TMA4267 Linear Statistical Models V2017 (L17)

Part 4: Design of Experiments

Mette Langaas

Department of Mathematical Sciences, NTNU

To be lectured: March 21, 2017

Today:

- Observational studies vs. designed experiments.
- Still linear regression, but now with k factors each with only 2 levels.
- Effect coding, orthogonal columns in design matrix.
- 2^k full factorial design.
- Simplified formulas for $\hat{\boldsymbol{\beta}}$, $\mathrm{Cov}(\hat{\boldsymbol{\beta}})$ and SSE.
- If time: from parameter estimated to main and interaction effects.

Part 4 is based on Tyssedal: Design of experiments note.

Design of experiments vs. observational studies

In this part of the course we are working with the linear regression model:

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon \text{ with } \varepsilon \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

and use results from Part 2 of the course.

Earlier in the course: both the design matrix $\boldsymbol{X}$ and the reponses Y were observed together in a randomly selected sample from a population.

- Munich rent index: rent prices vs. area, location, condition of bathroom, condition of kitchen, ...
- Lakes: pH level vs. content of SO_4, NO_3, latent Al, Ca, organic, position, area.
- Happiness: Happiness vs. love, money, sex and work.

Now: we choose (design) the experiment by specifying the design matrix $\boldsymbol{X}$ to be used to produce a sample, and then collecting reponses Y for this design matrix.

The pilot plant example - Version 1

At a pilot plant a chemical process is investigated.

- The outcome of the process is measured as chemical yield (in grams).
- Two quantitative variables (factors) were investigated:
  - Factor A: Temperature (in degrees C).
  - Factor B: Concentration (in percentage).

Experiment no.	Temperature	Concentration	Yield
1	160	20	60
2	180	20	72
3	160	40	54
4	180	40	68
	x_1	x_2	y

Regression with pilot plant data V1- original

```
> x1=c(160,180,160,180)
> x2=c(20,20,40,40)
> y=c(60,72,54,68)

> fitx=lm(y~x1*x2)
Coefficients:
(Intercept)           x1           x2        x1:x2
    -14.000        0.500       -1.100        0.005

> model.matrix(fitx)
  (Intercept)  x1 x2 x1:x2
1           1 160 20  3200
2           1 180 20  3600
3           1 160 40  6400
4           1 180 40  7200
```

Regression with pilot plant data V1- recoded

```
> # recode to -1 and 1
> z1=(x1-(max(x1)+min(x1))/2)/((max(x1)-min(x1))/2)
> z2=(x2-(max(x2)+min(x2))/2)/((max(x2)-min(x2))/2)
> fitz=lm(y~z1*z2)
Coefficients:
(Intercept)           z1           z2        z1:z2
       63.5          6.5         -2.5          0.5

> model.matrix(fitz)
  (Intercept) z1 z2 z1:z2
1           1 -1 -1     1
2           1  1 -1    -1
3           1 -1  1    -1
4           1  1  1     1
```

Regression with original and coded factors

Original: x_1 and x_2, gave estimated regression equation

$$\hat{y} = -14 + 0.5x_1 - 1.1x_2 + 0.005x_1 \cdot x_2$$

Coded: $z_1 = (x_1 - 170)/10$ and $z_2 = (x_2 - 30)/10$, gave estimated regression equation

$$\hat{y} = 63.5 + 6.5z_1 - 2.5z_2 + 0.5z_1 \cdot z_2$$

Can you compare these two results?

Regression with original and coded factors

Substitute $z_1 = (x_1 - 170)/10$ and $z_2 = (x_2 - 30)/10$ into the equation to get a estimated regression equation based on x_1 and x_2.

$$\begin{aligned}
\hat{y} &= 63.5 + 6.5 z_1 - 2.5 z_2 + 0.5 z_1 \cdot z_2 \\
&= 63.5 + 6.5\frac{x_1 - 170}{10} - 2.5\frac{x_2 - 30}{10} + 0.5\frac{x_1 - 170}{10} \cdot \frac{x_2 - 30}{10} \\
&= 63.5 - 6.5\frac{170}{10} + 2.5\frac{30}{10} + 0.5\frac{170 \cdot 30}{10 \cdot 10} \\
&+ x_1(6.5\frac{1}{10} - 0.5\frac{1}{10}\frac{30}{10}) + x_2(-2.5\frac{1}{10} - 0.5\frac{1}{10}\frac{170}{10}) \\
&+ 0.5\frac{1}{10}\frac{1}{10} x_1 \cdot x_2 \\
&= -14 + 0.5 x_1 - 1.1 x_2 + 0.005 x_1 \cdot x_2
\end{aligned}$$

Design of experiments (DOE) terminology

- Variables are called factors, and denoted A, B, C, ...
- We will only look at factors with two levels:
  - high, coded as $+1$ or just $+$, and,
  - low, coded as -1 or just $-$.
- In the pilot plant example we had two factors with two levels, thus $2 \cdot 2 = 4$ possible combinations. In general k factors with two levels gives 2^k possible combinations.

Standard notation for 2^2 experiment:

Experiment no.	A	B	AB	Level code	Response
1	-1	-1	1	1	y_1
2	1	-1	-1	a	y_2
3	-1	1	-1	b	y_3
4	1	1	1	ab	y_4
	z_1	z_2	z_{12}		y

Lima beans example

Experiment from Box, Hunter, Hunter, Statistics for Experimenters, page 321.

- A: depth of planting (0.5 inch or 1.5 inch)
- B: watering daily (once or twice)
- C: type of lima bean (baby or large)
- Y: yield

A	B	C	AB	AC	BC	ABC	Level code	Response
-	-	-	+	+	+	-	1	6
+	-	-	-	-	+	+	a	4
-	+	-	-	+	-	+	b	10
+	+	-	+	-	-	-	ab	7
-	-	+	+	-	-	+	c	4
+	-	+	-	+	-	-	ac	3
-	+	+	-	-	+	-	bc	8
+	+	+	+	+	+	+	abc	5
x_1	x_2	x_3	x_{12}	x_{13}	x_{23}	x_{123}		y

Main effects in DOE

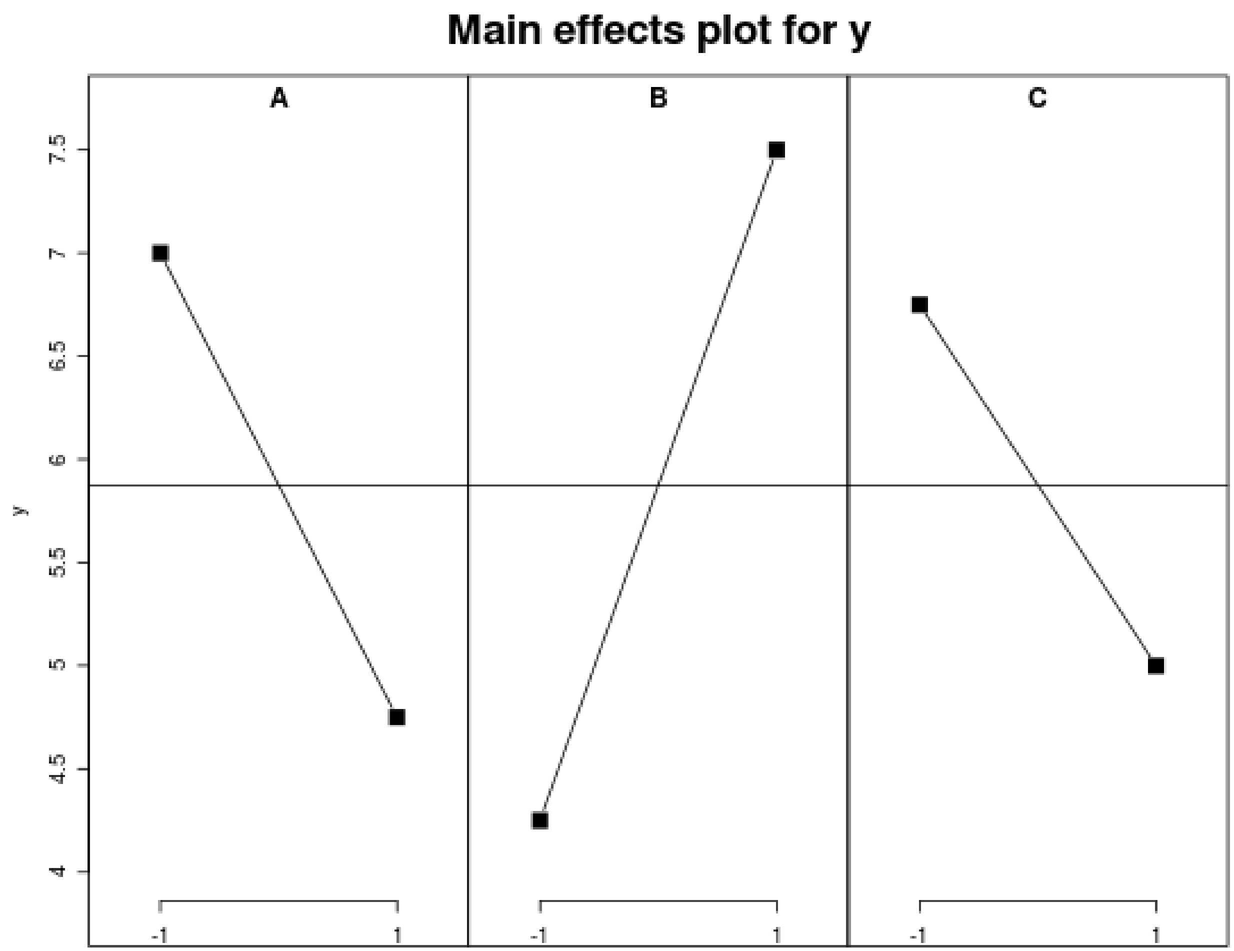
Main effect of A

$$\begin{aligned}\widehat{A} &= 2\hat{\beta}_1 \\ &= \frac{y_2 + y_4 + y_6 + y_8}{4} - \frac{y_1 + y_3 + y_5 + y_7}{4}\end{aligned}$$

Interpretation: mean response when A is high MINUS mean response when A is low.
Similarily, main effect of B

$$\begin{aligned}\widehat{B} &= 2\hat{\beta}_2 \\ &= \frac{y_3 + y_4 + y_7 + y_8}{4} - \frac{y_1 + y_2 + y_5 + y_6}{4}\end{aligned}$$

Interpretation: mean response when B is high MINUS mean response when B is low.



A	B	C	A:B	A:C	B:C	A:B:C
-2.25	3.25	-1.75	-0.75	0.25	-0.25	-0.25

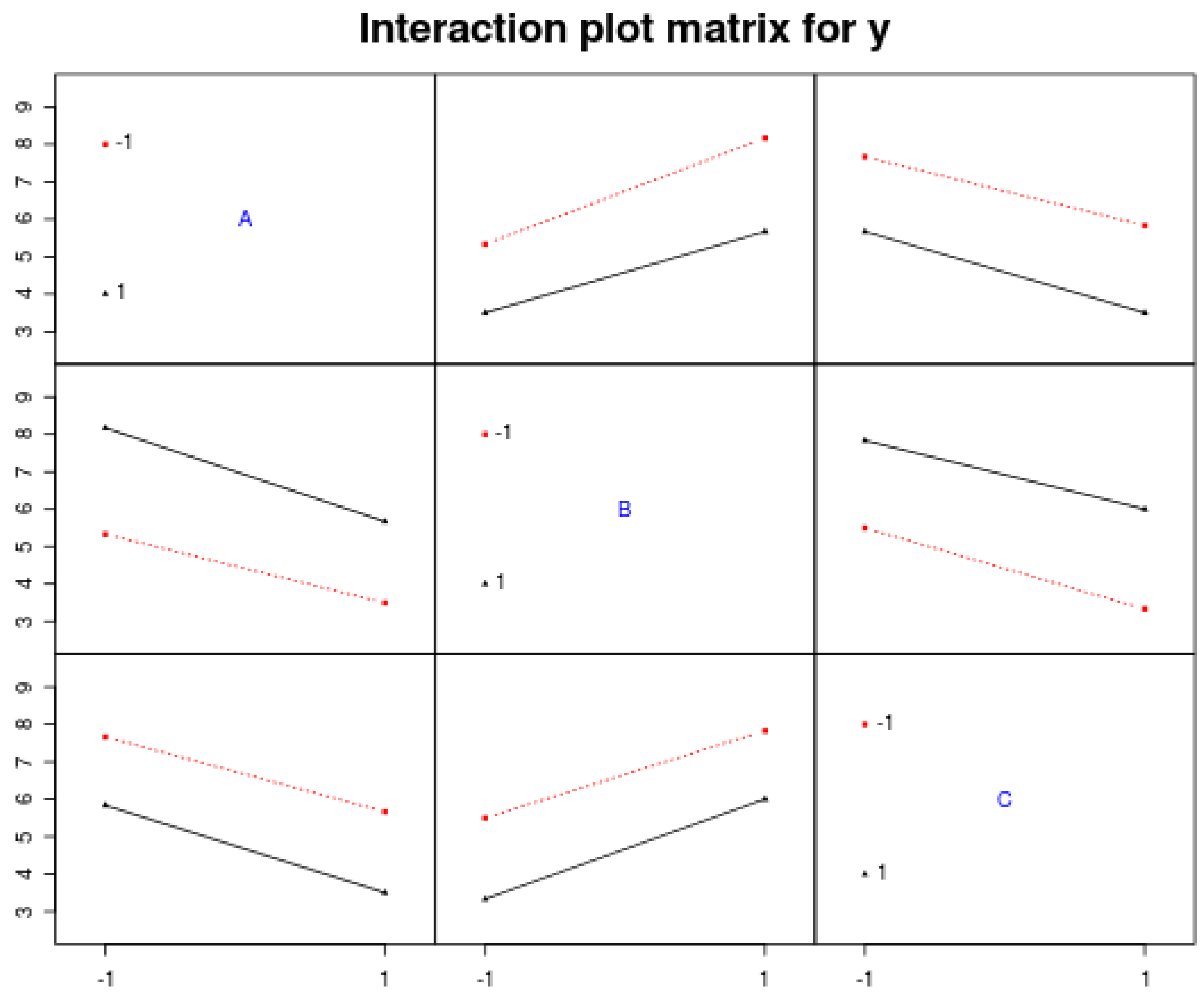
Explain the main effects in plain words!

A: depth (0.5 or 1), B: watering daily (once, twice), C: type (baby, large).

Interaction effect in DOE

- What is the terpretation in DOE associated with β_{12}?
- In DOE $2\hat{\beta}_{12}$ is denoted $\widehat{AB}$ and is called the *estimated interaction effect between A and B*.

$$
\begin{aligned}
\widehat{AB} &= 2\hat{\beta}_{12} \\
&= \frac{\text{estimated main effect of } A \text{ when } B \text{ is high}}{2} \\
&- \frac{\text{estimated main effect of } A \text{ when } B \text{ is low}}{2} \\
&= \frac{\text{estimated main effect of } B \text{ when } A \text{ is high}}{2} \\
&- \frac{\text{estimated main effect of } B \text{ when } A \text{ is low}}{2}
\end{aligned}
$$

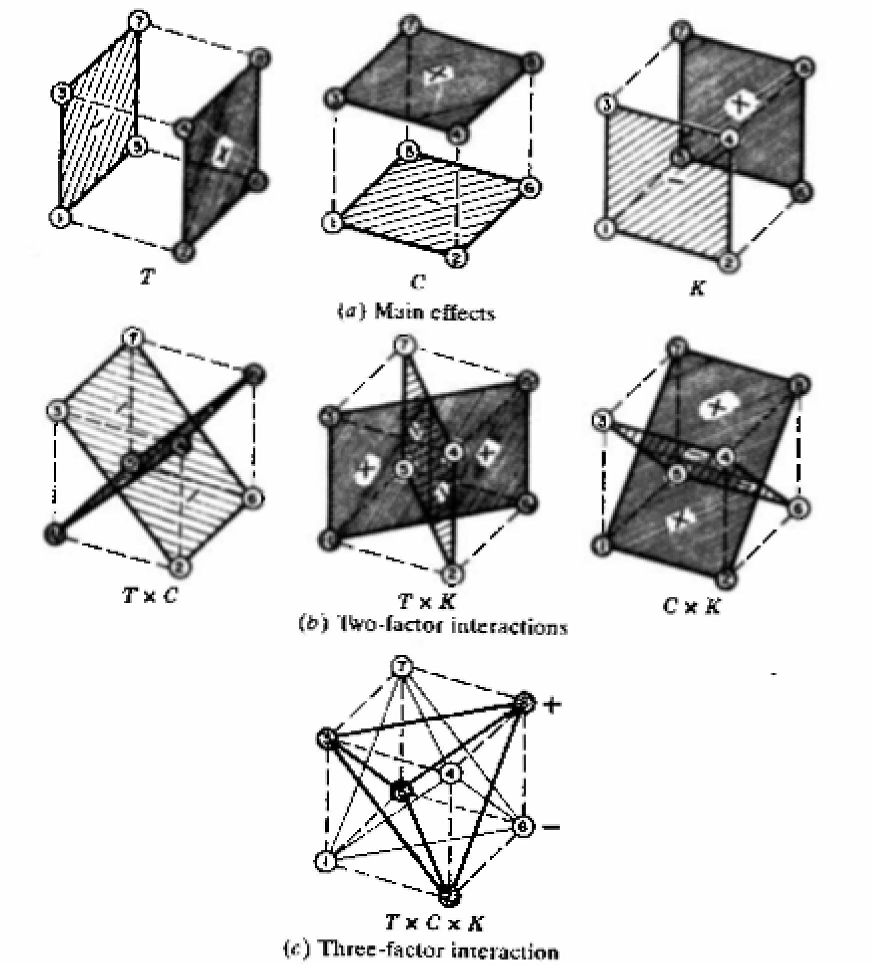



```
  A      B      C       A:B     A:C      B:C     A:B:C
-2.25  3.25  -1.75   -0.75   0.25    -0.25   -0.25
```

Interpretation of $\widehat{ABC}$

- $\widehat{ABC} = \frac{1}{2}\widehat{AB}$ interaction when C is at the high level - $\frac{1}{2}\widehat{AB}$ interaction when C is at the low level.
- Or, two other possible interpretation with swapped placed for A, B and C.
- And remember that $\widehat{AB} = \frac{1}{2}\widehat{A}$ main effect when B is at the high level - $\frac{1}{2}\widehat{A}$ main effect when B is at the low level.

Geometric interpretation of effects



2^k full factorial

- There are k factors: A, B, C, ..., and
- 2=each factor has two levels.
- There are 2^k possible experiments.
- We have in total 2^k parameters to be estimated:
  - 1 intercept
  - $k = \binom{k}{1}$ main effects: A, B, C, ...
  - $\binom{k}{2}$ two factor interactions: AB, AC, .., BC, BD,...
  - $\binom{k}{3}$ three factor interactions: ABC, ABD, ABE, ...
  - $\dots$
  - $\binom{k}{k} = 1$ k factor interaction.

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} \\
&+ \beta_{12} x_{12} + \cdots + \beta_{k-1,k} x_{k-1,k} \\
&+ \beta_{123} x_{123} + \cdots + \beta_{k-2,k-1,k} x_{k-2,k-1,k} \\
\cdots &+ \beta_{12\ldots k} x_{12\ldots k}
\end{aligned}
$$

TMA4267 L17
21.03.2017

Part 4: Design of Experiments (DOE) with 2^k factorial designs

Regression $Y = X\beta + \varepsilon$, $\varepsilon \sim N_n(0, \sigma^2 I)$

X: $n \times p$, intercept and k covariates

$\hat{\beta} = (X^T X)^{-1} X^T Y \sim N_p(\beta, \sigma^2 (X^T X)^{-1})$ + more!

And we used observational data.
Now: we design the experiment = choose X!

How should we choose X? Achieve some kind of optimality.

- minimize "$\mathrm{Var}(\hat{\beta})$" $= \mathrm{tr}(\sigma^2 (X^T X)^{-1})$
- minimize $\det(\mathrm{Cov}(\hat{\beta}))$

Our focus:

- maximize interpretability; e.g. by choosing X so that $\mathrm{Cov}(\hat{\beta}_j, \hat{\beta}_k) = 0 \iff (X^T X)$ is diagonal which we may achieve by choosing the columns of X to be orthogonal to each other.

We focus on 2^k factorial design

2: look at k factors each at 2 levels

Ex: Pilot plant

Y = yield

x_1 : A Temperature: 160, 180 $\xrightarrow{\text{recode}}$ −1, 1 z_1

x_2 : B Concentration: 20, 40 $\longrightarrow$ −1, 1 z_2

	A	B	AB	Y
1	−1	−1	1	y_1
2	1	−1	−1	y_2
3	−1	1	−1	y_3
4	1	1	1	y_4

A, B: standard order

AB: multiplying A and B

Observe that each factor column has $\sum_{i=1}^{n} x_{ij} = 0$,
and we also include an
intercept term with $\sum_{i=1}^{n} x_{i1} = n \Rightarrow \underset{4\times 4}{X}$

$\hat{\beta}$ and SSR will have simple formulas for this
full 2^2 design

full: do all combinations

2^k full factorial designs

Ex: Lima beans, $k=3$ factor at two levels.

All possible $2\cdot 2\cdot 2 = 2^3 = 8$ experiments performed

$$Y = \mathbb{X}\beta + \varepsilon, \quad \varepsilon \sim N_n(0, \sigma^2 I)$$

with $\mathbb{X}$ given as (8×8)

Intercept	A	B	C	AB	AC	BC	ABC
1	-1	-1	-1	1	1	1	-1
1	1	-1	-1	-1	-1	1	1
1	-1	1	-1	-1	1	-1	1
1	1	1	-1	1	-1	-1	-1
1	-1	-1	1	1	-1	-1	1
1	1	-1	1	-1	1	-1	-1
1	-1	1	1	-1	-1	1	-1
1	1	1	1	1	1	1	1

A, B, C: standard order

AB, AC, BC, ABC: by multiplying the relevant columns

Intercept, A, B, C: perform experiment

Hands-on:

1) Show that any two columns of $\mathbb{X}$ are orthogonal, $\sum_{i=1}^{n} x_{ij} x_{ik} = 0$

2) Show that $\sum_{i=1}^{n} x_{ij} = 0$ for all j except $\underset{\text{intercept column}}{j=0}$

3) And that $\sum_{i=1}^{n} x_{ij}^2 = n$.

Now: How does this (1+2+3) influence our formulas for

i) $\hat{\beta}$ ii) $\mathrm{Cov}(\hat{\beta})$ iii) $SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$

i) $\hat{\beta}$

$$\hat{\beta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T Y$$

$$(\mathbb{X}^T\mathbb{X}) = \begin{bmatrix} n & 0 & \cdots & 0 \\ 0 & \ddots & & \\ & & & n \end{bmatrix}$$

$$\{\mathbb{X}^T\mathbb{X}\}_{jk} = \sum_{i=1}^{n} x_{ij} x_{ik} = \begin{cases} 0 & j \neq k \\ n & j = k \end{cases}$$

$$\hat{\beta} = \begin{bmatrix} \frac{1}{n} & 0 & \cdots & 0 \\ 0 & \frac{1}{n} & & \\ \vdots & & \ddots & \\ 0 & & & \frac{1}{n} \end{bmatrix} \begin{bmatrix} \\ \mathbb{X} \\ \\ \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \underline{\underline{\Big\{ \frac{1}{n} \sum_{i=1}^{n} \overset{-1,1}{x_{ij}}\, y_i \Big\}_{\text{element } j}}}$$

4

$$\hat\beta_0 = \frac{1}{n}\sum_{i=1}^{n} 1\cdot y_i = \underline{\underline{\bar y}}$$

Observe that $\hat\beta_j$ is only dependent on x_{ij}, and not on x_{ik} $k \neq j$, so $\hat\beta_j$ will not change if we change the model. ← NEW now!

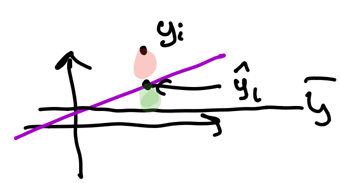
ii) $Cov(\hat\beta) = (X^TX)^{-1}\sigma^2$

$$= \begin{bmatrix} \frac{1}{n} & 0 \cdots & 0 \\ 0 & \frac{1}{n} & 0 \cdots \\ & 0 & & \frac{1}{n} \end{bmatrix} \sigma^2 \quad \text{so} \quad Var(\hat\beta_j) = \frac{1}{n}\sigma^2 \text{ for all } j = 0, \ldots, p$$

and $Cov(\hat\beta_j, \hat\beta_k) = 0$ for all $j \neq k$

↑

uncorrelated



iii) SST = SSE + SSR

$$\sum_{i=1}^{n}(y_i - \bar y)^2 = \sum_{i=1}^{n}(y_i - \hat y_i)^2 + \boxed{\sum_{i=1}^{n}(\hat y_i - \bar y)^2}$$

$\bar y = \hat\beta_0$ $\qquad$ $\hat y_i = \sum_{j=0}^{p-1} \hat\beta_j \cdot x_{ij}$

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n} \left(\sum_{j=0}^{p-1} \hat{\beta}_j \cdot x_{ij} - \hat{\beta}_0 \right)^2$$

$$= \sum_{i=1}^{n} \left(\hat{\beta}_0 + \sum_{j=1}^{p-1} \hat{\beta}_j x_{ij} - \hat{\beta}_0 \right)^2 = \sum_{i=1}^{n} \left(\sum_{j=1}^{p-1} \hat{\beta}_j x_{ij} \right)^2$$

$$(\hat{\beta}_1 \cdot x_{i1} + \hat{\beta}_2 x_{i2} + \cdots)(\hat{\beta}_1 \cdot x_{i1} + \hat{\beta}_2 \cdot x_{i2} + \cdots + \hat{\beta}_{p-1} x_{i,p-1})$$

$$= \sum_{i=1}^{n} \left(\hat{\beta}_1^2 x_{i1}^2 + \hat{\beta}_1 \hat{\beta}_2 x_{i1} \cdot x_{i2} + \cdots + \hat{\beta}_{p-1}^2 x_{i,p-1}^2 \right)$$

remember $\sum_{i=1}^{n} x_{ij} x_{ih} = 0$ for $j \neq h$

$\sum_{i=1}^{n} x_{ij}^2 = n$

so $\hat{\beta}_1 \cdot \hat{\beta}_2 \sum_{i=1}^{n} x_{i1} x_{i2} = 0$

etc.

$$= \sum_{i=1}^{n} \left(\hat{\beta}_1^2 x_{i1}^2 + \cdots + \hat{\beta}_{p-1}^2 x_{i,p-1}^2 \right) = n \cdot \sum_{j=1}^{p-1} \hat{\beta}_j^2$$

$$= \underbrace{n \cdot \hat{\beta}_1^2}_{\substack{SSR(x_1) \\ A}} + \underbrace{n \hat{\beta}_2^2}_{\substack{SSR(x_2) \\ B}} + \cdots + n \cdot \hat{\beta}_{p-1}^2$$

amount of variability due to each of the different covariates

6

TMA4267 Linear Statistical Models V2017 (L18)

Part 4: Design of Experiments
Full 2^k factorial designs
DOE Effects, estimating variability and performing inference
Compulsory DOE project

Mette Langaas

Department of Mathematical Sciences, NTNU

To be lectured: March 24, 2017

Last lesson - and today:

- Observational studies vs. designed experiments.
- Still linear regression, but now with k factors each with only 2 levels.
- Effect coding, orthogonal columns in design matrix.
- 2^k full factorial design.
- Simplified formulas for $\hat{\boldsymbol{\beta}}$, $\mathrm{Cov}(\hat{\boldsymbol{\beta}})$ and SSE.
- From parameter estimated to main and interaction effects.
- Inference.
- Compulsory exercise 4: the DOE project

Part 4 is based on Tyssedal: Design of experiments note.

Lima beans example

Experiment from Box, Hunter, Hunter, Statistics for Experimenters, page 321.

- A: depth of planting (0.5 inch or 1.5 inch)
- B: watering daily (once or twice)
- C: type of lima bean (baby or large)
- Y: yield

Research question: what is the combination of A, B, C giving the highest yield?

Design of experiments (DOE) terminology

- Variables are called factors, and denoted A, B, C, ...
- We will only look at factors with two levels:
  - high, coded as $+1$ or just $+$, and,
  - low, coded as -1 or just $-$.
- The lima beans example had three factors with two levels, thus $2^3 = 8$ possible combinations. In general k factors with two levels gives 2^k possible combinations.

Standard notation for 2^3 experiment (responses for lima beans included)

A	B	C	AB	AC	BC	ABC	Level code	Response
-	-	-	+	+	+	-	1	6
+	-	-	-	-	+	+	a	4
-	+	-	-	+	-	+	b	10
+	+	-	+	-	-	-	ab	7
-	-	+	+	-	-	+	c	4
+	-	+	-	+	-	-	ac	3
-	+	+	-	-	+	-	bc	8
+	+	+	+	+	+	+	abc	5
x_1	x_2	x_3	x_{12}	x_{13}	x_{23}	x_{123}		y

Results from last lecture: 2^k full factorial

Known from Part 2: $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ and $\text{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

- The design matrix is chosen so that the columns (containing -1 and 1) are orthogonal, and thus
  - $\sum_{i=1}^{n} x_{ij}x_{ik} = 0$ for all combinations of the columns of the design matrix $\boldsymbol{X}$.
  - $\sum_{i=1}^{n} x_{ij}^2 = n$.
- The orthogonal columns lead to that the following formulas are easy to interpret and calculate:
  - $\boldsymbol{X}^T\boldsymbol{X} =$ diagonal matrix with n on the diagonal.
  - $\hat{\boldsymbol{\beta}}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}Y_i$.
  - $\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{n}$.
  - $\text{Cov}(\hat{\beta}_j, \hat{\beta}_k) = 0$ for all $j \neq k$.
  - SSR$=\sum_{j=1}^{p-1} \hat{\beta}_j^2$.

See class notes for L17 for details on the derivation.

Lima beans example: full 2^3 factorial design

- A: depth of planting (0.5 inch or 1.5 inch)
- B: watering daily (once or twice)
- C: type of lima bean (baby or large)
- Y: yield

A	B	C	AB	AC	BC	ABC	Level code	Response
-	-	-	+	+	+	-	1	6
+	-	-	-	-	+	+	a	4
-	+	-	-	+	-	+	b	10
+	+	-	+	-	-	-	ab	7
-	-	+	+	-	-	+	c	4
+	-	+	-	+	-	-	ac	3
-	+	+	-	-	+	-	bc	8
+	+	+	+	+	+	+	abc	5
x_1	x_2	x_3	x_{12}	x_{13}	x_{23}	x_{123}		y

Write down the regression model with all possible interactions, and find $\hat{\beta}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} Y_i$ for the A and the AB columns.

Main effects in DOE

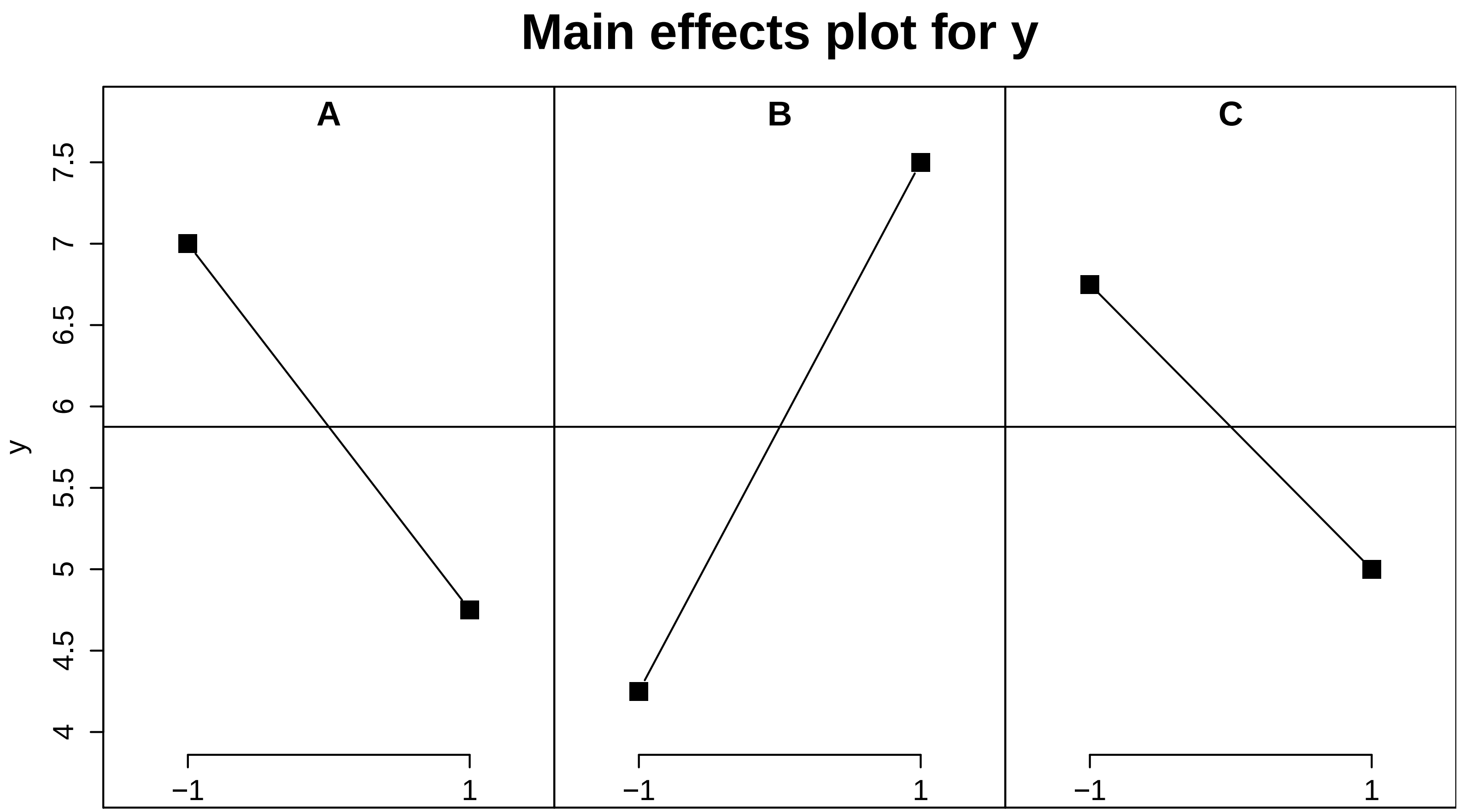
Main effect of A

$$\begin{aligned}\widehat{A} &= 2\hat{\beta}_1 \\ &= \frac{y_2 + y_4 + y_6 + y_8}{4} - \frac{y_1 + y_3 + y_5 + y_7}{4}\end{aligned}$$

Interpretation: mean response when A is high MINUS mean response when A is low.
Similarily, main effect of B

$$\begin{aligned}\widehat{B} &= 2\hat{\beta}_2 \\ &= \frac{y_3 + y_4 + y_7 + y_8}{4} - \frac{y_1 + y_2 + y_5 + y_6}{4}\end{aligned}$$

Interpretation: mean response when B is high MINUS mean response when B is low.




```
  A      B      C       A:B    A:C     B:C    A:B:C
-2.25  3.25  -1.75   -0.75   0.25    -0.25   -0.25
```

Explain the main effects in plain words!

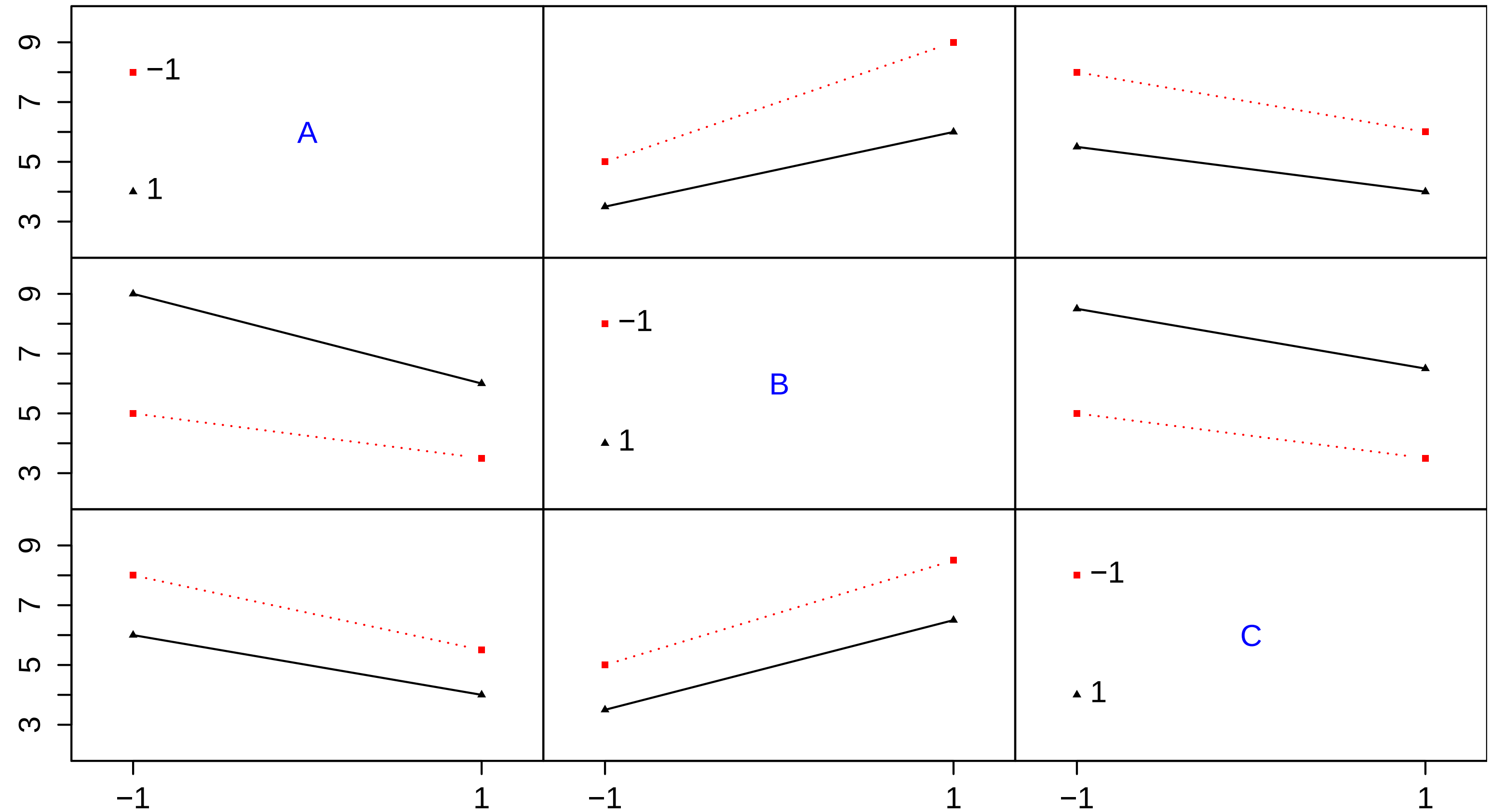
A: depth (0.5 or 1), B: watering daily (once, twice), C: type (baby, large).

Interaction effect in DOE

- What is the terpretation in DOE associated with β_{12}?
- In DOE $2\hat{\beta}_{12}$ is denoted $\widehat{AB}$ and is called the *estimated interaction effect between A and B*.

$$
\begin{aligned}
\widehat{AB} &= 2\hat{\beta}_{12} \\
&= \frac{\text{estimated main effect of } A \text{ when } B \text{ is high}}{2} \\
&- \frac{\text{estimated main effect of } A \text{ when } B \text{ is low}}{2} \\
&= \frac{\text{estimated main effect of } B \text{ when } A \text{ is high}}{2} \\
&- \frac{\text{estimated main effect of } B \text{ when } A \text{ is low}}{2}
\end{aligned}
$$

Interaction plot matrix for y




```
    A       B       C       A:B     A:C      B:C    A:B:C
-2.25    3.25   -1.75    -0.75    0.25    -0.25    -0.25
```

Interpretation of $\widehat{ABC}$

- $\widehat{ABC} = \frac{1}{2}\widehat{AB}$ interaction when C is at the high level - $\frac{1}{2}\widehat{AB}$ interaction when C is at the low level.
- Or, two other possible interpretation with swapped placed for A, B and C.
- And remember that $\widehat{AB} = \frac{1}{2}\widehat{A}$ main effect when B is at the high level - $\frac{1}{2}\widehat{A}$ main effect when B is at the low level.

R: DOE set-up for lima beans

```
> library(FrF2)
> plan <- FrF2(nruns=8,nfactors=3,randomize=FALSE)
creating full factorial with  8  runs ...
> plan
   A  B  C
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
class=design, type= full factorial
```

But, the experiment should be performed in *random order*. We use R to find the random order, and then we choose randomize=TRUE. I have used randomize=FALSE here because the y-values were easier to read in in standard order.

R: DOE add response

```
> y <- c(6,4,10,7,4,3,8,5)
> y
[1]  6  4 10  7  4  3  8  5
> plan <- add.response(plan,y)
> plan
   A  B  C  y
1 -1 -1 -1  6
2  1 -1 -1  4
3 -1  1 -1 10
4  1  1 -1  7
5 -1 -1  1  4
6  1 -1  1  3
7 -1  1  1  8
8  1  1  1  5
class=design, type= full factorial
```

R: DOE lm and effect

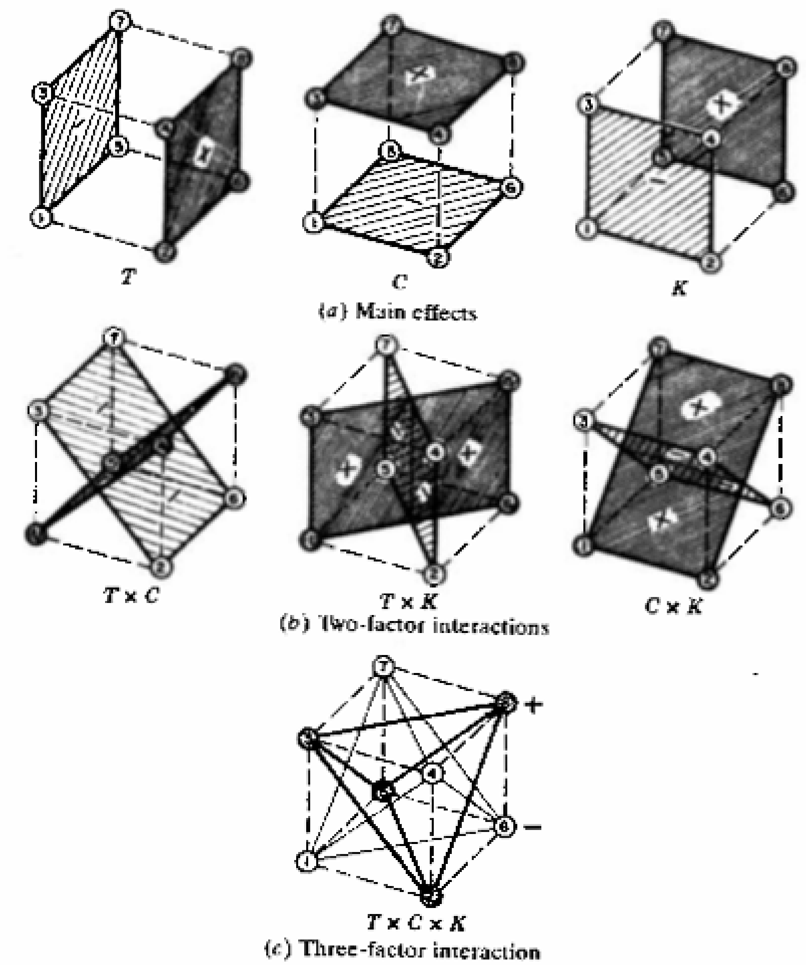
```
> lm3 <- lm(y~(.)^3,data=plan)
> MEPlot(lm3)
> IAPlot(lm3)
> effects <- 2*lm3$coeff
> effects
(Intercept) A1     B1     C1     A1:B1  A1:C1 B1:C1 A1:B1:C1
11.75       -2.25  3.25   -1.75 -0.75   0.25  -0.25 -0.25
```

2^k full factorial

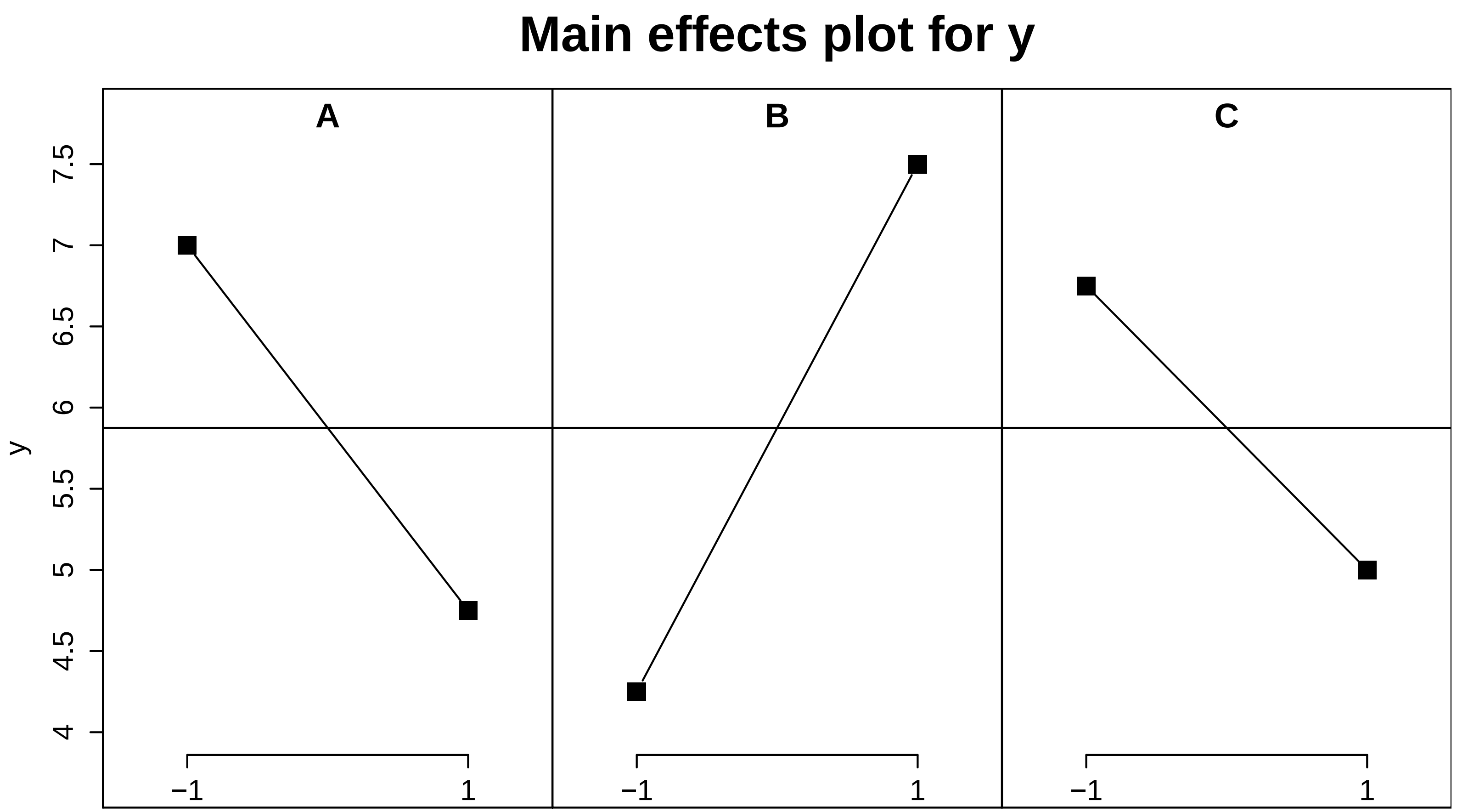
- There are k factors: A, B, C, ..., and
- 2=each factor has two levels.
- There are 2^k possible experiments.
- We have in total 2^k parameters to be estimated:
  - 1 intercept
  - $k = \binom{k}{1}$ main effects: A, B, C, ...
  - $\binom{k}{2}$ two factor interactions: AB, AC, .., BC, BD,...
  - $\binom{k}{3}$ three factor interactions: ABC, ABD, ABE, ...
  - $\cdots$
  - $\binom{k}{k} = 1$ k factor interaction.

$$\begin{aligned} Y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \\ &+ \beta_{12} x_{12} + \cdots + \beta_{k-1,k} x_{k-1,k} \\ &+ \beta_{123} x_{123} + \cdots + \beta_{k-2,k-1,k} x_{k-2,k-1,k} \\ \cdots & + \beta_{12\ldots k} x_{12\ldots k} + \varepsilon \end{aligned}$$

Geometric interpretation of effects



Lima beans: significant effects?




```
    A       B      C       A:B     A:C      B:C    A:B:C
-2.25   3.25  -1.75   -0.75    0.25    -0.25   -0.25
```

Lima beans: significant effects?

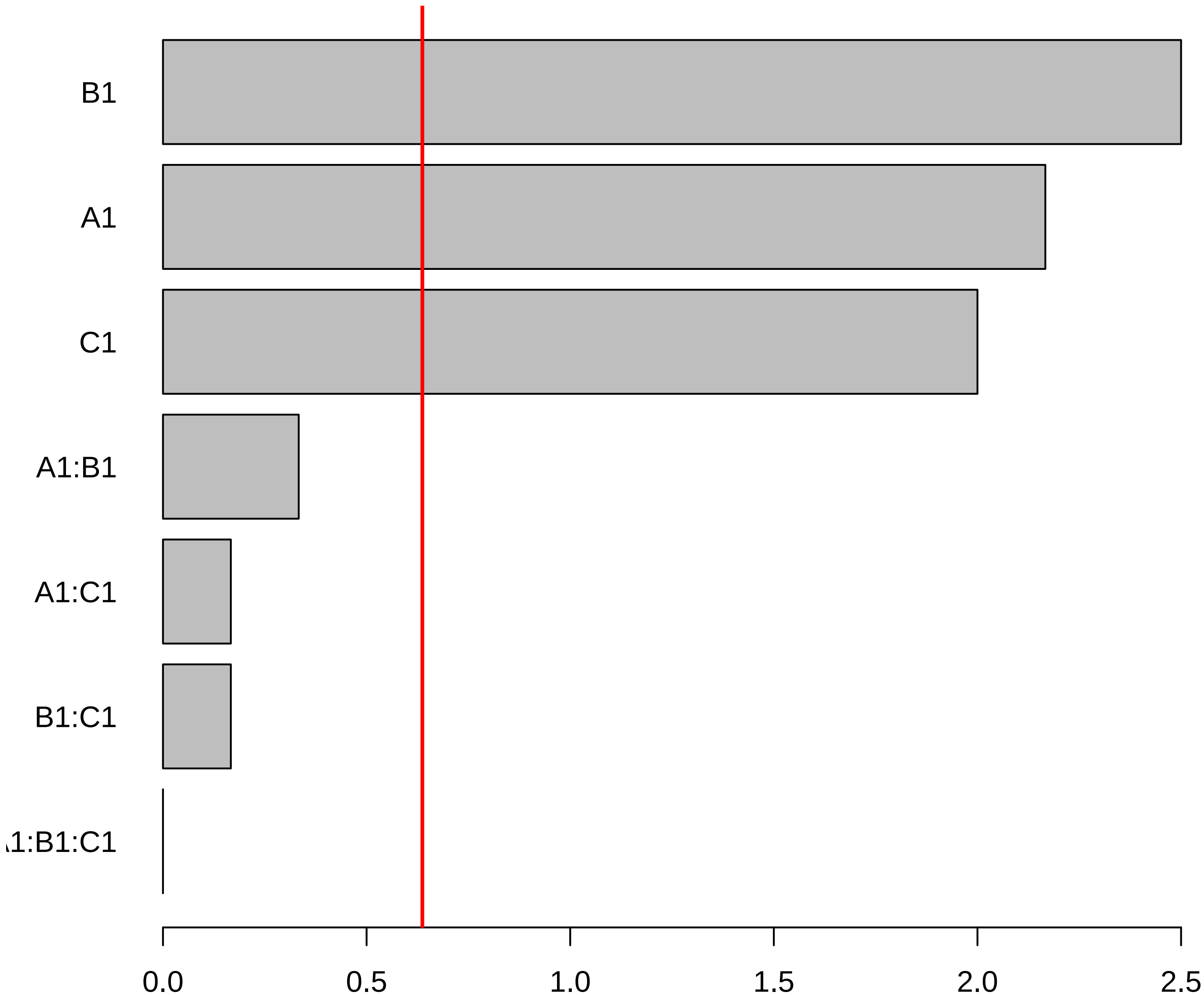
```
> summary(lm3)

Call:
lm.default(formula = y ~ (.)^3, data = plan)

Residuals:
ALL 8 residuals are 0: no residual degrees of freedom!

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.875         NA      NA       NA
A1            -1.125         NA      NA       NA
B1             1.625         NA      NA       NA
C1            -0.875         NA      NA       NA
A1:B1         -0.375         NA      NA       NA
A1:C1          0.125         NA      NA       NA
B1:C1         -0.125         NA      NA       NA
A1:B1:C1      -0.125         NA      NA       NA

Residual standard error: NaN on 0 degrees of freedom
Multiple R-squared:     1,Adjusted R-squared:   NaN
F-statistic:   NaN on 7 and 0 DF,  p-value: NA
```

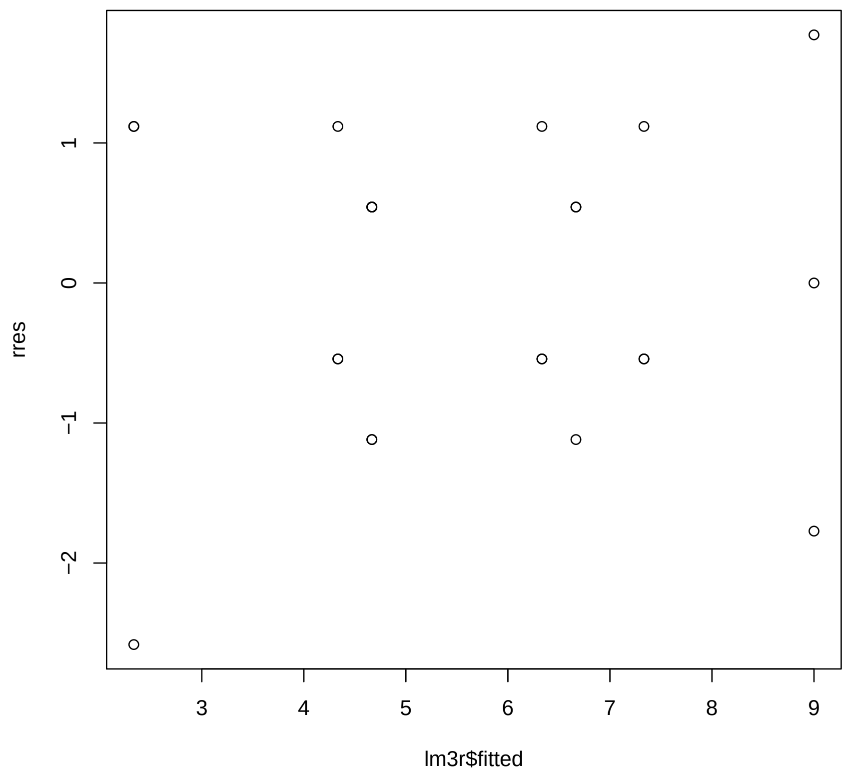
Estimation of σ^2

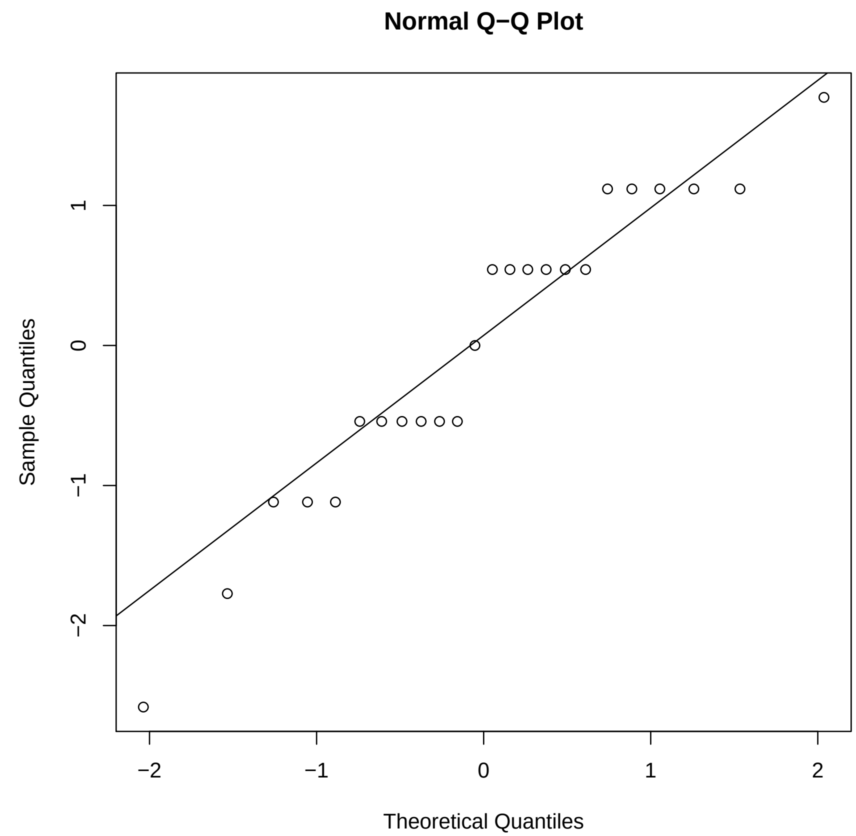
1. Perform replicates, estimate the full model and use s^2 from regression model.
2. Assuming specified higher order interactions are zero (changing the regression model).
3. If the two above is not possible: Lenth's Pseudo Standard Error (PSE).

Three factors in three full replicates

- Lima beans experiment from Box, Hunter, Hunter page 321.
  - A: depth of planting (0.5 inch or 1.5 inch)
  - B: watering daily (once or twice)
  - C: type of limabean (baby or large)
  - Y: yield
- $r = 3$: Performed in three full replicate experiments, i.e. three measurements for each combination of A, B and C.
- We then have $(r-1)2^3 = 2 \cdot 8 = 16$ degrees of freedom for estimating the error variance.
- Estimates follow automatically. Perform this for yourself. R code on course www-page.


B1
A1
C1
A1:B1
A1:C1
B1:C1
1:B1:C1
0.0
0.5
1.0
1.5
2.0
2.5


rres
lm3r$fitted
3
4
5
6
7
8
9
−2
−1
0
1


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−2
−1
0
1
2

ANOVA output: R

```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)    
A          1 28.167  28.167 52.0000 2.075e-06 ***
B          1 37.500  37.500 69.2308 3.319e-07 ***
C          1 24.000  24.000 44.3077 5.517e-06 ***
A:B        1  0.667   0.667  1.2308    0.2837    
A:C        1  0.167   0.167  0.3077    0.5868    
B:C        1  0.167   0.167  0.3077    0.5868    
A:B:C      1  0.000   0.000  0.0000    1.0000    
Residuals 16  8.667   0.542                      
```

Back to no extra replicates: Lima beans with only main effects

```
> lm1 <- lm(y~.,data=plan)
> summary(lm1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.8750     0.2165  27.135  1.1e-05 ***
A1           -1.1250     0.2165  -5.196  0.00653 **
B1            1.6250     0.2165   7.506  0.00169 **
C1           -0.8750     0.2165  -4.041  0.01559 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6124 on 4 degrees of freedom
Multiple R-squared: 0.9614,Adjusted R-squared: 0.9325
F-statistic: 33.22 on 3 and 4 DF,  p-value: 0.002755
> anova(lm1)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value   Pr(>F)
A          1 10.125  10.125  27.000 0.006533 **
B          1 21.125  21.125  56.333 0.001686 **
C          1  6.125   6.125  16.333 0.015585 *
Residuals  4  1.500   0.375
```

Back to no extra replicates: Assuming specified higher order interactions are zero
Result that is JUST a curiosity

- In general

$$\widehat{Effect}_j \sim N(Effect_j, \sigma^2_{effect})$$

- If we assume that the effect is zero ($\beta_j = 0$), then $\mathrm{E}(Effect_j) = 0$ and

$$\mathrm{E}(\widehat{Effect}_j^2) = \sigma^2_{effect}$$

- Thus $\widehat{Effect}_j^2$ is an unbiased estimator of σ^2_{effect} if $\beta_j = 0$.
- If several effects are assumed to be 0, we use the average of the $\widehat{Effect}_j^2$ to estimate σ^2_{effect}.

Lima beans estimated effects: full model

```
Estimated effects (2*coeff):
(Intercept)  A1    B1     C1    A1:B1   A1:C1   B1:C1  A1:B1:C1
  11.75    -2.25  3.25  -1.75   -0.75   0.25    -0.25   -0.25

Analysis of Variance Table
          Df Sum Sq Mean Sq F value Pr(>F)
A          1 10.125  10.125
B          1 21.125  21.125
C          1  6.125   6.125
A:B        1  1.125   1.125
A:C        1  0.125   0.125
B:C        1  0.125   0.125
A:B:C      1  0.125   0.125
Residuals  0  0.000
```

Lenth's PSE

Let $C_1, C_2, \ldots, C_m$ be estimated effects, e.g. $\hat{A}, \hat{B}, \widehat{AB}$, etc.

1. Order absolute values $|C_j|$ in increasing order.
2. Find the median of the $|C_j|$ and compute preliminary estimate

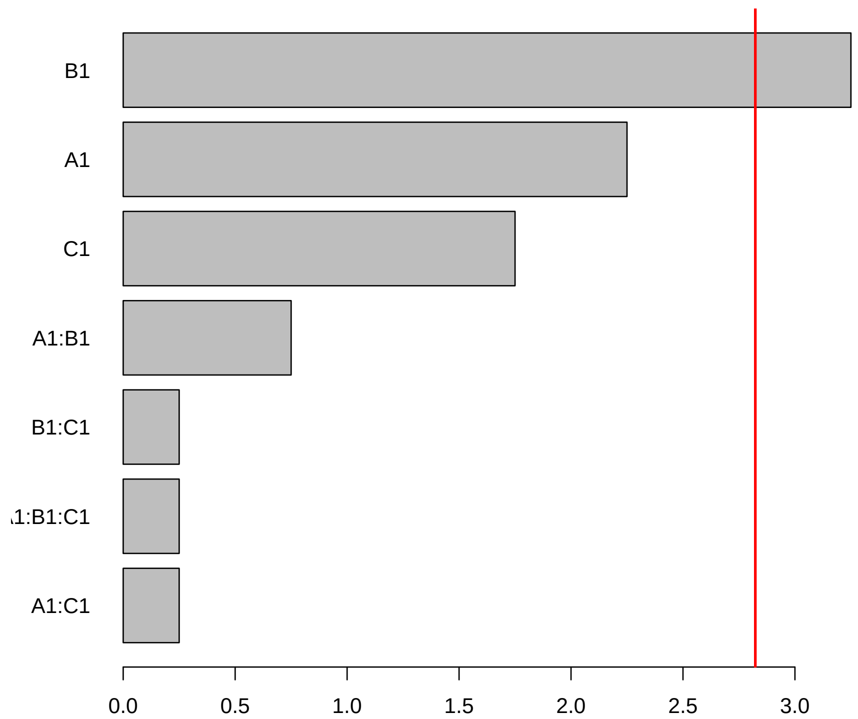
$$s_0 = 1.5 \cdot \text{median}_j |C_j|$$

3. Take out the effects C_j with $|C_j| \geq 2.5 \cdot s_0$ and find the median of the rest of the $|C_j|$. Then PSE is this median multiplied by 1.5, i.e.

$$\text{PSE} = 1.5 \cdot \text{median}\{|C_j| : |C_j| < 2.5 s_0\}$$

   and this is Lenth's estimate of σ_{effect}.
4. Lenth has suggested empirically that the degrees of freedom to be used with PSE is $m/3$ where m is the initial number of effects in the algorithm (intercept not included). Thus we claim as significant the effects for which $|C_j| > t_{\alpha/2, m/3} \cdot PSE$.

R: Pareto plot for Lima beans



Pareto plot: ordered histogram of absolute value of estimated effects, Length sign line added.

Which ν?

From the previous slide, connection between ν and your chosen estimation method for σ and σ_{effect}.

1. If you have performed the 2^k experiment r times, then $\nu = (r-1)2^k$.
2. If m effects (preferrable higher order interactions) are assumed to be zero, then $\nu = m$.
3. When Lenth's PSE is used, the degrees of freedom is

$$\nu = \frac{2^k - 1}{3}$$

where $2^k - 1$ is the number of effects in the model, while the 3 in the denominator has been found empirically by Lenth.

DOE workflow

1. Set up full factorial design with k factors in R, and
2. randomize the runs.
3. Perform experiments, and enter data into R.
4. Fit a full model (all interactions) - make Pareto-plot (with/without red line).
5. If you do not have replications, refit the data to a reduced model.
6. Assess model fit (residual plots, need transformations?).
7. Construct confidence intervals, assess significance.
8. Interpret you results (main and interaction plots).

Example compulsory project

“From a seed to a nice plant”

Factor	-	+
Seeds (A)	Broccoli Decicco	Sunflowers
Watering fluid (B)	Coffee	Water
Growth medium (C)	Soil	Cotton
Additional nutrients (D)	Without	With

Response: length of plant after 8 days of growing.

The experiments

StdOrder	RunOrder	CenterPt	Blocks	Seeds	Watering fluid	Growth medium	Additional nutrients	Length (response variable)
5	1	1	1	-1	-1	1	-1	0.1
2	2	1	1	1	-1	-1	-1	20.3
16	3	1	1	1	1	1	1	0.9
9	4	1	1	-1	-1	-1	1	0.2
15	5	1	1	-1	1	1	1	0.0
12	6	1	1	1	1	-1	1	6.9
6	7	1	1	1	-1	1	-1	1.1
1	8	1	1	-1	-1	-1	-1	11.7
10	9	1	1	1	-1	-1	1	5.9
13	10	1	1	-1	-1	1	1	0.0
4	11	1	1	1	1	-1	-1	23.3
8	12	1	1	1	1	1	-1	4.5
7	13	1	1	-1	1	1	-1	9.1
3	14	1	1	-1	1	-1	-1	12.2
14	15	1	1	1	-1	1	1	1.5
11	16	1	1	-1	1	-1	1	2.9

Full model

Estimated Effects and Coefficients for length (coded units)

Term	Effect	Coef
Constant		6,287
A	3,525	1,763
B	2,375	1,187
C	-8,275	-4,138
D	-8,000	-4,000
A*B	-0,675	-0,337
A*C	-3,825	-1,913
A*D	-0,500	-0,250
B*C	0,575	0,287
B*D	-1,600	-0,800
C*D	4,900	2,450
A*B*C	-0,875	-0,438
A*B*D	0,100	0,050
A*C*D	2,000	1,000
B*C*D	-1,650	-0,825
A*B*C*D	1,150	0,575

Full model

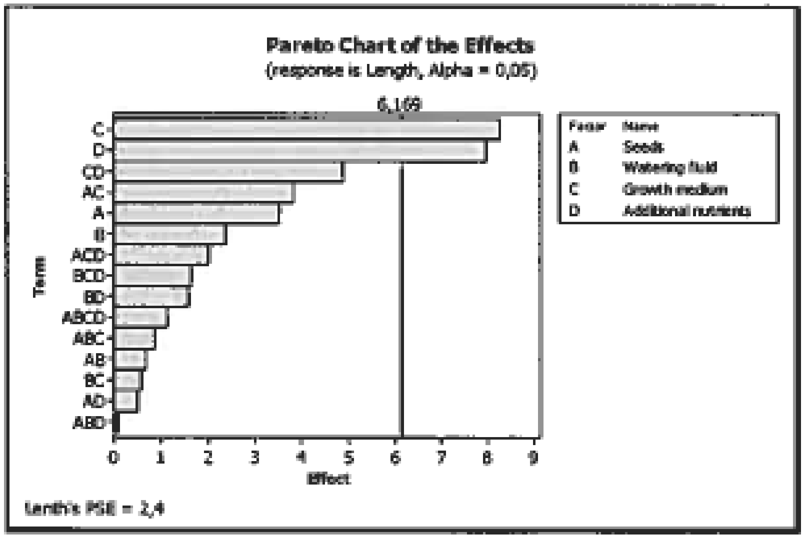


Figure 5.2 Pareto-chart of the effects with terms up to 4th order.

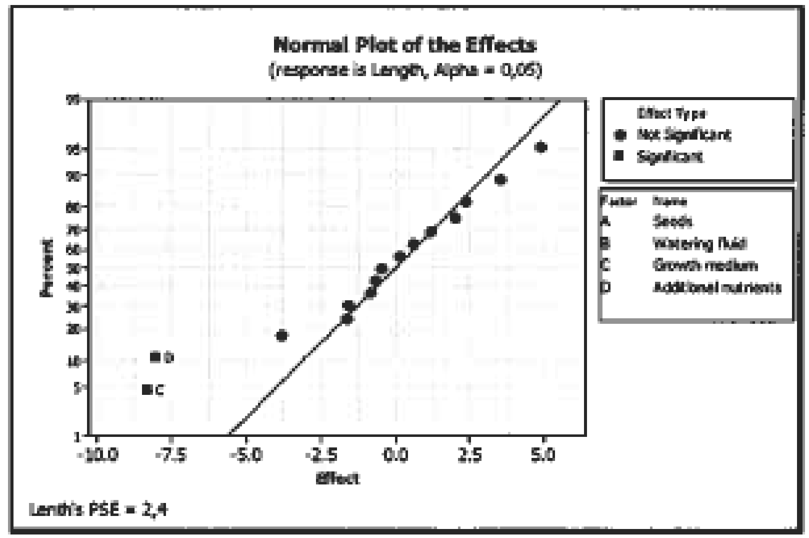


Figure 5.3 Normal plot of the effects with terms up to 4th order.

Inference

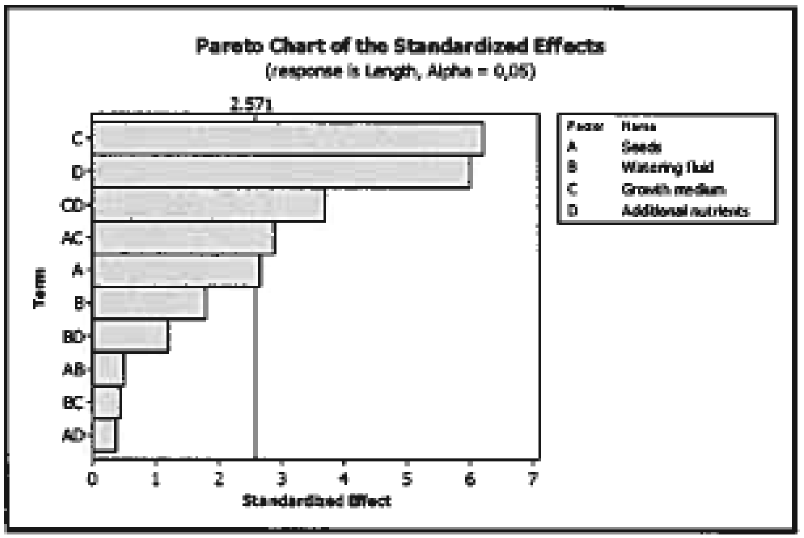


Figure 5.6 Pareto-chart of the effects with terms up to 2nd order.

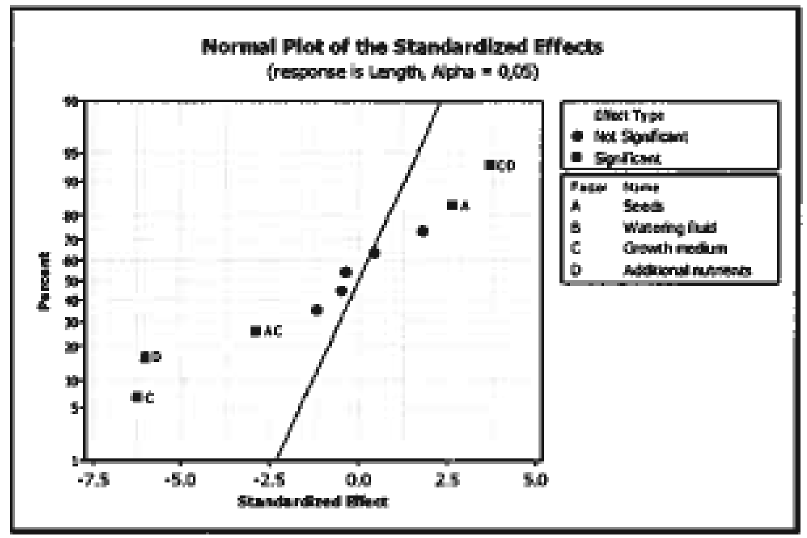


Figure 5.7 Normal plot of the effects with terms up to 2nd order.

A, C and D, AC and CD found to be significant.

Interpretation: Interaction plots

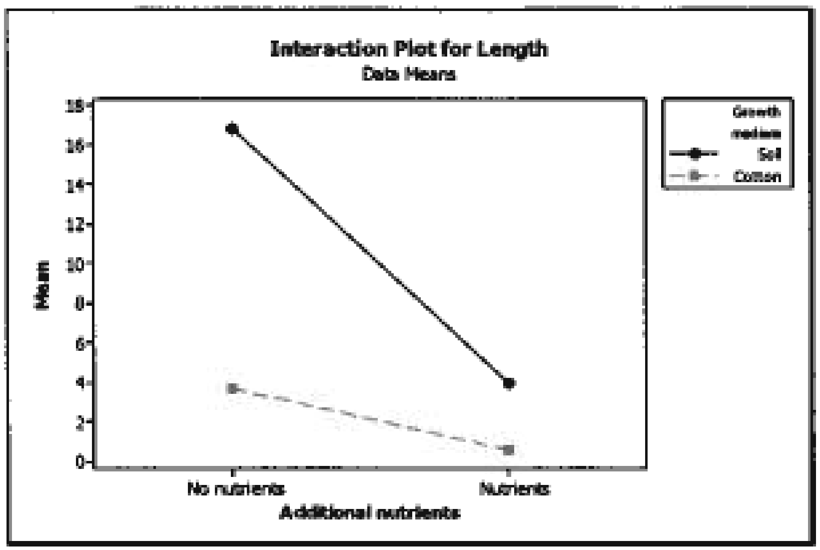


Figure 6.1 Interaction plot between growth medium and additional nutrients (CD).

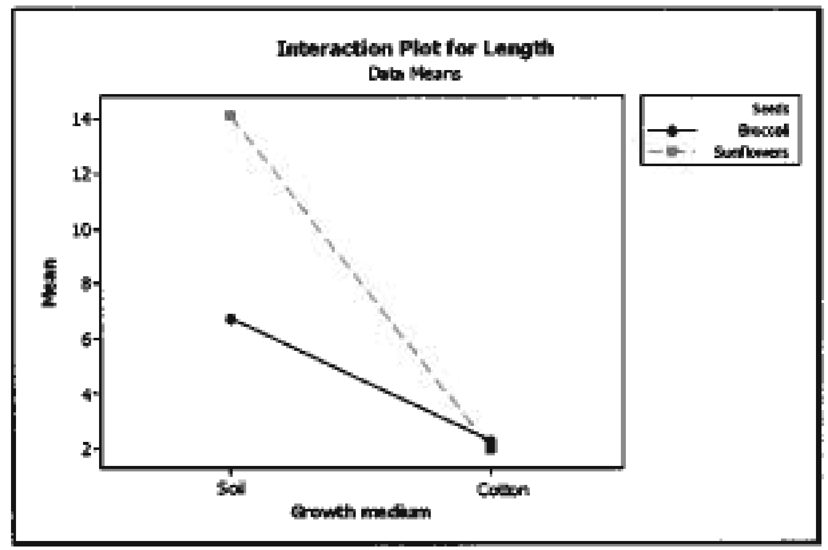


Figure 6.2 Interaction plot between seeds and growth medium (AC).

The practical issues (1)

- You may work alone, or in groups of two.
- You need to perform a multiple regression experiment consisting of 16 trials - that is, n=16 observations.
- The response that is measure should be continuous, so that the response itself or a transformation of the response in a regression model can be seen to be normally distributed. (It is also possible to assume that a response with at least 7 ordered categories can be seen as continuous.)
- You choose 3 or 4 factors with two levels each that might influence your response (it is possible to choose more factors, but then you need to do a so called fractional factorial design to be lectured soon).

The practical issues (2)

- If you choose 3 factors you need to perform all possible combinations of the 3 factors two times ($2 \cdot 2 \cdot 2 = 8$), if you choose 4 factors you need to perform all possible combinations only once ($2 \cdot 2 \cdot 2 \cdot 2 = 16$). If you choose more than 4 factors you need to study the "factional factorials" to find out which of the possible combinations you perform.
- A very important aspect of performing the 16 trials is that the trials should be independent and performed in a randomized order (why?). You use R to randomize the experiments for you.
- Each expermeent should be a complete new experiment - a genuine run replicate, unless you use blocking (not lectured yet). For example a block effect my be person or day.

Genuine run replicates

"When genuine run replicates are made under a given set of experimental conditions, the variation between the associated observations may be used to estimate the standard deviation of the effects. By *genuine* run replicated we mean that variation between runs made at the same experimental conditions is a reflection of the total variability afflicting runs made at different experimental conditions. This point requires careful consideration."
From Box, Hunter, Hunter (1978, 2005): "Statistics for Experimenters", Ch.10.6.

Genuine run replicates

Randomization of run order usually ensures that replicates are genuine. Pilot plant example: each run consists of

1. cleaning the reactor
2. inserting the appropriate catalyst carge
3. running the apparatus at at given temperature and a given feed concentration for 3 hrs to allow the process to settle down at the chosen experimental conditions, and
4. combining chemical analyses made on these samples.

A genuine run replicate must involve the taking of all these steps again. In particular, several chemical analyses from a single run would provide only an estimate of *analytical* variance, usually only a small part of the run-to-run variance.

From Box, Hunter, Hunter (1978, 2005): "Statistics for Experimenters", Ch.10.6.

The practical issues (3)

- After you have performed all 16 experiments you need to record the response and enter it into the experiment you have designed in R.
- Then you analyze the data, estimate effects, perform inference, check the model assumptions (RESIDUALS!), and explain your findings.

The report (1)

1. Describe the problem you want to study. Why is this interesting? What prior knowledge do you have? What do you want to achieve?
2. Selection of factors and levels: Which factors do you think are relevant to the problem described above? Which of these factors do you think is active/inert? Do you expect an interaction between some of the factors? Which levels should be used, and why do you think these are reasonable? How can you control that the factors really are at the desired level?
3. Selection of response variable: Which response variable will provide information about the problem described above? Are there several response variables of interest? How should the response be measured? What can you say about the accuracy of these measurements?

The report (2)

4. Choice of design: 2 k factorial, 2 k-p fractional factorial (resolution?)? Is it necessary or desirable to use a blocked design? Is it necessary or desirable with replicates?
5. Implementation of the experiment: Randomization. Describe any problems with the implementation.
6. Analysis of data: Calculation of effects and assessment of statistical significance. Use Lenth (not only), replicates or "setting some interactions to zero" to perform inference? Check the assumptions. RESIDUAL PLOTS!
7. Conclusion (explain main and interaction plots) and recommendations: Which conclusions can you draw from the experiment?

To get 10 points you need to have addressed all of these aspects in a correct manner! BUT - don't hand in more than 8 pages (included printout from R and plots)!

I don't want to collect data!

- Well, it is possible to instead analyse a observational data set (but talk to the lecturer first),
- or to perform a simulation experiment to investigate properties of the regression model.

Supervision?

- See course page - several possibilities until deadline for hand-in on Tuesday May 2.

Ex: Lime beans 2^3.

* = corrected on page 5

Write down the regression model with all possible interactions.

$$Y_i = \beta_0 + \underset{A}{\beta_1} \cdot x_{i1} + \underset{B}{\beta_2} \cdot x_{i2} + \underset{C}{\beta_3} \cdot x_{i3}$$

$$+ \underset{AB}{\beta_{12}}\, x_{i1} \cdot x_{i2} + \underset{AC}{\beta_{13}}\, x_{i1} \cdot x_{i3} + \underset{BC}{\beta_{23}}\, x_{i2} \cdot x_{i3}$$

$$+ \beta_{123}\, x_{i1} \cdot x_{i2} \cdot x_{i3} + \varepsilon_i$$

$$\hat{\beta}_1 = \frac{1}{8} \sum_{i=1}^{8} x_{i1} \cdot Y_i = \frac{1}{8}\left(-1 \cdot 6 + 1 \cdot 4 - 1 \cdot 10 + \dots + 1 \cdot 5\right)$$
$$= -1.125$$

$$\hat{\beta}_1 = \frac{1}{2} \underbrace{\frac{y_2 + y_4 + y_6 + y_8}{4}}_{\text{average of response when A is high}} - \frac{1}{2} \underbrace{\frac{y_1 + y_3 + y_5 + y_7}{4}}_{\text{average of response when A is low}}$$

Interpret $\hat{\beta}_1$: increase x_1 with one unit $\Rightarrow$ $\hat{y}$ increase with $\hat{\beta}_1$.

DOE Effects

For each β_j in the model (except β_0) we define an effect to be

$$Effect_j = 2 \cdot \beta_j$$

Why? β_j gives the change (in y) when x_{ij} goes from 0 to 1, while $Effect_j$ gives the change when x_{ij} goes from -1 to 1.

Thus: $\boxed{\hat{Effect}_j = 2 \cdot \hat{\beta}_j}$

This (unfortunately) means that

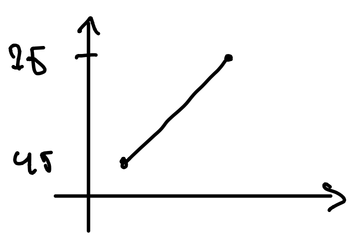
$$Var(\hat{Effect}_j) = Var(2 \cdot \hat{\beta}_j) = 4 \cdot \overbrace{Var(\hat{\beta}_j)}^{\frac{\sigma^2}{n}} = \underline{\underline{\frac{4\sigma^2}{n}}}$$

insert $\hat{\sigma}^2 \Rightarrow s^2_{Effect_j}$

WARNING:

DOE main effect: $2\hat{\beta}_j$ for A, B, C

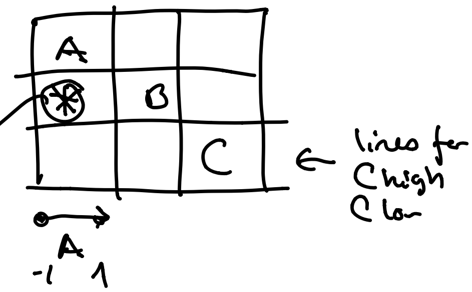
shown in main effects plot



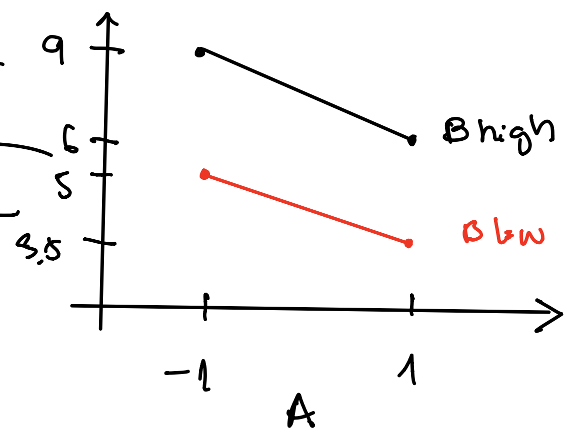
$\hat{B} = 3.25$

Interaction effect

Ex: $\hat{AB} = 2\hat{\beta}_{12}$



	Alow	Ahigh
Blow	5	3.5
Bhigh	9	6



$\hat{AB} = -0.75$

$$\hat{AB} = \frac{1}{2}(\text{est. main effect of } A \text{ when } B \text{ is high})$$

$$- \frac{1}{2}(\text{— " —————— low})$$

$$= \frac{1}{2}(6-9) - \frac{1}{2}(3.5-5) = -0.75$$

rather small effect

When the two lines in the interaction plot are parallel ⇒ there is <u>no</u> interaction effect.

⇒ DOE report (Ex 4) ← add explanation for one interaction effect.

Significant effects

$$H_0: \underset{2\beta_j}{\text{Effect}_j} = 0 \quad \text{vs} \quad H_1: \text{Effect}_j \neq 0$$

(or equivalently: $H_0: \beta_j = 0$ vs $H_1: \beta_j \neq 0$)

$$\text{Effect}_j = 2\beta_j$$

$$\widehat{\text{Effect}}_j = 2\hat{\beta}_j = \frac{2}{n}\sum_{i=1}^{n} x_{ij} \cdot y_i$$

$$E(\widehat{\text{Effect}}_j) = 2 \cdot E(\hat{\beta}_j) = 2\beta_j = \text{Effect}_j$$

$$Var(\widehat{\text{Effect}}_j) = 4 \cdot Var(\hat{\beta}_j) = 4 \cdot \frac{1}{n}\sigma^2 \equiv \sigma^2_{\text{effect}}$$

NBNB not dependent on j

$$\widehat{\text{Effect}}_j \sim N(\text{Effect}_j, \sigma^2_{\text{effect}})$$

If we have S^2_{effect} as an estimator for σ^2_{effect}, we might get

$$T_j = \frac{\widehat{\text{Effect}}_j - \text{Effect}_j}{S_{\text{effect}}} \sim t_\nu$$

ν is dependent on S^2_{effect}

95% CI: $\left[\widehat{\text{Effect}}_j \pm t_{\frac{\alpha}{2},\nu} \cdot S_{\text{effect}} \right]$

Hypothesis test: reject H_0 when

$$|t_j| = \left| \frac{\widehat{\text{Effect}}_j - 0}{s_{\text{effect}}} \right| > t_{\frac{\alpha}{2},\nu}$$

numerical value

$$|\widehat{\text{Effect}}_j| > t_{\frac{\alpha}{2},\nu} \cdot S_{\text{effect}}$$

---

1) Perform replication of a full 2^4 design. $\rightarrow$ use lm as before.

Pareto-plot: barplot (horizontal) of $\widehat{\text{Effect}}_j$ with red line at $\underline{t_{\frac{\alpha}{2},\nu} \cdot s_{\text{effect}}}$

Ex
Lima beans: 3 replicates of 8 observations $\Rightarrow \underline{n=24}$
Estimating 8 parameters (β_0 + A, B, C, AB, AC, BC, ABC)
$\Rightarrow n-p = 24-8 = 16 \leftarrow \underline{\nu = 16}$

Residual standard error in printout

$$s_{\text{effect}} = \sqrt{\frac{4}{n} \cdot \hat{\sigma}^2} = \sqrt{\frac{4}{24}} \cdot \underset{\underset{0.736}{\|}}{S} = 0.3$$

$$t_{\frac{0.05}{2},16} = 2.12$$

0.64
‖
red line
‖
$2.12 \cdot 0.3 = t_{\frac{\alpha}{2},\nu} \cdot s_{\text{effect}}$

2) Fit reduced model → read off:

Curiosity: S_{effect} can be calculated from the full model by $S_{effect}^2 = \frac{1}{m} \underbrace{\sum}_{\text{all Effects not part of model}} \text{Effects}^2$

3) Lenth's method.

TMA4267 Linear Statistical Models V2017 (L19)

Part 4: Design of Experiments
Blocking
Fractional factorial designs

Mette Langaas

Department of Mathematical Sciences, NTNU

To be lectured: March 28, 2017

DOE workflow

1. Set up full factorial design with k factors in R, and
2. randomize the runs.
3. Perform experiments, and enter data into R.
4. Fit a full model (all interactions).
5. If you do not have replications, look at Pareto plots and, use this to suggest at reduced model (if possible). Refit the reduced model.
6. Assess model fit (residual plots, need transformations?).
7. Assess significance.
8. Interpret you results (main and interaction plots).

Q: Randomization

Why do you need to randomize the order in which you perform the experiments?

To make the experiments

- A: random.
- B: robust to external factors.
- C: have constant variance.
- D: independent.

Vote at clicker.math.ntnu.no, TMA4267 classroom.

Genuine run replicates

"When genuine run replicates are made under a given set of experimental conditions, the variation between the associated observations may be used to estimate the standard deviation of the effects. By *genuine* run replicated we mean that variation between runs made at the same experimental conditions is a reflection of the total variability afflicting runs made at different experimental conditions. This point requires careful consideration."

From Box, Hunter, Hunter (1978, 2005): "Statistics for Experimenters", Ch.10.6.

Genuine run replicates

Randomization of run order usually ensures that replicates are genuine. Pilot plant example: each run consists of

1. cleaning the reactor
2. inserting the appropriate catalyst charge
3. running the apparatus at at given temperature and a given feed concentration for 3 hrs to allow the process to settle down at the chosen experimental conditions, and
4. combining chemical analyses made on these samples.

A genuine run replicate must involve the taking of all these steps again. In particular, several chemical analyses from a single run would provide only an estimate of *analytical* variance, usually only a small part of the run-to-run variance.

From Box, Hunter, Hunter (1978, 2005): "Statistics for Experimenters", Ch.10.6.

Pilot plant: A, B and C

A=Temperature, B=Concentration, C=Catalyst, Y=yield.

A	B	C	AB	AC	BC	ABC	Level code	Response
-	-	-	+	+	+	-	1	60
+	-	-	-	-	+	+	a	72
-	+	-	-	+	-	+	b	54
+	+	-	+	-	-	-	ab	68
-	-	+	+	-	-	+	c	52
+	-	+	-	+	-	-	ac	83
-	+	+	-	-	+	-	bc	45
+	+	+	+	+	+	+	abc	80
x_1	x_2	x_3	x_{12}	x_{13}	x_{23}	x_{123}		y

Blocking on ABC

Block 1 consists of experiments with ABC=-1.
Block 2 consists of experiments with ABC=1.

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
StdOrder	RunOrder	CenterPt	Blocks	A	B	C	ABC		Y	block effect
1	1	1	1	-1	-1	-1	-1	1	60	60
4	4	1	1	-1	1	1	-1	7	45	45
3	3	1	1	1	-1	1	-1	6	83	83
2	2	1	1	1	1	-1	-1	4	68	68
7	7	1	2	-1	-1	1	1	5	52	62
6	6	1	2	-1	1	-1	1	3	54	64
5	5	1	2	1	-1	-1	1	2	72	82
8	8	1	2	1	1	1	1	8	80	90

Blocking on ABC

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
StdOrder	RunOrder	CenterPt	Blocks	A	B	C	ABC		Y	block effect
1	1	1	1	-1	-1	-1	-1	1	60	60
4	4	1	1	-1	1	1	-1	7	45	45
3	3	1	1	1	-1	1	-1	6	83	83
2	2	1	1	1	1	-1	-1	4	68	68
7	7	1	2	-1	-1	1	1	5	52	62
6	6	1	2	-1	1	-1	1	3	54	64
5	5	1	2	1	-1	-1	1	2	72	82
8	8	1	2	1	1	1	1	8	80	90

- ABC is counfunded with the block effect. We can not separate these two effects from eachother.
- Suppose all values in block 2 is increased by 10 units.
  - Then the estimated effect of ABC will increase by 10.
  - But all other estimated effects remain unchanged - and these are the most important to estimate.

Original data			Added 10 to all obs in Block 2.		
Factorial Fit:			Factorial Fit:		
Y versus			"block effect" versus		
Block A B C			Block A B C		
Term	Effect	Coef	Term	Effect	Coef
Constant		64,250	Constant		69,250
Block		-0,250	Block		-5,250
A	23,000	11,500	A	23,000	11,500
B	-5,000	-2,500	B	-5,000	-2,500
C	1,500	0,750	C	1,500	0,750
A*B	1,500	0,750	A*B	1,500	0,750
A*C	10,000	5,000	A*C	10,000	5,000
B*C	0,000	0,000	B*C	0,000	0,000

2^3 with four blocks

We need two generators (columns) to define four blocks: the optimal choice is AB and AC

- Block 1: AB=AC=-1 (- -)
- Block 2: AB=-1, AC=1 (- +)
- Block 3: AB=1, AC=-1 (+ -)
- Block 4: AB=AC=1 (+ +)

Std order	A	B	C	AB	AC	BC	ABC
1	-	-	-	+	+	+	-
2	+	-	-	-	-	+	+
3	-	+	-	-	+	-	+
4	+	+	-	+	-	-	-
5	-	-	+	+	-	-	+
6	+	-	+	-	+	-	-
7	-	+	+	-	-	+	-
8	+	+	+	+	+	+	+

2^3 with AB and AC as generators

Std order	A	B	C	AB	AC	BC	ABC	Block
2	+	-	-	-	-	+	+	1
7	-	+	+	-	-	+	-	1
3	-	+	-	-	+	-	+	2
6	+	-	+	-	+	-	-	2
4	+	+	-	+	-	-	-	3
5	-	-	+	+	-	-	+	3
1	-	-	-	+	+	+	-	4
8	+	+	+	+	+	+	+	4

2^3 with AB and AC as generators

- Interaction effects AB and AC are confounded with the block effect, since they are the generators.
- Their product, $AB * AC = A^2BC = BC$, is alco confounded with the block effect (see that BC is constant within each block).
- Adding h_2 to block 2, h_3 to block 3 and h_4 to block 4 does not change the estimated main effects A, B, or C, and not the interaction effect ABC.
- However, AB will change with $2 \cdot h_3 + 2 \cdot h_4 - 2 \cdot h_2$, and we will NOT be able to separate the true AB effect from the block effect.

How to choose which blocks to be used for blocking?

- Idea: try to leave estimates for main effects and low order interaction unchanged by the blocking.
- Note: I=AA=BB=CC, where I is a column of 1's.
- How NOT to do this:
  - Find the blocks for a 2^3 experiment using generators ABC and AC.
  - The interaction between ABC and AC is ABC*AC=B.
  - This means chosing ABC and AC is not a good idea since then we can not trust our estimate of B.

Questions

Should you use a blocking factor in your compulsory project?
Do you understand the difference between blocking and repetition?

Box, Hunter, Hunter: Reactor example

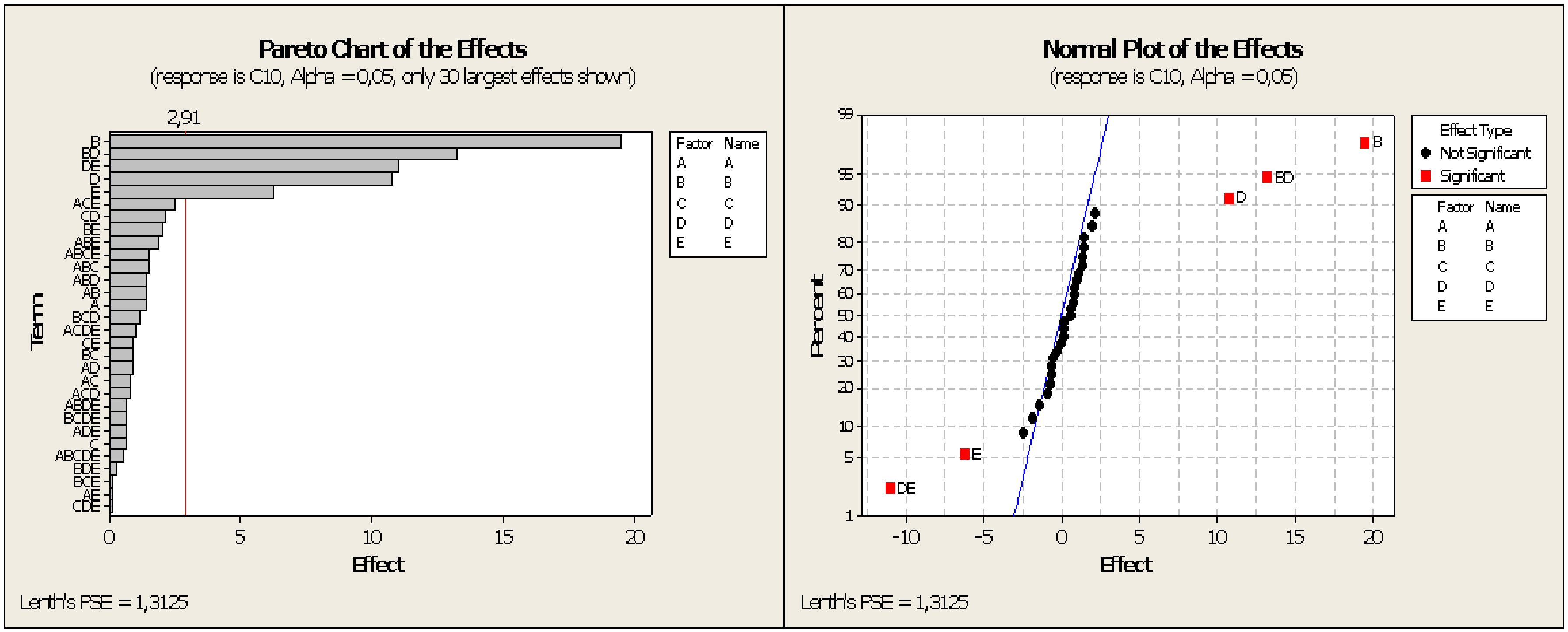
- A=feed rate (liters/min).
- B=Catalyst (%).
- C=Agitation rate (rpm).
- D=Temperature (deg C).
- E=Concentration (%).
- Response= (%) reacted.

Full factorial with $2^5 = 32$ experiments.
From Box, Hunter, Hunter (1978, 2005): "Statistics for Experimenters", Ch.12.2.

Reactor data: standard order

	A	B	C	D	E	y
1	-1	-1	-1	-1	-1	61
2	1	-1	-1	-1	-1	53
3	-1	1	-1	-1	-1	63
4	1	1	-1	-1	-1	61
5	-1	-1	1	-1	-1	53
6	1	-1	1	-1	-1	56
7	-1	1	1	-1	-1	54
8	1	1	1	-1	-1	61
9	-1	-1	-1	1	-1	69
10	1	-1	-1	1	-1	61
11	-1	1	-1	1	-1	94
12	1	1	-1	1	-1	93
13	-1	-1	1	1	-1	66
14	1	-1	1	1	-1	60
15	-1	1	1	1	-1	95
16	1	1	1	1	-1	98
17	-1	-1	-1	-1	1	56
18	1	-1	-1	-1	1	63
19	-1	1	-1	-1	1	70
20	1	1	-1	-1	1	65
21	-1	-1	1	-1	1	59
22	1	-1	1	-1	1	55
23	-1	1	1	-1	1	67
24	1	1	1	-1	1	65
25	-1	-1	-1	1	1	44
26	1	-1	-1	1	1	45
27	-1	1	-1	1	1	78
28	1	1	-1	1	1	77
29	-1	-1	1	1	1	49
30	1	-1	1	1	1	42
31	-1	1	1	1	1	81
32	1	1	1	1	1	82

Pareto and Normal plot



Redundancy

- The number of runs in a full 2^k factorial design increases geometrically when k is increased.
- E.g. $k = 7$ factors gives $2^7 = 128$ runs and we can estimate
  - $\binom{7}{1} = 7$ main effects
  - $\binom{7}{2} = 21$ 2nd order interactions
  - $\binom{7}{3} = 35$ 3rd order interactions
  - $\binom{7}{4} = 35$ 4th order interactions
  - $\binom{7}{5} = 21$ 5th order interactions
  - $\binom{7}{6} = 7$ 6th order interactions
  - $\binom{7}{7} = 1$ 7th order interactions

Redundancy (cont.)

- There is a hierarchy in absolute magnitude: the main effects tend to be larger than the 2nd order interactions, which tends to be larger than the 3rd order interactions, which ...
- At some point higher order interactions tend to become negligible and can be discarded.
- If many factors are introduced into a design, it often happens that some have *no* distinguishable effect at all.
- *Fractional factorial designs* exploit this redundancy!

Full 2^3 factorial experiment

How can we accomodate four factors here?

Std order	A	B	C	AB	AC	BC	ABC
1	-	-	-	+	+	+	-
2	+	-	-	-	-	+	+
3	-	+	-	-	+	-	+
4	+	+	-	+	-	-	-
5	-	-	+	+	-	-	+
6	+	-	+	-	+	-	-
7	-	+	+	-	-	+	-
8	+	+	+	+	+	+	+

Full 2^3 factorial experiment - turned into 4-factor experiment

Which effects are confounded?

	A	B	C	AB	AC	BC	D=ABC	ABD	ACD	BCD	ABCD
1	-	-	-	+	+	+	-	-	-	-	+
2	+	-	-	-	-	+	+	-	-	+	+
3	-	+	-	-	+	-	+	-	+	-	+
4	+	+	-	+	-	-	-	-	+	+	+
5	-	-	+	+	-	-	+	+	-	-	+
6	+	-	+	-	+	-	-	+	-	+	+
7	-	+	+	-	-	+	-	+	+	-	+
8	+	+	+	+	+	+	+	+	+	+	+

Half fraction of 2^4

- The design is called 2^{4-1}_{IV}.
- D=ABC is called the *generator* for the design.
- I=ABCD is called the *defining relation* for the design.
- The design is said to have *resolution IV*.
- The *alias structure* defines which effects are confounded:
  - A+BCD, B+ACD, C+ABD, D+ABC.
  - AB+CD, AC+BD, BC+AD.

What did we learn today?

- Why may experiments need to be performed in blocks? (Batches of raw material, performed on different days, different people performing the experiments.)

What did we learn today?

- Why may experiments need to be performed in blocks? (Batches of raw material, performed on different days, different people performing the experiments.)
- Should we also add a "block" effect if we perform repeated experiments? (Sometimes. If done by different people, or external factors have changed.)

What did we learn today?

- Why may experiments need to be performed in blocks? (Batches of raw material, performed on different days, different people performing the experiments.)
- Should we also add a "block" effect if we perform repeated experiments? (Sometimes. If done by different people, or external factors have changed.)
- Should then the block effect be a part of the regression model? (In most cases: yes!)

What did we learn today?

- Why may experiments need to be performed in blocks? (Batches of raw material, performed on different days, different people performing the experiments.)
- Should we also add a "block" effect if we perform repeated experiments? (Sometimes. If done by different people, or external factors have changed.)
- Should then the block effect be a part of the regression model? (In most cases: yes!)
- Why don't we want to perform a full factorial experiment, but a instead a fractional factorial? (If we have many factors we maybe not need to be able to estimate all possible interactions, and may accept that effects are confounded.)

What did we learn today?

- What is the easiest way to design a half-fraction of a 2^k factorial experiment? (Perform all the experiments where the highest order interaction =-1 or +1. E.g. for k=4 we may do 16 different experiments, and now we only do the 8 possible experiments where ABCD=+1=defining relation. This is the same as thinking that D=ABC=generator).

What did we learn today?

- What is the easiest way to design a half-fraction of a 2^k factorial experiment? (Perform all the experiments where the highest order interaction =-1 or +1. E.g. for k=4 we may do 16 different experiments, and now we only do the 8 possible experiments where ABCD=+1=defining relation. This is the same as thinking that D=ABC=generator).
- New words: *generator(s)*, *defining relation(s)*, *resolution*.

What did we learn today?

- What is the easiest way to design a half-fraction of a 2^k factorial experiment? (Perform all the experiments where the highest order interaction =-1 or +1. E.g. for k=4 we may do 16 different experiments, and now we only do the 8 possible experiments where ABCD=+1=defining relation. This is the same as thinking that D=ABC=generator).
- New words: *generator(s)*, *defining relation(s)*, *resolution*.
- Next time: more on interpreting "confounding", interpreting "resolution" and more fractional factorial experiments

Part 4: DoE

Performing a full 2^k factorial expr.

Two important aspects:

a) The run order is random, so that potential external factors are not confused/confounded with experimental factors.

b) Each experiment is a genuine run replicate, that is, reflects the total variability of the experiment.

Blocking

We will perform a 2^3 experiment, but have to use two batches of raw material ⇒ need to divide the 8 runs into two groups. What is the best way to do this?

Solution: use the ABC column to define the blocks, ABC is the block generator.

ABC = -1 : use batch 1
ABC = +1 : 2

The block "variable" will be a new regressor replacing the ABC factor.

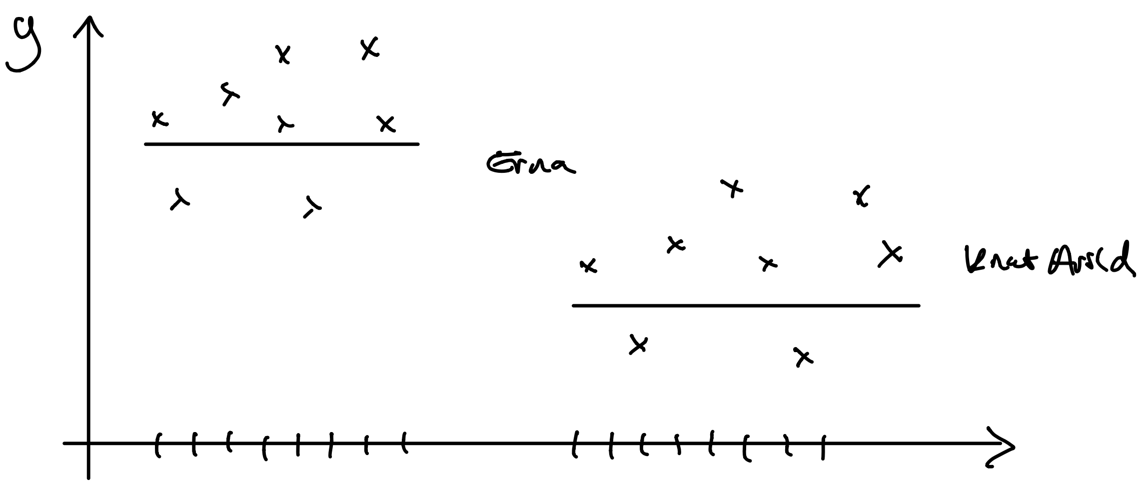
What would happen if I did not include the block as a regressor/covariate in the analysis?

$\Rightarrow$ SSE will be big.

Example: person as block

Erna and Knut Arild want to perform an 2^3 experiment together, and to get 16 obs. they will both conduct the same physical 2^3 experiments.

Should then a covariate telling who did each experiment be added to the regression model?



If the above figure gives a correct picture of the experiment NOT including a person covariate will make SSE very large, and will therefore give only non significant effects.

2^3 in four blocks

To divide the $2^3 = 8$ runs into 4 blocks we need two block generators. The best solution is to use AB and AC as generators for the blocks:

Block	AB	AC
1	−	−
2	−	+
3	+	−
4	+	+

← (run 2 and 7 in std. order)

Then the block effect will not be confounded with the main effect A, B, C or ABC interaction. But will be confounded with AB, AC and also

$$AB \cdot AC = \underset{\substack{\shortparallel \\ \text{col. with } +1 \\ \shortparallel \\ I}}{A^2} BC = BC$$

Q: What if ABC and BC were to be chosen as block generators?

$ABC \cdot BC = AB^2C^2 = A$ ← blocks will be confounded with the A effect

Fractional factorial designs

<u>Observation</u>: when the number of factors (k) is large, it may not be optimal to perform a full 2^k factorial design, due to the possible redundancy of the design.

higher order interactions tend to be smaller than lower order interactions

<u>Solution</u>: only perform a fraction of the full design!

$\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \ldots$

Now: move in the opposite direction to solve this.

We have a full 2^3 factorial design with factors A, B, C, but we also want to have a new factor D in the experiment.

Possible solution: let the ABC column define the levels of factor D.

1) $D = ABC$ is called the "generator" of the design.

2) $I = D \cdot D = D \cdot ABC = ABCD$ is called the "defining relation" of the design.

column of 1s

3) The number of letters (lenth) of the (shortest) defining relation is called the "resolution of the design", and is denoted by Roman numerals. Here: $\underline{\overline{IV}}$

4) Finaly: this is called a half-fraction of a 2^4 design

NOTATION $\quad 2^{4-1}_{\underline{IV}}$ design with $I = ABCD$ as defining relation and $D = ABC$ as generator.

We may perform 8 experiments and estimate 8 parameter

may use Lenth's method to assess significance

With 4 factors there are:

$4 = \binom{4}{1}$ main effect $\quad A, B, C, D$

$6 = \binom{4}{2}$ two-way interactions $\quad AB, AC, \ldots, CD$

$4 = \binom{4}{3}$ three-way interactions $\quad ABC, ABD, ACD, BCD$

$1 = \binom{4}{1}$ four-way interaction $\quad ABCD$

$4+6+4+1 = 15$ possible effect (+1 intercept)

Q: what can we estimate?

The "alias-structure" defines which effects are confounded.

<u>Obvious:</u> Since $D = ABC$, then D and ABC are confounded.

$\hat{D}$ may actually be $\hat{D} + \hat{ABC}$

<u>Method:</u> We want to find if any effects are confounded with A. We multiply A with the defining relation, $I = ABCD$.

1) Main effects:

$$A = A \cdot I = A \cdot ABCD = \overset{\overset{1s}{\|}}{A^2} BCD = BCD$$

that is A and BCD column are equal

$$B = B \cdot I = B \cdot ABCD = ACD$$

$$C = C \cdot I = C \cdot ABCD = ABD$$

$$D = D \cdot I = D \cdot ABCD = ABC$$

All main effects are confounded with 3-way interactions

$$l_A = A + BCD$$

we think that we estimate A, but we actually estimate A + BCD, BUT if 3-way interactions are small ⇒ ok!

2) 2-way:

$AB = AB \cdot I = AB \cdot ABCD = \underline{CD}$ $\qquad l_{AB} = AB + CD$

$AC =$ $\qquad l_{AC} = AC + BD$

$AD =$ $\qquad l_{AD} = AD + BC$

3) 3-way: already done ← main effects

4) $I = ABCD$ is confounded with intercept.

TMA4267 Linear Statistical Models V2017 (L20)

Part 4: Design of Experiments
Fractional factorial designs
Quiz with Kahoot!

Mette Langaas

Department of Mathematical Sciences, NTNU

To be lectured: March 30, 2017

What did we learn last lession?

- Why don't we want to perform a full factorial experiment, but a instead a fractional factorial? (If we have many factors we maybe not need to be able to estimate all possible interactions, and may accept that effects are confounded.)
- What is the easiest way to design a half-fraction of a 2^k factorial experiment? (Perform all the experiments where the highest order interaction =-1 or +1. E.g. for k=4 we may do 16 different experiments, and now we only do the 8 possible experiments where ABCD=+1=defining relation. This is the same as thinking that D=ABC=generator).
- New words:
  - *generator(s)*=how to generate the design,
  - *defining relation(s)*, found from the generators,
  - *resolution*=length of shortest defining relation,
  - *alias structure*=confounding pattern, found by multiplying each effect of interest with the defining relation.
- Today: more on interpreting "confounding", interpreting "resolution" and more fractional factorial experiments

Box, Hunter, Hunter: Reactor example

- A=feed rate (liters/min).
- B=Catalyst (%).
- C=Agitation rate (rpm).
- D=Temperature (deg C).
- E=Concentration (%).
- Response= (%) reacted.

Full factorial with $2^5 = 32$ experiments.
From Box, Hunter, Hunter (1978, 2005): "Statistics for Experimenters", Ch.12.2.

Half fraction with reactor example

- Instead of running a full factorial with $2^5 = 32$ experiments,
- we suggest running a half-fraction.
- We choose $I = ABCDE$ as the defining relation.

Reactor data: answer in groups

	A	B	C	D	E	y
1	-1	-1	-1	-1	-1	61
2	1	-1	-1	-1	-1	53
3	-1	1	-1	-1	-1	63
4	1	1	-1	-1	-1	61
5	-1	-1	1	-1	-1	53
6	1	-1	1	-1	-1	56
7	-1	1	1	-1	-1	54
8	1	1	1	-1	-1	61
9	-1	-1	-1	1	-1	69
10	1	-1	-1	1	-1	61
11	-1	1	-1	1	-1	94
12	1	1	-1	1	-1	93
13	-1	-1	1	1	-1	66
14	1	-1	1	1	-1	60
15	-1	1	1	1	-1	95
16	1	1	1	1	-1	98
17	-1	-1	-1	-1	1	56
18	1	-1	-1	-1	1	63
19	-1	1	-1	-1	1	70
20	1	1	-1	-1	1	65
21	-1	-1	1	-1	1	59
22	1	-1	1	-1	1	55
23	-1	1	1	-1	1	67
24	1	1	1	-1	1	65
25	-1	-1	-1	1	1	44
26	1	-1	-1	1	1	45
27	-1	1	-1	1	1	78
28	1	1	-1	1	1	77
29	-1	-1	1	1	1	49
30	1	-1	1	1	1	42
31	-1	1	1	1	1	81
32	1	1	1	1	1	82

- Which of the 32 experiments should be performed when $I = ABCDE$ is the defining relation? What is then the generator?
- What is the resolution for this design?
- Write down the aliasing pattern.

Resolution

A design is said to be of resolution R if no p-factor effect is aliased with an effect containing less than R-p factors.

A design of resolution

- III does not confound main effects with one another, but does confound main effects with two-factor interactions.
- IV does not confound main effects and two-factor interactions, but does confound two-factor interactions with other two-factor interactions.
- V does not confound main effects and two-factor interactions with each other, but does confound two-factor interactions with three-factor interactions and so on.

In general the resolution of a two-level factional design is *the length of the shortest word in the defining relation*.

Half fraction with reactor example: generator and defining relation

- Instead of running a full factorial with $2^5 = 32$ experiments,
- we suggest running a half-fraction.
- We choose $I = ABCDE$ as the defining relation.
- Alternative thinking:
  - Construct a full 2^4 design for A, B, C and D.
  - The column of signs for the ABCD interaction is written and used to define the levels for factor E.
  - This means $E = ABCD$ is the generator for the design, and $I = ABCDE$ is the defining relation.

R-code on course www-page.

Interpretation of confounding: example

Suppose there are three factors, A, B, C, for which we know the true effects and interaction effects:

$$\begin{aligned} A &= 8 \\ B &= 20 \\ C &= 2 \\ AB &= 4 \\ AC &= 2 \\ BC &= 6 \\ ABC &= 4 \end{aligned}$$

Also is known that average response is 70.

True regression model

The corresponding regression model is:

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 + \beta_{12} z_{12} + \beta_{13} z_{13} + \beta_{23} z_{23} + \beta_{123} z_{123} + \epsilon$$

where $z_{12} = z_1 z_2$, $z_{13} = z_1 z_3$, $z_{23} = z_2 z_3$, $z_{123} = z_1 z_2 z_3$, and where the coefficients β are half the corresponding effects, while $\beta_0 = 70$. The regression model is hence

$$y = 70 + 4z_1 + 10z_2 + z_3 + 2z_{12} + z_{13} + 3z_{23} + 2z_{123} + \epsilon$$

In the following we shall also for simplicity assume that the errors ϵ are 0. This makes it possible to compute the responses for any experiment for which the levels of A, B, C are specified.

Confounding example (cont.)

Assume now that a 2^{3-1} experiment is performed, with generator $C = AB$. And responses are computed using the true regression model (check!).

St. order		A	B	C=AB	AB	AC	BC	ABC	y
1	+	-	-	+	+	-	-	+	57
2	+	+	-	-	-	-	+	+	65
3	+	-	+	-	-	+	-	+	73
4	+	+	+	+	+	+	+	+	93
	Const.	z_1	z_2	z_3	z_{12}	z_{13}	z_{23}	z_{123}	
Coeff.	70	4	10	1	2	1	3	2	

Confounding example (cont.)

It is now seen that in all of these 4 experiments are

$$
\begin{aligned}
Const. &= z_{123} \\
z_1 &= z_{23} \\
z_2 &= z_{13} \\
z_3 &= z_{12}
\end{aligned}
$$

so for the performed experiment we may as well write the model as

$$
y = (\beta_0 + \beta_{123}) + (\beta_1 + \beta_{23})z_1 + (\beta_2 + \beta_{13})z_2 + (\beta_3 + \beta_{12})z_3
$$

Using that we know the values of the coefficients, the true model for the data is thus

$$
\begin{aligned}
y &= (70 + 2) + (4 + 3)z_1 + (10 + 1)z_2 + (1 + 2)z_3 \\
&= 72 + 7z_1 + 11z_2 + 3z_3
\end{aligned}
$$

Confounding example (cont.)

- Suppose now that we try to compute the main effect of A from our data. Apparently this will be

$$\ell_A = \frac{65 + 93}{2} - \frac{57 + 73}{2} = 79 - 65 = 14$$

  which is also found as twice the coefficient before z_1 in the regression model above.
- Similarly, the apparent interaction effect of B and C would be computed as

$$\ell_{BC} = \frac{-57 + 65 - 73 + 93}{2} = 14$$

  The truth (which is known to us) is, however, that $A = 8$ and $BC = 6$, so that it is the sum of A and BC which is 14.

This is what is meant by saying that the main effect of A and the interaction effect between B and C are *confounded* (mixed). The confounded effects are listed in R as the *alias structure*.

```
Factorial Fit: y versus A; B; C


Estimated Effects and Coefficients for y (coded units)


Term        Effect    Coef
Constant            72,000
A           14,000   7,000
B           22,000  11,000
C            6,000   3,000


Alias Structure
I + A*B*C
A + B*C
B + A*C
C + A*B
```

The bicycle example

TABLE 12.5. An eight-run experimental design for studying how time to cycle up a hill is affected by seven variables (I = 124, I = 135, I = 236, I = 1237).

run	seat up/down 1	dynamo off/on 2	handlebars up/down 3	gear low/medium 4 12	raincoat on/off 5 13	breakfast yes/no 6 23	tires hard/soft 7 123	time to climb hill (sec) y
1	−	−	−	+	+	+	−	69
2	+	−	−	−	−	+	+	52
3	−	+	−	−	+	−	+	60
4	+	+	−	+	−	−	−	83
5	−	−	+	+	−	−	+	71
6	+	−	+	−	+	−	−	50
7	−	+	+	−	−	+	−	59
8	+	+	+	+	+	+	+	88

From Box, Hunter, Hunter (1978, 2005): "Statistics for Experimenters", Ch.12.25

The bicycle example

- Set up a full factorial design in the three variables A, B, C.
- Use the generators: D=AB, E=AC, F=BC, G=ABC.
- Defining relations: I=ABD=ACE=BCF=ABCG.
- The design is of resolution III.
- It is a $1/16$ fraction of the full 2^7, and thus called 2_{III}^{7-4}.
- A design where every available contrast is associated with a factor is called a *saturated design*.

Using FrF2 in R, see file L20.R

```
> plan <- FrF2(nruns=8,nfactors=7,
generators=c("AB","AC","BC","ABC"),alias.info=2,randomize=FALSE)
> plan
   A  B  C  D  E  F  G
1 -1 -1 -1  1  1  1 -1
2  1 -1 -1 -1 -1  1  1
3 -1  1 -1 -1  1 -1  1
4  1  1 -1  1 -1 -1 -1
5 -1 -1  1  1 -1 -1  1
6  1 -1  1 -1  1 -1 -1
7 -1  1  1 -1 -1  1 -1
8  1  1  1  1  1  1  1
class=design, type= FrF2.generators
> summary(plan)
Call:
FrF2(nruns = 8, nfactors = 7, generators = c("AB", "AC", "BC",
    "ABC"), alias.info = 2, randomize = FALSE)
Experimental design of type  FrF2.generators
8  runs
Factor settings (scale ends):
   A  B  C  D  E  F  G
1 -1 -1 -1 -1 -1 -1 -1
2  1  1  1  1  1  1  1
Design generating information:
$legend
[1] A=A B=B C=C D=D E=E F=F G=G
$generators
[1] D=AB  E=AC  F=BC  G=ABC
Alias structure:
$main
[1] A=BD=CE=FG B=AD=CF=EG C=AE=BF=DG D=AB=CG=EF E=AC=BG=DF F=AG=BC=DE G=AF=BE=CD
```

Exam question on fractional factorials (K2014)

In a pilot study with four factors A, B, C and D, the 8 experiments listed below were run.

	A	B	C	D
1	-1	-1	-1	1
2	1	-1	-1	-1
3	-1	1	-1	-1
4	1	1	-1	1
5	-1	-1	1	1
6	1	-1	1	-1
7	-1	1	1	-1
8	1	1	1	1

What type of experiment is this?
What is the generator and the defining relation for the experiment?
What is the resolution of the experiment?
Write down the alias structure of the experiment.

Not covered: Response Surface Methods

Dates back to the 1950s, with popular book by Box and Draper.

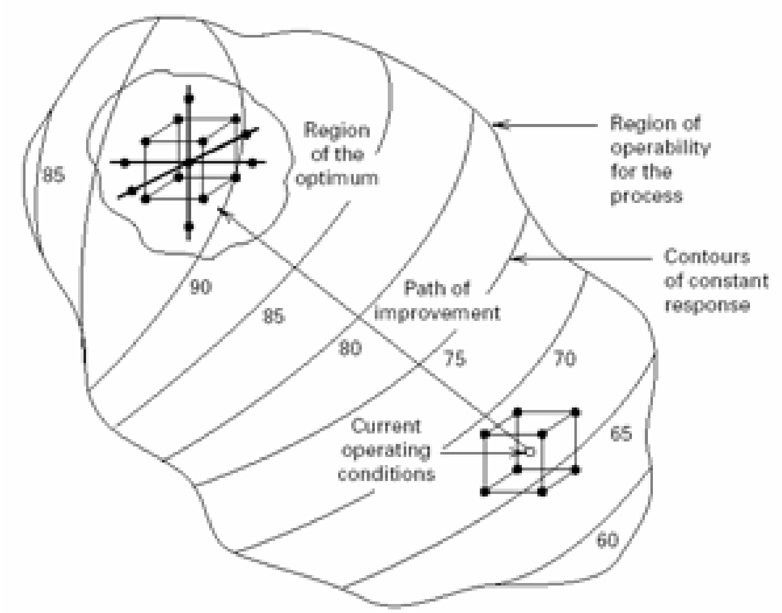


Figure 11-3 The sequential nature of RSM.

- The method performes sequential optimization, and can deal with several responses simultaneously.
- Central Composite Designs (CCD) and Box-Behnken Designs are two popular methods.
- John Tyssedal supervises 5th year project and master thesis in DOE.

https://onlinecourses.science.psu.edu/stat503/node/57

Final word about the DOE Compulsory Exercise 4

- If you want to have 4 factors and perform 16 runs see R-code named `https://www.math.ntnu.no/emner/TMA4267/2017v/RscriptDOEtreadmill.R`
- If you want to have 3 factors, but need a block effect - look at this code `https://www.math.ntnu.no/emner/TMA4267/2017v/DOE2in3withrepl.R`, because it is best to code the block with effect coding - FrFr use treatment coding - and then we don't have orthogonal columns and everything becomes difficult...

Summing up with Kahoot! quiz

kahoot.it

PART 4: DoE

30.03.2017 L20

TMA4267

Fractional factorial designs (cont.)

we focus on 2^k design

not all possible experiments are performed – but a fraction $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \ldots$

fraction: $\{1, \ldots, k-1\}$

$$2^{k-p}_{\text{resolution}: \{III, IV, V, \ldots\}}$$

Ex: reactor data

$I = ABCDE$ defining relation ← we do row 2, 3, 5, ..., 32

$2^4 \rightarrow E = ABCD$ generator ←

Either think to start with full factorial in A, B, C, D and then add E = ABCD, or just do the runs with ABCDE = 1 = I.

Resolution: V

Alias: $A \cdot I = BCDE$
$B \cdot I = ACDE$

$AB \cdot I = AB \cdot ABCDE = CDE$

⋮

→ see R-code L20.R

Interpretation of confounding

1) $y = 70 + 4z_1 + 10z_2 + 1z_3 + 2z_{12} + 3z_{23} + 2z_{123}$

2) $C = AB$, $A = BC$, $B = AC$, $ABC = I$

3) $y = 72 + 7z_1 + 11z_2 + 3z_3$

4) $\hat{A} = 2 \cdot \hat{\beta}_1 = 2 \cdot 7 \Rightarrow \hat{\beta}_1 = 7$
$\hat{BC} = 14 \Rightarrow \hat{\beta}_{23} = 7$ } but really $\hat{A} = 8$, $\hat{BC} = 6$

We think we estimate A, but really estimate
$\hat{A} + \hat{BC} = 14$.

Finally: bicycle example
Have 7 factors and perform 8 experiments $\Rightarrow$
2^{7-4} design. Think: full factorial in A, B, C and
add generators $D =$
$E =$
$F =$
$G =$

see L20.R for R-code.

Which of the following is NOT correct for a 2^k full factorial design matrix $\boldsymbol{X}$?

- A $\boldsymbol{X}$ only contains the numbers -1 and 1.
- B The sum of each column equals 1.
- C The columns of $\boldsymbol{X}$ are orthogonal.
- D $\boldsymbol{X}^T\boldsymbol{X}$ is a diagonal matrix.

$\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$, with $\varepsilon \sim N_n(0, \sigma^2 \boldsymbol{I})$

$\widehat{Effect}_j = 2 \cdot \frac{1}{n} \sum_{i=1}^{n} x_{ij} Y_i$.

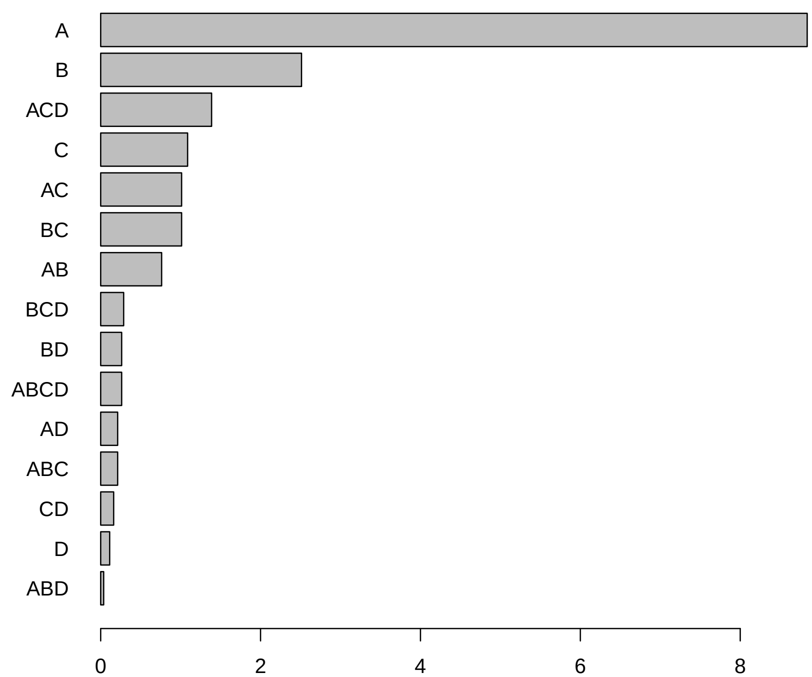
$\mathrm{Var}(\widehat{Effect}_j)$ equals

A σ^2

B $\frac{1}{n}\sigma^2$

C $\frac{2}{n}\sigma^2$

D $\frac{4}{n}\sigma^2$



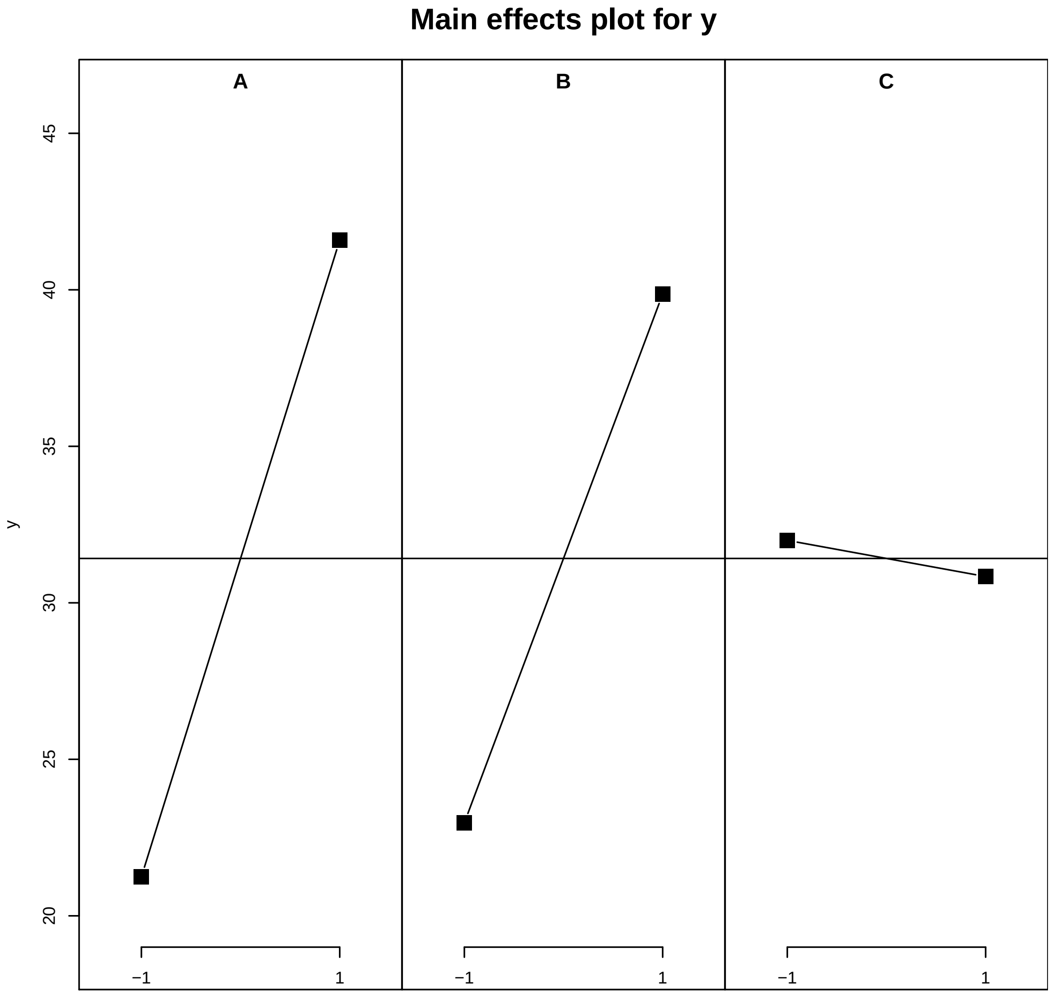
This plot is called

A Main effects plot

B Interaction effects plot

C Pareto plot

D Normal plot



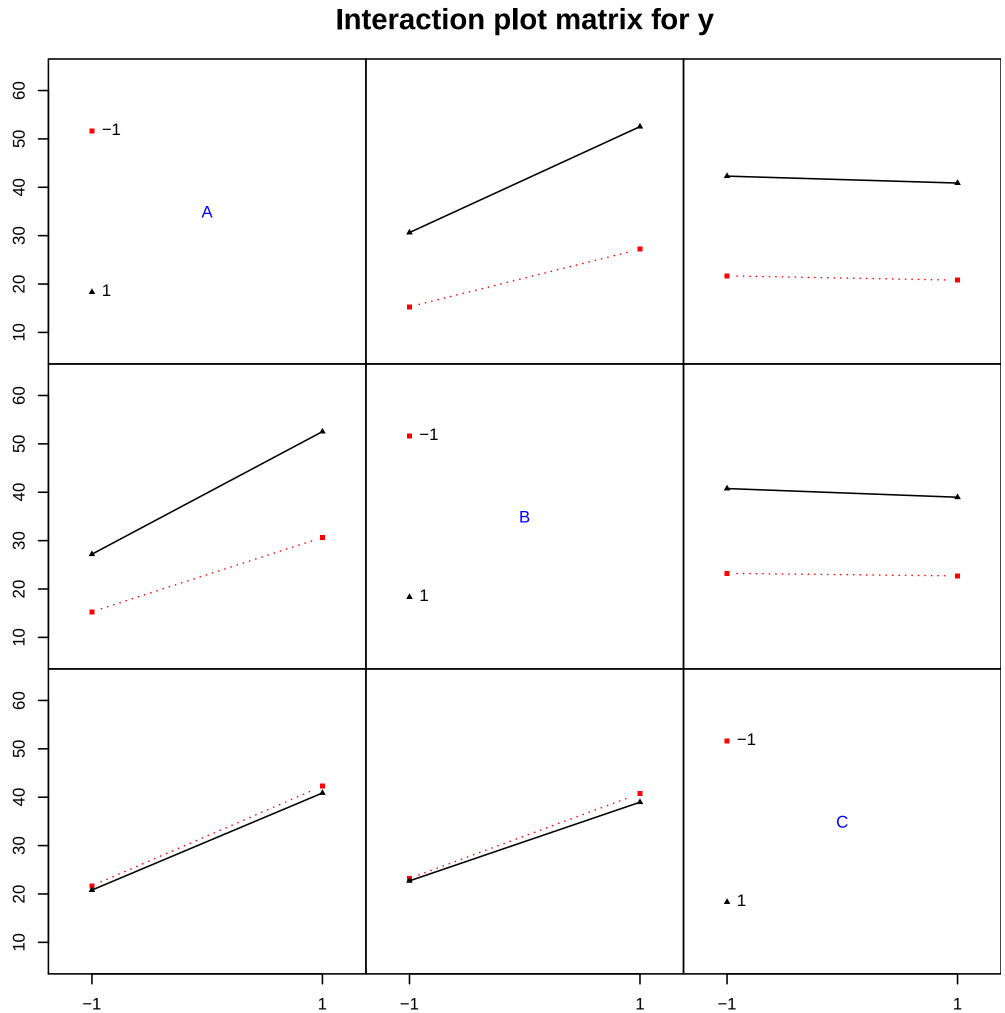
The estimated main effect of A is

A -20

B 0

C 5

D 20



Which of the estimated interaction effects AB, AC, BC is the largest?

A AB

B AC

C BC

Set up a full factorial design in the three variables A, B, C, and use generators: D=AB, E=AC, F=BC, G=ABC. What do you get?

A 2^{7-4}_{III}

B 2^{7-3}_{IV}

C 2^{7-4}_{IV}

D 2^{7-3}_{III}

For a design is of resolution III:

A Main effects are confounded with each other.

B Main effects are confounded with 2-way interactions.

C Main effects are confounded with 3-way interactions.

D Main effects are confounded with 4-way interactions.

Correct?

Are you sure you want to read the correct answers? Maybe try first?

Answers

Correct: BDCDAAB