

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 $\hat{\beta}$, $\text{Cov}(\hat{\beta})$ 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: $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$ and $\text{Cov}(\hat{\beta}) = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$.

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

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

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

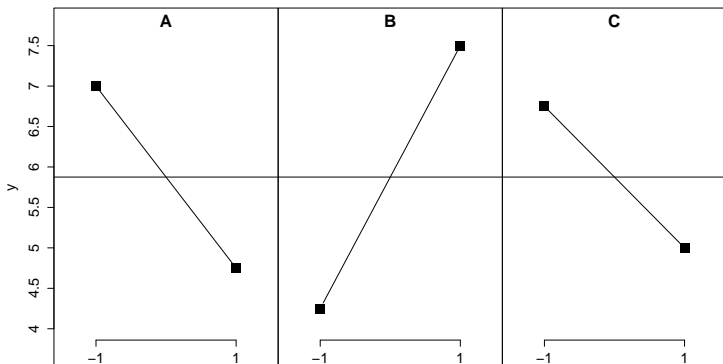
Interpretation: mean response when A is high MINUS mean response when A is low.

Similarly, main effect of B

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

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

Main effects plot 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

Explain the main effects in plain words!

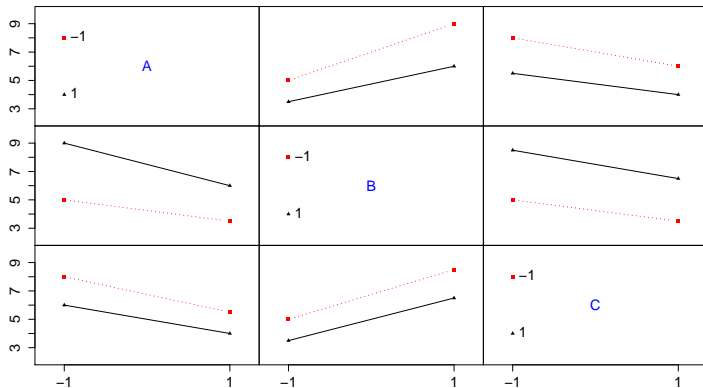
A: depth (0.5 or 1), B: watering daily (once, twice), C: type (baby, large).

Interaction effect in DOE

- ▶ What is the interpretation 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*.

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

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 places 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 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

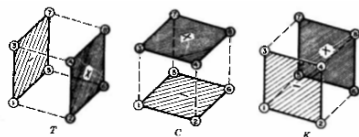
```
> lm3 <- lm(y~(.)^3,data=plan)
> MEPlot(lm3)
> IAPlot(lm3)
> effects <- 2*lm3$coeff
> effects
(Intercept) A1      B1      C1      A1:B1  A1:C1 B1:C1 A1:B1:C1
11.75      -2.25  3.25  -1.75 -0.75  0.25  -0.25  -0.25
```

2^k full factorial

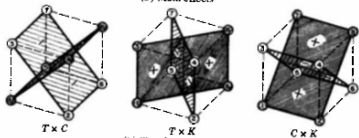
- ▶ There are k factors: A, B, C, ..., and
- ▶ 2=each factor has two levels.
- ▶ There are 2^k possible experiments.
- ▶ We have in total 2^k parameters to be estimated:
 - ▶ 1 intercept
 - ▶ $k = \binom{k}{1}$ main effects: A, B, C, ...
 - ▶ $\binom{k}{2}$ two factor interactions: AB, AC, .., BC, BD,...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

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

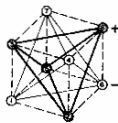
Geometric interpretation of effects



(a) Main effects

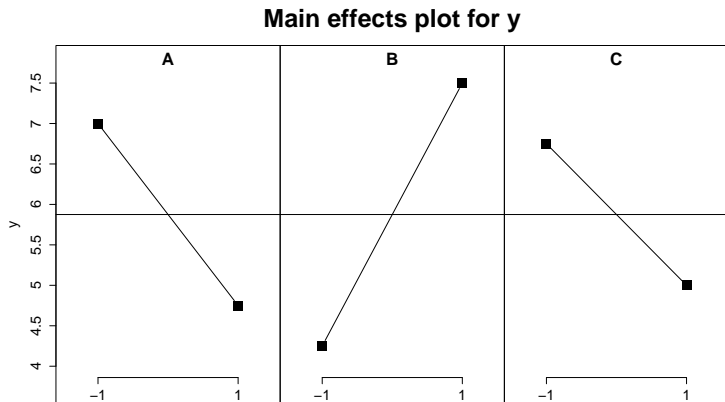


(b) Two-factor interactions



(c) Three-factor interaction

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

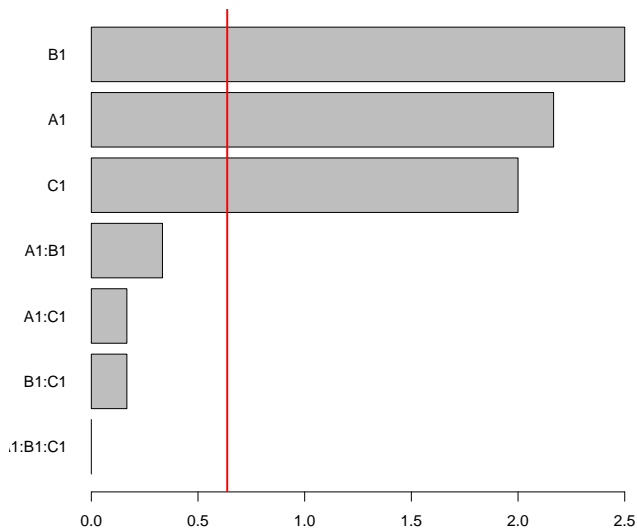
F-statistic: NaN on 7 and 0 DF, p-value: NA

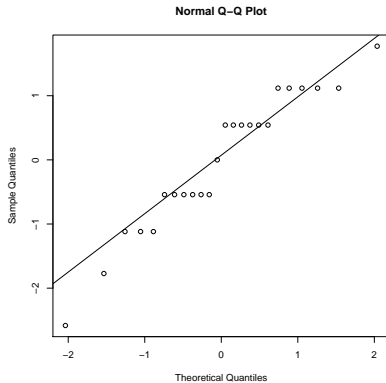
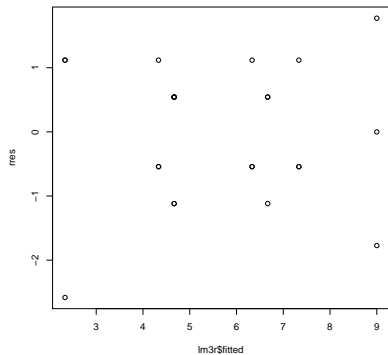
Estimation of σ^2

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](#).





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

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

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

$$E(\widehat{Effect}_j^2) = \sigma_{effect}^2$$

- ▶ Thus \widehat{Effect}_j^2 is an unbiased estimator of σ_{effect}^2 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_{effect}^2 .

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, \dots, 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

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

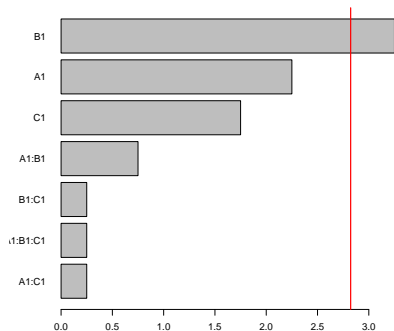
3. Take out the effects C_j with $|C_j| \geq 2.5 \cdot s_0$ and find the median of the rest of the $|C_j|$. Then PSE is this median multiplied by 1.5, i.e.

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

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 $|C_j| > t_{\alpha/2, m/3} \cdot \text{PSE}$.

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 (preferable 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

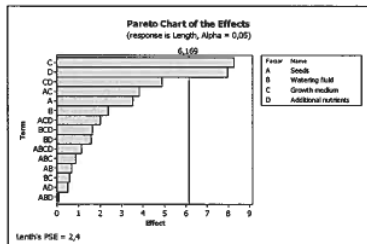


Figure 5.2 Pareto-chart of the effects with terms up to 4th order.

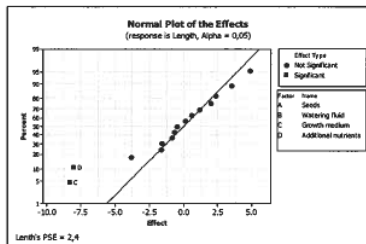


Figure 5.3 Normal plot of the effects with terms up to 4th order.

Inference

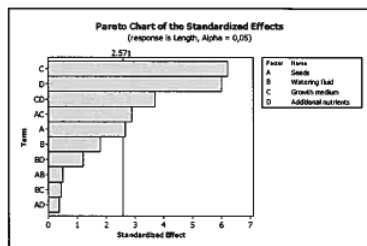


Figure 5.6 Pareto-chart of the effects with terms up to 2nd order.

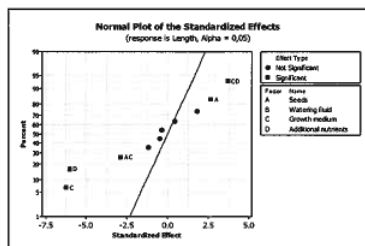


Figure 5.7 Normal plot of the effects with terms up to 2nd order.

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

Interpretation: Interaction plots

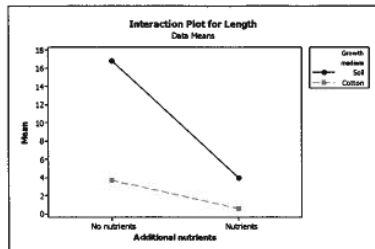


Figure 6.1 Interaction plot between growth medium and additional nutrients (CD).

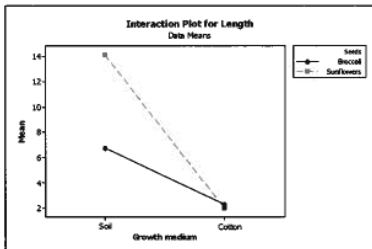


Figure 6.2 Interaction plot between seeds and growth medium (AC).

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 “fractional 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 experiment should be a complete new experiment - a genuine run replicate, unless you use blocking (not lectured yet). For example a block effect may 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 charge
3. running the apparatus at a 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.