NTNU
Norwegian University of
Science and Technology

TMA4267 Linear Statistical Models V2014 (19)

Design of experiments (note): full 2^k experiment (pages 1-14)

Mette Langaas

To be lectured: March 10, 2014
wiki.math.ntnu.no/emner/tma4267/2014v/start/

Outline DOE

- The full 2^k experiment (L19, note pages 1-14)
  - Coding, standard order.
  - Simple formulas for parameter estimates and SSR (due to orthogonality).
  - Main and interaction effects.
  - Lenths method, and other strategies for estimating σ^2.
  - External effect present when performing repetitions?
- Blocking in full 2^k experiments (L20, note pages 15-20)
- Fractions of 2^k experiments (L21, note pages 20-29)

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

DOE: 2^k full factorials (L19)

[DOE note on www-page, not book]

Ex: Lima beans, $k = 3$ factors

Study three factors, each at two levels. Perform all possible $2 \cdot 2 \cdot 2 = 2^3 = 8$ experiments.

	A	B	C	AB	AC	BC	ABC	y
1	-1	-1	-1	1	1	1	-1	y_1
2	1	-1	-1	-1	-1	1	1	y_2
3	-1	1	-1	-1	1	-1	1	y_3
4	1	1	-1	1	-1	-1	-1	y_4
5	-1	-1	1	1	-1	-1	1	y_5
6	1	-1	1	-1	1	-1	-1	y_6
7	-1	1	1	-1	-1	1	-1	y_7
8	1	1	1	1	1	1	1	y_8

↑ standard order

A, B, C: perform experiment

AB, AC, BC, ABC: multiply the relevant columns to get this

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

but with the coding given on the previous page, with an intercept column added:

$$\underset{n\times p}{\underline{X}} \qquad \begin{aligned} &n = 8 \\ &p = \#\text{ colums in design} = \overset{\text{A O C AB AC BC ABC}}{7} + \overset{\text{intercept}}{1} = 8 \end{aligned}$$

DOE coding w/intercept as first column

NB: β_0 is here used to denote the intercept

$$Y_i = \beta_0 + \overset{A}{\overline{\beta_1 X_{i1}}} + \overset{B}{\overline{\beta_2 X_{i2}}} + \cdots + \overset{ABC}{\overline{\beta_7 X_{i7}}} + \varepsilon_i$$

Work :

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

2) Show that $\sum_{i=1}^{\overset{8}{\overset{\|}{n}}} x_{ij} = 0$ $\forall j$ except $j=0$ (intercept)

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

Check: a) A and B column: $+\underline{1}-\underline{1}-\underline{1}+\underline{1}+\underline{1}-1-1+1=0$

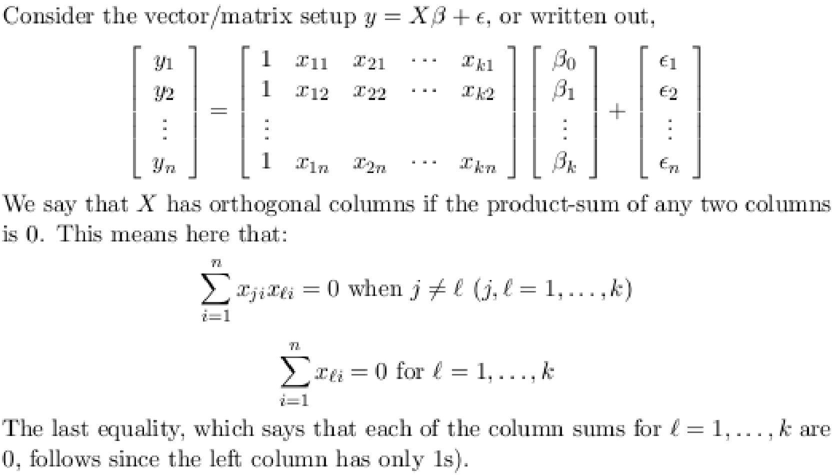
b) $-1\cdot 4 + 1\cdot 4 = 0$

c) $(-1)^2\cdot 4 + 1^2\cdot 4 = 8$

How does this influence our formulas from MLR:

i) $\hat{\beta}_j$, ii) $Var(\hat{\beta}_j)$ and iii) SSR

Orthogonality (WMMY notation)



Consider the vector/matrix setup $y = X\beta + \epsilon$, or written out,

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & & & & \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

We say that X has orthogonal columns if the product-sum of any two columns is 0. This means here that:

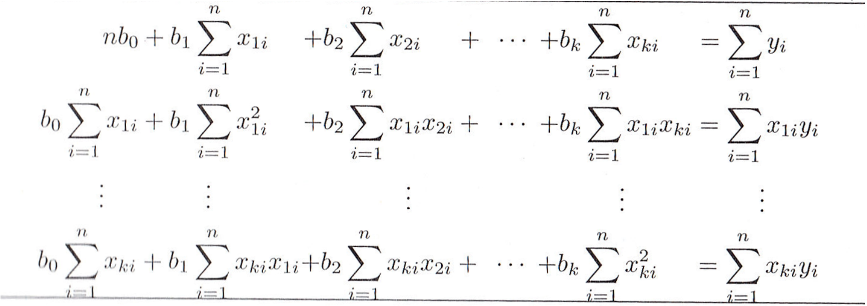
$$\sum_{i=1}^{n} x_{ji} x_{\ell i} = 0 \text{ when } j \neq \ell \ (j, \ell = 1, \ldots, k)$$

$$\sum_{i=1}^{n} x_{\ell i} = 0 \text{ for } \ell = 1, \ldots, k$$

The last equality, which says that each of the column sums for $\ell = 1, \ldots, k$ are 0, follows since the left column has only 1s).

General normal equations (WMMY notation)

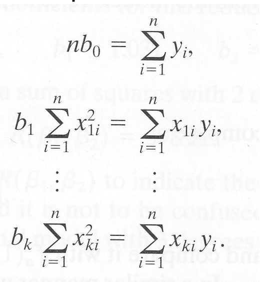
Normal equations for least squares estimation (without orthogonality)



$$
\begin{array}{lllllll}
nb_0 & + b_1 \sum_{i=1}^{n} x_{1i} & + b_2 \sum_{i=1}^{n} x_{2i} & + & \cdots & + b_k \sum_{i=1}^{n} x_{ki} & = \sum_{i=1}^{n} y_i \\
b_0 \sum_{i=1}^{n} x_{1i} & + b_1 \sum_{i=1}^{n} x_{1i}^2 & + b_2 \sum_{i=1}^{n} x_{1i} x_{2i} & + & \cdots & + b_k \sum_{i=1}^{n} x_{1i} x_{ki} & = \sum_{i=1}^{n} x_{1i} y_i \\
\vdots & \vdots & \vdots & & & \vdots & \vdots \\
b_0 \sum_{i=1}^{n} x_{ki} & + b_1 \sum_{i=1}^{n} x_{ki} x_{1i} & + b_2 \sum_{i=1}^{n} x_{ki} x_{2i} & + & \cdots & + b_k \sum_{i=1}^{n} x_{ki}^2 & = \sum_{i=1}^{n} x_{ki} y_i
\end{array}
$$

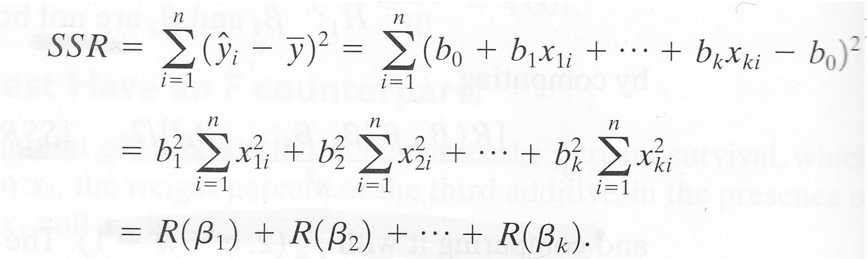
Orthogonality: normal equations (WMMY notation)

Normal equations for least squares estimation with orthogonality:



$$
\begin{aligned}
nb_0 &= \sum_{i=1}^{n} y_i, \\
b_1 \sum_{i=1}^{n} x_{1i}^2 &= \sum_{i=1}^{n} x_{1i} y_i, \\
\vdots \quad & \quad \vdots \\
b_k \sum_{i=1}^{n} x_{ki}^2 &= \sum_{i=1}^{n} x_{ki} y_i .
\end{aligned}
$$

Orthogonality: SSR (WMMY notation)



$$
\begin{aligned}
SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 &= \sum_{i=1}^{n}(b_0 + b_1 x_{1i} + \cdots + b_k x_{ki} - b_0)^2 \\
&= b_1^2 \sum_{i=1}^{n} x_{1i}^2 + b_2^2 \sum_{i=1}^{n} x_{2i}^2 + \cdots + b_k^2 \sum_{i=1}^{n} x_{ki}^2 \\
&= R(\beta_1) + R(\beta_2) + \cdots + R(\beta_k).
\end{aligned}
$$

i) $\hat{\beta} = (X^T X)^{-1} X^T Y$ $\qquad x_{i0} = 1 \;\; \forall i$ (intercept)

$r = (2^k - 1) \;\hat{=}\; 8$

$$= \begin{bmatrix} \frac{1}{\sum_{i=1}^{n} x_{i0}^2} & & & 0 \\ & \frac{1}{\sum_{i=1}^{n} x_{i1}^2} & & \\ & & \ddots & \\ 0 & & & \frac{1}{\sum_{i=1}^{n} x_{ir}^2} \end{bmatrix} \begin{bmatrix} \overset{x_{10}\,\downarrow}{1} & 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & & \cdots & x_{n1} \\ \vdots & & & & \\ x_{1r} & & \cdots & & x_{nr} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

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

$$\hat{\beta}_0 = \frac{1}{n} \sum_{i=1}^{n} 1 \cdot y_i = \underline{\underline{\bar{y}}}$$

Observe: $\hat{\beta}_j$ is independent of choice of model.

ii) $\operatorname{Var}(\hat{\beta}_j) = \operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} x_{ij} \cdot y_i\right)$

$$= \frac{1}{n^2} \underbrace{\sum_{i=1}^{n} x_{ij}^2}_{} \underbrace{\operatorname{Var}(y_i)}_{\sigma^2} = \frac{1}{n^2} \cdot n \cdot \sigma^2 = \underline{\underline{\frac{1}{n}\sigma^2}}$$

iii) $SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ and $\hat{\beta}_0 = \bar{y}$

$$\hat{y}_i = \sum_{j=0}^{r} \hat{\beta}_j \cdot x_{ij}$$

$$= \sum_{i=1}^{n} \left[\left(\sum_{j=0}^{r} \hat{\beta}_j x_{ij}\right) - \hat{\beta}_0\right]^2$$

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

$$= \sum_{i=1}^{n} \left(\sum_{j=1}^{r} \hat{\beta}_j x_{ij}\right)^2 = \sum_{j=1}^{r} \hat{\beta}_j^2 \underbrace{\sum_{i=1}^{n} x_{ij}^2}_{n}$$

since $\sum_{i=1}^{n} x_{ij} x_{ih} = 0$

$$= \underline{\underline{n \cdot \sum_{j=1}^{r} \hat{\beta}_j^2}}$$

$$\underbrace{n\hat{\beta}_1^2}_{SSR(x_1)} + \underbrace{n\hat{\beta}_2}_{} + \cdots + \underbrace{n\hat{\beta}_r^2}_{SSR(x_r)}$$

Main effects in DOE

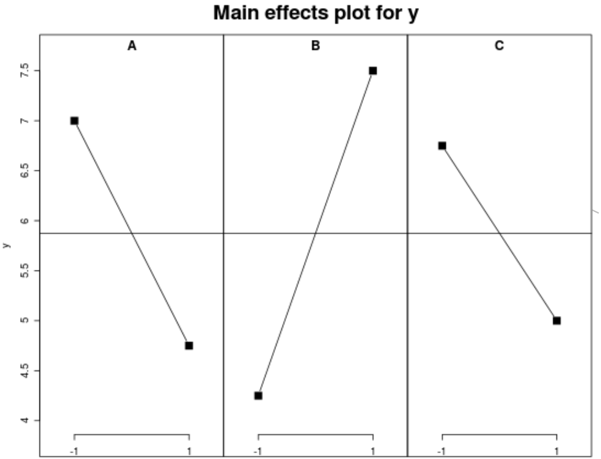
Main effect of A

$$\begin{aligned}\widehat{A} &= 2b_1\\ &= \frac{y_2+y_4+y_6+y_8}{4} - \frac{y_1+y_3+y_5+y_7}{4}\end{aligned}$$

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

$$\begin{aligned}\widehat{B} &= 2b_2\\ &= \frac{y_3+y_4+y_7+y_8}{4} - \frac{y_1+y_2+y_5+y_6}{4}\end{aligned}$$

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
```

DOE Effects

For each β_j in the moael equation (except $j=0$) we define an effect to be

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

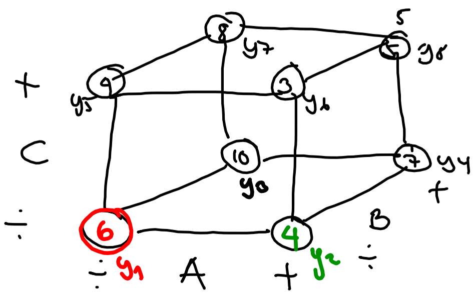
Why? β_j gives the change when x_{ij} goes from 0 to 1, or -1 to 0 etc, while Effect_j gives the change when x_{ij} goes from -1 to 1 (the renge of the experiment factors)

$$\widehat{\text{Effect}}_j = 2 \cdot \hat{\beta}_j$$

and

$$\text{Var}(\widehat{\text{Effect}}_j) = 4 \cdot \underbrace{\text{Var}(\hat{\beta}_j)}_{\frac{\sigma^2}{n}} = \underline{\underline{4 \frac{\sigma^2}{n}}} \overset{\text{DEF}}{=\!=} \sigma^2_{\text{effect}}$$

Cube plot of data [3 factor]



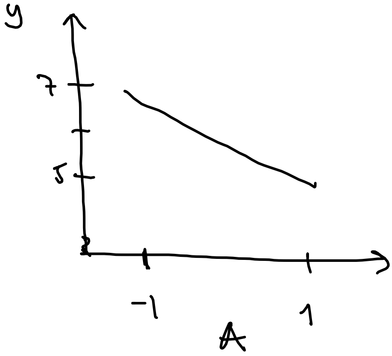
Main effects:

$$\hat{A} = 2\hat{\beta}_1 = 2 \cdot \frac{1}{n} \sum_{i=1}^{n} x_{i1} \cdot y_i$$

usual MLR formula

$$= \frac{2}{8}\left(\underbrace{(y_2+y_4+y_6+y_8)}_{\text{A high}} - \underbrace{(y_1+y_3+y_5+y_7)}_{\text{A low}}\right)$$

$$= \dots = -2,25$$



A high: average y is 4.75

A low: 7

Interection effect

$$\hat{AB} = 2\cdot\hat{\beta}_{12} = \frac{2}{8}\left(y_1 - y_2 - y_3 + y_4 + y_5 - y_6 - y_7 + y_8\right)$$

$$= \dots = \frac{1}{2}\left(\frac{y_4+y_8}{2} - \frac{y_3+y_7}{2}\right) - \frac{1}{2}\left(\frac{y_2+y_6}{2} - \frac{y_1+y_5}{2}\right)$$

$$= \underline{-0{,}75}$$

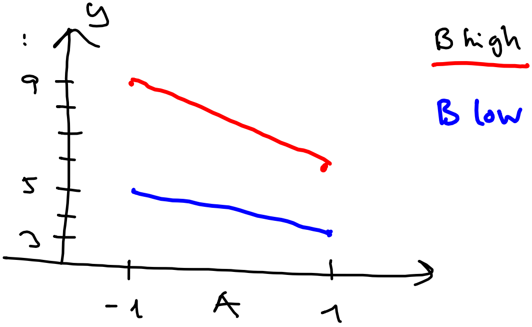
$\underline{A+ \quad B+}$; $= \underline{6}$

$\underline{A+ \quad B\div}$: $= \underline{3.5}$

$\underline{A\div \quad B+}$: $= \underline{9}$

$\underline{A\div \quad B\div}$ $\quad y_1, y_5 : \frac{1}{2}(y_1+y_5) = \underline{5}$

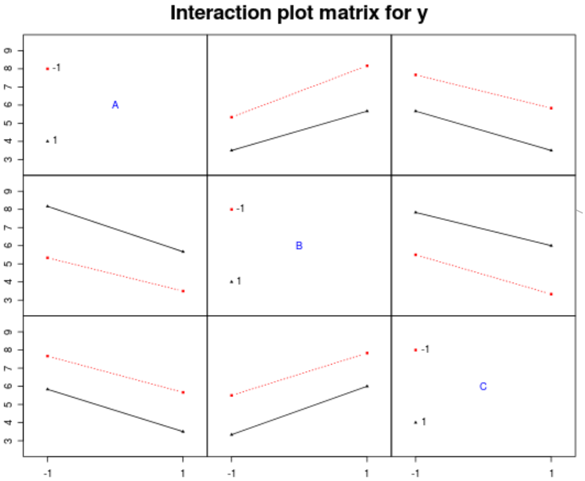
Interection plot :



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*.

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

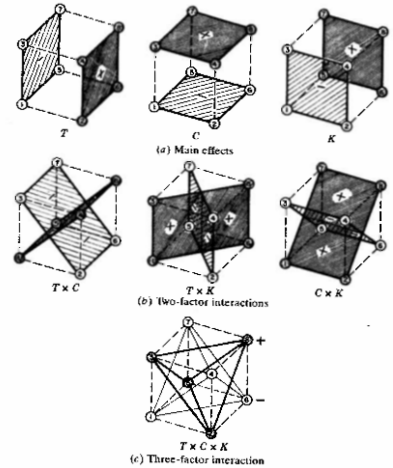


```
    A       B       C        A:B     A:C       B:C     A:B:C
-2.25   3.25   -1.75    -0.75   0.25     -0.25    -0.25
```

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



Important remark

- We will here denote the intercept by β_0.
- We will look at k dichotomous covariates, so we estimate $p = k + 1$ regression parameters.

R: DOE set-up

```
> 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
```

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
```

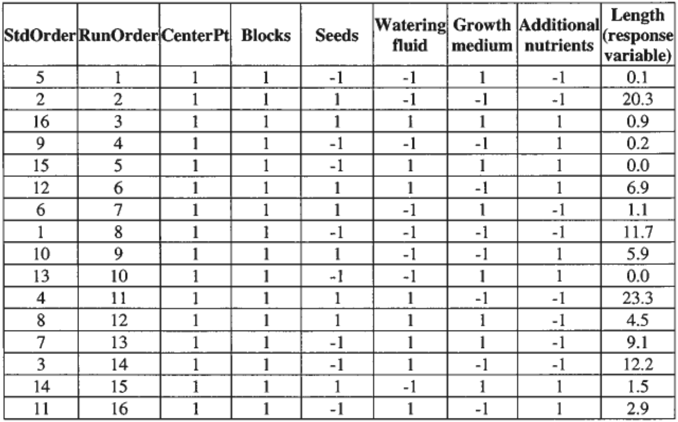
Example compulsory project

"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

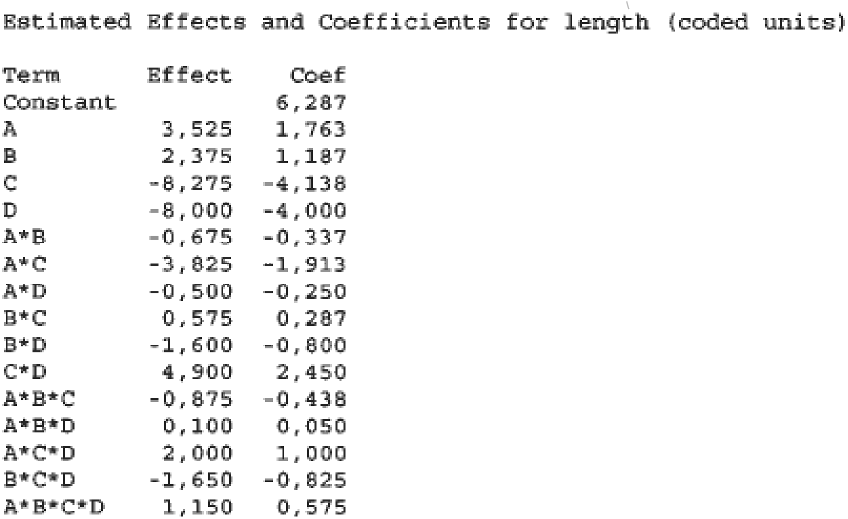
Response: length of plant after 8 days of growing.

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

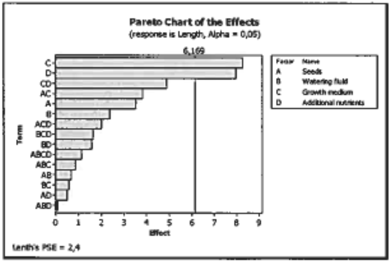


Figure 5.2 Pareto-chart of the effects with terms up to 4th order.

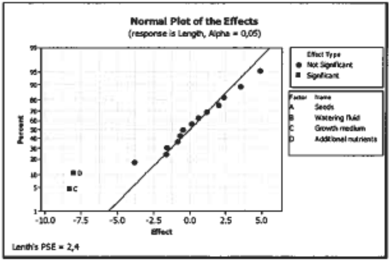


Figure 5.3 Normal plot of the effects with terms up to 4th order.

Inference

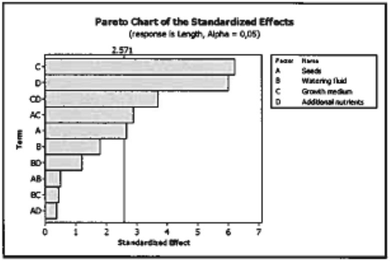


Figure 5.6 Pareto-chart of the effects with terms up to 2nd order.

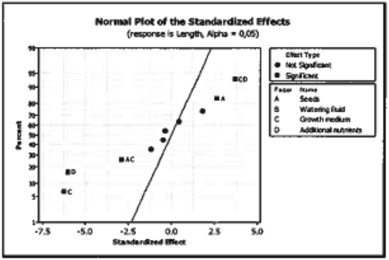


Figure 5.7 Normal plot of the effects with terms up to 2nd order.

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

Interpretation: Interaction plots

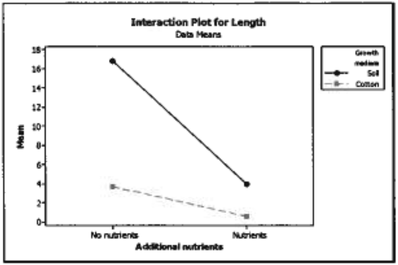


Figure 6.1 Interaction plot between growth medium and additional nutrients (CD).

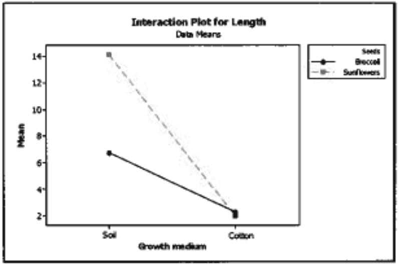


Figure 6.2 Interaction plot between seeds and growth medium (AC).

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 multippel 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 socalled fractional factorial design to be lectured).

The practical issues (2)

— If you choose 3 factors you need to perform all possible combinations of the 3 factors two times (2·2·2=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 experment 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.

The practical issues (3)

- After you have performed all 16 experiments you need to recored 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 and recommendations: Which conclusions can you draw from the experiment?

To get 20 points you need to have addressed all of these aspects in a correct manner! BUT - 10 pages is more than enough (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 MLR model.
- Example 1: Study how residuals look depending on the mismatch between the MLR model assumptions and the model you use?
- Example 2: Look into the original data of Galton on midparent and adult child height. Perfrom analyses separately for males and females. Several papers to read on this.

Supervision?

- Meeting with reference group is scheduled for Friday March 14 at 11.15.
- How would you like supervision of the project to be set up?
- Ideally, all projects should be discussed with the lecturer or TA before the data are collected - to avoid “obvious” complications.
- Current time slots: Monday 12-13 (meeting time lecturer), Wednesday 10-11 (supervision in 734 by TA). Need more time slots for week 12-14?