Aalto University
School of Chemical
Technology

Design of experiments

Mikko Mäkelä

Aalto University, School of Chemical Engineering
Department of Bioproducts and Biosystems
Espoo, Finland

Aalto University
School of Chemical
Technology

”The best time to plan an experiment is after you've done it.”
- Fisher

Introduce yourself

- Who are you?
- What do you do?
- Why are you here?
- Tell something funny about your name?

What to expect?

- Background and philosophy
- Theory
- Nomenclature
- Practical demonstrations and exercises

What not?

- Matrix algebra
- Statistical basics
- Detailed listing of possible designs

Intended learning outcomes

After the course you will be able to:

- Identify the basic principles of experimental design
- Use different programs for experimental design
- Recognise and use different design types
- Determine a suitable regression model based on design data
- Identify and apply different tools for model diagnostics

Course contents

Five sessions

- Introduction and factorial design
- Factorial design and diagnostics
- Central composite designs and optimization
- Mixture design and miscellaneous
- Practical groupwork

Requirements

Completed assignments and exam (pass/fail):

- Participation in all the sessions
- Given assignments and group work
- Course reader
- Individual exam (return by email)

Session 1

Introduction

- Why experimental design

Factorial design

- Design matrix
- Model equation = coefficients
- Residual
- Response contour

Some history

Originally by Fisher within agriculture and biology

- Fisher (1925) Statistical Methods for Research Workers (14th ed. reprint 1973: Hafner Publishing Company; New York)
- Fisher (1935) Design of Experiments (8th ed. reprint 1971: Hafner Publishing Company; New York)
- Box & Wilson (1951) On the experimental attainment of optimum conditions, *J Royal Stat Soc, Ser B*, 13, 1-45.
- Hill & Hunter (1966) A review of response surface methodology: a literature survey, *Technometrics*, 8, 571-590.

Experimental design or RSM?

Design →  → **Analysis**

Experimental design

- Planning and analysing experiments
- Emphasis on meaningful variation

Response Surface Methodology (RSM)

- Mathematical and statistical tools for the design and analysis of response surfaces
- Emphasis on optimization

Response surfaces

www.maanmittauslaitos.fi

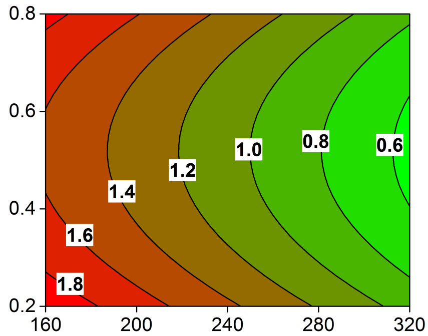
If the current location is known, a response surface provides information on

- Where to go
- How to get there
- Local maxima/minima

Is there a difference?

www.maanmittauslaitos.fi

vs.



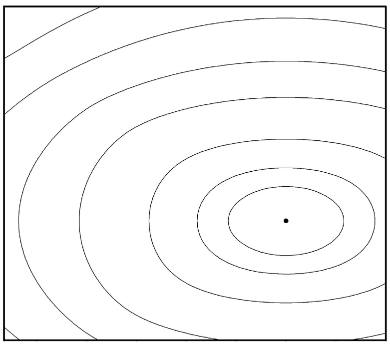
Research problem

$$A + B \xrightarrow{T,P} C$$

- A and B constant reagents
- C reaction product (response), to be maximized
- T and P reaction conditions (continuous factors), can be regulated

Response as a contour plot

P



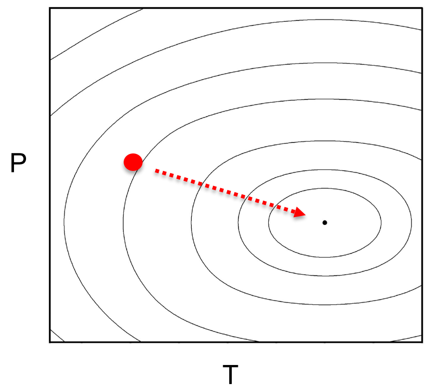
T

What kind of equation could describe C behaviour as a function of T and P?

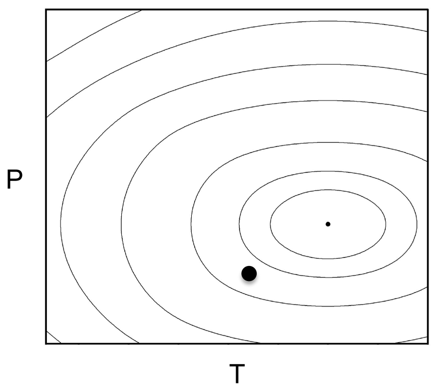
- C = f(T,P)

What else do we want to know?

- Which factors and interactions are important
- Positions of local optima (if they exist)
- Direction towards an optimum
- Surface and surface function around an optimum
- Statistical significance



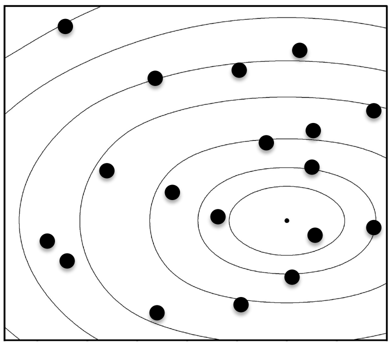
How can we do it?



The expert method

How can we do it?

P

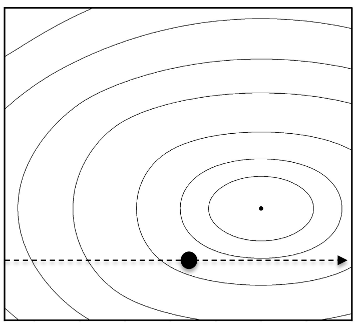


T

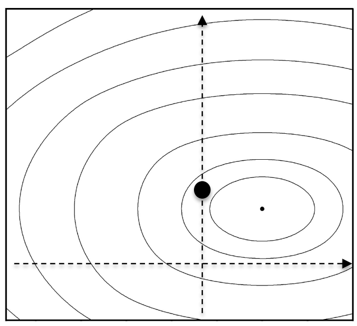
The PhD student method

How can we do it?

P fixed

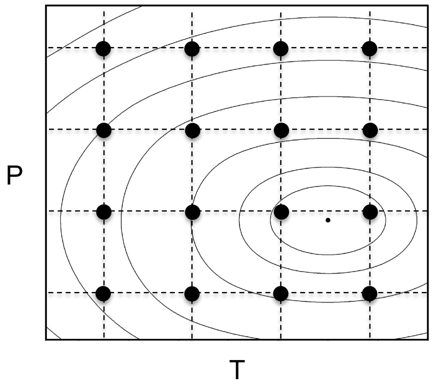


The classical method



T fixed

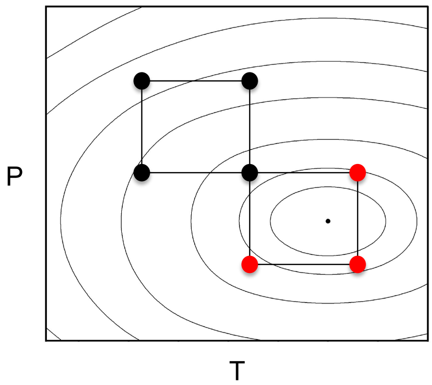
How can we do it?



The ”Soviet” method

- x^k possibilities with k factors on x levels
- 2 factors on 4 levels = 16 experiments

How can we do it?



The best method - factorial design

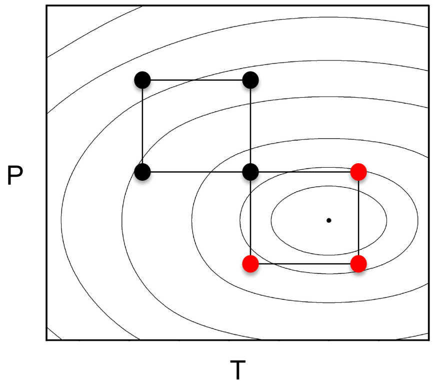
- ΔT, ΔP
- Factor interaction (diagonal)

Why experimental design?

- Reduce the number of experiments

→ Cost, time

- Extract maximal information
- Understand what happens
- Predict future behaviour



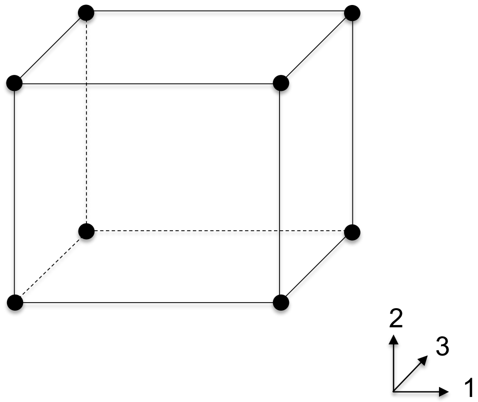
Challenges

- Multiple factors on multiple levels
- 6 factors on 3 levels, 3^6 experiments

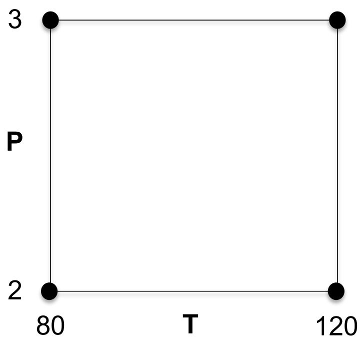
→ Only 2 levels

→ Discard factors

= SCREENING

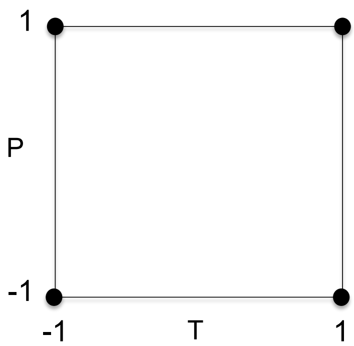


Factorial design



N:o	T	P
1	80	2
2	120	2
3	80	3
4	120	3

Factorial design

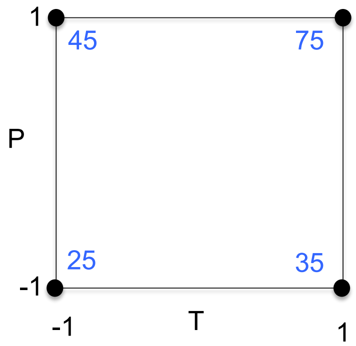


In coded levels:

N:o	T	T coded	P	P coded
1	80	-1	2	-1
2	120	1	2	-1
3	80	-1	3	1
4	120	1	3	1

The smallest possible full factorial design!

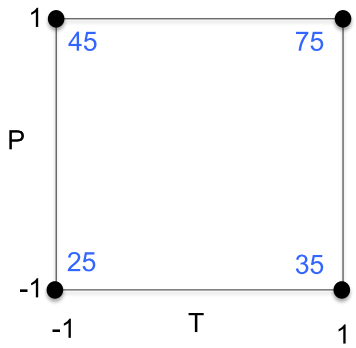
Factorial design



Design matrix:

N:o	T	P	C
1	-1	-1	25
2	1	-1	35
3	-1	1	45
4	1	1	75

Factorial design



Average T effect:

$$\mathrm{T} = \frac{75+35}{2} - \frac{45+25}{2} = 20$$

Average P effect:

$$\mathrm{P} = \frac{75+45}{2} - \frac{35+25}{2} = 30$$

Interaction (TxP) effect:

$$\mathrm{T{\times}P} = \frac{75+25}{2} - \frac{45+35}{2} = 10$$

Research problem

$$A + B \xrightarrow{T,P,K} C$$

- A and B constant reagents
- C reaction product (response), to be maximized
- T, P and K reaction conditions (continuous factors) at two different levels
- Number of experiments $2^3 = 8$ ($[\text{levels}]^{[\text{factors}]}$)

How to select proper factor levels?

Aalto University
School of Chemical
Technology

25.3.2019
27

Factorial design

First step

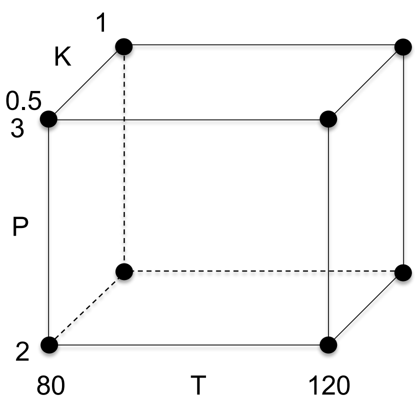
- Selection and coding of factor levels

→ Design matrix

T = [80, 120]

P = [2, 3]

K = [0.5, 1]



Aalto University
School of Chemical
Technology

25.3.2019
28

Factorial design

N:o	Order	T	P	K
1		-1	-1	-1
2		1	-1	-1
3		-1	1	-1
4		1	1	-1
5		-1	-1	1
6		1	-1	1
7		-1	1	1
8		1	1	1

Randomize!

Factorial design matrix

- Notice symmetry in diffent colums
- Inner product of two colums is zero
- E.g. **T'P** = 0

→ Orthogonality

Orthogonality

For a first-order orthogonal design, X'X is a diagonal matrix

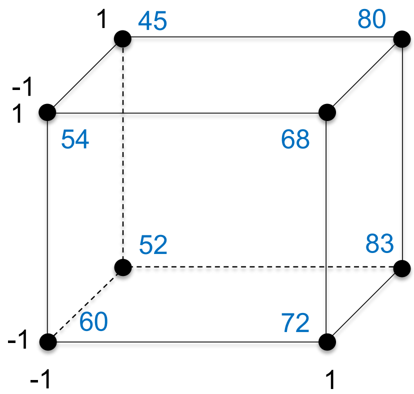
$$\mathbf{X} = \begin{bmatrix} -1 & -1 \\ 1 & -1 \\ -1 & 1 \\ 1 & 1 \end{bmatrix}, \mathbf{X}' = \begin{bmatrix} -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & 1 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & 1 \end{bmatrix} \begin{bmatrix} -1 & -1 \\ 1 & -1 \\ -1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}$$

If two columns are orthogonal, variable effects can be estimated independently

Factorial design

N:o	T	P	K	C
1	-1	-1	-1	60
2	1	-1	-1	72
3	-1	1	-1	54
4	1	1	-1	68
5	-1	-1	1	52
6	1	-1	1	83
7	-1	1	1	45
8	1	1	1	80



Empirical model

$y_c = f(T, P, K) + \varepsilon$

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e} \rightarrow \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ 1 & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \underbrace{\begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix}}_{} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

Measure **Choose** **Unknown! How to solve b?**

Least-squares regression

Minimize difference between measured and predicted values

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$
$$\rightarrow \mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\mathbf{b}$$

Cannot minimize a vector, minize the sum of squares (a scalar)

$$\frac{\partial \mathbf{e}'\mathbf{e}}{\partial \mathbf{b}} = \ldots = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{b} = 0$$
$$\rightarrow \mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

→ Least-squares estimate of **b**

Linearity

Some confusion about multiple linear regression and linearity

- Linear in coefficients or variables?

$$y_i = b_0 + b_1 x_i + \varepsilon_i$$

- Linear in both coefficients and variables

$$y_i = b_0 + b_1 x_i + b_2 x_i^2 + \varepsilon_i$$

- Linear in coefficients (with a fixed *x*, *y* a linear function of b_0, b_1 and b_2)

Factorial design

N:o	T	P	K	C
1	-1	-1	-1	60
2	1	-1	-1	72
3	-1	1	-1	54
4	1	1	-1	68
5	-1	-1	1	52
6	1	-1	1	83
7	-1	1	1	45
8	1	1	1	80

Model equation, main terms:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon_i$$

where

y_i denotes a response

x_i a factor or a variable (T, P or K)

β_i a coefficient

ε_i a residual

β_0 the mean term (average level)

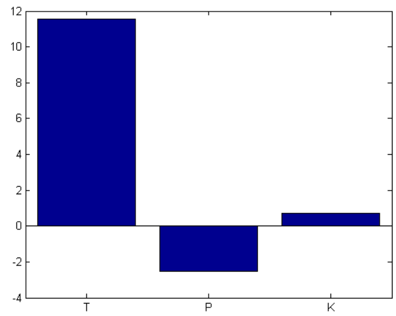
Factorial design

Model equation = coefficients

$$b = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 64.2 \\ 11.5 \\ -2.5 \\ 0.8 \end{bmatrix}$$

- b_0 average value (mean term)
- Large coefficient → important factor
- Interactions usually present

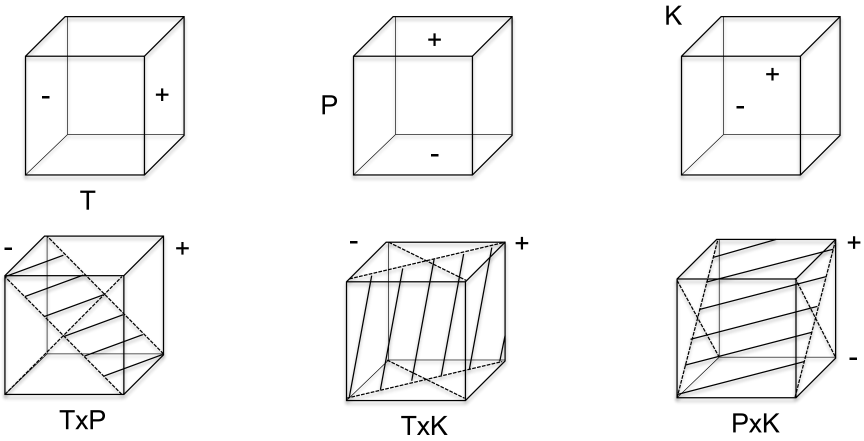
Due to coding, the coefficients are comparable!



Factorial design

N:o	T	P	K	TxP	TxK	PxK	TxKxP	C
1	-1	-1	-1		1		-1	60
2	1	-1	-1		-1		1	72
3	-1	1	-1		1		1	54
4	1	1	-1		-1		-1	68
5	-1	-1	1		-1		1	52
6	1	-1	1		1		-1	83
7	-1	1	1		-1		-1	45
8	1	1	1		1		1	80

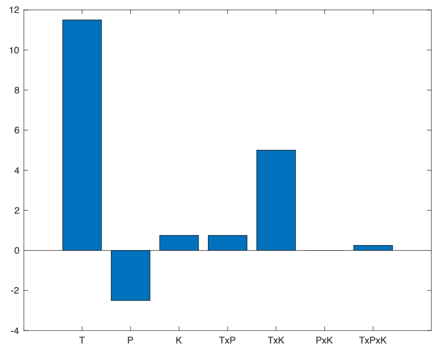
Factorial design



Factorial design

Model equation = coefficients

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \\ b_{12} \\ b_{13} \\ b_{23} \\ b_{123} \end{bmatrix} = \begin{bmatrix} 64.3 \\ 11.5 \\ -2.5 \\ 0.8 \\ 0.8 \\ 5.0 \\ 0 \\ 0.25 \end{bmatrix}$$



- Large interaction b_{13} (TxK)
- Important interaction, main effects cannot be removed

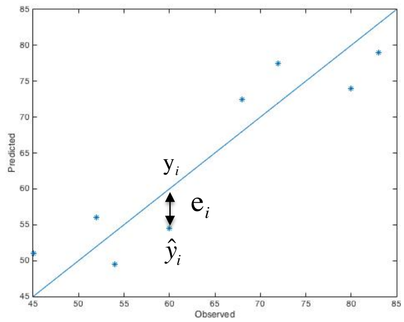
→ Which coefficients to include?

Factorial design

An estimate of standard error needed

- Replicates
- Model residual

$$\mathbf{e} = \mathbf{y} - \mathbf{Xb} = \mathbf{y} - \hat{\mathbf{y}}$$

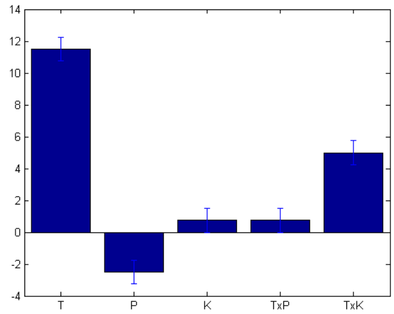


Factorial design

Error estimation allows significant testing

Remove insignificant coefficients

- Leave main effects
- Important interaction, main effect cannot be removed



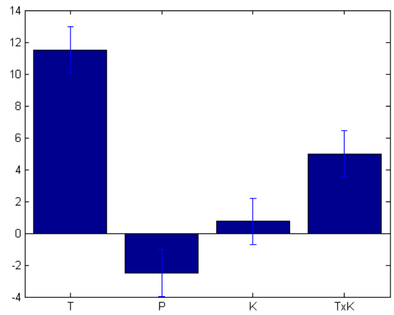
Factorial design

Error estimation allows significant testing

Remove insignificant coefficients

- Leave main effects
- Important interaction, main effect cannot be removed

Recalculate upon removal!



Factorial design

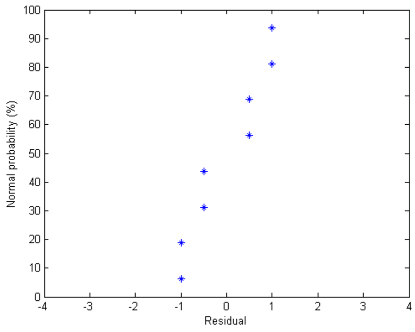
Model residuals

- Finding outliers
- If normally distributed

→ Random error

Several ways to present residuals

- Can a suggest response transformation



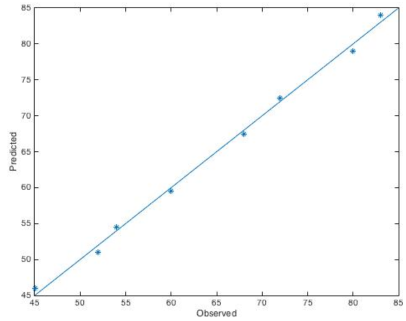
Factorial design

R^2 statistic

- Explained variation in measured response

$R^2 = 0.996$

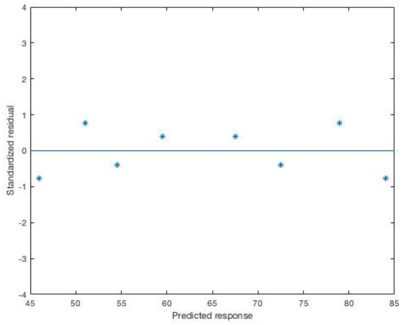
- 99.6% explained



Factorial design

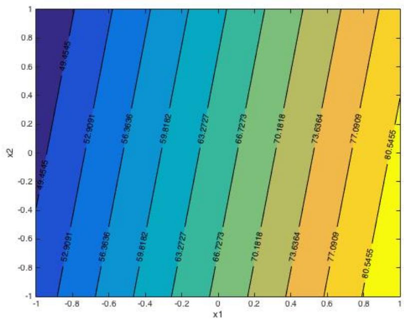
More things to look at

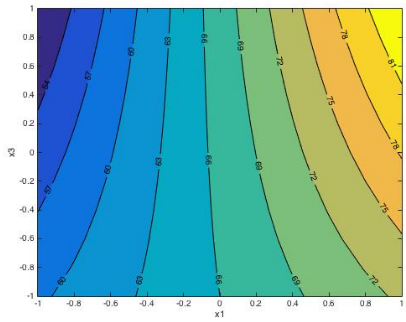
- Normal distribution of coefficients
- Residuals
- ANOVA



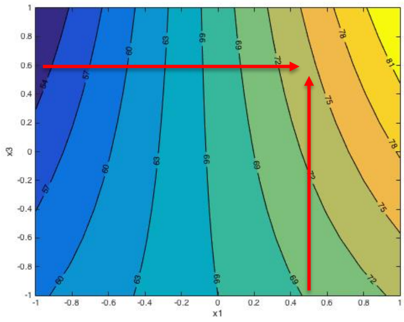
Factorial design

For three factors, 2D contours require one constant factor





Factorial design



Prediction

T = 110

P = 2 (min. level)

K = 0.9

Coded location

$\mathbf{x_m}$ = [1 0.5 -1 0.6 0.3]

Predicted response

y_m = 74.5

Session 1

Introduction

- Why experimental design

Factorial design

- Design matrix
- Model equation = coefficients
- Residual
- Response contour

Nomenclature

Factorial design
Screening
Design matrix
Model equation
Response
Effect (main/interaction)
Coefficient
Significance
Residual
Contour

Thank you!