Computerøvelser

Formålet med følgende opgaver er at sætte dig i stand til at lave variansanalyse i R, når der indgår flere faktorer i datasættet. Du bør opnå fortrolighed med følgende punkter

- Hvordan bruger jeg `lm()` til at specificere modelller, hvori der indgår flere faktorer?
- Hvordan undersøger jeg, om der er vekselvirkning i en tosidet variansanalyse herunder
  - brug af `anova()`
  - brug af `interaction.plot()`
  - variansanalyseskemaer og testskema på baggrund af R udskrifter
- Hvordan finder jeg parameterestimater under den additive model for tosidet variansanalyse?

Opgave 2.1: Tosidet variansanalyse i R.

Denne opgave er baseret på opgave 1 fra *Eksamen i statistisk forsøgsplanlægning, 27. maj 2005.*

I et kostforsøg afprøvedes effekten af tre forskellige diæter på forsøgspersoners energiomsætning (kJ/24 timer), som blev målt efter at personen havde været en periode på en blandt tre forskellige diæter. For hver diæt indgik 3 mænd og 3 kvinder i forsøget, der således totalt talte 18 forsøgspersoner.

a) Indlæs datasættet fra Excel-arket `FP270505.xls` eller fra tekstfilen `FP270505.txt` i R og gem det som `kost`.

b) Datasættet `kost` indeholder variablene `SEX` og `DIET`. Benyt om nødvendigt kommandoen `factor()` til at lave numeriske variable om til faktorer.

c) Diskuter effekten af kommandoen `SEX:DIET`. Benyt `is.factor()` til at undersøge, om resultatet af R opfattes som en faktor eller en numerisk variabel.

Den statistiske model

$$Y_i = \gamma(\texttt{SEX} \times \texttt{DIET}_i) + e_i, \tag{1}$$

hvor $e_1, \ldots, e_{18}$ er uafhængige og normalfordelte $\sim N(0, \sigma^2)$, beskriver modellen, hvor der er vekselvirkning mellem `SEX` og `DIET`

d) Benyt kommandoen `lm()` til at fitte den statistiske model (1) og gem resultatet som `model`. Udskriv `model` på skærmen.

e) Hvad er parameterestimatet for værdien af energiomsætningen for mænd, som har fået diæt nr. 3 og for kvinder, som har fået diæt nr. 2?

f) Hvad er estimatet for spredningen ($s = \hat{\sigma}$) ?

g) Diskuter effekten af kommandoen `interaction.plot(DIET, SEX, energioms)`. Hvordan skal resultatet fortolkes?

h) Hvilken statistisk model fittes med `model1<-lm(energioms~SEX+DIET)`?

i) Hvad er parameterestimatet for værdien af energiomsætningen for mænd, som har fået diæt nr. 3 og for kvinder, som har fået diæt nr. 2?

j) Fit den ensidede variansanalysemodel svarende til hver af faktorerne `SEX` og `DIET` og gem resultaterne som `modelS` hhv. `modelD`.

k) Diskuter på baggrund af kommandoerne `anova(model1,model)`, `anova(modelD,model1)` og `anova(modelS,model1)` hvilket slutmodel den tosidede variansanalyse af `kost` fører frem til.

Opgave 2.2: Trefaktorforsøg i R.

I denne opgave betragtes datasættet fra lærebogens exercise 3.4. Datasættet findes på ugeplanen for uge 2 under navnet `Ex34.xls` (Excel) eller `Ex34.txt` (flad tekstfil) og indeholder faktorerne temperatur (`TEMP`), lucerne meal (`LUC`) og Pseudemonas-ADP (`ADP`) samt den målte variable `mineral`.

a) Indlæs datasættet i R og gem det som `terbuthyl`.

b) Undersøg, hvad kommandoen `table(TEMP,LUC)` gør.

c) Benyt kommandoen `table()` til at undersøge om datasættet sammer fra et balanceret trefaktorforsøg.

d) Benyt `factor()` til at sikre dig, at alle variable har den rigtige type (numerisk/faktor).

Vi skal i det følgende betragte den fulde model for tresidet variansanalyse givet ved

$$Y_i = \delta(\texttt{TEMP} * \texttt{LUC} * \texttt{ADP}_i) + e_i, \tag{2}$$

hvor $e_1, \ldots, e_{16}$ er uafhængige og normalfordelte $N(0, \sigma^2)$. Nedenfor diskuteres, hvordan man kan specificere modeller med tre faktorer i R samt teste for reduktion i modellen.

e) Estimer modellen ved at benytte kommandoen

```
lm(mineral~TEMP*LUC*ADP,data=terbuthyl)
```

og gem resulatet som `modelA`.

f) Benyt `lm()` til at fitte en model, hvor du på højreside kun indsætter vekselvirkning mellem de tre faktorer specificeret som `TEMP:LUC:ADP`.

```
lm(mineral~TEMP:LUC:ADP,data=terbuthyl)
```

Gem resultatet som `modelB`.

g) Både `modelA` og `modelB` estimerer modellen beskrevet ved (2). Sammenlign `modelA` og `modelB` ved brug af kommandoen `summary()` og diskuter, hvordan man skal fortolke parameterestimaterne for de to forskellige måder at beskrive den samme model på.

Det viser sig, at man kan reducere modellen (2) til

$$Y_i = \phi(\texttt{TEMP} * \texttt{ADP}_i) + e_i, \tag{3}$$

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

h) Reduktionen fra (2) til (3) bør foretages i 4 trin ved brug af `anova()`. Anfør F-teststørrelser og p-værdi for hvert af disse test.

i) Find parameterestimaterne i hver af de fire grupper givet ved faktoren `TEMP * ADP` under slutmodellen.

j) Find estimatet for spredningen under slutmodellen (3).

k) Find LSD-værdien for faktoren `TEMP * ADP`.

Teoretiske øvelser

Formålet med de teoretiske øvelser nedenfor er, at man lærer at lave variansanalyse med flere faktorer og herunder bliver fortrolig med brugen af faktordiagrammer. Du bør desuden klart kunne redegøre for forskellen mellem additive modeller og modeller med vekselvirkninger for forsøg med to faktorer.

Opgave 2.3: Additiv model for tosidet variansanalyse

Ved et dyrkningsforsøg ønskes effekten af fire forskellige gødningstyper undersøgt. Datasættet er indlæst i R og ser ud som følger

```
markforsog
```

```
##    NITROGEN udbytte
## 1         C    70.3
## 2         C    72.5
## 3         C    79.0
## 4         C    86.2
## 5         A    75.5
## 6         A    63.0
## 7         A    65.4
## 8         A    67.7
## 9         N    85.2
## 10        N    80.5
## 11        N    83.6
## 12        N    92.3
## 13        K    35.7
## 14        K    39.6
## 15        K    45.5
## 16        K    50.5
```

Derefter er kørt en analyse i R, som giver udskriften

```
model1<-lm(udbytte~NITROGEN-1,data=markforsog)
summary(model1)
```

```
Call:
lm(formula = udbytte ~ NITROGEN - 1)

Residuals:
   Min     1Q Median     3Q    Max
-7.125 -4.600 -1.000  3.731  9.200

Coefficients:
```

```
          Estimate Std. Error t value Pr(>|t|)
NITROGENA   67.900      3.041   22.33 3.85e-11 ***
NITROGENC   77.000      3.041   25.32 8.76e-12 ***
NITROGENK   42.825      3.041   14.08 7.99e-09 ***
NITROGENN   85.400      3.041   28.08 2.58e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.083 on 12 degrees of freedom
Multiple R-Squared: 0.9944,     Adjusted R-squared: 0.9925
F-statistic: 531.6 on 4 and 12 DF,  p-value: 2.176e-13
```

a) Opskriv den statistiske model svarende til udskriften ovenfor.

b) Angiv estimaterne for parametrene i modellen.

c) Angiv et 95 %-konfidensområde for estimatet svarende til gødningstyperne C og N.

d) Angiv LSD-værdien for sammenligning mellem to grupper og diskuter, om der er forskel på effekten af gødningstyperne C og N.

Dyrkningsforsøget er i virkeligheden fortaget på fire forskellige marker, hvilket fremgår af det fuldstændige datasæt, som er anført nedenfor.

```
markforsog

##    NITROGEN MARK udbytte
## 1         C    1    70.3
## 2         C    2    72.5
## 3         C    3    79.0
## 4         C    4    86.2
## 5         A    1    75.5
## 6         A    2    63.0
## 7         A    3    65.4
## 8         A    4    67.7
## 9         N    1    85.2
## 10        N    2    80.5
## 11        N    3    83.6
## 12        N    4    92.3
## 13        K    1    35.7
## 14        K    2    39.6
## 15        K    3    45.5
## 16        K    4    50.5
```

I forsøget indgår således faktorerne `MARK, NITROGEN` og `MARK` $\times$ `NITROGEN`.

e) Opskriv faktordiagrammet hørende til forsøget.

f) Kan man ved en sædvanlig tosidet variansanalyse teste, om der er vekselvirkning mellem `MARK` og `NITROGEN`

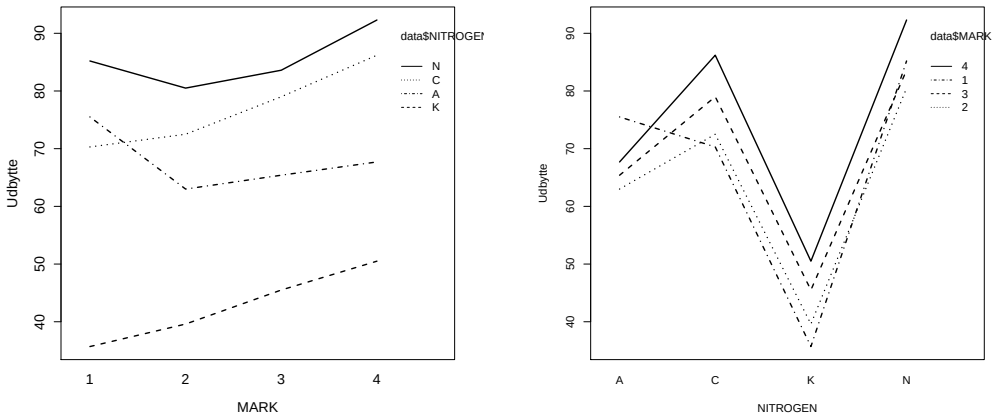

Figure 1: Interaction plots for faktorerne `MARK` og `NITROGEN`.

g) Benyt de to *interaction plots* på figur 1 til at diskutere, om der er vekselvirkning mellem `MARK` og `NITROGEN`.

Analysen af data fortsættes som følger

```
> model<-lm(udbytte~NITROGEN+MARK)
> summary(model)

Call:
lm(formula = udbytte ~ NITROGEN + MARK)

Residuals:
    Min      1Q  Median      3Q     Max
-6.0938 -2.0688  0.4437  1.8125  9.2062

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   66.294      3.254  20.370 7.72e-09 ***
NITROGENC      9.100      3.479   2.616  0.02801 *
NITROGENK    -25.075      3.479  -7.207 5.04e-05 ***
NITROGENN     17.500      3.479   5.030  0.00071 ***
MARK2         -2.775      3.479  -0.798  0.44564
MARK3          1.700      3.479   0.489  0.63680
MARK4          7.500      3.479   2.156  0.05948 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.92 on 9 degrees of freedom
Multiple R-Squared: 0.9517,     Adjusted R-squared: 0.9195
F-statistic: 29.57 on 6 and 9 DF,  p-value: 1.969e-05
```

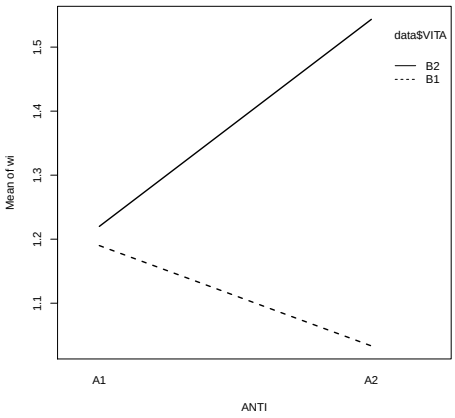

Figure 2: Interaction plots for faktorerne `ANTI`(antibiotika) og `VITA`(vitamin).

h) Angiv den statistiske model, som svarer til `model` i R-programmet ovenfor.

i) Undersøg, ved at beregne en relevant LSD-værdi, om der i denne model er forskel på gødningstype C og N.

NB: Det kan vises, at både `MARK` og `NITROGEN` har signifikant indflydelse på udbyttet, således at modellen beskrevet i spm. *h*) faktisk også er den relevante slutmodel på den statistiske analyse.

Opgave 2.4: Tosidet variansanalyse med vekselsvirkning

Løs opgave 3.2 fra lærebogen, dog med den undtagelse, at spørgsmål 1 ændres til:

1. Hvad er det, som er optegnet på figur 2.

Ved løsning af delspørgsmål 2-5 kan nedenstående R-udskrift benyttes. For de interesserede kan datasættet hentes på ugeplanen for uge 2 under navnet `Ex32.xls` (Excel) eller `Ex32.txt` (flad tekstfil).

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

> summary(model)

Call:
lm(formula = wi ~ ANTI * VITA)
```

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.110000 -0.025000  0.003333  0.016667  0.110000

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.19000    0.03496  34.039 6.06e-10 ***
ANTIA2        -0.15667    0.04944  -3.169 0.013220 *
VITAB2         0.03000    0.04944   0.607 0.560818
ANTIA2:VITAB2  0.48000    0.06992   6.865 0.000129 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06055 on 8 degrees of freedom
Multiple R-Squared: 0.9336,     Adjusted R-squared: 0.9087
F-statistic: 37.48 on 3 and 8 DF,  p-value: 4.659e-05

> model1<-lm(wi~ANTI+VITA)

> summary(model1)

Call:
lm(formula = wi ~ ANTI + VITA)

Residuals:
    Min      1Q  Median      3Q     Max
-0.1533 -0.1100 -0.0350  0.1217  0.2300

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.07000    0.07493  14.280 1.73e-07 ***
ANTIA2       0.08333    0.08652   0.963   0.3606
VITAB2       0.27000    0.08652   3.121   0.0123 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1499 on 9 degrees of freedom
Multiple R-Squared: 0.5423,     Adjusted R-squared: 0.4406
F-statistic: 5.333 on 2 and 9 DF,  p-value: 0.02968
```

```
> modelA<-lm(wi~ANTI)

> summary(modelA)

Call:
lm(formula = wi ~ ANTI)

Residuals:
    Min      1Q  Median      3Q     Max
-0.2883 -0.1533 -0.0050  0.1292  0.2717

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.20500    0.08375  14.388 5.21e-08 ***
ANTIA2       0.08333    0.11844   0.704    0.498
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2051 on 10 degrees of freedom
Multiple R-Squared: 0.04717,    Adjusted R-squared: -0.04811
F-statistic: 0.495 on 1 and 10 DF,  p-value: 0.4977

> modelV<-lm(wi~VITA)

> summary(modelV)

Call:
lm(formula = wi ~ VITA)

Residuals:
     Min       1Q   Median       3Q      Max
-0.19167 -0.11417 -0.04667  0.14583  0.18833

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.11167    0.06096  18.236 5.28e-09 ***
VITAB2       0.27000    0.08621   3.132   0.0107 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1493 on 10 degrees of freedom
Multiple R-Squared: 0.4952,     Adjusted R-squared: 0.4447
F-statistic: 9.809 on 1 and 10 DF,  p-value: 0.01066
```

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

Model 1: wi ~ ANTI + VITA
Model 2: wi ~ ANTI * VITA
  Res.Df      RSS Df Sum of Sq      F    Pr(>F)
1      9 0.202133
2      8 0.029333  1  0.172800 47.127 0.0001290 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> anova(modelA,model1)
Analysis of Variance Table

Model 1: wi ~ ANTI
Model 2: wi ~ ANTI + VITA
  Res.Df     RSS Df Sum of Sq      F  Pr(>F)
1     10 0.42083
2      9 0.20213  1   0.21870 9.7376 0.01231 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> anova(modelV,model1)
Analysis of Variance Table

Model 1: wi ~ VITA
Model 2: wi ~ ANTI + VITA
  Res.Df      RSS Df Sum of Sq      F Pr(>F)
1     10 0.222967
2      9 0.202133  1  0.020833 0.9276 0.3606
```