SD2 - uge 1, torsdag

Anne Petersen

Opgave 2.1 fra dokument på Absalon

Vi starter med at indlæse data og attache det, så vi har lettere adgang til variablene:

```
setwd("C:/Users/Anne/Dropbox/Arbejde/STATforLIFE2/uge1")
kost <- read.table("FP270505.txt", header=T)
attach(kost)
kost
```

```
##    SEX DIET energioms
## 1    M    1     12389
## 2    M    1     13519
## 3    M    1     12470
## 4    M    2     10002
## 5    M    2     12166
## 6    M    2     11268
## 7    M    3     12190
## 8    M    3     12175
## 9    M    3     12063
## 10   F    1      8566
## 11   F    1     10744
## 12   F    1      9827
## 13   F    2      8422
## 14   F    2     10858
## 15   F    2     10524
## 16   F    3      9161
## 17   F    3      9130
## 18   F    3      9214
```

Vi har to kategoriske variable, `SEX` og `DIET`, og vi vil gerne have, at R forstår, at de begge er faktorer. Da `SEX` har niveauer angivet ved bogstaver (F/M), sker dette automatisk:

```
is.factor(SEX)
```

```
## [1] TRUE
```

men fordi `DIET` er angivet med tal, må vi selv sørge for at gemme den som en faktorvariabel:

```
is.factor(DIET)
```

```
## [1] FALSE
```

```
DIET <- factor(DIET)
is.factor(DIET)
```

```
## [1] TRUE
```

Nu kigger vi på produktfaktoren `SEX:DIET`, dvs. faktoren som fremkommer ved at betragte alle kombinationer af de to faktorer `SEX` og `DIET`:

```
is.factor(SEX:DIET)
```

```
## [1] TRUE
```

```
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
## [18] F:3
## Levels: F:1 F:2 F:3 M:1 M:2 M:3
```

R vælger (heldigvis) at fortolke et produkt af to faktorer som en ny faktor - og det var lige det, vi gerne ville have.

Vi fitter nu en tosidet variansanalysemodel med produktfaktoren som forklarende variabel. Bemærk at vi også inddrager de marginale effekter af hver af de forklarende variable (også kendt som hovedeffekterne) jf. det hierarkiske princip (som I kender fra SD1):

```
model <- lm(energioms ~ SEX:DIET + SEX + DIET)
#Alternativ måde at skrive samme model på:
model <- lm(energioms ~ SEX*DIET)
```

Og vi kigger på modelresultaterne:

```
summary(model)
```

```
##
## Call:
## lm(formula = energioms ~ SEX * DIET)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1512.7  -261.9    39.0   472.7  1031.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9712.3      501.4  19.369 2.03e-10 ***
## SEXM          3080.3      709.1   4.344 0.000955 ***
## DIET2          222.3      709.1   0.314 0.759260
## DIET3         -544.0      709.1  -0.767 0.457825
## SEXM:DIET2   -1869.7     1002.9  -1.864 0.086911 .
## SEXM:DIET3    -106.0     1002.9  -0.106 0.917568
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 868.5 on 12 degrees of freedom
## Multiple R-squared:  0.7765, Adjusted R-squared:  0.6834
## F-statistic: 8.341 on 5 and 12 DF,  p-value: 0.001327
```

Vi ser at referencegruppen er SEX=F, DIET=1 (fordi det er de niveauer der "mangler" i `summary()`-outputtet). Vi kan aflæse at parameterestimatet for mænd, som har fået diæt 3 er

$$9712.3 + 3080.3 - 544.0 - 106.0 = 12142.6$$

og det kunne vi også få R til at bestemme vha. `predict()`-funktionen:

```
predict(model, new=data.frame(SEX="M", DIET="3"))
```

```
##        1
## 12142.67
```

Vi aflæser parameterestimatet for kvinder, som har fået diæt 2:

$$9712.3 + 222.3 = 9934.6$$

eller vha. R:

```
predict(model, new=data.frame(SEX="F", DIET="2"))
```

```
##        1
## 9934.667
```

Vi aflæser estimatet for spredningen, s, (som står under "residual standard error") til at være

$$s = 868.5$$

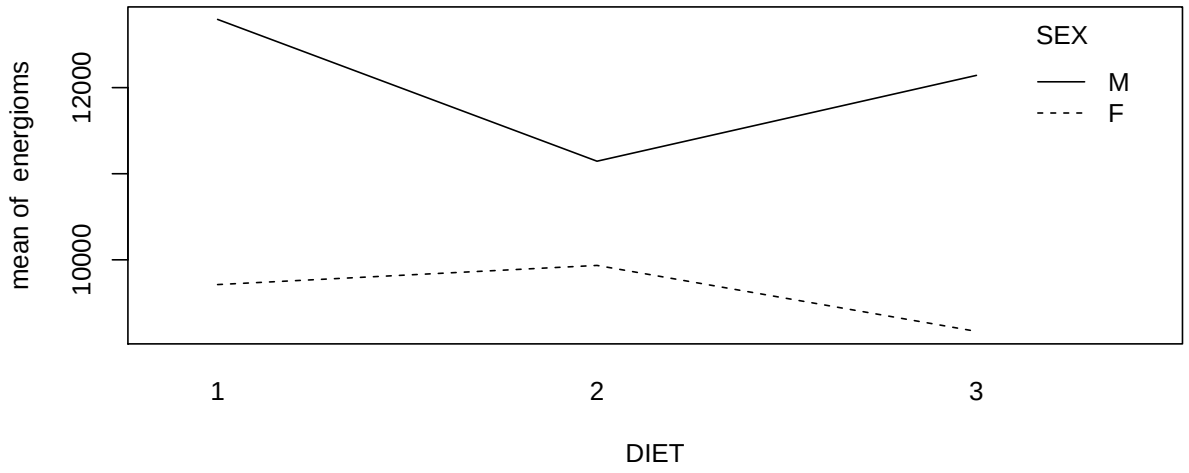
og vi kan selvfølgelig også trække tallet direkte ud af `summary()`-outputtet, hvis vi hellere vil det:

```
summary(model)$sigma
```

```
## [1] 868.4979
```

Vi betragter nu et interaktionsplot (vekselvirkningsplot):

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



Vi ser at mænd generelt har højere energiomsætning end kvinder. Det ser desuden ud til at de forskellige kosttypers påvirkning af energiomsætningen er forskellig for de to køn, fx. er mænds energiomsætning lavest på kost 2 mens kvinders her er højest. Vi ser desuden, at der er større udsving i energiomsætningen for mænd på tværs af kosttyperne end for kvinder. Alt i alt tyder det altså på, at der godt kunne være en vekselvirkningseffekt. Men det må vi hellere undersøge mere stringent vha. tests.

Vi fitter derfor en additiv model med effekter af `SEX` og `DIET` og kigger på den:

```
add_model <- lm(energioms ~ DIET + SEX)
summary(add_model)
```

```
##
## Call:
## lm(formula = energioms ~ DIET + SEX)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1748.89  -306.86   -33.89   392.24  1528.89
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10041.6      442.9  22.671 1.95e-12 ***
## DIET2         -712.5      542.5  -1.313     0.21
## DIET3         -597.0      542.5  -1.100     0.29
## SEXM          2421.8      442.9   5.468 8.29e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 939.6 on 14 degrees of freedom
## Multiple R-squared:  0.6949, Adjusted R-squared:  0.6295
## F-statistic: 10.63 on 3 and 14 DF,  p-value: 0.0006623
```

Vi ser at referencegruppen stadig er kvinde og kost nr. 1. Og vi kan aflæse, at parameterestimatet for mænd der har fået kost 3 er

$$10041.6 - 597.0 + 2421.8 = 11866.4$$

og for kvinder, som har fået kost 2 får vi

$$10041.6 - 712.5 = 9329.1$$

eller vi kan gør det vha. R:

```
predict(add_model, new=data.frame(SEX=c("M","F"),
                                  DIET=c("3", "2")))
```

```
##         1         2
## 11866.389  9329.111
```

Vi fitter nu ensidede variansanalysemodeller med hhv. `SEX` og `DIET` som forklarende variabel:

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

Vi kan nu teste modellerne op mod hinanden. Vi starter med den mest komplicerede model, dvs. modellen med vekselvirkningseffekten, og tester den op mod modellen, som kun har to additive effekter:

```
anova(add_model, model)
```

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

Bemærk, at vi tester hypotesen om ingen effekt af vekselvirkningen. Vi finder $p = 0.1542 > 0.05$ og accepterer altså hypotesen. Der er dermed ingen signifikant effekt af vekselvirkningen, og vi kan arbejde videre med den forsimplede model, `add_model`. Men lad os se, om vi kan gøre det endnu simplere - måske er det ikke nødvendigt at inddrage begge de forklarende variable? Vi tester det:

```
anova(modelS, add_model)
```

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

```
anova(modelD, add_model)
```

```
## Analysis of Variance Table
##
## Model 1: energioms ~ DIET
## Model 2: energioms ~ DIET + SEX
##   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
```

Vi ser, at der ikke er nogen signifant effekt af `DIET` (vi får $p = 0.3948$, når vi fjerner denne effekt). Der er derimod en signfikant effekt af `SEX` i den additive model ($p = 8.291 \cdot 10^{-5}$). Vi går altså videre med `modelS`. Lad os lige tjekke, om der stadig er en signifikant effekt af `SEX` i denne model:

```
anova(modelS)
```

```
## Analysis of Variance Table
##
## Response: energioms
##           Df   Sum Sq  Mean Sq F value    Pr(>F)
## SEX        1 26392534 26392534  29.917 5.142e-05 ***
## Residuals 16 14114980   882186
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#Alternativ måde at få samme resultat:
anova(lm(energioms ~ 1), modelS)
```

```
## Analysis of Variance Table
##
## Model 1: energioms ~ 1
## Model 2: energioms ~ SEX
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1     17 40507514
## 2     16 14114980  1  26392534 29.917 5.142e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

og vi ser, at der er en signifikant effekt af `SEX` i denne model, så vi kan ikke reducere modellen yderligere. Altså er vores slutmodel `modelS`, dvs. modellen

$$Y_i = \alpha_{\mathrm{SEX}(i)} + e_i$$

med notation som i opgavebeskrivelsen.