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:

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

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.

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

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

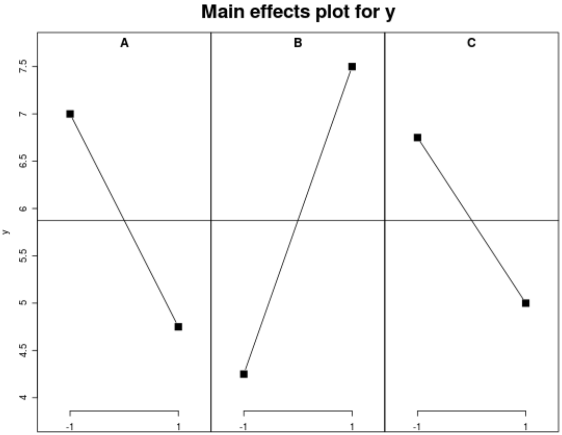
Main effect of A

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

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

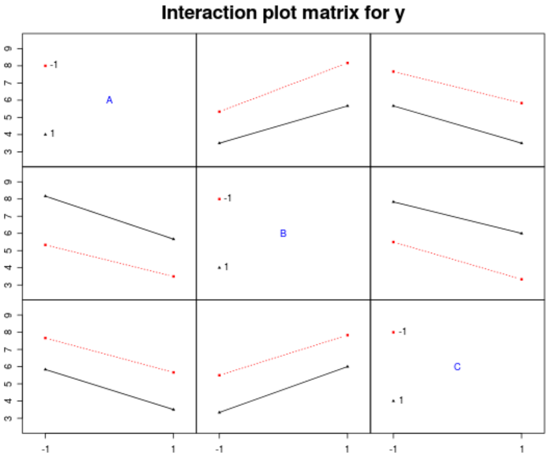
Explain the main effects in plain words!

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*.

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

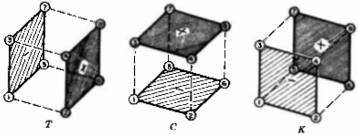



```
  A       B      C        A:B     A:C       B:C     A:B:C
-2.25   3.25  -1.75    -0.75    0.25     -0.25    -0.25
```

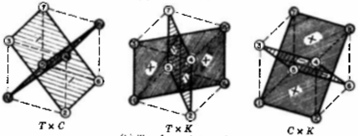
Interpretation of $\widehat{ABC}$

- $\widehat{ABC} = \frac{1}{2}\widehat{AB}$ 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.

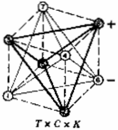
Geometric interpretation of effects



(a) Main effects



(b) Two-factor interactions



(c) Three-factor interaction

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.

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