TMA4267 Linear Statistical Models
Part 4: Design of Experiments (DOE)
Solutions to recommended exercise 6 - V2017

February 20, 2017

Problem 1: Exam V2015, 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 $2 \cdot (-1/4) = -1/2$.

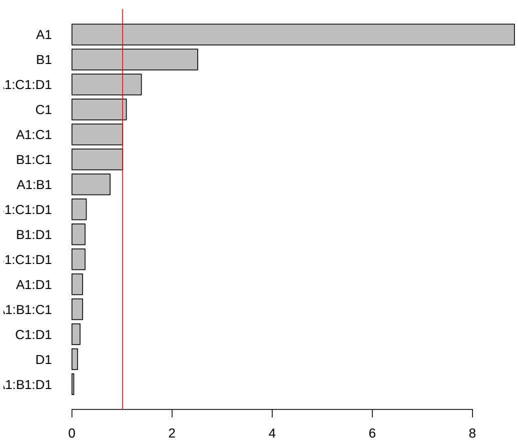
Problem 2: Factorial experiments

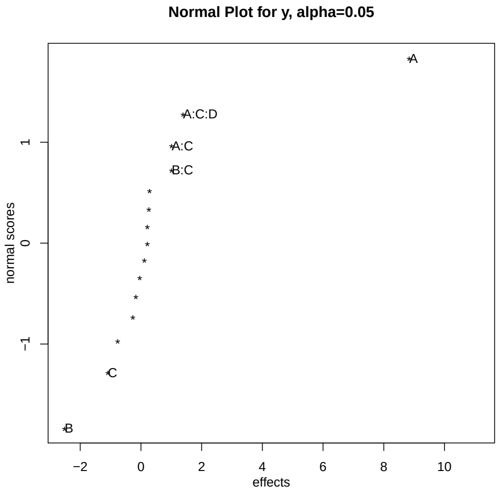
a)

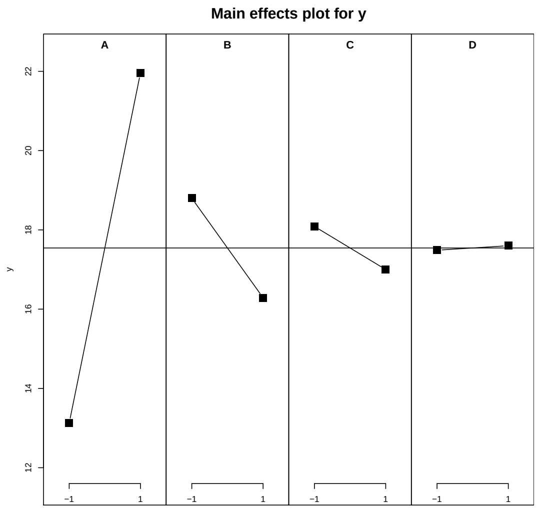
```
> library(FrF2)
> plan <- FrF2(nruns=16,nfactors=4,randomize=FALSE)
creating full factorial with  16  runs ...
> plan
    A  B  C  D
1  -1 -1 -1 -1
2   1 -1 -1 -1
3  -1  1 -1 -1
4   1  1 -1 -1
5  -1 -1  1 -1
6   1 -1  1 -1
7  -1  1  1 -1
8   1  1  1 -1
9  -1 -1 -1  1
10  1 -1 -1  1
11 -1  1 -1  1
12  1  1 -1  1
13 -1 -1  1  1
14  1 -1  1  1
15 -1  1  1  1
16  1  1  1  1
class=design, type= full factorial
```

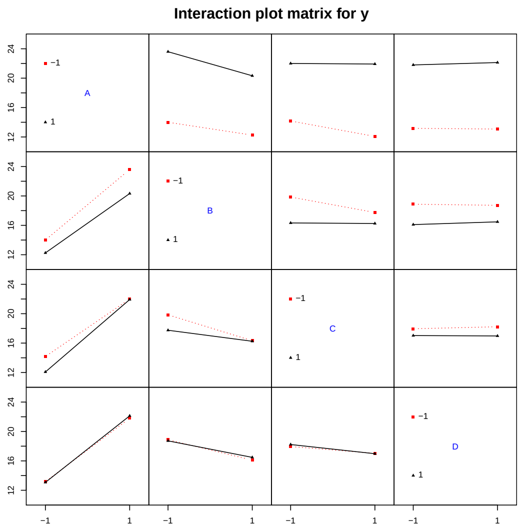
```
> y <- c(14.6,24.8,12.3,20.1,13.8,22.3,12.0,20.0,16.3,23.7,13.5,19.4,11.3,23.6,11.2,21.8)
> plan <- add.response(plan,y)
> plan
    A  B  C  D    y
1  -1 -1 -1 -1 14.6
2   1 -1 -1 -1 24.8
3  -1  1 -1 -1 12.3
4   1  1 -1 -1 20.1
5  -1 -1  1 -1 13.8
6   1 -1  1 -1 22.3
7  -1  1  1 -1 12.0
8   1  1  1 -1 20.0
9  -1 -1 -1  1 16.3
10  1 -1 -1  1 23.7
11 -1  1 -1  1 13.5
12  1  1 -1  1 19.4
13 -1 -1  1  1 11.3
14  1 -1  1  1 23.6
15 -1  1  1  1 11.2
16  1  1  1  1 21.8
class=design, type= full factorial
> lm4 <- lm(y~(.)^4,data=plan)
> effects <- 2*lm4$coeff
> 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
> anova(lm4) # to see the seqSS mentioned in the solutions to d)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value Pr(>F)
A          1 312.406 312.406
B          1  25.251  25.251
```

```
C           1   4.731   4.731
D           1   0.051   0.051
A:B         1   2.326   2.326
A:C         1   4.101   4.101
A:D         1   0.181   0.181
B:C         1   4.101   4.101
B:D         1   0.276   0.276
C:D         1   0.106   0.106
A:B:C       1   0.181   0.181
A:B:D       1   0.006   0.006
A:C:D       1   7.701   7.701
B:C:D       1   0.331   0.331
A:B:C:D     1   0.276   0.276
Residuals   0   0.000
Warning message:
In anova.lm(lm4) :
  ANOVA F-tests on an essentially perfect fit are unreliable
> DanielPlot(lm4)
> barplot(sort(abs(effects[-1]),decreasing=FALSE),las=1,horiz=TRUE)
> MEPlot(lm4)
> IAPlot(lm4)
```









$$
\begin{aligned}
\hat{A} &= 8.84 \\
\hat{B} &= -2.51 \\
\hat{C} &= -1.09 \\
\hat{D} &= 0.11 \\
\vdots\ &\ \vdots \\
\widehat{ABCD} &= -0.262
\end{aligned}
$$

From the Pareto and Daniel plots it looks like A and B are the most important factors.

b) The corresponding regression model is

$$
\begin{aligned}
Y = \beta_0 &+ \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 + \beta_4 z_4 && (1)\\
&+ \beta_{12} z_1 z_2 + \beta_{13} z_1 z_3 + \beta_{14} z_1 z_4 && (2)\\
&+ \beta_{23} z_2 z_3 + \beta_{24} z_2 z_4 + \beta_{34} z_3 z_4 && (3)\\
&+ \beta_{123} z_1 z_2 z_3 + \beta_{124} z_1 z_2 z_4 + \beta_{134} z_1 z_3 z_4 && (4)\\
&+ \beta_{234} z_2 z_3 z_4 + \beta_{1234} z_1 z_2 z_3 z_4 + \epsilon && (5)
\end{aligned}
$$

And the estimated effects are of the kind

$$\hat{A} = 2\hat{\beta}_1 \tag{6}$$

where $\hat{\beta}_1$ is the least squares estimator of β_1. Same goes for the other effects.

c) In the analysis in a) we have 16 equations and 16 coefficients to estimate. Therefore there are no degrees of freedom left to estimate the variance. If we assume that the variance is known it is possible to make inference about the effects. For factor A we have:

$$
\left.
\begin{aligned}
&\hat{A} = \tfrac{1}{8}(-Y_1 + Y_2 - \cdots - Y_{15} + Y_{16}) \\
&\mathrm{Var}(\hat{A}) = \tfrac{1}{64} 16\sigma^2 = \tfrac{\sigma^2}{4}
\end{aligned}
\right\} \Rightarrow \hat{A} \sim N\left(\mu_A, \frac{\sigma^2}{4}\right)
$$

95 % confidence interval for μ_A:

$$\hat{A} \pm z_{0.025}\frac{\sigma}{2} = (6.88, 10.80)$$

95 % confidence interval for μ_B:

$$\hat{B} \pm z_{0.025}\frac{\sigma}{2} = (-4.47, -0.5)$$

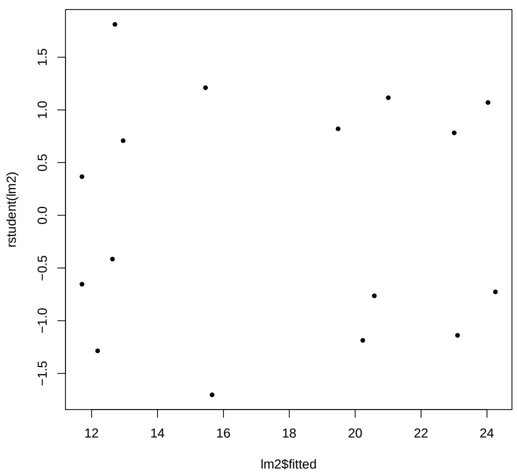
```
> nruns <- 16
> sigma <- 2
> sigmaeff <- sqrt(4*sigma^2/nruns)
> sigmaeff
[1] 1
```

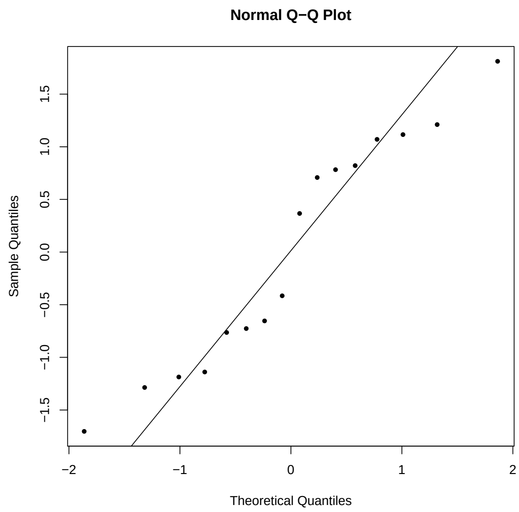
```
> CIefflower <- effects-1.96*sigmaeff
> CIeffupper <- effects+1.96*sigmaeff
> cbind(effects,CIefflower,CIeffupper)
            effects CIefflower CIeffupper
(Intercept) 35.0875    33.1275    37.0475
A1           8.8375     6.8775    10.7975
B1          -2.5125    -4.4725    -0.5525
C1          -1.0875    -3.0475     0.8725
D1           0.1125    -1.8475     2.0725
A1:B1       -0.7625    -2.7225     1.1975
A1:C1        1.0125    -0.9475     2.9725
A1:D1        0.2125    -1.7475     2.1725
B1:C1        1.0125    -0.9475     2.9725
B1:D1        0.2625    -1.6975     2.2225
C1:D1       -0.1625    -2.1225     1.7975
A1:B1:C1     0.2125    -1.7475     2.1725
A1:B1:D1    -0.0375    -1.9975     1.9225
A1:C1:D1     1.3875    -0.5725     3.3475
B1:C1:D1     0.2875    -1.6725     2.2475
A1:B1:C1:D1 -0.2625    -2.2225     1.6975
```

d) If there are good reasons to assume that the 3- and 4-factor interactions are 0, we have enough degrees of freedom to estimate the variance.

```
> 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

> anova(lm2)
Analysis of Variance Table
```

```
Response: y
          Df  Sum Sq Mean Sq  F value    Pr(>F)
A          1 312.406 312.406 183.9168 3.906e-05 ***
B          1  25.251  25.251  14.8653   0.01193 *
C          1   4.731   4.731   2.7850   0.15602
D          1   0.051   0.051   0.0298   0.86971
A:B        1   2.326   2.326   1.3691   0.29470
A:C        1   4.101   4.101   2.4141   0.18096
A:D        1   0.181   0.181   0.1063   0.75756
B:C        1   4.101   4.101   2.4141   0.18096
B:D        1   0.276   0.276   0.1623   0.70373
C:D        1   0.106   0.106   0.0622   0.81300
Residuals  5   8.493   1.699
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> effects <- lm2$coeff
> plot(lm2$fitted,rstudent(lm2),pch=20)
> qqnorm(rstudent(lm2),pch=20)
> qqline(rstudent(lm2))
```





We see that the estimator for σ^2 is now:

$$s^2 = MSE = 1.699$$

from the anova printout above, look for Mean Square Residuals.

It is also possible to see this from the print-out from anova under a)

$$s^2 = \frac{SS_{ABC} + \cdots + SS_{BCD} + SS_{ABCD}}{5} = \frac{0.181 + 0.006 + 7.701 + 0.331 + 0.276}{5} = 1.699$$

where 8.22 is 3-way Seq SS (sum of the first 4 numbers above), and 0.276 is 4-way Seq SS from the full analysis in section a). The variance of the effects is thus estimated by

$$s^2_{effect} = \frac{4s^2}{n} = 0.425$$

We can also obtain this estimate of σ^2_{effect} directly by using the estimated effects

$$s^2_{effect} = \frac{\widehat{ABC}^2 + \cdots + \widehat{BCD}^2 + \widehat{ABCD}^2}{5} = \frac{0.213^2 + 0.038^2 + 1.387^2 + 0.288^2 + 0.263^2}{5} = 0.425$$

Now we can do a T-test or an equivalent F-test to decide which of the effects are significant. This may be read off directly from the anova(lm2) printout above.

We use the results and do an F-test:

$$F_A = \frac{MSA}{MSE} = \frac{n\hat{A}^2/4}{1.699} = \frac{312.4}{1.699} = 184$$

$$F_B = \frac{MSB}{MSE} = \frac{25.251}{1.699} = 14.87,$$

and get the p-values:

$$p = P(F_{1,5} > 183.9) \approx 0$$

$$p = P(F_{1,5} > 14.87) = 2P(T_5 > 3.85) = 0.012$$

Either use the t-distribution since

$$F_{1,\nu} = T_\nu^2$$

or use the F-distribution directly.

We conclude that both A and B are significant at level 0.05.

e

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

We see that ABCD is the only effect confounded with the block effect

f To perform the experiment in four blocks, we need two generators. Choosing ABC and AD as generators gives

$$ABC \cdot AD = BCD \tag{7}$$

$$ABC \cdot BCD = AD \tag{8}$$

$$AD \cdot BCD = ABC \tag{9}$$

And we see that the effects confounded with the blocks are ABC, BCD and AD. This design avoids main efffects being confounded with the block effect.

FrF2 choose default a different blocking (design2 below), but can be forced to choose the same as above (design3 below).

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

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        15.1.4      1  1  1  1 -1
2      2        14.1.3      1  1  1 -1  1
3      3         4.1.2      1 -1 -1  1  1
4      4         1.1.1      1 -1 -1 -1 -1
  run.no run.no.std.rp Blocks  A  B  C  D
5      5        11.2.4      2  1 -1  1 -1
6      6        10.2.3      2  1 -1 -1  1
7      7         8.2.2      2 -1  1  1  1
8      8         5.2.1      2 -1  1 -1 -1
   run.no run.no.std.rp Blocks  A  B  C  D
9       9         7.3.2      3 -1  1  1 -1
10     10         6.3.1      3 -1  1 -1  1
11     11        12.3.4      3  1 -1  1  1
12     12         9.3.3      3  1 -1 -1 -1
   run.no run.no.std.rp Blocks  A  B  C  D
13     13        16.4.4      4  1  1  1  1
14     14        13.4.3      4  1  1 -1 -1
15     15         2.4.1      4 -1 -1 -1  1
16     16         3.4.2      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
> design.info(design2)$aliased.with.blocks
$aliased.with.blocks
[1] "AB"

> design3 <-FrF2(16,4,blocks=c("ABC","AD"),alias.block.2fis=TRUE)
> summary(design3)
Call:
FrF2(16, 4, blocks = c("ABC", "AD"), alias.block.2fis = TRUE)

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

> design.info(design3)$aliased.with.blocks
[1] "AD"
```

Problem 3: Process development

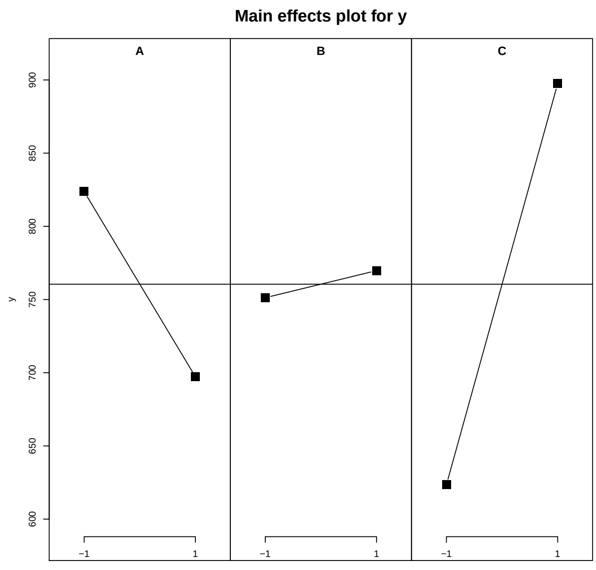
Run	A	B	C	Response
1	-1	-1	-1	550
2	1	-1	-1	669
3	-1	1	-1	633
4	1	1	-1	642
5	-1	-1	1	1037
6	1	-1	1	749
7	-1	1	1	1075
8	1	1	1	729

Intercept	A	B	C	AB	AC	BC	ABC
760.50	-126.5	*	274.0	-42.0	-190.5	-9.5	13.0

a) Let y_i be the response in run i.

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

The main effects plot for B shows that the mean B response at the low level is at 751.25, and going from the low to the high level the mean B response increaes 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 s^2 be the estimated variance in the regression model (estimate for σ^2). Due to the orthogonality of the DOE design all estimated standard deviations are $s/\sqrt{n}$ where

$n = 16$. From the printout we see that $S = 47.46$ (residual standard error) and Std.Error is then $47.46/4 = 11.865$ for all regression coefficients.

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

The Estimate is the estimated regression coefficient, the Std.Error is the estimated standard deviation of the regression coefficient, the t-value is the value of the t-statistics (see below), the p-value is from the test described below.

The t-statistic: Estimate/Std.Error=3.688/11.865=0.311.
H_0: The coefficient for the covariate B is zero, H_1: different from zero. A p-value of 0.76 means that we do not reject H_0 at significance level 0.05 and assume that the B coefficient is zero - and can be removed from the model.

What are the significant covariates in the model? Significant covariates are A, C and AC (and the intercept).

c) Since we have an orthogonal design the presence of factors othogonal to A and C does not change the parameter estimates for the regression coefficients in the model. But, the regression model is important for the estimation of the error variance σ^2 and the Std.Error will then change with the change in the model.

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 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.

A low and C low: $\hat{y} = 776.062 + 50.812 - 153.062 - 76.812 = 597$.
A low and C high: $\hat{y} = 776.062 + 50.812 - 153.062 + 76.812 = 1056.75$.
A high and C low: $\hat{y} = 776.062 - 50.812 - 153.062 + 76.812 = 649$.
A high and C high: $\hat{y} = 776.062 - 50.812 + 153.062 - 76.812 = 801.5$.

Calculate a 95% prediction interval for the etch rate based on your chosen levels for A and C. Since we have an othogonal design, the covariance matrix for the regression coefficients will be diagonal (all correlations are zero). The formula for the prediction interval with covariates $\boldsymbol{x_0}$ is

$$[\boldsymbol{x}_0^T\boldsymbol{B} \pm t_{n-k-1}(\frac{\alpha}{2})\sqrt{(1+\boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0)s^2}]$$

The covariate vector is $x_0 = (1, -1, 1, -1)$ for the intercept, A at low and C at high and thus AC at low level. $\boldsymbol{B}$ is the vector of regression coefficients for the intercept, A, C and AC, thus $(776.06, -50.81, 153.06, -76.81)$. The matrix $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ is a diagonal matrix with 1/16 on the diagonal. s is read off the printout as 41.96. The t critical value is $t_{16-3-1}(0.25) = t_{12}(0.25) = 2.18$.

$\boldsymbol{x}_0^T\boldsymbol{B} = y_0 = 1056.75$ and we add
$2.18 \cdot \sqrt{1 + (1, -1, 1, -1)diag(1/16)(1, -1, 1, -1)} \cdot 41.96 = 2.18 \cdot \sqrt{1 + 4/16} \cdot 41.96 = 102.3$. The interval is then $[954.45, 1159.05]$.

d) We now assume that in a pilot study with three factors only runs 1, 4, 6 and 7 from the table in the start of this problem were performed.

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

For 2^5 experiments, we have five factors: A B C D and E. The requirements are as follows: no main effect and the two-factor interactions involved factor A: AB, AC, AE, AD, and AE should not be confounded with the block-effects.

For DOE blocks, there is no general method for how to choose the blocking factors. However, in this problem, as we can see, no two-factor interactions involved with A should be confounded. This can give us the first impression that we only use interactions involved with B, C, D and E for blocking. Let B1, B2, and B3 be threse block generators. (Remember that we may produce 8 blocks from three block generators by letting the block be defined by the 8 combinations of -1 and 1 for the three block generators, see page 16 of the DOE note).

For instance we can try with B1=BC, B2=CD, B3=DE, which gives us
B1B2=BD
B1B3=BCDE
B2B3=CE
B1B2B3=BE

Similarly, blocking factors such as B1=BD, B2=CE, B3=CD also satisfies the requirement, you can check by yourself.

We may think about the factor A now:
B1=ABC, B2=ACD, B3=ADE will also satisfy the requirements since
B1B2=BD
B1B3=BCDE
B2B3=CE
B1B2B3=ABE.

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

Problem 5: Design resolution

a) D=ABC, I=ABCD. Therefore the resolution is IV. The resolution is the length of the shortes defining relation.

b) E=ABC
F=ABD
G=ACD
H=BCD
which gives I=ABCE=ABDF=ACDG=BCDH

The additional words are obtained from
I^2= CDEF=BDEG=ADEH=BCFG=ACFH=ABGH

I^3=AEFG=BEFH=CEGH=DFGH
I^4=ABCDEFGH

None of the words have shorter length than four which means that the design is of resolution IV.

c) With B1=AB we get that the two-factor interactions CE, DF and GH also are confounded with the block effect in addition to some four-factor interactions and a six-factor interaction.

d) It is possible to investigate 16 factors in 32 runs and still have a resolution IV design. This can be seen as follows. A fold-over of a resolution III design becomes a resolution IV design. In 16 runs it is possible to construct a resolution III design in 15 factors. Adding a column of plus 1's and then do the folding gives us a resolution IV design for 16 factors in 32 runs.