Norwegian University of Science and Technology
Department of Mathematical Sciences

TMA4267 Linear statistical models
Recommended exercises 9–10 – solutions

Problem 1 Exam 2015 Spring, Problem 2

a) The least squares estimator of $\boldsymbol{\beta}$ is in general $(X^\mathrm{T}X)^{-1}X^\mathrm{T}\boldsymbol{Y}$. Since the columns of X are orthogonal, $X^\mathrm{T}X$ is diagonal with $\boldsymbol{x}_j^\mathrm{T}\boldsymbol{x}_j$ as entry (j,j), where $\boldsymbol{x}_j$ denotes the jth column of X. So $(X^\mathrm{T}X)^{-1}$ is diagonal with $1/(\boldsymbol{x}_j^\mathrm{T}\boldsymbol{x}_j)$ as entry (j,j). The jth row of $(X^\mathrm{T}X)^{-1}X^\mathrm{T}$ is then $\boldsymbol{x}_j^\mathrm{T}/(\boldsymbol{x}_j^\mathrm{T}\boldsymbol{x}_j)$, and the jth entry of the estimator $\boldsymbol{x}_j^\mathrm{T}\boldsymbol{Y}/(\boldsymbol{x}_j^\mathrm{T}\boldsymbol{x}_j)$.

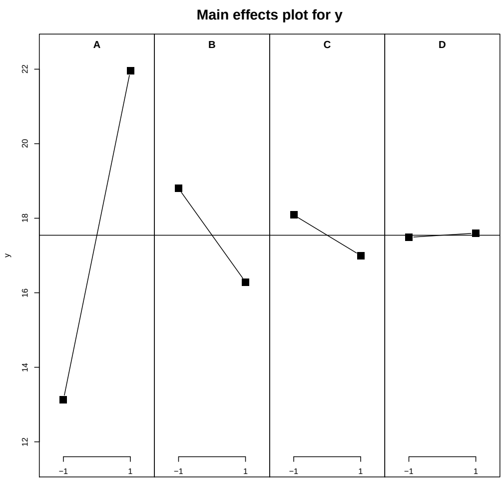
b) The interaction vector is $(1 \quad -1 \quad -1 \quad 1)^\mathrm{T}$. By the above, the coefficient estimate is $(1 \quad -1 \quad -1 \quad 1)(6 \quad 4 \quad 10 \quad 7)^\mathrm{T}/4 = (6-4-10+7)/4 = -1/4$. The estimate of the effect is $2\cdot(-1/4) = -1/2$.

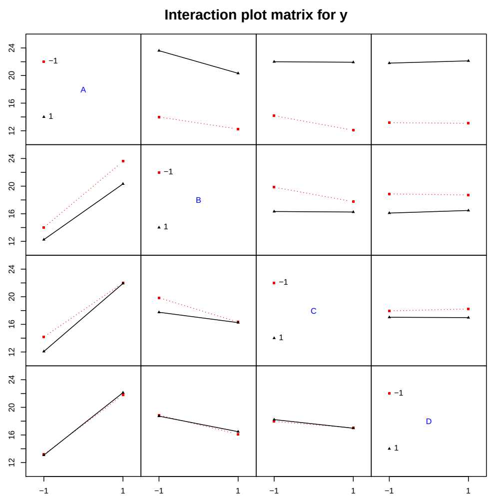
Problem 2 Factorial experiments

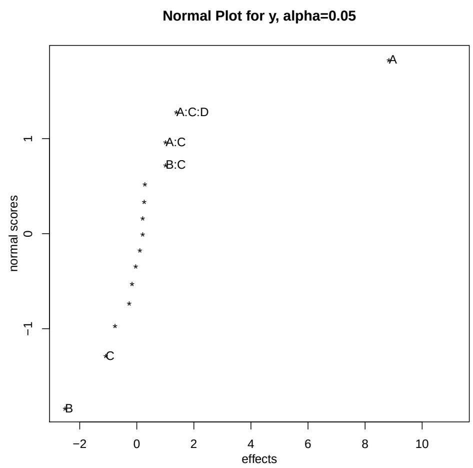
a) Output of `summary(lm4)` and of effects, followed by plots:

```
> summary(lm4)

Call:
lm.default(formula = y ~ .^4, data = plan)

Residuals:
ALL 16 residuals are 0: no residual degrees of freedom!

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.54375         NA      NA       NA
A1           4.41875         NA      NA       NA
B1          -1.25625         NA      NA       NA
C1          -0.54375         NA      NA       NA
D1           0.05625         NA      NA       NA
A1:B1       -0.38125         NA      NA       NA
A1:C1        0.50625         NA      NA       NA
A1:D1        0.10625         NA      NA       NA
B1:C1        0.50625         NA      NA       NA
B1:D1        0.13125         NA      NA       NA
C1:D1       -0.08125         NA      NA       NA
A1:B1:C1     0.10625         NA      NA       NA
A1:B1:D1    -0.01875         NA      NA       NA
A1:C1:D1     0.69375         NA      NA       NA
```

```
B1:C1:D1      0.14375         NA      NA       NA
A1:B1:C1:D1 -0.13125         NA      NA       NA

Residual standard error: NaN on 0 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:    NaN
F-statistic:   NaN on 15 and 0 DF,  p-value: NA

> 2*lm4$coeff
(Intercept)          A1          B1          C1          D1       A1:B1       A1:C1
    35.0875      8.8375     -2.5125     -1.0875      0.1125     -0.7625      1.0125
      A1:D1       B1:C1       B1:D1       C1:D1    A1:B1:C1    A1:B1:D1    A1:C1:D1
     0.2125      1.0125      0.2625     -0.1625      0.2125     -0.0375      1.3875
   B1:C1:D1 A1:B1:C1:D1
     0.2875     -0.2625
```







b) The corresponding regression model is

$$
\begin{aligned}
Y = \beta_0 &+ \beta_{\mathrm{A}} A + \beta_{\mathrm{B}} B + \beta_{\mathrm{C}} C + \beta_{\mathrm{D}} D \\
&+ \beta_{\mathrm{AB}} AB + \beta_{\mathrm{AC}} AC + \beta_{\mathrm{AD}} AD + \beta_{\mathrm{BC}} BC + \beta_{\mathrm{BD}} BD + \beta_{\mathrm{CD}} CD \\
&+ \beta_{\mathrm{ABC}} ABC + \beta_{\mathrm{ABD}} ABD + \beta_{\mathrm{ACD}} ACD + \beta_{\mathrm{BCD}} BCD + \beta_{\mathrm{ABCD}} ABCD + \epsilon,
\end{aligned}
$$

with A, B, C and D being the covariates, taking values ± 1, and β_j the coefficients.

c) In the model in a), the column space of the design matrix is the entire $\mathbb{R}^{16}$, meaning that we have a perfect fit of the data. The standard deviation estimates are all based on the error sum of squares, SSE, which is zero. (In addition, the unbiased estimator of the error term variance will have zero in the denominator.)
If we assume that the variance is known it is possible to make inference about the effects. We know from the theory for two-level factorial designs that $\hat{\beta}_i \sim N(\beta_i, \sigma^2/n)$, where n

is the number of observations. Thus $1-\alpha = P\Big(-z_{\alpha/2} < (\hat{\beta}_i - \beta_i)/(\sigma/\sqrt{n}) < z_{\alpha/2}\Big) = P\Big(\hat{\beta}_i - z_{\alpha/2}\sigma/\sqrt{n} < \beta_i < \hat{\beta}_i + z_{\alpha/2}\sigma/\sqrt{n}\Big)$, where $z_{\alpha/2}$ is the upper α-quantile of $N(0,1)$. Thus confidence intervals for the effects have bounds $2\hat{\beta}_i \pm 2z_{\alpha/2}\sigma/\sqrt{n} = 2\hat{\beta}_i \pm 1.960$ with $n = 16,\ \ \sigma = 2$ and $\alpha = 0.05$.

```
> cbind(2*lm4$coeff-qnorm(.975), 2*lm4$coeff+qnorm(.975))
                 [,1]      [,2]
(Intercept) 33.127536 37.047464
A1           6.877536 10.797464
B1          -4.472464 -0.552536
C1          -3.047464  0.872464
D1          -1.847464  2.072464
A1:B1       -2.722464  1.197464
A1:C1       -0.947464  2.972464
A1:D1       -1.747464  2.172464
B1:C1       -0.947464  2.972464
B1:D1       -1.697464  2.222464
C1:D1       -2.122464  1.797464
A1:B1:C1    -1.747464  2.172464
A1:B1:D1    -1.997464  1.922464
A1:C1:D1    -0.572464  3.347464
B1:C1:D1    -1.672464  2.247464
A1:B1:C1:D1 -2.222464  1.697464
```

The main effects of A and B are the ones significantly different from zero (zero is not in the confidence interval).

d) To assume that three-way and four-way interactions are zero, is the same as omitting them from the model.

```
> lm2 <- lm(y~.^2, data=plan)
> summary(lm2)

Call:
lm.default(formula = y ~ .^2, data = plan)

Residuals:
      1       2       3       4       5       6       7       8       9      10
-1.0562  0.7687 -0.3313  0.6188  1.0937 -0.8062  0.2938 -0.5813  0.8437 -0.5562
     11      12      13      14      15      16
 0.5438 -0.8312 -0.8812  0.5938 -0.5063  0.7937

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.54375    0.32583  53.844 4.18e-08 ***
A1           4.41875    0.32583  13.562 3.91e-05 ***
B1          -1.25625    0.32583  -3.856   0.0119 *
C1          -0.54375    0.32583  -1.669   0.1560
```

```
D1           0.05625    0.32583   0.173   0.8697
A1:B1       -0.38125    0.32583  -1.170   0.2947
A1:C1        0.50625    0.32583   1.554   0.1810
A1:D1        0.10625    0.32583   0.326   0.7576
B1:C1        0.50625    0.32583   1.554   0.1810
B1:D1        0.13125    0.32583   0.403   0.7037
C1:D1       -0.08125    0.32583  -0.249   0.8130
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.303 on 5 degrees of freedom
Multiple R-squared:  0.9765,     Adjusted R-squared:  0.9296
F-statistic: 20.81 on 10 and 5 DF,  p-value: 0.001849
```

An estimate of the variance of the error is $\hat{\sigma}^2 = 1.303^2 = 1.70$. An estimate of the variance of the coefficient estimators is $\hat{\sigma}^2/16 = 0.106$. An estimate of the variance of the effect estimators is $4\hat{\sigma}^2/16 = 0.42$.

We also know from the theory of two-level factorial designs that the estimated variance of the coefficient estimators is the mean of the squares of the omitted coefficient estimates from the full model of a), $\big(0.10625^2 + (-0.01875)^2 + 0.69375^2 + 0.14375^2 + (-0.13125)^2\big)/5 = 0.106$ again.

At the 0.05 level, the significant effects are the main effects of A and B.

e)

```
> design1 <- FrF2(16, 4, blocks=2, randomize=FALSE)
> summary(design1)
Call:
FrF2(16, 4, blocks = 2, randomize = FALSE)

Experimental design of type  FrF2.blocked
16  runs
blocked design with  2  blocks of size  8

Factor settings (scale ends):
   A  B  C  D
1 -1 -1 -1 -1
2  1  1  1  1

Design generating information:
$legend
[1] A=A B=B C=C D=D

$`generators for design itself`
[1] full factorial

$`block generators`
[1] ABCD
```

```
no aliasing of main effects or 2fis  among experimental factors

Aliased with block main effects:
[1] none

The design itself:
  run.no run.no.std.rp Blocks  A  B  C  D
1      1         2.1.1      1 -1 -1 -1  1
2      2         3.1.2      1 -1 -1  1 -1
3      3         5.1.3      1 -1  1 -1 -1
4      4         8.1.4      1 -1  1  1  1
5      5         9.1.5      1  1 -1 -1 -1
6      6        12.1.6      1  1 -1  1  1
7      7        14.1.7      1  1  1 -1  1
8      8        15.1.8      1  1  1  1 -1
   run.no run.no.std.rp Blocks  A  B  C  D
9       9         1.2.1      2 -1 -1 -1 -1
10     10         4.2.2      2 -1 -1  1  1
11     11         6.2.3      2 -1  1 -1  1
12     12         7.2.4      2 -1  1  1 -1
13     13        10.2.5      2  1 -1 -1  1
14     14        11.2.6      2  1 -1  1 -1
15     15        13.2.7      2  1  1 -1 -1
16     16        16.2.8      2  1  1  1  1
class=design, type= FrF2.blocked
NOTE: columns run.no and run.no.std.rp  are annotation,  not part of the data frame
```

$ABCD$ is the only effect confounded with the block effect.

f)

```
> design2 <- FrF2(16, 4, blocks=4, alias.block.2fis=TRUE, randomize=FALSE)
> summary(design2)
Call:
FrF2(16, 4, blocks = 4, alias.block.2fis = TRUE, randomize = FALSE)

Experimental design of type  FrF2.blocked
16  runs
blocked design with  4  blocks of size  4

Factor settings (scale ends):
   A  B  C  D
1 -1 -1 -1 -1
2  1  1  1  1

Design generating information:
$legend
[1] A=A B=B C=C D=D
```

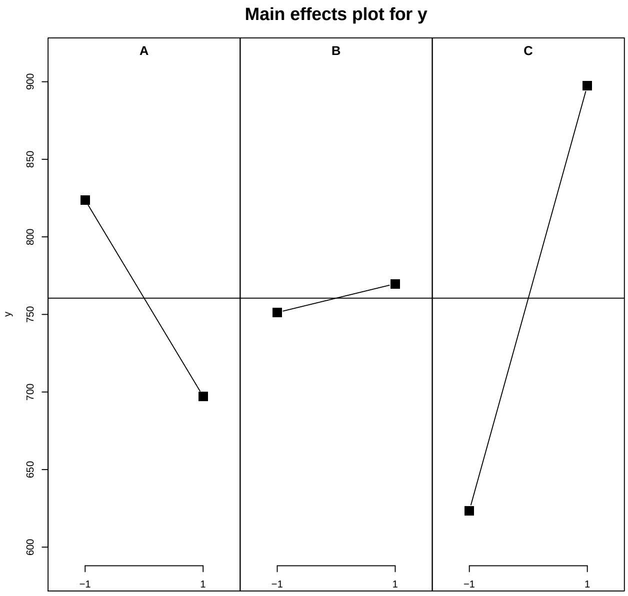
```
$`generators for design itself`
[1] full factorial

$`block generators`
[1] ACD BCD


no aliasing of main effects or 2fis  among experimental factors

Aliased with block main effects:
[1] AB

The design itself:
  run.no run.no.std.rp Blocks  A  B  C  D
1      1         1.1.1      1 -1 -1 -1 -1
2      2         4.1.2      1 -1 -1  1  1
3      3        14.1.3      1  1  1 -1  1
4      4        15.1.4      1  1  1  1 -1
  run.no run.no.std.rp Blocks  A  B  C  D
5      5         5.2.1      2 -1  1 -1 -1
6      6         8.2.2      2 -1  1  1  1
7      7        10.2.3      2  1 -1 -1  1
8      8        11.2.4      2  1 -1  1 -1
   run.no run.no.std.rp Blocks  A  B  C  D
9       9          6.3.1      3 -1  1 -1  1
10     10          7.3.2      3 -1  1  1 -1
11     11          9.3.3      3  1 -1 -1 -1
12     12         12.3.4      3  1 -1  1  1
   run.no run.no.std.rp Blocks  A  B  C  D
13     13          2.4.1      4 -1 -1 -1  1
14     14          3.4.2      4 -1 -1  1 -1
15     15         13.4.3      4  1  1 -1 -1
16     16         16.4.4      4  1  1  1  1
class=design, type= FrF2.blocked
NOTE: columns run.no and run.no.std.rp  are annotation,  not part of the data frame
```

ACD and BCD were suggested as block generators, so the four combinations of values of ACD and BCD are used to identify the four blocks. Now the third-order interactions ACD and BCD are confounded with the block effects, and also the second-order interaction $ACD{\cdot}BCD = AB$. But no main effects are confounded with the block effects.

Problem 3 Process development – from Exam TMA4255 2012 Summer

a) The coefficient estimate for B is $\hat{\beta}_{\mathrm{B}} = \frac{1}{n}\sum_{i=1}^{8} B_i y_i$, where y_i is the response in run i, and the effect estimate $\hat{B} = 2\hat{\beta}_{\mathrm{B}}$. In the language of two-level factorial designs:

$$\begin{aligned}\hat{B} &= \text{mean response when } B \text{ is high} - \text{mean response when B is low}\\ &= \tfrac{1}{4}(y_3 + y_4 + y_7 + y_8) - \tfrac{1}{4}(y_1 + y_2 + y_5 + y_6)\\ &= \tfrac{1}{4}(633 + 642 + 1075 + 729) - \tfrac{1}{4}(550 + 669 + 1037 + 749)\\ &= 769.75 - 751.25 = 18.5.\end{aligned}$$



The main effects plot for B shows that the mean response at the low B level is at 751.25, and going from the low to the high B level, the mean response increases with 18.5 to 769.75. The increase from the low to the high mean level of B is the B main effect.

b) The "Std. Error" column gives the estimated standard deviation of the regression coefficients. Let $\hat{\sigma}^2$ be the estimate of the variance σ^2 of the regression model. Due to the orthogonality of the DOE design, all estimated standard deviations are $\hat{\sigma}/\sqrt{n}$, where $n = 16$. From the printout we see that $\hat{\sigma} = 47.46$ (residual standard error) and Std.Error is then $47.46/\sqrt{16} = 11.865$ for all regression coefficients.

The estimated effect for B is by definition twice the estimated coefficient for B.

The t statistic is the coefficient estimate divided by its estimated standard error. For B is is $3.688/11.865 = 0.311$. The p-value given is for the test of the null hypothesis that the coefficient for the covariate B is zero, against the alternative that it is different from zero. A p-value of 0.76 implies that we do not reject the null hypothesis at significance level 0.05.

At the 005 level, the significant covariates are A, C and AC (and the intercept).

c) Since we have an orthogonal design, coefficient estimates will stay unchanged in any submodel. But the regression model has influence on the estimate of the error variance σ^2 and thus on estimates of coefficient estimator standard deviations.

Just looking at the estimated coefficients in the reduced model we see that the etching rate will increase with C and decrease with A. This would suggest to keep A at the low level and C at the high level. The interaction effect between A and C is negative, so with A at low level and C at high level the net effect is positive.

We may also calculate the estimated response (predictions) with the four combinations of A and C, which confirms that A low and C high is optimal:

$$\begin{aligned}
A \text{ low and } C \text{ low: } \hat{y} &= 776.062 + 50.812 - 153.062 - 76.812 = 597 \\
A \text{ low and } C \text{ high: } \hat{y} &= 776.062 + 50.812 - 153.062 + 76.812 = 1056.75 \\
A \text{ high and } C \text{ low: } \hat{y} &= 776.062 - 50.812 - 153.062 + 76.812 = 649 \\
A \text{ high and } C \text{ high: } \hat{y} &= 776.062 - 50.812 + 153.062 - 76.812 = 801.5
\end{aligned}$$

A $100(1-\alpha)\%$ prediction interval for a new response of an observation having covariates $\boldsymbol{x}_0$ has bounds $\boldsymbol{x}_0^{\mathrm{T}}\hat{\boldsymbol{\beta}} \pm t_{\alpha/2}\hat{\sigma}\sqrt{1+\boldsymbol{x}_0^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}\boldsymbol{x}_0}$ (see *Recommended exercises 7*, Problem 1). Here, $\boldsymbol{x}_0 = (1 \quad -1 \quad 1 \quad -1)^{\mathrm{T}}$ (intercept, A low, C high and thus AC low). $\hat{\boldsymbol{\beta}} = (776.06 \quad -50.81 \quad 153.06 \quad -76.81)^{\mathrm{T}}$ is the vector of coefficient estimates. Because of the orthogonal design, $(X^{\mathrm{T}}X)^{-1}$ is $\frac{1}{16}I$ with I a 4×4 identity matrix. The error standard deviation estimate $\hat{\sigma} = 41.69$ is read off the printout. With $\alpha = 0.05$ we find the critical value $t_{0.025} = 2.179$ (df $= n - 4 = 12$).

Then $\boldsymbol{x}_0^{\mathrm{T}}\hat{\boldsymbol{\beta}} \pm t_{\alpha/2}\hat{\sigma}\sqrt{1+\boldsymbol{x}_0^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}\boldsymbol{x}_0} = \boldsymbol{x}_0^{\mathrm{T}}\hat{\boldsymbol{\beta}} \pm t_{\alpha/2}\hat{\sigma}\sqrt{1+\boldsymbol{x}_0^{\mathrm{T}}\boldsymbol{x}_0/16} = 776.06 - (-50.81) + 153.06 - (-76.81) \pm 2.179 \cdot 41.69\sqrt{1+4/16} = 1056.7 \pm 101.6$, and we get the prediction interval $(955, 1158)$.

d) This is a half fraction of a 2^3 experiment, thus a 2^{3-1} experiment. The generator for the design is $AB = -C$, and the defining relation is thus $I = -ABC$. The alias stucture is: $A = -BC$, $B = -AC$, $C = -AB$. The defining relation has three letters, and thus this is a resolution III experiment.

Problem 4 Blocking

We first try using BC, CD and DE as block generators, thereby avoiding A. Then the block effects are confounded by the three mentioned two-factor interactions, and also with $BC \cdot CD = BD$, $BC \cdot DE = BCDE$, $CD \cdot DE = CE$, and $BC \cdot CD \cdot DE = BE$. The requirements are thus met.

There are other choices, e.g. block generators BD, CE and CD also satisfy the requirements. You can check by yourself.

But, actually also block generators including A may work: Let ABC, ACD and ADE be block generators. They are confounded by these three three-factor interactions, and with $ABC \cdot ACD = BD$, $ABC \cdot ADE = BCDE$, $ACD \cdot ADE = CE$, and $ABC \cdot ACD \cdot ADE = ABE$.

You can actually find many other choices that satisfy the requirements.

Problem 5 Fractional factorial design

a) $D = ABC$, so $1 = ABCD$, and the resolution (the minimum number of factors in the defining relation, which can consist of several equalities in general) is IV.

b) We have generators $E = ABC$, $F = ABD$, $G = ACD$ and $H = BCD$. Then we have the defining relation

$$\begin{aligned}
1 &= ABCE && \text{(first generator)} \\
&= ABDF = CDEF && \text{(second generator and product with previous)} \\
&= ACDG = BDEG \\
&\quad = BCFG = AEFG && \text{(third generator and product with previous)} \\
&= BCDH = ADEH = ACFH = BEFH = ABGH = CEGH = DFGH \\
&\quad = ABCDEFG && \text{(fourth generator and product with previous)}
\end{aligned}$$

The minimum length of the words is four, which means that the design is of resolution IV.

c) With AB a blocking factor, the two-factor interactions CE, DF and GH are confounded with the block effect in addition to some four-factor interactions and a six-factor interaction.