Vejl. besvarelse af opgave 2.1

a) Indlæsning af data. Ved at brug af `attach` sikres, at vi kan benytte variabelnavnene `DIET,` `SEX` og `energioms` i R ved senere analyser.

```
> kost<-read.table(file="FP270505.txt",header=T)
> kost
   SEX DIET energioms
1    M    1     12389
2    M    1     13519
            ...
[ more datalines here ]
            ...
17   F    3      9130
18   F    3      9214
> attach(kost)
```

b) Lav `DIET` om til en faktor.

```
> DIET ### not a factor
 [1] 1 1 1 2 2 2 3 3 3 1 1 1 2 2 2 3 3 3
Levels: 1 2 3
> DIET<-factor(kost$DIET)
> DIET ### now a factor
 [1] 1 1 1 2 2 2 3 3 3 1 1 1 2 2 2 3 3 3
Levels: 1 2 3
```

c) Kommandoen `SEX:DIET` laver en ny faktor (produktfaktoren) med et antal niveauer svarende til alle kombinationer af `SEX` og `DIET`.

```
> SEX:DIET
 [1] M:1 M:1 M:1 M:2 M:2 M:2 M:3 M:3 M:3 F:1 F:1 F:1 F:2 F:2 F:2 F:3 F:3 F:3
Levels: F:1 F:2 F:3 M:1 M:2 M:3
> is.factor(SEX:DIET)
[1] TRUE
```

d) Den statistiske model (1) kan f.eks. fittes som anført nedenfor.

```
> model<-lm(energioms~SEX:DIET)
> model

Call:
lm(formula = energioms ~ SEX:DIET)

Coefficients:
```

```
(Intercept)   SEXF:DIET1   SEXM:DIET1
    12142.7      -2430.3        650.0
 SEXF:DIET2   SEXM:DIET2   SEXF:DIET3
    -2208.0       -997.3      -2974.3
 SEXM:DIET3
         NA
```

Samme model kan fittes ved at skrive et af følgende udtryk, men bemærk at fortolkningen af parameterestimaterne afhænger af, hvilket udtryk der vælges.

```
> model<-lm(energioms~SEX*DIET)
> model<-lm(energioms~SEX+DIET+SEX*DIET)
> model<-lm(energioms~SEX:DIET-1)
```

e) De efterspurgte estimater bliver som følger (-se ovenfor)

$$\begin{aligned} \hat{\gamma}(\texttt{M,3}) &= 12142.7 \\ \hat{\gamma}(\texttt{F,2}) &= 12142.7 - 2208.0 = 9934.7 \end{aligned}$$

f) Estimat for residualspredning $\hat{\sigma} = 868.5$ fås ved kommandoen

```
> summary(model)

Call:
lm(formula = energioms ~ SEX + DIET + SEX * DIET)
               ...
    [more output here]
               ...
Residual standard error: 868.5 on 12 degrees of freedom
```

g) Kommandoen

```
> interaction.plot(DIET,SEX,energioms)
```

laver et interaction-plot, der kan benyttes til grafisk at vurdere, om der er en vekselvirkning mellem `DIET` og `SEX`. Det er ret svært alene på baggrund af plottet, at sige om vekselvirkningen er signifikant, men hvis man ved et test har eftervist en vekselvirkning, kan plottet ofte benyttes til at forklare *retningen* af denne.

h) Kommandoen `model1`**<-lm(energioms**`~SEX`+`DIET`**)** fitter den additive model for tosidet variansaanslyse:

$$Y_i = \alpha(\texttt{SEX}_i) + \beta(\texttt{DIET}_i) + e_i, \quad e_i \sim N(0, \sigma^2).$$

i) Estimaterne for den additive model udskrives på skærmen:

```
> model1<-lm(energioms~SEX+DIET)
> model1

Call:
lm(formula = energioms ~ SEX + DIET)

Coefficients:
(Intercept)         SEXM        DIET2        DIET3
    10041.6       2421.8       -712.5       -597.0
```

Estimatet for `SEX=M` og `DIET=3` bliver

$$10041.6 + 2421.8 - 597.0 = 11866.4.$$

Estimatet for `SEX=F` og `DIET=2` bliver

$$10041.6 - 712.5 = 9329.1.$$

j) Ensidede variansanalysemodeller fittes:

```
> modelS<-lm(energioms~SEX)
> modelD<-lm(energioms~DIET)
```

k) Først testes om vi kan se bort fra vekselvirkningen `SEX:DIET`.

```
> anova(model1,model)
Analysis of Variance Table

Model 1: energioms ~ SEX + DIET
Model 2: energioms ~ SEX + DIET + SEX * DIET
  Res.Df      RSS Df Sum of Sq      F Pr(>F)
1     14 12360169
2     12  9051464  2   3308705 2.1933 0.1542
```

Modelreduktionen godkendes: $F = 2.1933, p = 0.1542$. Dernæst testes om vi kan fjerne hovedeffekten af `SEX`.

```
> anova(modelD,model1)
Analysis of Variance Table

Model 1: energioms ~ DIET
Model 2: energioms ~ SEX + DIET
  Res.Df      RSS Df Sum of Sq      F    Pr(>F)
1     15 38752703
2     14 12360169  1  26392534 29.894 8.291e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Testet forkastes: $F = 29.894, p < 0.0001$. I stedet testes om vi kan se bort fra effekten af `DIET`.

```
> anova(modelS,model1)
Analysis of Variance Table

Model 1: energioms ~ SEX
Model 2: energioms ~ SEX + DIET
  Res.Df      RSS Df Sum of Sq      F Pr(>F)
1     16 14114980
2     14 12360169  2   1754811 0.9938 0.3948
```

Modelreduktionen godkendes: $F = 0.9938, p = 0.3948$. Vores slutmodel bliver den ensidede variansanalysemodel

$$Y_i = \alpha(\texttt{SEX}_i) + e_i, \quad e_i \sim N(0, \sigma^2).$$

Parameterestimater og konfidensintervaller fås som følger. Bemærk, at vi genfitter modellen uden intercept, således at parameterestimaterne kan aflæses direkte fra R-udskriften.

```
> modelSny<-lm(energioms~SEX-1)
> summary(modelSny)

Coefficients:
     Estimate Std. Error t value Pr(>|t|)
SEXF   9605.1      313.1   30.68 1.21e-15 ***
SEXM  12026.9      313.1   38.41  < 2e-16 ***

Residual standard error: 939.2 on 16 degrees of freedom

> confint(modelSny)
         2.5 %   97.5 %
SEXF  8941.406 10268.82
SEXM 11363.183 12690.59
```

Vejl. besvarelse af opgave 2.2

a) Benyt f.eks. kommandoen

```
> terbuthyl<-read.table(file="Ex34.txt",header=T)
> attach(terbuthyl)
> terbuthyl
```

b) Kommandoen optæller hvor mange gange hver af de fire kombinationer af `TEMP` og `LUC` optræder i datasættet.

c) Kommandoen

```
> table(TEMP,LUC,ADP)
, , ADP = 0
```

```
    LUC
TEMP 0 1
  10 2 2
  20 2 2

, , ADP = 1

    LUC
TEMP 0 1
  10 2 2
  20 2 2
```

viser, at hver kombation af de tre faktorer optræder netop to gange i datasættet.

d) Løsning:

```
> TEMP<-factor(TEMP)
> LUC<-factor(LUC)
> ADP<-factor(ADP)
```

e)-g) For at sikre dig selv, at du har forstået output, kan du for eksempel prøve at udregne estimatet hørende til gruppen `TEMP`=10, `LUC`=1 og `ADP`=1 baseret på output for `modelA` hhv. `modelB`.

```
> modelA

Call:
lm(formula = mineral ~ TEMP * LUC * ADP)

Coefficients:
     (Intercept)            TEMP20              LUC1              ADP1
           2.520             0.980            -0.240             2.865
     TEMP20:LUC1       TEMP20:ADP1         LUC1:ADP1  TEMP20:LUC1:ADP1
           0.275            -0.660            -0.075            -0.265

> modelB<-lm(mineral~TEMP:LUC:ADP)
> modelB

Call:
lm(formula = mineral ~ TEMP:LUC:ADP)

Coefficients:
     (Intercept)  TEMP10:LUC0:ADP0  TEMP20:LUC0:ADP0  TEMP10:LUC1:ADP0
           5.400            -2.880            -1.900            -3.120
TEMP20:LUC1:ADP0  TEMP10:LUC0:ADP1  TEMP20:LUC0:ADP1  TEMP10:LUC1:ADP1
          -1.865            -0.015             0.305            -0.330
TEMP20:LUC1:ADP1
              NA
```

h) Der er lidt valgfrihed mht. i hvilken rækkefølge, man fjerner de forskellige faktorer.

```
> model1<-lm(mineral~TEMP*ADP+TEMP*LUC+LUC*ADP)

> anova(model1,modelB) ### TEMP:ADP:LUC fjernes: F=0.3609, p=0.5646
Analysis of Variance Table

Model 1: mineral ~ TEMP * ADP + TEMP * LUC + LUC * ADP
Model 2: mineral ~ TEMP:LUC:ADP
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1      9 0.40671
2      8 0.38915  1   0.01756 0.3609 0.5646

> model2<-lm(mineral~TEMP*ADP+TEMP*LUC)

> anova(model2,model1) ### LUC:ADP fjernes: F=0.9528, p=0.3545
Analysis of Variance Table

Model 1: mineral ~ TEMP * ADP + TEMP * LUC
Model 2: mineral ~ TEMP * ADP + TEMP * LUC + LUC * ADP
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     10 0.44976
2      9 0.40671  1   0.04306 0.9528 0.3545

> model3<-lm(mineral~TEMP*ADP+LUC)

> anova(model3,model2) ## TEMP:LUC fjernes: F=0.4515, p=0.5169
Analysis of Variance Table

Model 1: mineral ~ TEMP * ADP + LUC
Model 2: mineral ~ TEMP * ADP + TEMP * LUC
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     11 0.47007
2     10 0.44976  1   0.02031 0.4515 0.5169

> model4<-lm(mineral~TEMP*ADP)

> anova(model4,model3) ## LUC fjernes: F=3.9818, p=0.07136
Analysis of Variance Table

Model 1: mineral ~ TEMP * ADP
Model 2: mineral ~ TEMP * ADP + LUC
  Res.Df     RSS Df Sum of Sq      F  Pr(>F)
1     12 0.64022
2     11 0.47007  1   0.17016 3.9818 0.07136 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

i)-j) Parameterestimater fås som følger. Bemærk, at vi genfitter modellen uden intercept, således at parameterestimaterne kan aflæses direkte fra R-udskriften.

```
> model4ny<-lm(mineral~TEMP:ADP-1)
> summary(model4ny)

Call:
lm(formula = mineral ~ TEMP:ADP - 1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
TEMP10:ADP0   2.4000     0.1155   20.78 8.91e-11 ***
TEMP20:ADP0   3.5175     0.1155   30.46 9.84e-13 ***
TEMP10:ADP1   5.2275     0.1155   45.26 8.82e-15 ***
TEMP20:ADP1   5.5525     0.1155   48.08 4.29e-15 ***

Residual standard error: 0.231 on 12 degrees of freedom
```

k) Residualspredningen under slutmodellen estimateres til $s = 0.231$ og estimatet har 12 frihedsgrader. Da der er fire observationer for hver kombination af `TEMP:ADP` bliver LSD-værdien for sammenligning af grupperne givet ved produktfaktoren `TEMP*ADP`

$$\text{LSD} = t_{0.975,12} \cdot s \cdot \sqrt{\frac{1}{4} + \frac{1}{4}} \approx 0.3559.$$

Vejl. besvarelse af opgave 2.3

a) Den statistiske model er en ensidet variansanalyse:

$$Y_i = \alpha(\texttt{Nitrogen}_\texttt{i}) + \texttt{e}_\texttt{i},$$

hvor $e_1, \ldots, e_{16}$ er uafhængige og normalfordelte $N(0, \sigma^2)-$fordelte.

b) Estimaterne for effekten af gødningstyperne aflæses af første søjle i taballen øverst på opgaveformuleringens side 5: For type A bliver estimatet 67.900. Estimat for spredning: $s = \hat{\sigma} = 6.083$.

c) På kompendiets side 28 formel (3.8) er angivet, hvordan man bestemmer konfidensintervallerne for estimaterne i en ensidet variansanalyse. For gødningstype C og N fås konfidensintervaller

$$\begin{array}{ll} \texttt{NitrogenC} & 77.000 \pm t_{0.975,12} \cdot s/\sqrt{4} \approx 77.000 \pm 6.627 \\ \texttt{NitrogenN} & 85.400 \pm t_{0.975,12} \cdot s/\sqrt{4} \approx 85.400 \pm 6.627. \end{array}$$

d) På kompendiets side 27 er angivet formlen for LSD-værdien for sammenligning af grupperne og vi får

$$t_{0.975,12} \cdot s \cdot \sqrt{\frac{2}{4}} \approx 9.372.$$

Denne værdi sammenholdes med forskellen mellem parameterestimater, som er $85.4000 - 77.000 = 8.400$ og dermed *mindre* end LSD-værdien. Den ensidede variansanalysemodel viser altså ikke forskel på effekterne af gødningstyperne C og N.

e) Faktordiagrammet hørende til forsøget bliver:

$$
\begin{array}{ccccc}
 & & & \mathtt{MARK}^{4}_{3} & \\
\mathtt{I}^{16}_{0} & \mathtt{MARK} \times \mathtt{NITROGEN}^{16}_{9} & & & \mathtt{0}^{1}_{1} \\
 & & & \mathtt{NITROGEN}^{4}_{3} &
\end{array}
$$

f) Da der kun er en gentagelse af forsøget for hver af de 16 kombinationer af produktfaktoren `MARK` $\times$ `NITROGEN`, er det ikke muligt at teste vekselvirkningen. Man må i stedet tage udgangspunkt i den additive model.

g) Da kurverne på de to interaction plots (næsten) er parallele tyder det på, at der ikke er vekselvirkning mellem `MARK` og `NITROGEN`.

h) R-programmet tager her udgangspunkt i den additive model med faktorerne `MARK` og `NITROGEN`, som formelt kan anføres ved:

$$Y_i = \alpha(\mathtt{MARK_i}) + \beta(\mathtt{NITROGEN_i}) + \mathtt{e_i},$$

hvor $e_1, \ldots, e_{16}$ er uafhængige og normalfordelte $N(0, \sigma^2)$.

i) LSD-værdierne for sammenligning mellem effekten af to gødningstyper er angivet på side 34 i kompendiet. Således finder vi, at

$$LSD_{\mathtt{NITROGEN}} = t_{0.975,9} \cdot s \cdot \sqrt{\frac{2}{4}} \approx 2.262 \cdot 4.92 \cdot \sqrt{\frac{1}{2}} \approx 7.870.$$

Bemærk, at både estimatet for spredningen og antallet af frihedsgrader har ændret sig i forhold til i spm. d), siden vi nu tager udgangspunkt i den additive model med faktorerne `MARK` og `NITROGEN`. Begge dele kan dog findes på R-udskriften på side 6 i opgaveformuleringen. Bemærk, at forskellen på 8.400 mellem parameterestimaterne for gødningstyperne C og N i dette spørgsmål *overskrider* LSD-værdien. Hvis vi tager udgangpunkt i modellen, hvor begge (relevante) faktorer inddrages finder vi mao. en forskel på de to typer gødning. I spm. d) "drukner" forskellen mellem gødningstyper i den variation, der er mellem de enkelte marker, og som vi ikke tager højde for i den statistiske analyse.

Vejl. besvarelse af opgave 2.4

1) Figur 2 er et interaction plot til undersøgelse af vekselvirkningen mellem faktorerne antibiotika (`ANTI`) og niveau for vitamin B_{12} (`VITA`). For hvert niveau af `VITA` er gennemsnittet af vægtforøgelsen tegnet op imod niveauet af `ANTI`.

2) Man benytter den fulde tosidede variansanalysemodel til at beskrive data. Således beskrives vægtforøgelsen (Y_i) for de enkelte grise som:

$$Y_i = \alpha(\mathtt{ANTI} \times \mathtt{VITA_i}) + \mathtt{e_i},$$

hvor $e_1, \ldots, e_{12}$ er uafhængige og normalfordelte $N(0, \sigma^2)$. Modellen er i opgaveformuleringen fitted med R-kommandoen

```
model<-lm(wi~ANTI*VITA).
```

3) Analyse af den tosidede variansanalysemodel er beskrevet i kompendiets kapitel 3.2 og man kan med fordel opsummere analysen i et skema som nedenfor:

Faktor	SS_e	df_e	Test	F	p-værdi
`ANTI` $\times$ `VITA`	0.02933	8			
`ANTI` $+$ `VITA`	0.1728	1	Fjern vekselvirk.	47.1273	1.29e-04

Da testet bliver forkastet, er det meningsløst at begynde at teste hovedvirkningerne væk.

4) Da vi i spm. 3) ikke kan fjerne vekselvirkningen, bliver vores slutmodel den fulde tofaktormodel beskrevet under spm. 2). Parameterestimaterne kan aflæses af R-udskrifterne i opgaveformuleringen efter kommandoen

```
summary(model).
```

Udfordringen ligger i at finde ud af, hvordan R vælger at parametrisere modellen. Interceptet svarer her til gruppen `ANTI` = `A1` og `VITA` = `B1`, og samtlige parameterestimater bliver:

`ANTI`	`VITA`	Parameterestimat
A1	B1	1.1900
A2	B1	1.0333(=1.1900-0.1567)
A1	B2	1.2200(=1.1900+0.0300)
A2	B2	1.5433(=1.1900-0.1567+0.0300+0.4800)

Estimatet for spredningen er $s = \hat{\sigma} = 0.06055$ og længden af 95%−konfidensintervaller for estimaterne af middelværdierne i de 4 grupper givet ved `ANTI`×`VITA` bliver:

$$t_{0.975,8} \cdot s/\sqrt{3} = 2.306 \cdot 0.0605/\sqrt{3} \approx 0.0806.$$

5) Forsøget viser, at antibiotika har en positiv på grisenes vækst, men kun hvis der samtidig tilføres B_{12} vitamin til kosten. Rent faktisk ses en svagt signifikant ($p = 0.01322$) negativ effekt af antibiotika for gruppen af grise, som ikke tilføres B_{12} med kosten.