Eksamen i Statistisk Dataanalyse 2, 9. april 2015

Vejledende besvarelse

Opgave 1

1. Det vil naturligt at inddrage in-growth core (`incore`) og klimakammer (`octagon`) som tilfældige faktorer i modellen. Desuden bør de 4 faktorer `depth, co2, temp` og `drought` samt deres vekselvirkninger indgå i modellen med systematisk model. Den statistiske model kan udtrykkes som

$$\texttt{DW}_i = \gamma(\texttt{depth} \times \texttt{co2} \times \texttt{temp} \times \texttt{drought}_i) + A(\texttt{octagon}_i) + B(\texttt{incore}_i) + e_i,$$

   hvor

   - $A(1), \dots, A(12)$ er uafhængige og $\sim N(0, \sigma_A^2)$
   - $B(1), \dots, B(96)$ er uafhængige og $\sim N(0, \sigma_B^2)$
   - $e_1, \dots, e_{192}$ er uafhængige og $\sim N(0, \sigma^2)$

2. Igennem hele opgaven er kovariansstrukturen givet som for udgangsmodellen (dvs. at vi inkluderer tilfældige effekter af `octagon` og `incore`). Med udgangspunkt i hintet starter vi med at teste, om udgangsmodellen kan reduceres til

$$\texttt{DW}_i = \gamma(\texttt{depth} \times \texttt{co2} \times \texttt{temp}_i) + A(\texttt{octagon}_i) + B(\texttt{incore}_i) + e_i,$$

   svarende til, at vi fjerner effekten af `drought`. Det konstateres, at effekten af `drought` kan fjernes ($L.Ratio = 7.792, p = 0.454$).

   Dernæst konstateres, at vi kan fjerne trefaktorvekselvirkningen $\texttt{depth} \times \texttt{co2} \times \texttt{temp}$ ($L.Ratio = 1.275, p = 0.259$), således at modellen reduceres til

$$\texttt{DW}_i = \alpha(\texttt{co2} \times \texttt{temp}_i) + \beta(\texttt{depth} \times \texttt{temp}_i) + \gamma(\texttt{depth} \times \texttt{co2}_i) + A(\texttt{octagon}_i) + B(\texttt{incore}_i) + e_i.$$

   Herfra er der (som altid) flere muligheder for rækkefølgen i forbindelsen med modelreduktion. Man kunne vælge at fjerne effekten af $\texttt{temp} \times \texttt{co2}$ ($L.Ratio = 0.329, p = 0.566$), således at vi har reduceret modellen til

$$\texttt{DW}_i = \beta(\texttt{depth} \times \texttt{temp}_i) + \gamma(\texttt{depth} \times \texttt{co2}_i) + A(\texttt{octagon}_i) + B(\texttt{incore}_i) + e_i.$$

   På baggrund af det approksimative likelihood ratio test kan man fjerne effekten af $\texttt{temp} \times \texttt{depth}$ ($L.Ratio = 3.398, p = 0.065$), hvorved modellen reduceres til

$$\texttt{DW}_i = \beta(\texttt{temp}_i) + \gamma(\texttt{depth} \times \texttt{co2}_i) + A(\texttt{octagon}_i) + B(\texttt{incore}_i) + e_i.$$

Modellen kan ikke reduceres yderligere idet hverken vekselvirkningen $\texttt{depth} \times \texttt{co2}$ ($L.Ratio = 18.135, p < 0.0001$) eller hovedeffekten `temp` ($L.Ratio = 15.666, p = 0.0001$) kan fjernes.

Slutmodellen bliver

$$\texttt{DW}_i = \beta(\texttt{temp}_i) + \gamma(\texttt{depth} \times \texttt{co2}_i) + A(\texttt{octagon}_i) + B(\texttt{incore}_i) + e_i.$$

Vi kan tænke på den systematiske del af modellen som en additiv model med faktorerne `temp` og $\texttt{depth} \times \texttt{co2}$.

3. Slutmodellen fra delspørgsmål refittes med `method='REML'` som anført i `modfinal` i R-udskriften nedenfor. Estimatet i dybden 0-5cm for en kontrolprøve bliver 28.798 hvilken kan aflæses som (`Intercept`) i `summary(modfinal)`.

   Variansestimaterne for de tilfældige effekter bliver

$$\sigma_A^2 = 17.794 \quad \sigma_B^2 = 71.007 \quad \sigma^2 = 238.133.$$

4. Et estimat for forskellen mellem en kontrolprøve og en prøve, hvor alle klimafaktorer er modificerede kan bestemmes ved brug af `estimable()`-funktionen. Estimat + 95 %-konfidensinterval bliver $16.08[4.56 - 27.60]$.

5. Opgaven er bevidst lidt uklart formuleret her: det kræves blot at tørvægten `DW` modelleres som en lineær funktion af dybden (`depth`), men der siges ikke noget om, at nogle af de øvrige klimafaktorer behøver indgå i modellen. Derfor er der mange rigtige svar på dette delspørgsmål. Den simpleste løsning er kun at inkludere dybde som en kovariat i den systematiske del af modellen.

   En anden vigtig pointe i opgaven er, at man bør indse, at vi har gentagne målinger over dybde. Derfor er det nærliggende at modellere den serielle korrelation inden for in-growth core ved f.eks. en Diggle-model. En mulighed er derfor at benytte modellen

$$\texttt{DW}_i = \alpha + \beta \cdot \texttt{depth}_i + A(\texttt{octagon}_i) + B(\texttt{incore}_i) + D_i + e_i.$$

   hvor D_i'erne er beskrevet som for Diggle-modellen i kompendiets kapitel 10.3.

Eksempel på R-kode som kunne være brugt til løsning af opgave 1

```
### indlaes data og lav variable om til faktorer
data1<-read.table("data/data1.txt",header=T,sep="\t",dec=".")
head(data1,18)

##    octagon incore co2 temp drought depth       DW
## 1        1      1   0    0       0   0-5 15.34994
## 2        1      1   0    0       0  Ohor 14.76958
## 3        1      2   0    0       0   0-5 34.86580
## 4        1      2   0    0       0  Ohor 33.61352
```

```
## 5         1       3   0    0       1   0-5 20.18645
## 6         1       3   0    0       1  0hor 21.39042
## 7         1       4   0    0       1   0-5 26.37573
## 8         1       4   0    0       1  0hor 10.69521
## 9         1       5   0    1       1   0-5 21.25855
## 10        1       5   0    1       1  0hor 14.26028
## 11        1       6   0    1       1   0-5 28.51257
## 12        1       6   0    1       1  0hor 11.20451
## 13        1       7   0    1       0   0-5 42.91128
## 14        1       7   0    1       0  0hor 32.08564
## 15        1       8   0    1       0   0-5 33.85831
## 16        1       8   0    1       0  0hor 29.02986
## 17        2       9   1    0       0   0-5 35.28254
## 18        2       9   1    0       0  0hor 17.82535

data1$co2<-factor(data1$co2)
data1$temp<-factor(data1$temp)
data1$drought<-factor(data1$drought)
data1$incore<-factor(data1$incore)
data1$octagon<-factor(data1$octagon)

### fit af udgangsmodel ###
library(nlme)
m0<-lme(DW~co2*drought*temp*depth
        ,random=~1|octagon/incore,data=data1,method="ML")

### modelreduktion ###

m1<-lme(DW~co2*temp*depth
        ,random=~1|octagon/incore,data=data1,method="ML")
anova(m1,m0) ### test for om drought kan fjernes fra modellen

##    Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## m1     1 11 1659.049 1694.882 -818.5247
## m0     2 19 1667.258 1729.150 -814.6289 1 vs 2 7.791592  0.4541

m2<-lme(DW~co2*temp+temp*depth+co2*depth
        ,random=~1|octagon/incore,data=data1,method="ML")
anova(m2,m1) ### test for om 3-faktorvekselvirkning kan fjernes

##    Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## m2     1 10 1658.324 1690.899 -819.1620
## m1     2 11 1659.049 1694.882 -818.5247 1 vs 2 1.274694  0.2589

m3a<-lme(DW~temp*depth+co2*depth
        ,random=~1|octagon/incore,data=data1,method="ML")
anova(m3a,m2) ### test for effekt af temp x co2
```

```
##      Model df      AIC      BIC    logLik   Test   L.Ratio p-value
## m3a      1  9 1656.653 1685.970 -819.3264
## m2       2 10 1658.324 1690.899 -819.1620 1 vs 2 0.3288819  0.5663

m3b<-lme(DW~co2*temp+co2*depth
        ,random=~1|octagon/incore,data=data1,method="ML")
anova(m3b,m2) ### test for effekt af temp x depth

##      Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## m3b      1  9 1659.722 1689.039 -820.8609
## m2       2 10 1658.324 1690.899 -819.1620 1 vs 2 3.397849  0.0653

m3c<-lme(DW~co2*temp+temp*depth
        ,random=~1|octagon/incore,data=data1,method="ML")
anova(m3c,m2) ### test for effekt af co2 x depth

##      Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## m3c      1  9 1675.052 1704.370 -828.5261
## m2       2 10 1658.324 1690.899 -819.1620 1 vs 2 18.72815  <.0001

m4a<-lme(DW~temp+co2*depth
        ,random=~1|octagon/incore,data=data1,method="ML")
anova(m4a,m3a) ### test for effekt af temp x depth

##      Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## m4a      1  8 1658.051 1684.111 -821.0254
## m3a      2  9 1656.653 1685.970 -819.3264 1 vs 2 3.397849  0.0653

m4b<-lme(DW~temp*depth+co2
        ,random=~1|octagon/incore,data=data1,method="ML")
anova(m4b,m3a) ### test for effekt af co2 x depth

##      Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## m4b      1  8 1673.381 1699.441 -828.6905
## m3a      2  9 1656.653 1685.970 -819.3264 1 vs 2 18.72815  <.0001

m5a<-lme(DW~temp+co2+depth
        ,random=~1|octagon/incore,data=data1,method="ML")
anova(m5a,m4a) ### test for effekt af co2 x depth

##      Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## m5a      1  7 1674.185 1696.988 -830.0927
## m4a      2  8 1658.051 1684.111 -821.0254 1 vs 2 18.13463  <.0001

m5b<-lme(DW~co2*depth
        ,random=~1|octagon/incore,data=data1,method="ML")
anova(m5b,m4a) ### test for effekt af temp
```

```
##     Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## m5b     1  7 1671.717 1694.520 -828.8586
## m4a     2  8 1658.051 1684.111 -821.0254 1 vs 2 15.66639   1e-04

### genfitter slutmodel med REML-estimation

modfinal<-lme(DW~temp+co2*depth
        ,random=~1|octagon/incore,data=data1,method="REML")
summary(modfinal)$tTable

##                    Value Std.Error DF   t-value      p-value
## (Intercept)     28.79834  3.374327 94  8.534543 2.385749e-13
## temp1           11.60920  2.814199 83  4.125225 8.752660e-05
## co21            24.16948  4.337316 10  5.572450 2.365801e-04
## depth0hor      -12.89209  3.149955 94 -4.092784 9.006758e-05
## co21:depth0hor -19.69404  4.454710 94 -4.420948 2.635506e-05

VarCorr(modfinal)

##             Variance     StdDev
## octagon =   pdLogChol(1)
## (Intercept)  17.79447     4.218349
## incore =    pdLogChol(1)
## (Intercept)  71.00654     8.426538
## Residual    238.13324    15.431566

### alternativt kan slutmodellen fittes med lmer()-funktionen

library(lme4)

## Loading required package: Matrix
## Loading required package: Rcpp
##
## Attaching package: 'lme4'
##
## The following object is masked from 'package:nlme':
##
##    lmList

modfinalalt<-lmer(DW~temp+co2*depth+(1|incore)+(1|octagon),data=data1)
modfinalalt

## Linear mixed model fit by REML ['lmerMod']
## Formula: DW ~ temp + co2 * depth + (1 | incore) + (1 | octagon)
##    Data: data1
## REML criterion at convergence: 1622.547
```

```
## Random effects:
##  Groups   Name        Std.Dev.
##  incore   (Intercept)  8.427
##  octagon  (Intercept)  4.218
##  Residual             15.432
## Number of obs: 192, groups:  incore, 96; octagon, 12
## Fixed Effects:
##    (Intercept)           temp1            co21       depthOhor
##          28.80           11.61           24.17          -12.89
## co21:depthOhor
##         -19.69

### udtraekker estimater for forskel
kontrol.Ohor<-c(1,0,0,1,0)
modified.Ohor<-c(1,1,1,1,1)
difference<-modified.Ohor-kontrol.Ohor
est<-rbind(kontrol.Ohor,modified.Ohor,difference)
library(gmodels)
estimable(modfinal,est,conf.int=0.95)

##                Estimate Std. Error  t value DF     Pr(>|t|)  Lower.CI
## kontrol.Ohor  15.90625   3.374327 4.713904 94 8.408525e-06  9.206448
## modified.Ohor 31.99090   3.374327 9.480673 10 2.585798e-06 24.472425
## difference    16.08464   5.170303 3.110967 10 1.104248e-02  4.564489
##                Upper.CI
## kontrol.Ohor  22.60606
## modified.Ohor 39.50937
## difference    27.60479
```

Opgave 2

1. Det vil være naturligt at udføre forsøget, som et fuldstændigt randomiseret blokforsøg. Inden for hver af de 4 blokke allokeres de 4 behandlinger til de 4 jordlodder ved lodtrækning. Det vil være naturligt, at lade `mark` indgå som en tilfældig effekt, således at den statistiske model bliver

$$Y_i = \gamma(\mathtt{A_i} \times \mathtt{B_i}) + \mathtt{A}(\mathtt{mark_i}) + \mathtt{e_i},$$

hvor $A(1),\ldots,A(4)$ er uafhængige $\sim N(0,\sigma^2_{\mathtt{mark}})$ og $e_1,\ldots,e_{16}$ er uafhængige $\sim N(0,\sigma^2)$. Et faktordiagram hørende til modellen ser ud som følger

$$
\begin{tikzcd}
& {[\mathtt{mark}^4_3]} \arrow[drr] & & \\
{[\mathtt{I}^{16}_9]} \arrow[ur] \arrow[dr] & & \mathtt{A}^2_1 \arrow[r] & \mathtt{0}^1_1 \\
& \mathtt{A}\times\mathtt{B}^4_1 \arrow[ur] \arrow[r] & \mathtt{B}^2_1 \arrow[ur] &
\end{tikzcd}
$$

2. Forsøget er et 2^n-te forsøg med 3 faktorer på hver 2 niveauer. Der er 4 blokke i forsøgsplanen, men det ses at præcis de samme 4 behandlinger forekommer på blok 1+4 hhv 2+3. Ved brug af skemaet nedenfor følger det af kompendiets sætning 9.6(?), at vekselvirkningen $\mathtt{A}\times\mathtt{B}$ er konfunderet med blok, når man betragter et par af marker (f.eks. 1+2 eller 3+4).

A	B	C	sum A+B	mark 1+4	mark 2+3	sum A+B+C	ny	plan
1	1	1	2	x		3	x	
1	1	2	2	x		4		x
1	2	1	3		x	4		x
1	2	2	3		x	5	x	
2	1	1	3		x	4		x
2	1	2	3		x	5	x	
2	2	1	4	x		5	x	
2	2	2	4	x		6		x

Ved den foreslåede forsøgsplan er vekselvirkningen $\mathtt{A}\times\mathtt{B}$ konfunderet 2 gange. Dette medfører, at styrken vil være lav ved test for effekten af pågældende vekselvirkning.

En alternativ forsøgplan opnås, hvis man bruger partiel konfundering og i stedet konfunderer to forskellige effekter på hvert par af marker. Man kunne f.eks. vælge at konfundere trefaktorvekselvirkningen $\mathtt{A}\times\mathtt{B}\times\mathtt{C}$ på to af markerne. I skemaet ovenfor er vist, hvilke 4 behandlinger der i givet fald skulle optræde på to af markerne. Partiel konfundering vil give en højere styrke ved test af vekselvirkningen $\mathtt{A}\times\mathtt{B}$ på bekostning af en lavere styrke ved test af trefaktorvekselvirkningen.

Man kunne også som alternativ forsøgsplan forslå at udføre forsøget som et splitplot-forsøg med f.eks. `A` som helplot-faktor og $\mathtt{B}\times\mathtt{C}$ som delplot-faktor. Umiddelbart virker dette som en dårligere forsøgsplan end den foreslåede, da det i princippet svarer til at dobbeltkonfundere hovedeffekten af `A` med `mark`. Der kunne imidlertid være praktiske hensyn som gjorde, at det var svært at variere den ene behandlingsfaktor inden for `mark`. I dette tilfælde kunne et split-plot forsøgsdesign være en fornuftig løsning.

3. På baggrund af oplysningerne i opgaveteksten identificeres følgende størrelser:

- antal blokke $v_B = 10$
- antal behandlinger per blok (blokstørrelse) $r_B = 4$
- antal behandlinger $v_T = 10$

Ved brug af det matematiske relationer i kompendiets sætning 9.6 konstateres, at hver behandling må optræde

$$r_T = \frac{v_B \cdot r_B}{v_T} = \frac{10 \cdot 4}{10} = 4$$

gange, hvis forsøget skal være et balanceret ufuldstændigt blokforsøg. Desuden skal hvert par af behandlinger mødes

$$\lambda = \frac{r_T(r_B - 1)}{v_T - 1} = \frac{4 \cdot (4-1)}{10-1} = \frac{12}{9} \approx 1.33$$

gange inden for samme blok i forsøgsplanen. Da λ *ikke* er et helt tal, kan der ikke findes et balanceret ufuldstændigt blokforsøg af den ønskede størrelse.

Opgave 3

1. Variablene `sex`, `gruppe` og deres vekselvirkning indgår som faktorer i modellen. Desuden indgår `age` og baselinemålingen `benpres.0` som kovariater i modellen, der kan opskrives som

   $$\texttt{benpres.6}_i = \alpha(\texttt{gruppe} \times \texttt{sex}_i) + \beta \cdot \texttt{benpres.0}_i + \gamma \cdot \texttt{age}_i + e_i,$$

   hvor $e_1, \ldots, e_{147}$ er uafhængige $N(0, \sigma^2)$. På baggrund af residualplottet virker antagelsen om varianshomogenitet rimelig. QQ-plottet viser, at de standardiserede residualer med rimelighed kan beskrives ved en standard-normalfordeling.

2. Estimat for 50 årig kvinde i kontrolgruppen med baselineværdien `benpres.0=100`

   $$100 \cdot 0.8761 + 50 \cdot (-0.2221) + 27.6090 = 104.1140$$

   Estimat for 50 årig mand i kontrolgruppen med baselineværdien `benpres.0=100`

   $$100 \cdot 0.8761 + 50 \cdot (-0.2221) + 27.6090 + 23.9697 - 16.5259 = 111.5578$$

3. På baggrund af `R`-udskriften konstateres, at man kan fjerne effekten af `age` ($F = 2.3561, p = 0.127$), svarende til modellen (`m1`)

   $$\texttt{benpres.6}_i = \alpha(\texttt{gruppe} \times \texttt{sex}_i) + \beta \cdot \texttt{benpres.0}_i + e_i.$$

   Dernæst konkluderes, at man kan fjerne vekselvirkningen mellem `sex` og `gruppe` fra modellen ($F = 3.2564, p = 0.07327$), svarende til modellen (`m3`)

   $$\texttt{benpres.6}_i = \alpha(\texttt{gruppe}_i) + \gamma(\texttt{sex}_i) + \beta \cdot \texttt{benpres.0}_i + e_i.$$

   På baggrund af et `summary()` af modellen `m3` konstateres, at ingen af de øvrige hovedeffekter kan testes væk

   - hovedeffekt af `benpres.0`: $t = 18.801, p < 0.0001$
   - hovedeffekt af `sex`: $t = -4.94, p < 0.0001$
   - hovedeffekt af `gruppe`: $t = 3.706, p = 0.0003$

Slutmodellen udtryker, at der er en lineær sammenhæng mellem baseline muskelstyrken (`benpres.0`) og muskelstyrken efter 6 måneder. Hældningen estimeres til $\hat{\beta} = 0.884$ (-og denne er i øvrige uafhængig af både `sex` og `gruppe`).

'Skæringen' svarende til kvinder i interventionsgruppen `gruppe=I,sex=F` estimeres til 29.83. Forskellen mellem mænd og kvinders muskelstyrke estimeres til 16.31, mens muskelstyrken i kontrolgruppen (`gruppe=K`) ligger 14.14 lavere end muskelstyrken i interventionsgruppen. Residualspredningen estimeres til $\hat{\sigma} = 17.3$.

4. I `R`-udskriften er `estimable()` benyttet til at udtrække forskellige linearkombinationer af parametrene i slutmodellen (`m3`). Svarende til kombinationen `est5` finder vi, at muskelstyrken ved 6 måneder for 50 årige mænd interventionsgruppen med baseline muskelstyrken 100 er

$$134.6 \quad [126.5 - 142.6] \quad (\text{estimat+95 \%-konfidensinterval}).$$