Reeksamen i Statistik 2, 23. august 2018

Vejledende besvarelse

Opgave 1

1. Ifølge EH Korollar 9.43 er X regulært normalfordelt hvis og kun hvis variansen Σ er invertibel. Det ses at determinanten af Σ er lig

$$\frac{1}{2}\{1\cdot 2\cdot 1+1\cdot(-1)\cdot 0+0\cdot 1\cdot(-1)-1\cdot(-1)\cdot(-1)-1\cdot 1\cdot 1-0\cdot 2\cdot 0\}=0,$$

hvorfor Σ ikke er invertibel. Alternativt kan bemærkes at række 1 i Σ er lig med summen af række 2 og 3, hvorfor Σ ikke er invertibel. Det konkluderes at X følger en singulær normalfordeling.

2. Det følger af EH Lemma 9.47 at Y er normalfordelt med varians $\Sigma_Y = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$. M. Der er tale om en regulær normalfordeling og af EH Sætning 9.42 konkluderes, at den tilhørende præcision er givet ved

$$\langle x,y\rangle = x^T\Sigma_Y^{-1}y = x^T \cdot \underbrace{2\begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}}_{:=\Sigma_Y^{-1}} \cdot y.$$

Tætheden for Y i $y=(1,1)$ bliver ifølge EH Sætning 9.20

$$f(y)=\frac{(det\Sigma_Y^{-1})^{1/2}}{(2\pi)^{2/2}}\exp(-\frac{1}{2}y^T\Sigma_Y^{-1}y)=\frac{4^{1/2}}{2\pi}\exp(-\frac{1}{2}4)\approx 0.0431.$$

3. Det følger EH Sætning 9.47 og 9.48 at X_1 og X_3 er uafhængige og at begge variable er normalfordelte med middelværdi 0 og varians $1/2$. Dermed er $\tilde{X}_1=\sqrt{2}X_1$ og $\tilde{X}_3=\sqrt{2}X_3$ uafhængige og standardnormalfordelte. Dermed er $\tilde{X}_1^2+\tilde{X}_3^2=2X_1^2+2X_3^2$ per definition χ^2-fordelte med 2 frihedsgrader. Det er netop dette udtryk som fremkommer, når man udregner matrixproduktet fra opgaveformuleringen.

En alternativ løsning består i at bemærke, at $Z=(X_1,X_3)^T$ er regulært normalfordelt med middelværdi $(0,0)^T$ og varians $\Sigma_Z=\frac{1}{2}I_2$. Dermed gælder ifølge EH Sætning 9.29 at $||Z||^2_{\Sigma_Z^{-1}}=Z^T(\frac{1}{2}I_2)^{-1}Z$ χ^2-fordelt med 2 frihedsgrader. Opgaven løses nu ved at indse, at $||Z||^2_{\Sigma_Z^{-1}}$ er identisk med matrixproduktet i opgaveformuleringen.

Opgave 2

1. Med $\beta = (\beta_1, \beta_2)^T$ bliver designmatricen

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \\ 1 & 6 \end{pmatrix}.$$

   Maksimaliseringsestimateren i den lineære normale model findes ved brug af EH Korollar 10.21. Vi finder, at

$$\hat{\beta} = (A^T A)^{-1} A^T W = (-0.7333, 2.1142)^T$$

   og

$$\hat{\sigma}^2 = \frac{||W - A\hat{\beta}||^2}{6} = 0.184127.$$

2. Tilsvarende giver EH Korollar 10.21 at $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (A^T A)^{-1})$ og at $\hat{\sigma}^2$ er χ^2-fordelt med $6-2=4$ frihedsgrader og skalaparamater $\sigma^2/6$. Spredningen på estimatet for β_2 kan estimeres ved at indsætte et estimat for σ i udtrykket $\sigma^2(A^T A)^{-1}$ for variansen. Benyttes den centrale estimator $\tilde{\sigma}^2 = \frac{6}{4}\hat{\sigma}^2$ så estimeres standard error for estimatet på β_2 til

$$se(\hat{\beta}_2) = \sqrt{\frac{6}{4} \cdot 0.184127 \cdot 0.05714286} = 0.1256.$$

   Grænserne for et 95 % - konfidensinterval kan beregnes som $2.1142 \pm 2.7764 \cdot 0.1256$, hvor 2.7764 angiver 97.5 % - fraktilen i en t-fordeling med $6-2=4$ frihedsgrader. Konfidensintervallet bliver $[1.765 - 2.463]$.

Følgende R kode er benyttet til den vejledende besvarelse af opgave 2, men alle beregninger kan ret let foretages i hånden.

Først beregnes maksimaliseringsestimaterne

```
W <- matrix(ncol = 1, data = c(1,4,6,7,10,12))
A <- cbind(1, 1:6)
bhat <- solve(t(A)%*%A)%*%t(A)%*%W
bhat # estimat for beta

##            [,1]
## [1,] -0.7333333
## [2,]  2.1142857

shat2 <- sum((W - A%*%bhat)^2)/6
shat2 # estimat for sigma^2

## [1] 0.184127
```

Dernæst bestemmes (det estimeres værdier) for parametrene i fordelingen af maksimaliserings-estimatorerne.

```
bhatvar <- shat2*6/4 * solve(t(A)%*%A)
bhatvar
```

```
##            [,1]        [,2]
## [1,]  0.2393651 -0.05523810
## [2,] -0.0552381  0.01578231
```

```
bhatse <- sqrt(diag(bhatvar))
bhatse
```

```
## [1] 0.4892495 0.1256277
```

```
bhat2ci <- bhat[2] + c(-1, 1) * bhatse[2] * qt(0.975, 6 - 2)
bhat2ci
```

```
## [1] 1.765487 2.463084
```

Det er også muligt at lade `lm()`-funktionen foretage beregningerne.

```
summary(lm(W ~ A - 1))
```

```
##
## Call:
## lm(formula = W ~ A - 1)
##
## Residuals:
##        1        2        3        4        5        6
## -0.38095  0.50476  0.39048 -0.72381  0.16190  0.04762
##
## Coefficients:
##    Estimate Std. Error t value Pr(>|t|)
## A1  -0.7333     0.4892  -1.499    0.208
## A2   2.1143     0.1256  16.830 7.31e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5255 on 4 degrees of freedom
## Multiple R-squared:  0.9968,Adjusted R-squared:  0.9952
## F-statistic: 624.4 on 2 and 4 DF,  p-value: 1.019e-05
```

```
confint(lm(W ~ A - 1))
```

```
##        2.5 %    97.5 %
## A1 -2.091708 0.6250411
## A2  1.765487 2.4630841
```

Opgave 3

1. Det fremgår af tabellen over forsøgsdesignet i opgaveformuleringen, at designet er sammenhængende og at faktorerne M og D opfylder *balanceligningen* fra EH Lemma 13.11. Dermed er faktorerne geometrisk ortogonale. For et sammenhængende design er minimum af faktorerne blot den konstante faktor. Dermed bliver dimensionen af det additive underrum
$$dim(L_M + L_D) = dimL_M + dimL_D - dimL_1 = 2 + 2 - 1 = 3.$$

2. F teststørrelsen for test af den additive hypotese kan ifølge EH formel (10.31) udtrykkes som
$$F = \frac{(||P_{M\times D}||^2 - ||P_{M+D}X||^2)/1}{(||X||^2 - ||P_{M\times D}X||^2)/(24-4)}.$$
For tosidet variansanalyse med geometrisk ortogonale faktorer kan vi benytte EH formel (13.5) til at beregne
$$||P_{M+D}X||^2 = ||P_M X||^2 + ||P_D X||^2 - ||P_1 X||^2,$$
hvor alle tre størrelser på højresiden fremgår af faktorstrukturdiagrammet i opgaveformuleringen. Vi finder at $||P_{M+D}X||^2 = 429918.342$ og dermed, at
$$F = \frac{(429921.699 - 429918.342)/1}{(431533.199 - 429921.699)/(24-4)} = 0.0417.$$
Under hypotesen om at der ikke er nogen vekselvirkning vil F teststørrelsen følge en F-fordeling med $(1,20)$ frihedsgrader. Vi finder således den tilsvarende P-værdi $= 0.840$. Vi kan således ikke forkaste hypotesen om, at der ikke er vekselvirkning mellem medikament og dosis.

3. De estimerede middelværdier bliver

M	D	E[X]
A	lav	99.507
A	hoj	134.814
B	lav	135.419
B	hoj	168.894
0	0	97.733

4. Data fra USB-nøglen indlæses i R. Teststørrelsen bliver $F = 0.0434$ med tilhørende P-værdi $= 0.8361$. Der lader således ikke til at være en vekselvirkning mellem dosis og behandling.

5. Det er helt legalt blot at kigge på antallet af estimater for middelværdistrukturen i R udskriften, når man skal argumentere for, at dimensionen af den additive model er 4.

   Ønsker man at regne mere formelt på tingene kan benyttes, at
$$dim(L_M + L_D) = dimL_M + dimL_D - dimL_{M\wedge D}.$$

Udfordringen ligger i at minimum $M \wedge D$ her er en faktor på 2 niveauer, som blot holder styr på om målingen stammer fra en person, som har modtaget et medikament (dvs. $M = A$ eller $M = B$) eller ej (dvs. $M = 0$). Minimum kan umiddelbart aflæses ud fra antalstabellen til beskrivelse af det fulde forsøgsdesign som findes i opgaveformuleringen. Indsættes i ovenstående formel fås nu, at

$$dim(L_M + L_D) = dimL_M + dimL_D - dimL_{M \wedge D} = 3 + 3 - 2 = 4.$$

Det virker ikke rimeligt at reducere modellen yderligere ved at se bort fra effekten af M ($F = 89.143, p < 0.0001$) eller faktoren D ($F = 69.519, p < 0.0001$).

6. Som altid er der ikke entydighed omkring valget af designmatrix ved parametrisering af det additive middelværdiunderrum. Nedenfor angives estimaterne fra en et parametrisering som benytter kontrolgruppen som reference (estimat: 97.733, 95 %-KI: [93.3 – 102.1]). Gives den lave dosis af A øges estimatet med 2.079 (95 % - KI: [-5.8-10.0]), mens den tilsvarende effekt for medikament B estimeres til 37.534 (95 % - KI: [30.979048 44.088322]). Gives i stedet den høje dosis øges estimatet med 34.086 (95 % - KI: [25.8-42.4]) uanset medikament (da vi betragter en additiv model!).

   På baggrund af konfidensintervallerne for estimaterne i den valgte parametrisering kan vi umiddelbart konkludere: i) at medikament A ikke har effekt i den lave dosis, ii) at medikament B *har* effekt i den lave dosis, iii) der er effekt af at bruge høj dosis i stedet for lav dosis.

   Afhængigt at den valgte parametrisering kan andre aspekter af effekten af de forskellige behandlingskombinationer undersøges. For at få fuldt point for delopgaven er det nok, at der udtrækkes relevante konklusioner fra estimater og konfidensintervaller fra mindst en fornuftig parametrisering af modellen.

Følgende R-kode er benyttet i forbindelse med den vejledende besvarelse

```
data2 <- read.table("stat2_2018_aug_opg3.txt", header = T)
```

```
head(data2)

##   M   D         x
## 1 A lav  95.77016
## 2 A lav  84.50122
## 3 A lav  99.35571
## 4 A lav 102.70881
## 5 A lav 117.35284
## 6 A lav  97.35289
```

```
mod0 <- lm(x ~ M:D - 1, data = data2)

### Delopgave 4: test af vekselvirkning

mod1 <- lm(x ~ M + D, data = data2)
anova(mod1, mod0) # F = 0.0434, P = 0.8361

## Analysis of Variance Table
##
## Model 1: x ~ M + D
## Model 2: x ~ M:D - 1
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     36 2707.4
## 2     35 2704.1  1    3.3564 0.0434 0.8361

### Delopgave 5: yderligere modelreduktion

mod2a <- lm(x ~ M, data = data2)
anova(mod2a, mod1) # F = 69.519, p < 0.0001

## Analysis of Variance Table
##
## Model 1: x ~ M
## Model 2: x ~ M + D
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     37 7935.7
## 2     36 2707.4  1    5228.3 69.519 6.297e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

mod2b <- lm(x ~ D, data = data2)
anova(mod2b, mod1) # F = 89.143, p < 0.0001

## Analysis of Variance Table
##
## Model 1: x ~ D
## Model 2: x ~ M + D
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     37 9411.5
## 2     36 2707.4  1    6704.1 89.143 2.819e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
### Delopgave 6: estimater fra additive model

summary(mod1)

##
## Call:
## lm(formula = x ~ M + D, data = data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.4196  -5.6774  -0.4192   6.5403  17.5406
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   97.733      2.168  45.079  < 2e-16 ***
## MA             2.079      3.892   0.534    0.596
## MB            37.534      3.232  11.613 9.87e-14 ***
## Dhoj          34.086      4.088   8.338 6.30e-10 ***
## Dlav              NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.672 on 36 degrees of freedom
## Multiple R-squared:  0.8954,Adjusted R-squared:  0.8867
## F-statistic: 102.8 on 3 and 36 DF,  p-value: < 2.2e-16

confint(mod1)

##                 2.5 %     97.5 %
## (Intercept) 93.336023 102.129992
## MA          -5.813555   9.972101
## MB          30.979048  44.088322
## Dhoj        25.794768  42.376833
## Dlav               NA         NA

# Konklusion:
# A, lav: ingen effekt (2.079)
# B, lav: signifikant effekt (37.534)
# Forskel p<U+00E5> hoj og lav: signifikant (34.086)

mod1alt <- lm(x ~ D + M , data = data2)
summary(mod1alt)

##
## Call:
## lm(formula = x ~ D + M, data = data2)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.4196  -5.6774  -0.4192   6.5403  17.5406
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   97.733      2.168  45.079  < 2e-16 ***
## Dhoj          71.619      4.336  16.517  < 2e-16 ***
## Dlav          37.534      3.232  11.613 9.87e-14 ***
## MA           -35.454      3.755  -9.442 2.82e-11 ***
## MB                NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.672 on 36 degrees of freedom
## Multiple R-squared:  0.8954,Adjusted R-squared:  0.8867
## F-statistic: 102.8 on 3 and 36 DF,  p-value: < 2.2e-16
```

Opgave 4

1. Alle kombinationer af P og V er afprøvet netop een gang i forsøget. Derfor er disse to faktorer geometriske ortogonale med trivielt minimum (= den konstante faktor). Da B er grovere end P følger det umiddelbart af regneregler for minimum, at alle øvrige faktorer i designet er geometrisk ortogonale, og at minimumskonstruktion ikke fører til nye faktorer.

2. Den væsentlige udfordring er at lade faktorstrukturdiagrammet afspejle, at B er grovere end P. Det udfyldte diagram med dimensioner ser ud som følger

$$\begin{tikzcd} & P_6^8 \arrow[r] & B_1^2 \arrow[dr] & \\ I_{18}^{32} \arrow[ur] \arrow[dr] & & & 1_1^1 \\ & B \times V_3^8 \arrow[r] \arrow[uur] & V_3^4 \arrow[ur] & \end{tikzcd} \tag{1}$$

3. Modellen kan udtrykkes ved at vektoren $X = (X_i)_{i \in I}$ bestående af udbyttet på de enkelte parceller er normalfordelt på $\mathbb{R}^{32}$ med $\xi = EX \in L_{B \times V}$ og varians $VX = \sigma^2 + \nu_1^2 B_1$, hvor B_1 er effektmatricen hørende til parret $(\mathtt{P}, \mathtt{1})$.

   Den totale varians på udbyttet bliver $\sigma^2 + \nu_1^2$ og kovariansmatricen for de 4 målinger på samme plot kan udtrykkes som

$$\begin{pmatrix} \sigma^2 + \nu_1^2 & \nu_1^2 & \nu_1^2 & \nu_1^2 \\ \nu_1^2 & \sigma^2 + \nu_1^2 & \nu_1^2 & \nu_1^2 \\ \nu_1^2 & \nu_1^2 & \sigma^2 + \nu_1^2 & \nu_1^2 \\ \nu_1^2 & \nu_1^2 & \nu_1^2 & \sigma^2 + \nu_1^2 \end{pmatrix}$$

4. Følgende R kode kan benyttes til at estimere parametrene i den ønskede varianskomponentmodel i R

```
data4 <- read.table("stat2_2018_aug_opg4.txt", header = T)
```

```
head(data4)
```

```
##   P  B     V udbytte
## 1 1 Ja  Lami 52.3836
## 2 1 Ja  Lofa 49.6232
## 3 1 Ja Salka 49.6232
## 4 1 Ja  Zita 49.7334
## 5 2 Ja  Lami 55.4953
## 6 2 Ja  Lofa 52.7372
```

```r
library(lme4)

## Loading required package: Matrix

m0 <- lmer(udbytte ~ V * B + (1|P), data = data4)
VarCorr(m0)

##  Groups   Name        Std.Dev.
##  P        (Intercept) 1.1926
##  Residual             1.8243
```

Variansestimaterne bliver $\hat{v}_1 = 1.1926$ og $\hat{\sigma} = 1.8243$.

5. Middelværdiestimaterne hørende til de ønskede kombinationer af B og V fremgår af følgende R-udskrift, hvor de ønskede grupper optræder i linjerne $6, 1, 5$.

```r
m0alt <- lmer(udbytte ~ V : B - 1 + (1|P), data = data4)
coef(summary(m0alt))

##             Estimate Std. Error  t value
## VLami:BJa   55.64332   1.089786 51.05893
## VLofa:BJa   49.89000   1.089786 45.77962
## VSalka:BJa  54.11810   1.089786 49.65937
## VZita:BJa   52.82257   1.089786 48.47058
## VLami:BNej  61.21607   1.089786 56.17255
## VLofa:BNej  54.89210   1.089786 50.36960
## VSalka:BNej 57.70762   1.089786 52.95316
## VZita:BNej  58.52837   1.089786 53.70629
```

6. Konfidensintervaller for middelværdierne i de otte gruppe givet ved $B \times V$ kan beregnes i R ved bruge af `confint()`-funktionen.

```r
confint(m0alt)

## Computing profile confidence intervals ...

##                 2.5 %    97.5 %
## .sig01       0.000000  2.197756
## .sigma       1.219952  2.159005
## VLami:BJa   53.716425 57.570225
## VLofa:BJa   47.963100 51.816900
## VSalka:BJa  52.191200 56.045000
## VZita:BJa   50.895675 54.749475
## VLami:BNej  59.289175 63.142975
## VLofa:BNej  52.965200 56.819000
## VSalka:BNej 55.780725 59.634525
## VZita:BNej  56.601475 60.455275
```

For at vi kan udtale os om effekten af bayleton for sorten `Lami` benyttes en ny parametrisering af modellen, hvoraf forskellene mellem grupperne $\{Ja, Lami\}$ og $\{Nej, Lami\}$ direkte kan aflæses.

```
m0altny <- lmer(udbytte ~ V + B:V - 1 + (1|P), data = data4)
coef(summary(m0altny))

##               Estimate Std. Error   t value
## VLami        55.643325   1.089786 51.058933
## VLofa        49.890000   1.089786 45.779618
## VSalka       54.118100   1.089786 49.659369
## VZita        52.822575   1.089786 48.470581
## VLami:BNej    5.572750   1.541191  3.615873
## VLofa:BNej    5.002100   1.541191  3.245608
## VSalka:BNej   3.589525   1.541191  2.329060
## VZita:BNej    5.705800   1.541191  3.702203

confint(m0altny)

## Computing profile confidence intervals ...

##                   2.5 %    97.5 %
## .sig01        0.0000000  2.197756
## .sigma        1.2199519  2.159005
## VLami        53.7164246 57.570225
## VLofa        47.9630996 51.816900
## VSalka       52.1911996 56.045000
## VZita        50.8956746 54.749475
## VLami:BNej    2.8477013  8.297799
## VLofa:BNej    2.2770513  7.727149
## VSalka:BNej   0.8644763  6.314574
## VZita:BNej    2.9807513  8.430849
```

Forskellen mellem grupperne aflæses til 5.572750 $[95-\%KI : 2.847701-8.297799]$. Da konfidensintervallet *ikke* indholder værdien 0, så lader det til at sprøjtning har nogen effekt på udbyttet for sorten `Lami`. Effekten af at sprøjte med bayleton kan i øvrigt genfindes for alle fire sorten, hvilken kunne uddybes ved en mere systematisk analyse af datasættet.