Università degli Studi di Padova

Department of Industrial Engineering

Design of Experiments
Lesson #8

Academic year 2025-2026

Prof. Pierantonio Facco

CAPE-Lab, Computer-Aided Process Engineering Laboratory

Email: pierantonio.facco@unipd.it

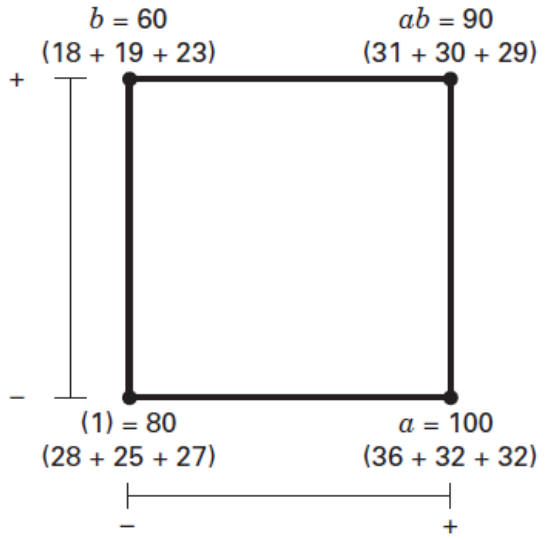
URL: https://research.dii.unipd.it/capelab/

Recap with an example: catalytic reaction

Catalytic reaction

- Investigation the effect of the concentration of the reactant and the amount of the catalyst on the conversion (yield) in a chemical process
- Objective: to determine if adjustments to either of these two factors in such a way as to increase the yield
- Experimentation:
  - factor A: **reactant concentration**
    - levels of interest: 15 and 25%
  - factor *B*: **catalyst quantity**
    - 1 or 2 pounds
  - experiments are replicated three times ($n = 3$)
  - the order in which the runs are made is random, so this is a **completely randomized experiment**

Factor A	Factor B	Treatment Combination	Replicate I	Replicate II	Replicate III
−	−	*A* low, *B* low	28	25	27
+	−	*A* high, *B* low	36	32	32
−	+	*A* low, *B* high	18	19	23
+	+	*A* high, *B* high	31	30	29



Main effects and interaction

- The main effects and the interaction are calculated in the following manner:

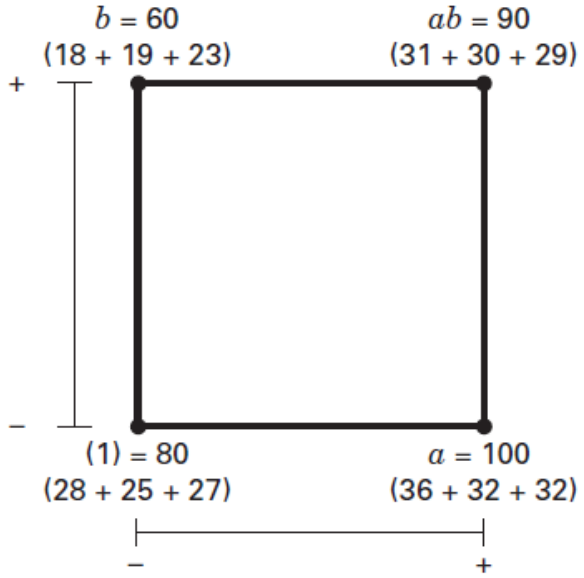
$$A = \frac{1}{2n}\{[ab - b] + [a - (1)]\}$$
$$= \frac{1}{2n}[ab + a - b - (1)]$$

$$B = \frac{1}{2n}\{[ab - a] + [b - (1)]\}$$
$$= \frac{1}{2n}[ab + b - a - (1)]$$

$$AB = \frac{1}{2n}\{[ab - b] - [a - (1)]\}$$
$$= \frac{1}{2n}[ab + (1) - a - b]$$

Calculation of the main effects

- Results:
  - the effect of *A* (reactant concentration) is positive
    - increasing *A* from the low level (15%) to the high level (25%) will increase the yield
  - the effect of *B* (catalyst) is negative
    - this suggests that increasing the amount of catalyst added to the process will decrease the yield
  - the interaction effect appears to be smaller in relation to the main effects



$$A = \frac{1}{2(3)}(90 + 100 - 60 - 80) = 8.33$$

$$B = \frac{1}{2(3)}(90 + 60 - 100 - 80) = -5.00$$

$$AB = \frac{1}{2(3)}(90 + 80 - 100 - 60) = 1.67$$

Sum of squares

- The sum of squares can be calculated as follows:
  - they give an idea of the relative importance of the factors and how they are related to the noise

$$SS_A = \frac{[ab + a - b - (1)]^2}{4n}$$

$$SS_B = \frac{[ab + b - a - (1)]^2}{4n}$$

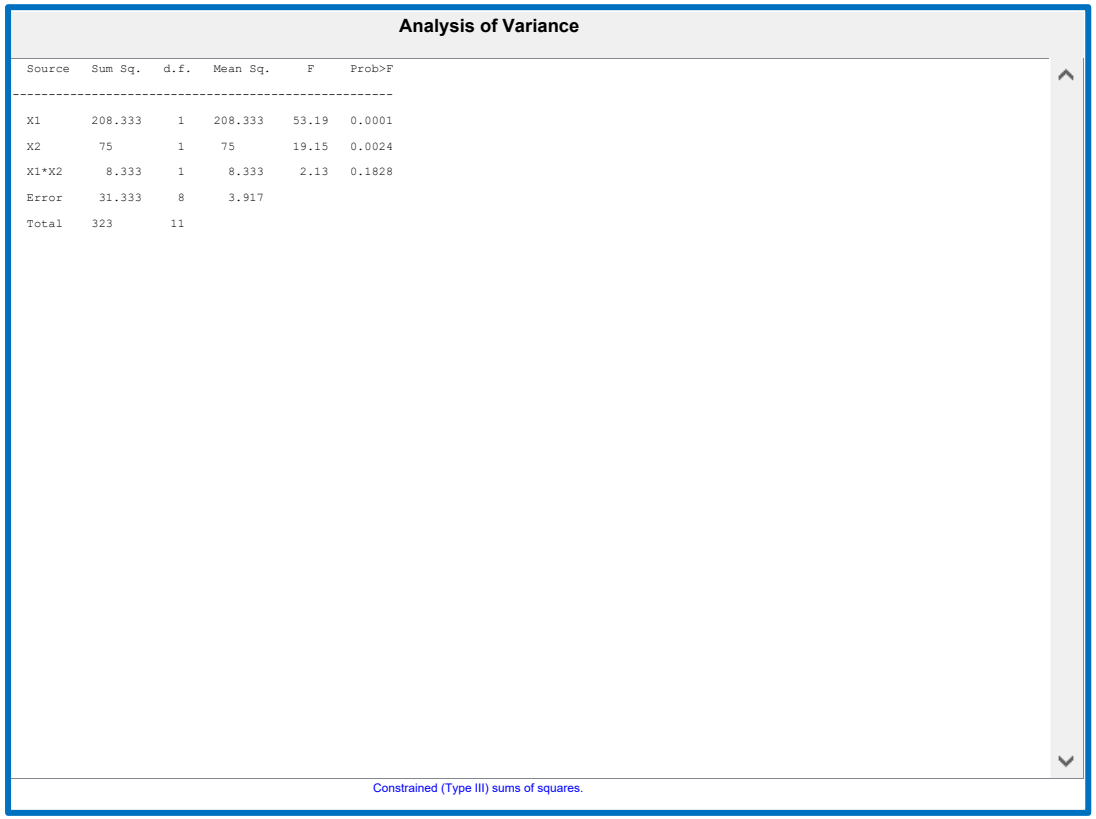
$$SS_{AB} = \frac{[ab + (1) - a - b]^2}{4n}$$

$$SS_T = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{n} y_{ijk}^2 - \frac{y_{...}^2}{4n}$$

$$SS_E = SS_T - SS_A - SS_B - SS_{AB}$$

Added value with respect to ANOVA

- The **analysis of variance** can generally be used to confirm this interpretation
- It is always important to examine the **magnitude and direction of the factor** effects to determine which variables are likely to be important (not provided by ANOVA)
  - the main effect are statistically significant
  - the interaction is not significant



Regression model

- In a 2^K factorial design, it is easy to express the results of the experiment in terms of a **regression model**
  - the regression model approach is much more natural and intuitive
- The regression model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

  - x_1 is a **coded variable** that represents the factor A (reactant concentration)
  - x_2 is a coded variable that represents the factor B (amount of catalyst)
  - β_i's are regression coefficients
- Coefficients are **estimated in a least-squares manner**:
  - the intercept $\hat{\beta}_o$ is the grand average of all the observations
  - the regression coefficients $\hat{\beta}_i$ are one-half the corresponding factor effect estimates
    - the regression coefficients are one-half the effect estimate of each factor because a regression coefficient measures the effect of a one-unit change in x on the mean of y, and the effect estimate is based on a two-unit change

Model predictions $\hat{y}$ and residuals

- The regression model can be used to obtain the **prediction** (or estimation or fitted value) $\hat{y}$ within the design domain:

$$\hat{y} = \hat{\beta}_o + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_{12} x_1 x_2$$

- The **residuals** should be **normally distributed**, and no pattern should be evident in their relationship with time, factors, fitted observations:
  - test of normality of the residuals and the standardized residuals
  - visual inspection of the plots of the residuals against:
    - time
    - fitted variables
    - factors
- The regression models are affected by **uncertainty**:
  - parameters uncertainty:

$$\hat{\beta}_v - t_{\alpha,N-V-1}\sqrt{\sigma^2 c_{vv}} \leq \beta_v \leq \hat{\beta}_v + t_{\alpha,N-V-1}\sqrt{\sigma^2 c_{vv}}$$

  - prediction uncertainty:

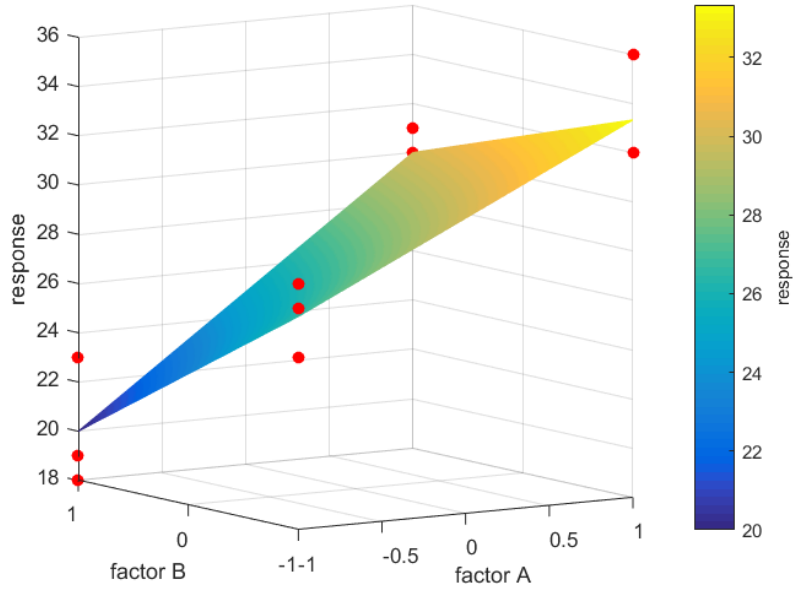
$$\hat{y}(\mathbf{x}_n) - t_{\alpha,N-V}\sqrt{\hat{\sigma}^2 \mathbf{x}_n^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_n} \leq \boldsymbol{\mu}_y(\mathbf{x}_n) \leq \hat{y}(\mathbf{x}_n) + t_{\alpha,N-V}\sqrt{\hat{\sigma}^2 \mathbf{x}_n^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_n}$$

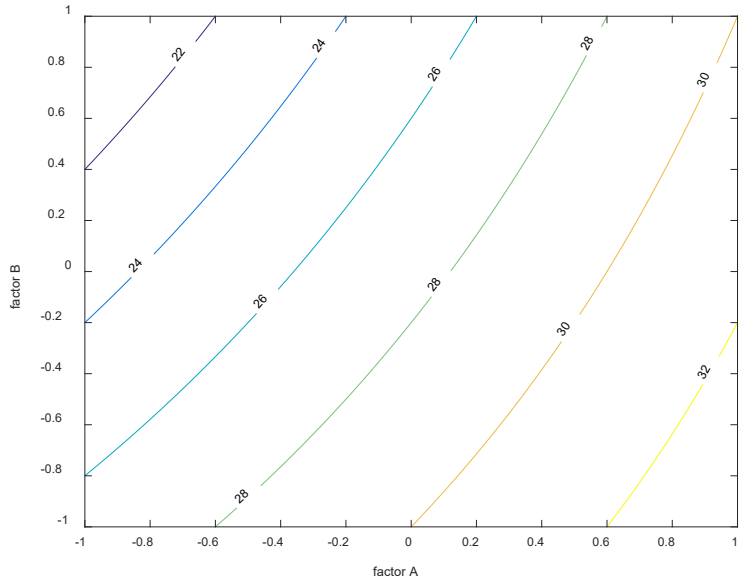
Response surface

- The regression model is used to generate a **response surface**, namely a representation of the **functional dependence of the response from the factors**
- The model adequacy is usually observed in the parameters that judge the **fitting performance** of the model
- The models' coefficients are utilized to understand the **importance of the factors, their interactions** (and even higher order combinations) **and their directions**
  - in depth study on the **coefficient uncertainty** give information on the factors'/interactions' importance and their reliability
- Thanks to response surfaces the experimenter can find the indications on the direction of potential **improvement** for the process and the system and the conditions that optimize it

Catalytic reaction: response surface

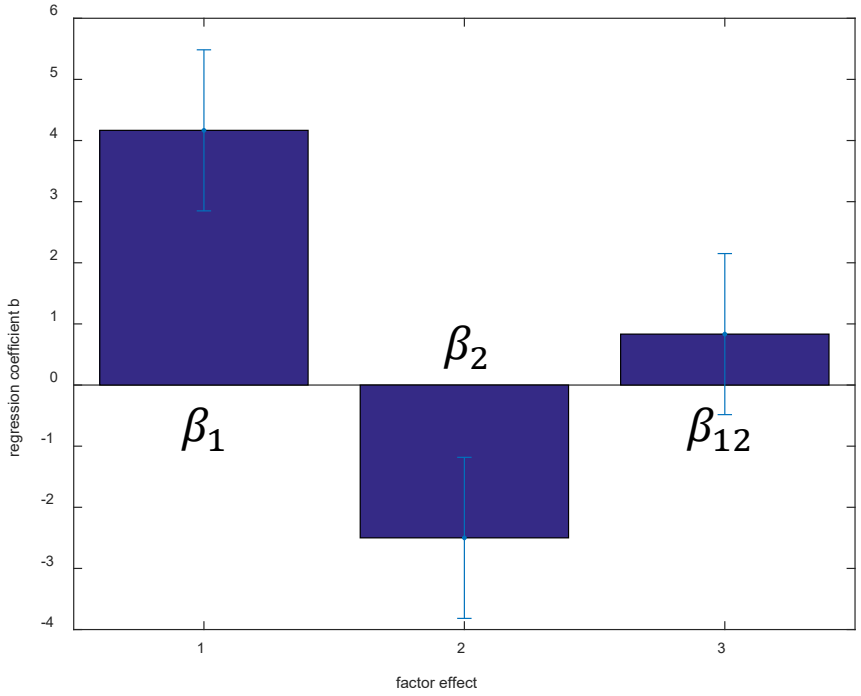
- The response surface and the contour plot show:
  - models fits very well the available data:
    - $R^2 > 90\%$
    - the surface is close to the measured points
  - the best fitting surface is a plane:
    - limited curvature indicates that limited interaction is present
    - the maximum of the yield is in correspondence of high values of factor A (reactant concentration) and low values of factor B (content of catalyst)





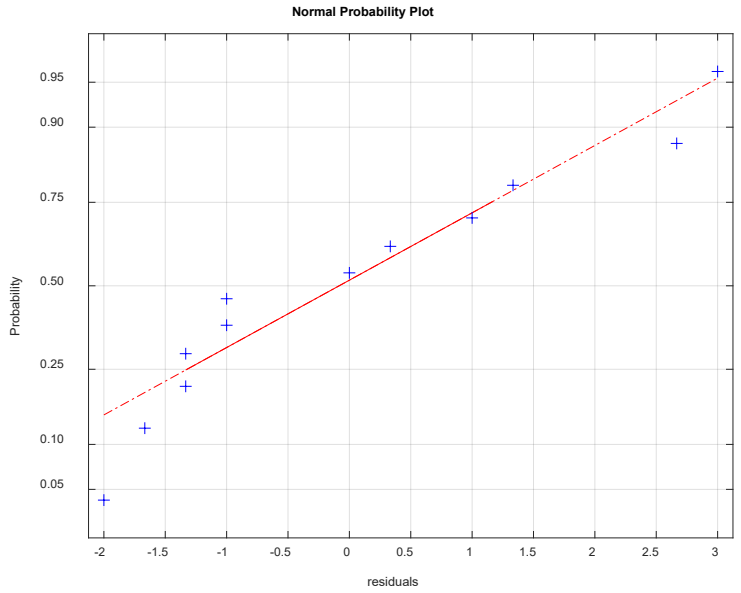
Catalytic reaction: regression coefficients

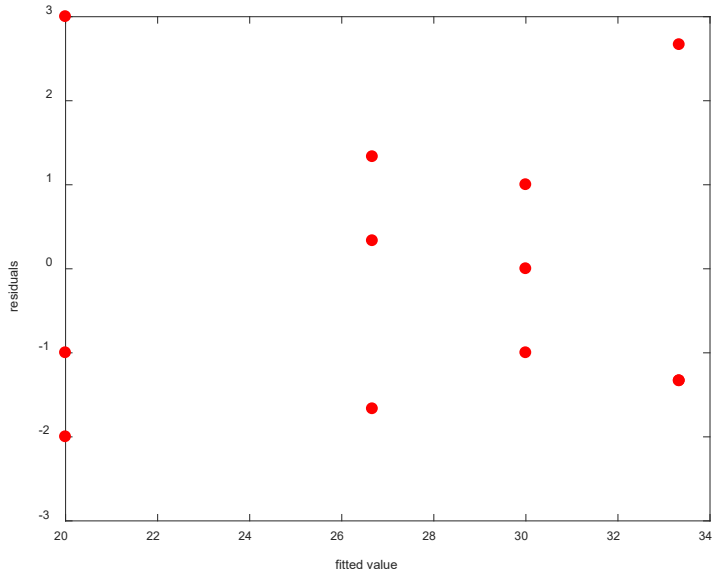
- The regression coefficients show that:
  - the highest effect is the one of the reactant concentration
    - this contribution is a positive one, in the sense that the product yield increases with increasing values of the reactant
  - the second contribution to the product yield is the one of the catalyst
    - the catalyst has a negative contribution on the response variable
  - the interaction contribution is the lowest one and the uncertainty (error bars) show that it is not reliably high

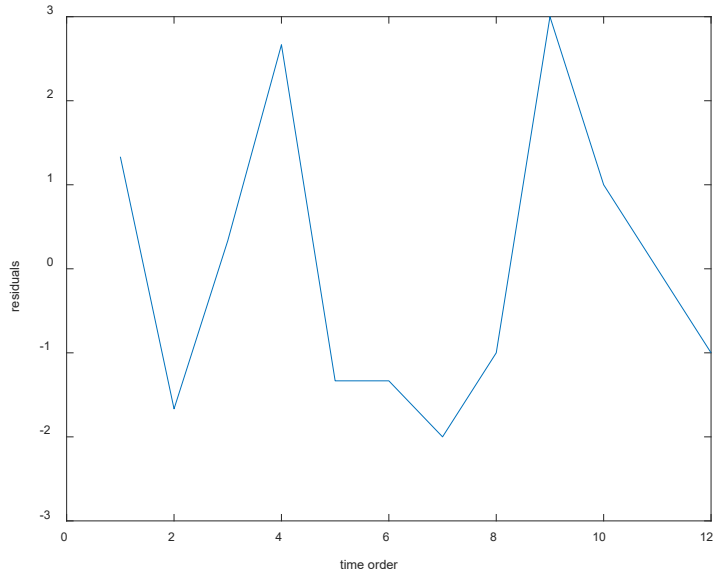


Catalytic reaction: residuals analysis

- No major concerns are related to the residuals:
  - normality seems to be evident
  - no patterns are present:
    - in time
    - against fitted values



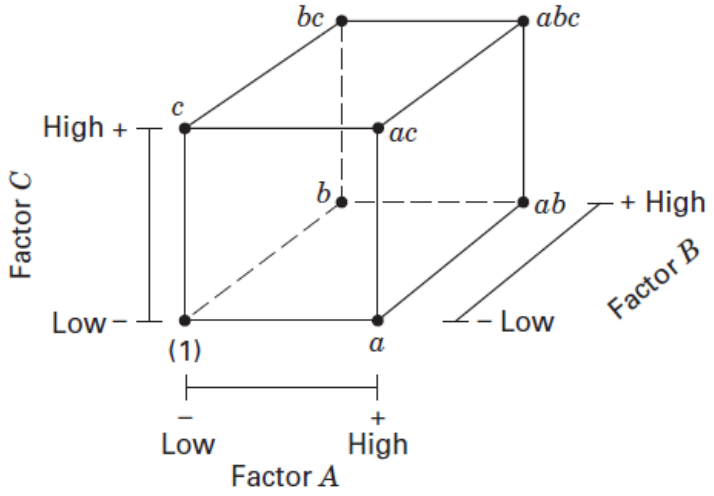




Full-factorial designs in higher dimension

High-dimension full-factorial designs

- The methodology of experiment data modelling can be (easily) transferred to the higher dimensional spaces:
  - **larger number of factors**
- Suppose that 3 factors, *A*, *B*, and *C*, each at two levels, are of interest:
  - the design is called a 2^3 **factorial design**
  - the eight treatment combinations be displayed geometrically as a cube



Run	Factor A	B	C
1	–	–	–
2	+	–	–
3	–	+	–
4	+	+	–
5	–	–	+
6	+	–	+
7	–	+	+
8	+	+	+

Degrees of freedom

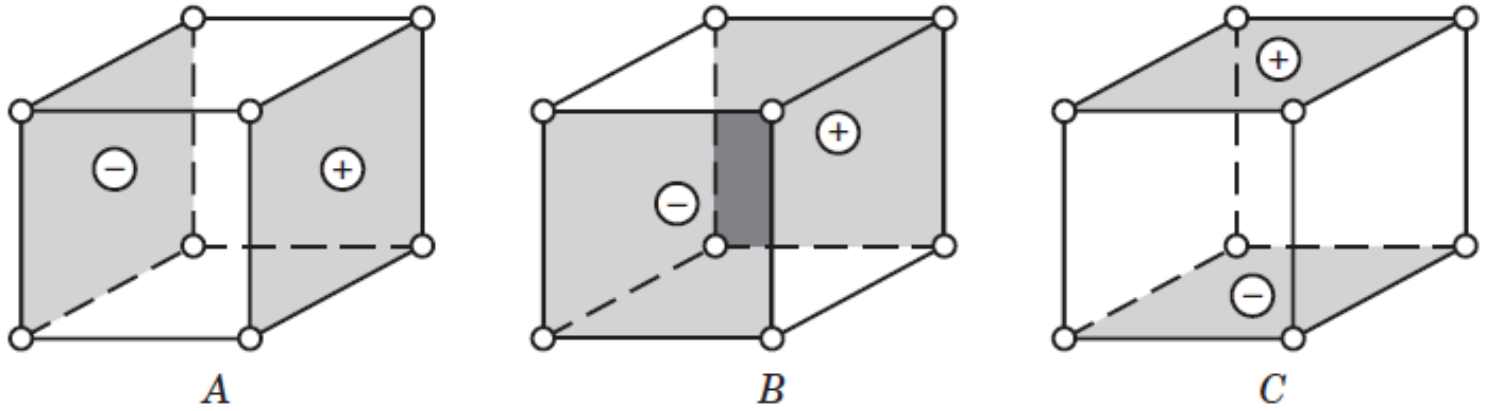
- There are seven degrees of freedom between the eight treatment combinations in the 2^3 design
  - 3 degrees of freedom are associated with the main effects of *A, B*, and *C*
  - 3 degrees of freedom are associated with interactions: *AB, AC*, and *BC*
  - 1 degree of freedom is associated with the complex interaction *ABC*

Main effects

- Four effects are related to A:
  - one is for low levels of both B and C
  - one is low levels of B and high level of C
  - one is high levels of B and low level of C
  - one is high levels of both B and C

$$A = \frac{1}{4n}[a + ab + ac + abc - (1) - b - c - bc]$$

- This equation can also be developed as a **contrast** between the four treatment combinations in the right face of the cube
- This can be extended to effects B and C

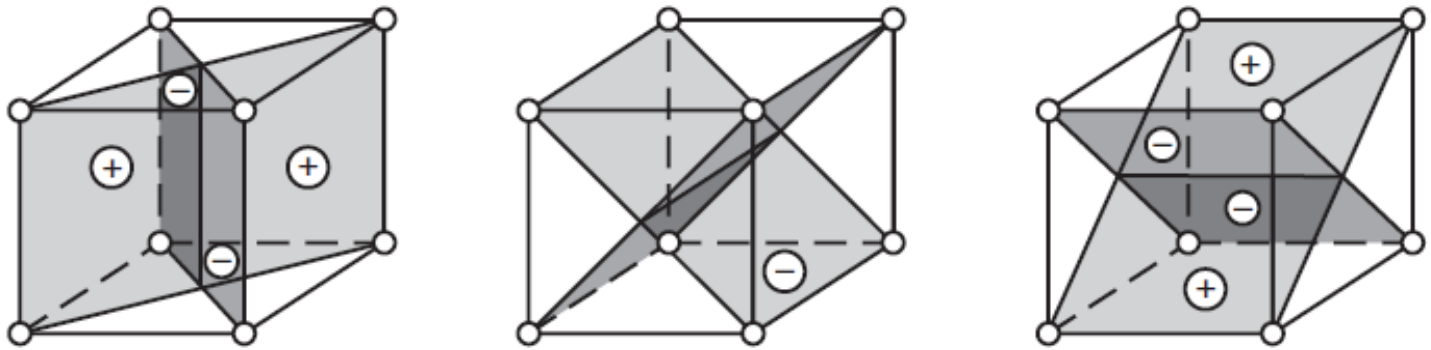


Two-factors interactions

- The two-factor interaction effects may be computed easily
- A measure of the *AB* interaction is the difference between:
  - the average *A* effects at the two levels of *B*
  - by convention one-half of this difference is called the *AC* interaction

$$AC = \frac{1}{4n}[(1) - a + b - ab - c + ac - bc + abc]$$

  - the *AC* interaction is the difference in averages between runs on two diagonal planes in the cube

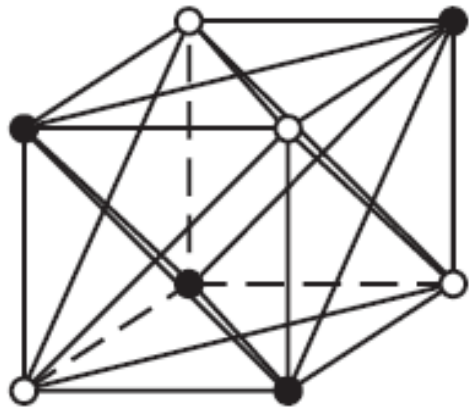


Three-factors interactions

- The *ABC* interaction is defined as the average difference between the *AB* interaction at the two different levels of *C*
  - the *ABC* interaction is the difference in two averages
  - if the runs in the two averages are isolated, they define the vertices of the two tetrahedra that comprise the cube

$$ABC = \frac{1}{4n}\{[abc - bc] - [ac - c] - [ab - b] + [a - (1)]\}$$

$$= \frac{1}{4n}[abc - bc - ac + c - ab + b + a - (1)]$$



ABC

● = + runs

○ = – runs

General 2^K designs

- The methods of analysis that is presented may be generalized to the case of a 2^K **factorial design**, that is, a design with K factors each at two levels
- The statistical model for a 2^K design would include K main effects, two-factor interactions, three-factor interactions, and so on …, and one K-factor interaction
  - the complete model would contain $(2^K - 1)$ effects for a 2^K design
- The treatment combinations may be written in **standard order** by introducing the factors one at a time:
  - for example, the standard order for a 24 design is (1), *a, b, ab, c, ac, bc, abc, d, ad, bd, abd, cd, acd, bcd*, and *abcd*.

General DoE data analysis procedure

1. estimate factor effects

- determine sign and magnitude
- obtain preliminary information regarding:
  - what are important factors?
  - do factors have positive or negative effect on y?
  - are non-linearities present?

2. build initial model

- response surface through MLR
  - check model accuracy through determination coefficient
  - check adjusted determination coefficient

3. perform statistical testing

- are all the regressors meaningful?
  - check p-values for all the regressors

4. refine model

- remove any non-significant variables from the model

5. analyze residuals

- check for model adequacy and assumptions
  - model refinement after residual analysis

6. interpret results

- study regression parameters
- visualize response surface and contour plot interpretation

Single-replicate full factorial designs

Single replicate designs

- For even a moderate number of factors, the total number of treatment combinations in a 2^K factorial design is large
  - for example:
    - a 2^5 design has 32 treatment combinations
    - a 2^6 design has 64 treatment combinations, etc…
- Because resources are usually limited, *the number of replicates that the experimenter can employ may be restricted*
- Very frequently resources only allow a **single replicate (unreplicated factorials)** of the design to be run, unless the experimenter is willing to omit some of the original factors:
  - an obvious risk is that the model may be fitted to noise
  - if the response y is highly variable, misleading conclusions may result from the experiment
- The single-replicate strategy is often used in *screening experiments* when there are **relatively many factors under consideration**:
  - since one can never be entirely certain in such cases that the experimental error is small, a good practice in these types of experiments is to *spread out the factor levels aggressively*
  - there is *no internal estimate of error* (or “pure error”)
  - one approach to the analysis is to assume that certain high-order interactions are negligible and combine their mean squares to estimate the error
    - this is an appeal to the **sparsity of effects principle**:
      - most systems are dominated by some of the main effects and low order interactions
      - most high-order interactions are negligible

Analyzing unreplicated factorial designs

- Occasionally real high-order interactions occur
- The use of an error mean square obtained by pooling high-order interactions is inappropriate in these cases
- A simple way to overcome this problem:
  - examining a **normal probability plot of the estimates of the effects**
    - the effects that are negligible are normally distributed, with mean zero and variance σ^2 and will tend to fall along a straight line on this plot
    - significant effects will have nonzero means and will not lie along the straight line
    - the preliminary model will be specified to contain those effects that are apparently nonzero, based on the normal probability plot
    - the apparently negligible effects are combined as an estimate of error

Example: filtration rate

- A chemical product is produced in a **pressure vessel**
- Objective: experiments are carried out in the pilot plant to study how the factors influence the filtration rate of the product
- Experimentation:
  - 4 factors:
    - temperature T (factor A)
    - pressure P (factor B)
    - concentration of formaldehyde $C_{formaldehyde}$ (factor C)
    - stirring rate SR (factor D)
  - 2^4 factorial factors are
    - each factor is present at two levels
    - single replicate experiments

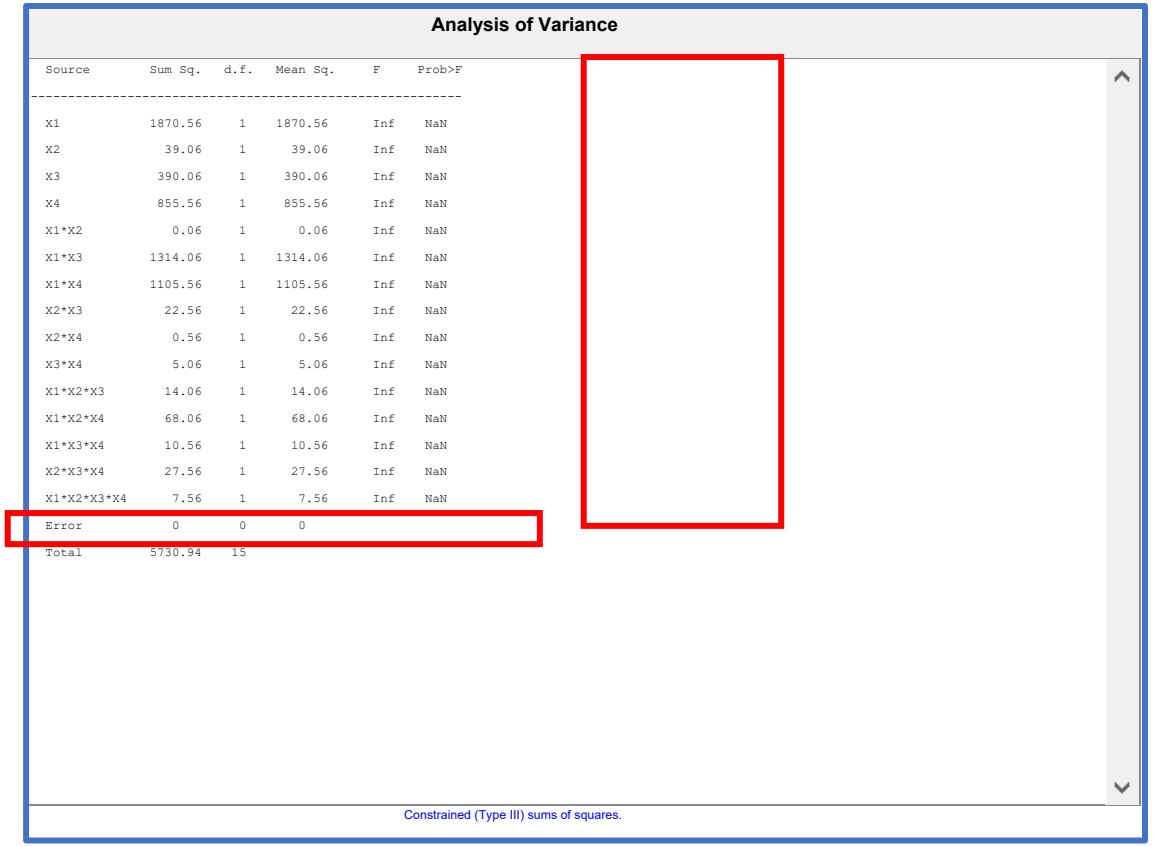
Design and responses

- Available data:

Factor				Filtration Rate (gal/h)
A	*B*	*C*	*D*	
−	−	−	−	45
+	−	−	−	71
−	+	−	−	48
+	+	−	−	65
−	−	+	−	68
+	−	+	−	60
−	+	+	−	80
+	+	+	−	65
−	−	−	+	43
+	−	−	+	100
−	+	−	+	45
+	+	−	+	104
−	−	+	+	75
+	−	+	+	86
−	+	+	+	70
+	+	+	+	96

N-way ANOVA for single replicate experiments

- The only available information is related to the effects of the factors:
  - **no estimation of the SSE is available**, nor the significance of the effects



Analysis of Variance

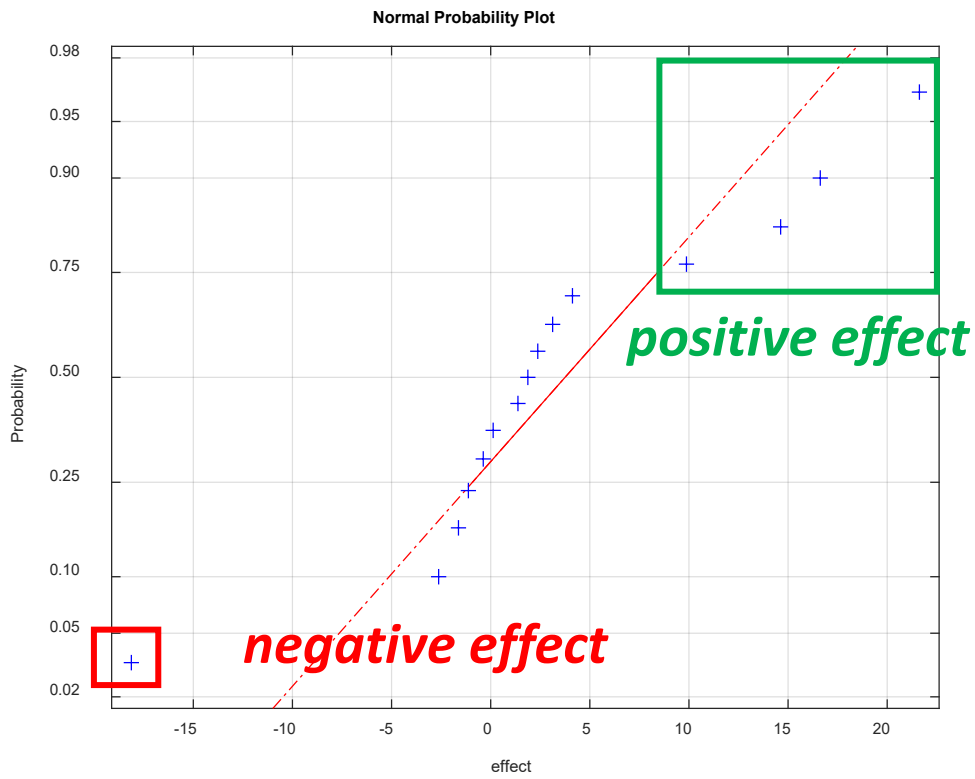
Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
X1	1870.56	1	1870.56	Inf	NaN
X2	39.06	1	39.06	Inf	NaN
X3	390.06	1	390.06	Inf	NaN
X4	855.56	1	855.56	Inf	NaN
X1*X2	0.06	1	0.06	Inf	NaN
X1*X3	1314.06	1	1314.06	Inf	NaN
X1*X4	1105.56	1	1105.56	Inf	NaN
X2*X3	22.56	1	22.56	Inf	NaN
X2*X4	0.56	1	0.56	Inf	NaN
X3*X4	5.06	1	5.06	Inf	NaN
X1*X2*X3	14.06	1	14.06	Inf	NaN
X1*X2*X4	68.06	1	68.06	Inf	NaN
X1*X3*X4	10.56	1	10.56	Inf	NaN
X2*X3*X4	27.56	1	27.56	Inf	NaN
X1*X2*X3*X4	7.56	1	7.56	Inf	NaN
Error	0	0	0		
Total	5730.94	15			

Constrained (Type III) sums of squares.

Analysis of the normality of the effects

- The effects that deviate from the normality are the most significant

Model Term	Effect Estimate
A	21.625
B	3.125
C	9.875
D	14.625
AB	0.125
AC	−18.125
AD	16.625
BC	2.375
BD	−0.375
CD	−1.125
ABC	1.875
ABD	4.125
ACD	−1.625
BCD	−2.625
ABCD	1.375



Main effect and interactions

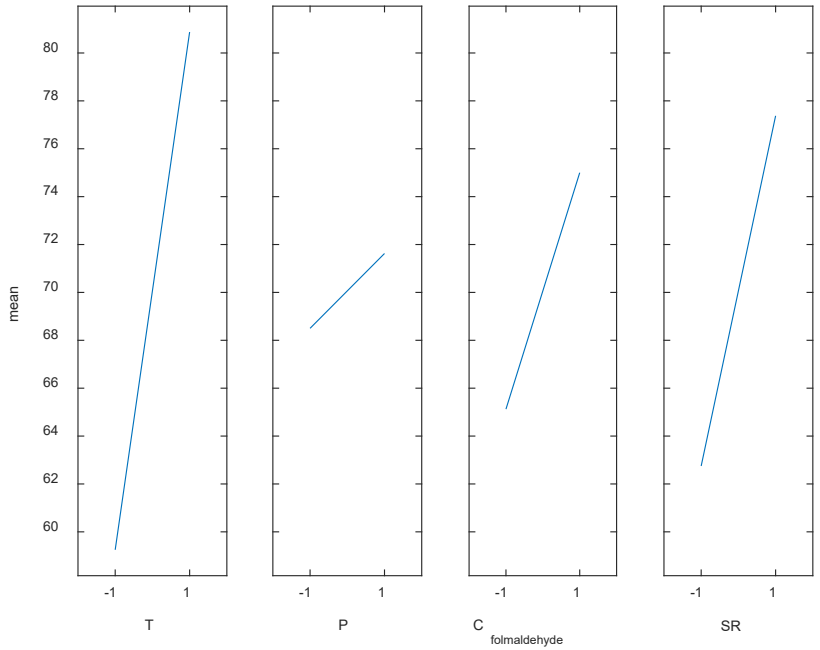
- The main effect acting on the process are related to:
  - temperature
  - formaldehyde concentration
  - stirring rate

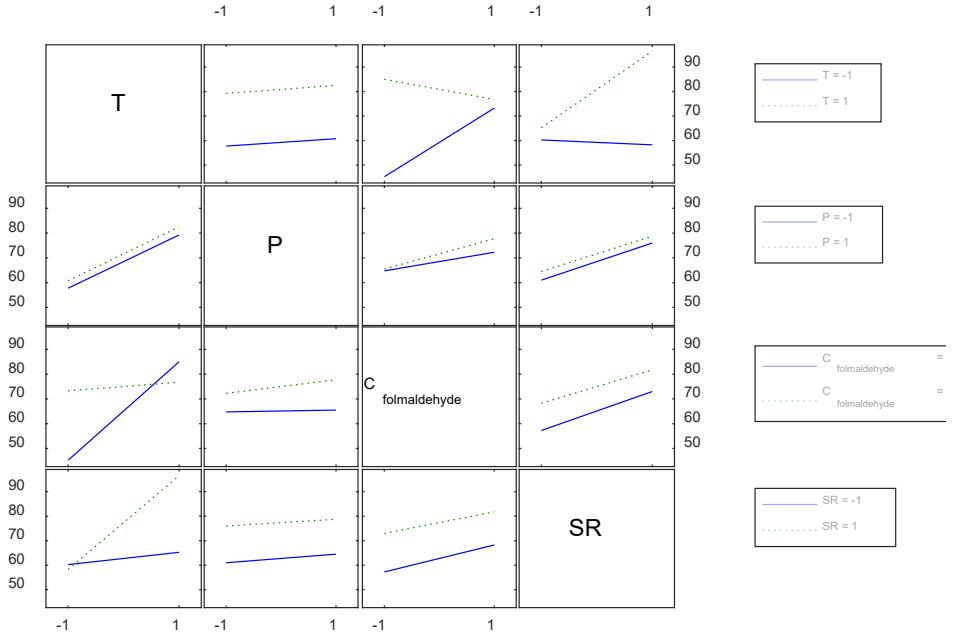
positive effect

- The most important interactions are related to:
  - temperature and formaldehyde concentration
  - temperature and stirring rate

negative effect

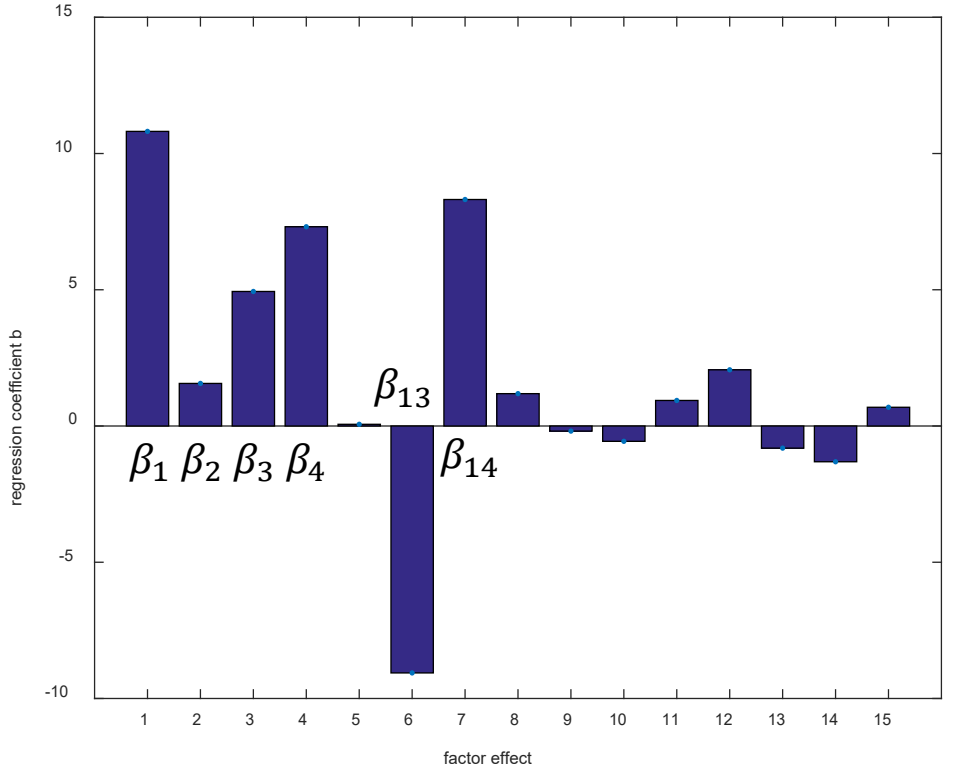
positive effect





Response hyper-surface approach

- The analysis of the regression coefficients confirms all the outcomes obtained with N-way ANOVA on both the main effects and the interactions:
  - increasing T, $C_{formaldhyde}$ and SR increases the filtration rate
  - the interaction T - $C_{formaldhyde}$ has a negative effect on the filtration rate
  - the interaction T - SR has a positive effect on the filtration rate



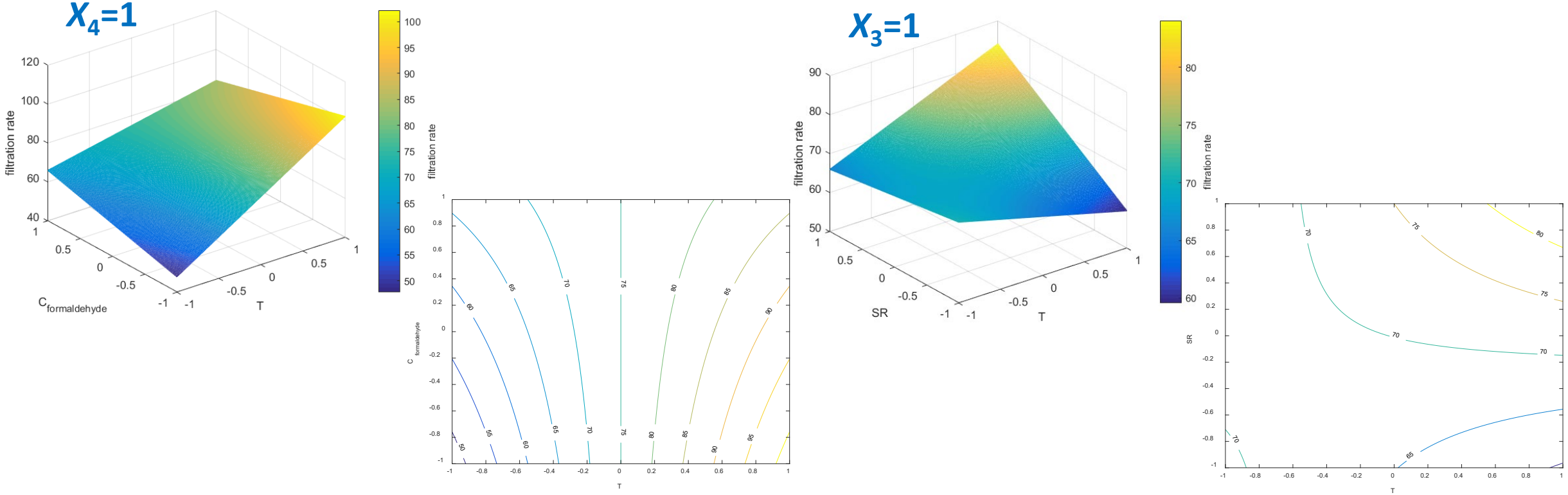
Model update

- The model is updated to the form:

$$y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \beta_4 x_4 + \beta_{13} x_1 x_3 + \beta_{14} x_1 x_4$$

- The results can be summarized in response surfaces:
  - high values of temperature and stirring rate and low values of formaldehyde concentration increase the filtration rate

X_4=1



X_3=1

D-optimal design

- The 2^K factorial designs are **D-optimal designs**:
  - the least squares estimation of the model parameters minimizes the variance of the model regression coefficients:

$$\mathbf{B} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

  - the coefficients are one half of the effect estimates
  - minimizing the variance of the regression coefficients means maximizing the determinant of the **information matrix** $\mathbf{X}^{\mathrm{T}}\mathbf{X}$

The issue of nonlinearity

- A potential issue in the use of two-level factorial designs is the **assumption of linearity** in the factor effects
  - perfect linearity is unnecessary, and the 2^K system will work quite well even when the linearity assumption holds only very approximately
  - if **interaction terms** are added to a main effect or first-order model, then the model is capable of representing some curvature in the response
- **Second-order response surface model**: in some situations, the curvature in the response function will not be adequately modeled by interactions and a logical model to consider is:

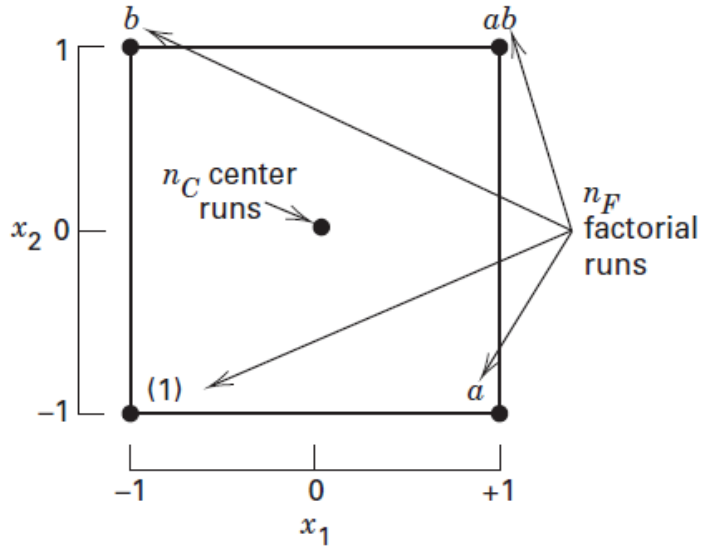
$$y = \beta_0 + \sum_{i=1}^{K} \beta_i x_i + \sum_{i<j} \beta_{ij} x_i x_j + \sum_{i=1}^{K} \beta_{ii} x_i^2 + \varepsilon$$

  - where the ii represent pure second-order or **quadratic effects**

Adding center points

- The method for considering potential second order effects consists of adding **center points** to the 2^K design

- Replicating center points in a 2^K factorial:
  - provide protection against **curvature** from second-order effects
  - allow an **independent estimate of error** to be obtained
  - n_C replicates run at the points $x_k = 0$ (for $k = 1, 2, \ldots, K$)
  - center points do not affect the usual effect estimates in a 2^K design
  - when we add center points, we assume that the K factors are **quantitative**



- For example, consider a 2^2 design:
  - one observation at each of the factorial points (n_F in total)
  - n_C observations at the center point $(0, 0)$
  - let $\bar{y}_{\mathrm{F}}$ be the average of the four runs at the four factorial points, and $\bar{y}_{\mathrm{C}}$ be the average of the n_C runs at the center point
  - compare to the error mean square to test for the significance of the curvature
    - if the difference is small, then the center points lie on or near the plane passing through the factorial points ➡ there is no quadratic curvature
    - if $(\bar{y}_{\mathrm{F}} - \bar{y}_{\mathrm{C}})$ is large ➡ quadratic curvature is present
    - a single-degree-of-freedom **sum of squares for pure quadratic curvature** is given by:

$$SS_{\text{Pure quadratic}} = \frac{n_F n_C (\bar{y}_F - \bar{y}_C)^2}{n_F + n_C}$$

Use of the center points

- Suggestions:

1. consider using the *current operating conditions* (or recipe) as the center point in the design:
   - at least some of the runs in the experiment are going to be performed under familiar conditions for the operating personnel
2. provide a rough check of whether *anything "unusual" occurred* during the experiment when the center point corresponds to the usual operating recipe:
   - the center point responses should be very similar to the responses observed historically
3. consider running the replicates at the center point in **nonrandom order**
   - one or two center points at or near the beginning of the experiment
   - one or two near the middle
   - one or two near the end

   - spreading the center points out, the experimenter has a rough check on the process **stability**
4. sometimes experiments must be conducted in situations where there is little or no prior information. In these cases, running two or three center points as the *first few runs* in the experiment can be very helpful to provide a **preliminary estimate of variability**
   - if the magnitude of the variability seems reasonable, continue
   - if larger than anticipated variability is observed, stop
5. usually, center points are employed when all design factors are **quantitative**
   - sometimes there will be one or more qualitative or categorical variables and several quantitative ones

Generating full factorial designs

- Once the number of levels and the number of factor are selected, it is easy to generate the **experimental matrix** for a full-factorial design:
  - information can be found in tables in the main textbooks
  - it can be easily built by hand
  - in Matlab® the command are:
    - for 2^K factorial designs: `ff2n`
    - for general full factorial designs: `fullfact`

… per sempre a fianco a me!