair of columns has all 9 possible rows the same number of time (i.e., once).

Given data from a 3^k factorial experiment we can fit the two level model

$$\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \sum_{j=1}^{k} \beta_{jj} x_{ij}^2 + \sum_{1 \leqslant j < j' \leqslant k} \beta_{jj} x_{ij} x_{ij'}$$

by least squares. We might prefer to center the pure quadratics

$$\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \sum_{j=1}^{k} \beta_{jj} (x_{ij}^2 - 2/3) + \sum_{1 \leqslant j < j' \leqslant k} \beta_{jj} x_{ij} x_{ij'}$$

to make them orthogonal to the intercept term.

Table 12.2 shows a 3^{4-2} fractional factorial experiment. It is known as an 'orthogonal array' because each pair of columns has all 9 possible combinations of variables the same number of times. We will see much more about orthogonal

arrays later. The top of Table 12.2 shows a construction in modular arithmetic. We will see more of that construction too.

When using an orthogonal array, be sure to randomize the levels. That is, there are 6 possible ways to map the levels 0, 1 and 2 of the array onto the experimental levels -1, 0 and 1 and one of those should be chosen at random. An independent randomization should be made for each column. It would be a very poor practice to just subtract 1 from each entry in the array. The run order should also be randomized.

These 3^{k-p} designs can also be run in blocks whose size is a power of 3.

There is an extensive selection of three level designs here: `http://neilsloane.com/oadir/#3_2`. For a comprehensive account of orthogonal arrays, see Hedayat et al. (1999).

12.3 Central composite designs

In the ***central composite design*** of Box and Wilson (1951) we begin with a two level design with values $\{-1,1\}^k$, ordinarily a 2^{k-p} fractional factorial, then add some number n_0 of center points $(0,0,0,\dots,0)$ and then $2k$ "star points" varying one factor at a time. These take the form $(\pm\alpha,0,0,\dots,0)$, $(0,\pm\alpha,0,\dots,0)$, $(0,0,\pm\alpha,\dots,0)$, and so on, up to $(0,0,0,\dots,\pm\alpha)$, for some $\alpha>0$. The points are then used in random order, sometimes within blocks.

In setting up a central composite design we have to choose our three components: the two level design to use, the number of center points, and the value α for the star points.

The analysis is usually a quadratic linear regression. Given a point $\boldsymbol{x}\in\mathbb{R}^d$ we form a vector of features

$$f(\boldsymbol{x}) = (1, x_1,\dots,x_d, x_1x_2,\dots,x_{d-1}x_d, x_1^2,\dots,x_d^2)^{\mathsf{T}} \in \mathbb{R}^p$$

for $p = 1+d+d(d-1)/2+d = 1+d(d+3)/2$ and fit by least squares. That is

$$\hat\beta = (F^{\mathsf{T}}F)^{-1}F^{\mathsf{T}}Y$$

where $F\in\mathbb{R}^p$ has i'th row $\boldsymbol{f}_i = f(\boldsymbol{x}_i)$ and $Y\in\mathbb{R}^n$ with i'th element Y_i.

In choosing the two-level experiment, we will want all quadratic and cross terms to be estimable. That is, $F^{\mathsf{T}}F$ should be invertible, which we can easily check before commencing to experiment. Naively that would be solved by using a resolution V experiment that keeps the cross terms uncounded with each other. By the time the center points and star points are included, the true condition becomes much more subtle. See Wu and Hamada (2011, Chapter 9.7) for a very careful exposition. For instance, one can use what they call resolution III* designs. Those have resolution III with no words of length exactly four. That is one cannot have $ABCD=\pm I$. They also point out that one can use Plackett-Burman (i.e., Hadamard) points for the two level design. In dimension $k=2$, even a 2^{2-1} experiment plus center points and axial points can make the model estimable.

$2k$ points OAAT $\pm\alpha \times e_j$
n_0 center points
https://www.jstor.org/stable/2983966
https://www.google.com/search?q=central+composite+design

It is also common to code the pure quadratic features as $x_{ij}^2 - (1/n)\sum_{i'=1}^n x_{i'j}^2$ to make them orthogonal to the intercept. This also makes them orthogonal to the linear terms because, breaking the design into its three parts we find that

$$\begin{aligned}\sum_i (x_{ij}^2 - \overline{x_j^2}) x_{ij'} &= \sum_i x_{ij}^2 x_{ij'} \qquad \text{from } \sum_i x_{ij'} = 0\\ &= \sum_{i \in \text{Factorial}} x_{ij'} + (\alpha * 0) + (-\alpha * 0)\\ &= 0.\end{aligned}$$

Exercise: show that $x_{ij}^2 - \overline{x_j^2}$ is orthogonal to $x_{ij}x_{ij'}$ for $j' \neq j$ and to $x_{ij'}x_{ij''}$ when no two of j, j' and j'' are equal.

One very tricky issue is choosing the value of α. Taking $\alpha = 1$ is convenient because it keeps all factors at three levels. We will have a subset of a 3^k factorial experiment. Another choice is to take $\alpha = \sqrt{k}$ because this keeps $\|\boldsymbol{x}_i\|^2 = k$ on the star points just like it is for the factorial points. This is called a "spherical design" because then both the star and factorial points are embedded within a sphere of radius $\sqrt{k}$. When we choose a spherical design we need some zero points or else $\sum_{j=1}^k x_{ij}^2 = k$ for all i and we will have a singular matrix F. Exercise: is this exactly the same problem that we saw with a centerpoint design and the parameter γ or is it different?

If we choose $\alpha = \sqrt{k}$ then we might find that values $x_{ij} \in \pm\sqrt{k}$ are too far from the region of interest even though they are exactly the same distance from the center as the factorial points are. The issue stems from factor sparsity. If x_1 is a very important variable, much more so than the others, then taking $x_{i1} = \pm\sqrt{k}$ represents a much more consequential change than just taking everything in $\{-1, 1\}$. Something is too far from the region of interest if the quadratic model that serves over $[-1, 1]^k$ does not extrapolate well there. Perhaps changing a geometric parameter for a transistor by that much turns it into a diode. It is even possible that operating at $x_{ij} = \pm\sqrt{k}$ is unsafe if x_j represents temperature or pressure. This sort of non-statistical issue can be much more important than designing to reduce $\text{var}(\hat{\beta})$ and it requires the input of a domain expert.

One possible way to choose α is to obtain orthogonality, that is a diagonal matrix $F^{\mathsf{T}}F \in \mathbb{R}^{p\times p}$. If the pure quadratic terms are centered then it remains possible that they are not orthogonal to each other. There is one value of α that makes them orthogonal. After some algebra, this is

$$\alpha = \Big(\frac{QF}{4}\Big)^{1/4}$$

for $Q = [(F+T)^{1/2} - F^{1/2}]^2$ where F is number of factorial observations and $T = 2k + n_0$. This then makes $\text{corr}(\hat{\beta}_{jj}, \hat{\beta}_{j'j'}) = 0$.

Another way to choose α is to obtain ***rotatability***:

$$\mathrm{var}(\hat{\mathbb{E}}(Y\,|\,\boldsymbol{x})) = f(\boldsymbol{x})^{\mathsf{T}}(F^{\mathsf{T}}F)^{-1}f(\boldsymbol{x})\sigma^2 = g(\|\boldsymbol{x}\|)$$

for some function $g(\cdot)$. Now the statistical information when predicting $\mathbb{E}(Y\,|\,\boldsymbol{x})$ depends only on $\|\boldsymbol{x}\|$ and not on the angle between $\boldsymbol{x}$ and any of the coordinate axes. There is not a strong motivation for choosing rotatability. Rather it is a tie-breaker condition when choices are otherwise equal. Also, when factor sparsity is present then sparse vectors $\boldsymbol{x}$ such as $(1,0,0,\ldots,0)$ and $(0,1,0,\ldots,0)$ should be more important than others of the form $(\cos(\theta),\sin(\theta),0,\ldots,0)$ for arbitrary $0<\theta<\pi/2$.

Box and Draper (1987) include some blocking schemes for central composite designs. In a blocked analysis we would use indictor variables taking the value 1 in a given block and zero outside of that block. It is then important to have those indicator variables be orthogonal to the linear, quadratic and mixed terms in the second order regression model. NIST shows some examples of central composite designs in blocks at `https://www.itl.nist.gov/div898/handbook/pri/section3/pri3364.htm`. If that link doesn't work well, look for section 5.3.3.6.4 entitled "Blocking a response surface design" in their online engineering statistics handbook.

12.4 Box-Behnken designs

The second major class of response surface designs are the Box-Behnken designs from Box and Behnken (1960). Those designs have three levels 0 and ±1. A small example for factors A, B and C looks like this:

A	B	C
±1	±1	0
0	±1	±1
±1	0	±1
0	0	0

The first row denotes a 2^2 factorial in A and B with factor C held at zero. There follow two similar rows with A and then B held at zero. Finally there is a row representing n_0 runs at $\boldsymbol{x}=0$. If, for instance $n_0=3$ then this experiment would have 15 runs. According to Box and Behnken (1960), "The exact number of center points is not critical". Geometrically this design has 12 points on the surface of the unit cube $[-1,1]^3$, one at the midpoint of each of the 12 edges connecting the 8 vertices within six faces. There are also center points.

We can recognize the strategy in this design. There is a balanced incomplete block structure designating some number $r<k$ of the factors that take values ±1 while the remaining $k-r$ factors are held at zero. Then some number of center points are added.

Table 12.3 shows another Box-Behnken design, this time for 4 factors. It is arranged in three blocks each of which has its own center point.

A	B	C	D
±1	±1	0	0
0	0	±1	±1
0	0	0	0
±1	0	0	±1
0	±1	±1	0
0	0	0	0
±1	0	±1	0
0	±1	0	±1
0	0	0	0

Table 12.3: A schematic for a Box-Behnken design in four factors with three blocks and one center point per block.

Exercise: are the block indicator variables for the Box-Behnken design in 12.3 orthogonal to the regression model? If we change it to $n_0 = 4$ are the orthogonal? Since there are three block variables you can drop the intercept column.

Like 3^{k-p} designs Box-Behnken designs can easily handle categorical variables at 3 levels. The tabled designs in the literature and textbooks involve only modest dimensions k. For large k, Box-Behnken designs use many more runs than parameters. While that may be useful in some settings, in others there is a premium on minimizing n by taking it just slightly larger than the number of parameters.

12.5 Uses of response surface methodology

One of the main uses is to fit a quadratic model, and estimate the direction of steepest ascent. Then, assuming that larger $\mathbb{E}(Y \mid \boldsymbol{x})$ is better move the region of interest in the direction of apparent improvement and run another experiment. The approach features a 'human in the loop' deciding which variables to explore and how at each iteration of the experiment. This is called ***evolutionary operation*** by Box (1957).

There is a large related field of stochastic optimization that takes possibly noisy data and uses it to decide where next to sample with the goal of finding an optimum. See Kushner and Yin (2003) and Spall (2003). Those methods often emphasize iterations taking one or two new data points at each round.

12.6 Optimal designs

This section is based primarily on Atkinson et al. (2007). We will pick $\boldsymbol{x}_i \in \mathbb{R}^k$ and then compute features $f(\boldsymbol{x}_i) \in \mathbb{R}^p$ (such as for a quadratic regression). Then our model is $Y_i = f(\boldsymbol{x}_i)^\mathsf{T}\beta + \varepsilon_i$ we estimate β by $\hat\beta = (F^\mathsf{T}F)^{-1}F^\mathsf{T}Y$

and our accuracy is described by $\mathrm{var}(\hat\beta) = (F^\mathsf{T}F)^{-1}\sigma^2 \in \mathbb{R}^{p\times p}$. When we predict Y for a given $\boldsymbol{x}$ then we may use $\widehat{Y}(\boldsymbol{x}) = f(\boldsymbol{x})^\mathsf{T}\hat\beta$ with $\mathrm{var}(\widehat{Y}(\boldsymbol{x})) = f(\boldsymbol{x})^\mathsf{T}(F^\mathsf{T}F)^{-1}f(\boldsymbol{x})\sigma^2$.

We will want to choose $\boldsymbol{x}_i$ in some way that $\mathrm{var}(\hat\beta)$ is small and there are many ways that a matrix can be considered small. Before that however it is worth noting that in this setting $\mathrm{var}(\hat\beta)$ does not depend on the true β! The fact that β does not appear in the formula $(F^\mathsf{T}F)^{-1}\sigma^2$ is an enormous simplification, that we usually take for granted. Otherwise the best design for learning β would depend on the unknown β and then the design problem would be intrinsically circular. This actually happens for logistic regression which we consider briefly below. Once again, finding that a symbol ***is not*** in a formula is quite special.

Before getting started we should rule out three approaches. First, we don't want to solve the problem by letting $n \to \infty$. That's expensive and may be wasteful. We want to be as efficient as we can with the n that we can afford. Second, we don't want to solve it by letting $\sigma^2 \to 0$. We want to be as efficient as we can with a given quality of measurement. Finally, while small $\mathrm{var}(\hat\beta)$ corresponds to large $F^\mathsf{T}F$, we don't want to solve the problem by letting $\|\boldsymbol{f}_i\| \to \infty$ where $\boldsymbol{f}_i = f(\boldsymbol{x}_i)$. The linear model is only approximate and so we need to keep $\boldsymbol{x}_i$ in or near the region of interest. Also, the extreme $\|\boldsymbol{x}_i\|$ that we would ordinarily need for extreme $\|\boldsymbol{f}_i\|$ may be expensive or unsafe.

We formulate the design problem by fixing n and requiring $\boldsymbol{x}_i \in \mathcal{X} \subset \mathbb{R}^d$. Depending on the problem we might require $\boldsymbol{x}_i \in \mathcal{X} = \{\boldsymbol{x} \in \mathbb{R}^d \mid \|\boldsymbol{x}\| \leqslant 1\}$ or $\boldsymbol{x}_i \in \mathcal{X} = \{\boldsymbol{x} \in \mathbb{R}^d \mid \max_j |x_j| \leqslant 1\}$.

There are numerous notions of optimality, including A-optimality, D-optimality, E-optimality, G-optimality and I-optimality. Using

$$\mathrm{var}(\hat\beta) = (F^\mathsf{T}F)^{-1}\sigma^2 = \frac{\sigma^2}{n}M^{-1} \quad \text{where} \quad M \equiv \frac{F^\mathsf{T}F}{n} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{f}_i\boldsymbol{f}_i^\mathsf{T}$$

we can describe the optimality criteria via the matrix M.

Perhaps the most famous and widely used notion is ***D-optimality***. A D-optimal design minimizes $\det(F^\mathsf{T}F)^{-1}\sigma^2$ (sometimes called the generalized variance of $\hat\beta$) or equivalently it maximizes $\det(M)$. This is the product of the eigenvalues of M. The 'D' stands for determinant.

A-optimality with 'A' for average refers to minimizing $\sum_j \mathrm{var}(\hat\beta_j)$ or maximizing the sum or average of the eigenvalues of M^{-1}. E-optimality with 'E' for extreme refers to minimizing $\max_j 1/\lambda_j$ where λ_j are the eigenvalues of M.

G-optimality refers to minimizing $\max_{\boldsymbol{x}\in\mathcal{X}} \mathrm{var}(f(\boldsymbol{x})^\mathsf{T}\hat\beta)$ where $\mathrm{var}(f(\boldsymbol{x})^\mathsf{T}\hat\beta) \propto f(\boldsymbol{x})^\mathsf{T}M^{-1}f(\boldsymbol{x})$. This notion goes back to Smith (1918) for polynomial regression, which is the first optimal design paper.

I-optimal design also called V-optimal refers minimizing $\int_{\mathcal{X}} \mathrm{var}(f(\boldsymbol{x})^\mathsf{T}\hat\beta)g(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ where $g(\cdot) \geqslant 0$ measures interest level. It could be a distribution but does not have to be. One could also minimize $\int_{\mathcal{X}'} \mathrm{var}(f(\boldsymbol{x})^\mathsf{T}\hat\beta)g(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ where the set $\mathcal{X}'$ includes extrapolations to points that are not in $\mathcal{X}$. D_A optimality refers to minimizing $\det(\mathrm{var}(A^\mathsf{T}\hat\beta))$ for some matrix A.

These definitions raise the question "which optimality is best?" To address that we need to consider ***design measures***. In a design measure, we generalize from $M(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = (1/n)\sum_{i=1}^n f(\boldsymbol{x}_i)f(\boldsymbol{x}_i)^\mathsf{T}$ to

$$M(\mu) = \int_{\mathcal{X}} f(\boldsymbol{x})f(\boldsymbol{x})^\mathsf{T}\mu(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \mathbb{E}(f(\boldsymbol{x})f(\boldsymbol{x})^\mathsf{T}), \qquad \boldsymbol{x} \sim \mu$$

for a distribution μ. Then instead of finding the best list of n points in $\mathcal{X}$ we relax to problem to just seeking the optimal distribution μ. While $M(\mu)$ is written above as an integral as if μ were continuous, μ can also be discrete as all we need is the expectation. In fact, most of the solutions we see will involve discrete distributions μ. Given an optimal or near optimal design measure μ we can then pick $\boldsymbol{x}_i$ to approximate μ.

Atkinson et al. (2007) consider a problem of quadratic regression through origin. The model is $\mathbb{E}(Y \mid x) = \beta_1 x + \beta_2 x^2$ with $0 \leqslant x \leqslant 1$. That is, $\mathcal{X} = [0, 1]$. This is not a model we would often want to fit, but it is an excellent small illustration of how optimal design works. They find that to maximize the D-optimality condition $\det(\mathbb{E}(f(\boldsymbol{x})f(\boldsymbol{x})^\mathsf{T}))$ under $\boldsymbol{x} \sim \mu$ one should choose $\mu(\sqrt{2}-1) = 1/\sqrt{2}$ and $\mu(1) = 1 - 1/\sqrt{2}$. This optimal design measure only uses two different points x. Because the fraction of data at $\sqrt{2}-1$ is not a rational number we cannot actually find a finite set of points with empirical distribution μ. Instead, we would approximate it taking roughly $n/\sqrt{2}$ observations at $\boldsymbol{x} = 1/\sqrt{2}$ and the rest at $\boldsymbol{x} = 1$.

For a design measure D-optimality maximizes $-\log\det(M(\mu))$, If $\mu = \sum_{i=1}^n \omega_i 1_{\boldsymbol{x}_i}$ then for fixed $\boldsymbol{x}_i$, we get a convex function $\sum_{i=1}^n \omega_i \boldsymbol{f}_i\boldsymbol{f}_i^\mathsf{T}$ of $(\omega_1, \ldots, \omega_n)$ because $\log(\det(\cdot))$ is a convex operation on matrices. Some algorithms use a large n and then get most ω_i equal to zero or close to it. It is typical for the optimal design measure to have p points $\boldsymbol{x}_i$ with positive probabililty where p is the number of parameters in the regression model. This leaves us with no way to estimate σ^2 or test whether a model with more than p parameters would be suitable. Perhaps that is not surprising. Criteria like D-optimality do not include either variance estimation or testing goodness of fit.

The ***general equivalence theorem*** is an important result of Kiefer and Wolfowitz (1960). It is that D-optimality and G-optimality are equivalent for design measures that continuously weight the same set of $\boldsymbol{x}_i$.

A special property of D-optimality is ***equivariance***. If we change units from meters to centimeters, the D-optimal designs scale accordingly and we would run exactly the same set of physical experiments either way. Generally replacing $\boldsymbol{f}_i$ by $A \times \boldsymbol{f}_i + b$ the new optimal points have $\tilde{f}_i = Af_i^* + b$ and the same ω_i.

Optimal designs are not always good enough to use because they focus only on variance and might require awkward input combinations instead of for instance using just three levels of a continuous factor which could be convenient for implementation. In other words, we might not have been able to put all of the important criteria into the objective function or encoded all of the desired constraints. We can however compare a design such as a central composite or Box-Behnken to the optimal design and see how close to 100% efficiency we get.

Optimal design can also help us when we have a more complicated constraint region $\mathcal{X}$ to work with than the usual cubes and balls. It is not always possible to do the global optimization that we would want in optimal design.

Box and Draper (1987) are skeptical about the optimal design approaches sometimes called "alphabetic optimality". The work some examples that optimize mean squared error taking account of bias and variance. The bias comes from the possibility that the model is a polynomial of higher order than the one fit by the response surface model. They find that the optimal designs for mean squared error are similar to those we would get optimizing for bias squared and they can be quite different from what we would find optimizing just for variance like the alphabetic optimality designs do. They 'match moments' over the region, making $(1/n)\sum_{i=1}^n \boldsymbol{f}_i\boldsymbol{f}_i^{\mathsf{T}} = \mathbb{E}(\boldsymbol{f}\boldsymbol{f}^{\mathsf{T}})$.

Now we turn briefly to logistic regression. It is not a pre-requisite for this course but many readers will have encountered it. Logistic regression is for binary responses $Y \in \{0, 1\}$. The model relating Y to $x \in \mathbb{R}$ is

$$\Pr(Y = 1 \,|\, X = x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \equiv p(x; \beta)$$

and in general it uses $\Pr(Y = 1 \,|\, \boldsymbol{x}) = \exp(\beta^{\mathsf{T}} f(\boldsymbol{x}))/(1 + \exp(f(\boldsymbol{x})^{\mathsf{T}}\beta))$. The likelihood function is

$$L(\beta) = \prod_{i=1}^n p(x_i; \beta)^{Y_i}(1 - p(x_i; \beta))^{1-Y_i}$$

and $\hat{\beta}$ is optained by maximizing $\log(L(\beta))$ numerically.

The design problem is about where to take x_i. If we were to set $n/2$ of the $x_i = \infty$ (or as large as possible) and $n/2$ of the $x_i = -\infty$ (or as small as possible) we would not ordinarily get the best design. We might well get all $Y_i = 1$ at one extreme and all $Y_i = 0$ at the other, with no idea of the shape of the $\Pr(Y = 1 \,|\, \boldsymbol{x})$ curve (and a degenerate log likelihood as well). It turns out that the optimal design for estimating β takes $n/2$ observations at a point x with $p(x; \beta) \approx 0.15$ and $n/2$ observations and x with $p(x; \beta) \approx 0.85$. This design has the same number of distinct design points as parameters. Those design points depend on the true β. A starting point in this literature is Chaloner and Larntz (1989).

12.7 Mixture designs

Suppose that $x_{ij} \geqslant 0$ is proportion of input j in used in observation i, with $\sum_{j=1}^J x_{ij} = 1$. In some settings Y_i depends mostly on the proportions and not the absolute levels of the inputs. For instance, when mixing paints the ratios will matter more than the absolute amounts to the color of the final produce, assuming that the mixing has been done well. In many recipes, proportions matter much more than absolute amounts.

For $k = 3$ we then have an experimental region defined by an equilateral triangle with corners $(1, 0, 0)$, $(0, 1, 0)$ and $(0, 0, 1)$. In general, our input space is a simplex $\{\boldsymbol{x} \in [0, 1]^k \mid \sum_{j=1}^k x_j = 1\}$ with intrinsic dimension $k - 1$ embedded in the k dimensional unit cube. This different shape motivates different experimental designs. See the book by Cornell (0002).

The changed space also has consequences for polynomial models. The first order model is

$$\mathbb{E}(Y \,|\, \boldsymbol{x}) = a_0 + \sum_{j=1}^k a_j x_j = \sum_{j=1}^k (a_0 + a_j) x_j \equiv \sum_{j=1}^k b_j x_j,$$

where $b_j = a_0 + a_j$. We don't need an intercept term. The Second order model is

$$\mathbb{E}(Y \,|\, \boldsymbol{x}) = \sum_{j=1}^k b_j x_j + \sum_{j=1}^{k-1} \sum_{j'=j+1}^k b_{jj'} x_j x_{j'}.$$

We don't need pure quadratic terms because $x_1^2 = x_1(1 - x_2 - x_3 - \ldots - x_k)$, which can be written as the linear term and some cross terms, and of course, the same holds for all x_j.

Bibliography

Atkinson, A., Donev, A., and Tobias, R. (2007). *Optimum experimental designs, with SAS*. Oxford University Press.

Box, G. E. (1957). Evolutionary operation: A method for increasing industrial productivity. *Journal of the Royal Statistical Society: Series C*, 6(2):81–101.

Box, G. E. and Behnken, D. W. (1960). Some new three level designs for the study of quantitative variables. *Technometrics*, 2(4):455–475.

Box, G. E. and Draper, N. R. (1987). *Empirical model-building and response surfaces.* John Wiley & Sons, New York.

Box, G. E. and Wilson, K. B. (1951). On the experimental attainment of optimum conditions. *Journal of the Royal Statistical Society: Series B*, 13(1):1–38.

Chaloner, K. and Larntz, K. (1989). Optimal bayesian design applied to logistic regression experiments. *Journal of Statistical Planning and Inference*, 21(2):191–208.

Cornell, J. A. (20002). *Experiments with mixtures: designs, models, and the analysis of mixture data*. John Wiley & Sons, New York, third edition.

Hedayat, A. S., Sloane, N. J. A., and Stufken, J. (1999). *Orthogonal arrays: theory and applications*. Springer Science & Business Media, New York.

Khuri, A. I. and Mukhopadhyay, S. (2010). Response surface methodology. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(2):128–149.

Kiefer, J. and Wolfowitz, J. (1960). The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366.

Kushner, H. and Yin, G. G. (2003). *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer-Verlag, New York.

Myers, R. H., Khuri, A. I., and Carter, W. H. (1989). Response surface methodology: 1966–l988. *Technometrics*, 31(2):137–157.

Myers, R. H., Montgomery, D. C., and Anderson-Cook, C. M. (2016). *Response surface methodology: process and product optimization using designed experiments*. John Wiley & Sons, New York, fourth edition.

Smith, K. (1918). On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations. *Biometrika*, 12(1/2):1–85.

Spall, J. C. (2003). *Introduction to stochastic search and optimization: estimation, simulation, and control*. John Wiley & Sons, Hoboken, NJ.

Wu, C. J. and Hamada, M. S. (2011). *Experiments: planning, analysis, and optimization*, volume 552. John Wiley & Sons.