Eksamen i Statistisk Dataanalyse 2, 14. april 2016

Vejledende besvarelse

Opgave 1

1. Faktoren `subj` bør indgå med tilfældig effekt i modellen, og `eksp` (= `size` × `position`) som systematisk effekt. Da hver person har udført hvert eksperiment 10 gange, har vi målinger nok til også at inkludere vekselvirkningen `subj` × `eksp` i modellen som tilfældig effekt. Man kan rent faktisk også diskutere om vekselvirkninger mellem `subj` og `position` hhv `size` bør inddrages (som tilfældige effekter) i modellen, men det forventes ikke, at man gør dette. I princippet kan man også argumentere for, at de 10 replikationer for hver person af hvert eksperiment kan ses som gentagne målinger, og at den serielle korrelation kunne modelleres med f.x. en Diggle-model. Imidlertid er der ikke nogle variable i datasættet, som gør det muligt at afgøre, hvilken replikation der svarer til de enkelte datalinjer i datasættet. Derfor vil forsøg på at bruge Diggle modellen typisk føre til en fejlmeddelelse i R.

   I lyset af den foregående diskussion bliver en fornuftig udgangsmodel således (hvor `eksp` alternativt kan benyttes i stedet til at betegne vekselvirkningen)

$$\mathtt{v}_i = \gamma(\mathtt{size} \times \mathtt{position}_i) + A(\mathtt{subj}_i) + B(\mathtt{subj} \times \mathtt{eksp}_i) + e_i,$$

   hvor

   - $A(1), \ldots, A(10)$ er uafhængige og $\sim N(0, \sigma_A^2)$
   - $B(1,1), \ldots, B(10,15)$ er uafhængige og $\sim N(0, \sigma_B^2)$
   - $e_1, \ldots, e_{1500}$ er uafhængige og $\sim N(0, \sigma^2)$.

   Det vil også være naturlig at knytte en kommentar til besvarelsen om, at man har undersøgt om der er varianshomogenitet ved at kigge på f.x. residualplot.

2. Igennem hele opgaven er kovariansstrukturen givet som for udgangsmodellen (dvs. at vi inkluderer tilfældige effekter af `subj` og `subj` × `eksp`). Vi starter med at teste, om udgangsmodellen kan reduceres til

$$\mathtt{v}_i = \alpha(\mathtt{size}_i) + \beta(\mathtt{position}_i) + A(\mathtt{subj}_i) + B(\mathtt{subj} \times \mathtt{eksp}_i) + e_i,$$

   svarende til, at vi fjerner effekten af vekselvirkningen `size` × `position`. Det konstateres, at vekselvirkningen kan fjernes ($L.Ratio = 5.831, p = 0.666$).

   Dernæst konstateres, at vi hverken fjerne hovedeffekten af `position` ($L.Ratio = 93.47, p < 0.0001$) eller `size` ($L.Ratio = 374.31, p < 0.0001$).

Slutmodellen bliver den additive model

$$\mathtt{v}_i = \alpha(\mathtt{size}_i) + \beta(\mathtt{position}_i) + A(\mathtt{subj}_i) + B(\mathtt{subj} \times \mathtt{eksp}_i) + e_i,$$

hvor parameterestimater med tilhørende 95 %-konfidensintervaller bliver

$$\begin{aligned}
\hat{\alpha}(\mathtt{M}) + \hat{\beta}(\mathtt{15}) &= 1.167[1.059 - 1.275] & \hat{\alpha}(\mathtt{M}) + \hat{\beta}(\mathtt{22.5}) &= 1.157[1.049 - 1.265] \\
\hat{\alpha}(\mathtt{M}) + \hat{\beta}(\mathtt{30}) &= 1.120[1.012 - 1.228] & \hat{\alpha}(\mathtt{M}) + \hat{\beta}(\mathtt{37.5}) &= 1.086[0.978 - 1.194] \\
\hat{\alpha}(\mathtt{M}) + \hat{\beta}(\mathtt{45}) &= 1.061[0.952 - 1.169] & \hat{\alpha}(\mathtt{S}) - \hat{\alpha}(\mathtt{M}) &= -0.192[(-0.209) - (-0.174)] \\
\hat{\alpha}(\mathtt{T}) - \hat{\alpha}(\mathtt{M}) &= 0.183[0.165 - 0.200] & \hat{\sigma}_A &= 0.170[0.107 - 0.271] \\
\hat{\sigma}_B &= 0.033[0.026 - 0.041] & \hat{\sigma} &= 0.093[0.090 - 0.097].
\end{aligned}$$

Den additive model udtrykker, at effekten af forhindringens højde (`size`) på den maksimale hastighed (`v`) er uafhængig at placeringen (`position`) af forhindringen (og omvendt!).

De første 5 estimater ovenfor angiver den forventede maksimale hastighed for de 5 forskellige placeringer (`position`), når der benyttes en forhindring med `size = M`. Benyttes i stedet en forhindring med `size = S` ændres den maksimale hastighed med -0.192 m/s, mens en forhindring med `size = T` øger den maksimale hastighed med 0.183 m/s. Begge disse ændringer er signifikante.

Ved passende omparametriseringer kan man få estimaterne ud på en måde, så man kan diskutere, hvordan den maksimale hastighed ændres i takt med placeringen (`position`).

Ud fra estimater og 95 %-konfidensintervaller for variansparametrene ses, at ingen af varianskomponenterne kan sættes til 0. Hovedparten af den uforklarede variation skyldes variation mellem personer (mere præcist 74.8 %), mens residualvariation udgør 22.4 %. Kun 2.8 % af variationen kan tilskrives vekselvirkningen `subj` $\times$ `eksp`.

3. Vi tester her den additive model

$$\mathtt{v}_i = \alpha(\mathtt{size}_i) + \beta(\mathtt{position}_i) + A(\mathtt{subj}_i) + B(\mathtt{subj} \times \mathtt{eksp}_i) + e_i,$$

mod en model, hvor der er en lineær sammenhæng mellem `position` og maksimal hastighed. Bemærk, at da vi tester imod den additive model (for tosidet variansanalyse), så skal hældningen være den samme for hvert niveau af faktoren `size`. Den reducerede model bør derfor skrives som

$$\mathtt{v}_i = \alpha(\mathtt{size}_i) + \beta \cdot \mathtt{position}_i + A(\mathtt{subj}_i) + B(\mathtt{subj} \times \mathtt{eksp}_i) + e_i.$$

Det viser sig, at man godt kan antage, at der er en lineær sammenhæng mellem maksimal hastighed og placering ($L.Ratio = 3.009, p = 0.390$).

Det er ikke en del af opgaven at reducere modellen yderligere. For en god ordens skyld bemærkes dog, at hverken `size` eller `position` (svarende til hypotesen $H_0 : \beta = 0$) kan fjernes fra modellen.

4. Den letteste løsning på opgaven er at bemærke, at slutmodellen udtrykker at forskellen i maksimal hastighed når der benyttes en stor hhv. en middelstor forhindring er uafhængig

af værdien af placeringen (`position`). Derfor bliver den ønskede forskel 0.183 m/s [95 %-KI: 0.165-0.200] svarende til estimatet for $\alpha(\texttt{T}) - \alpha(\texttt{M})$.

En noget mere besværlig (men helt korrekt!) løsning er at benytte `estimable`-funktionen til at udtrække estimater og 95 %-konfidensintervaller for de to forskelle.

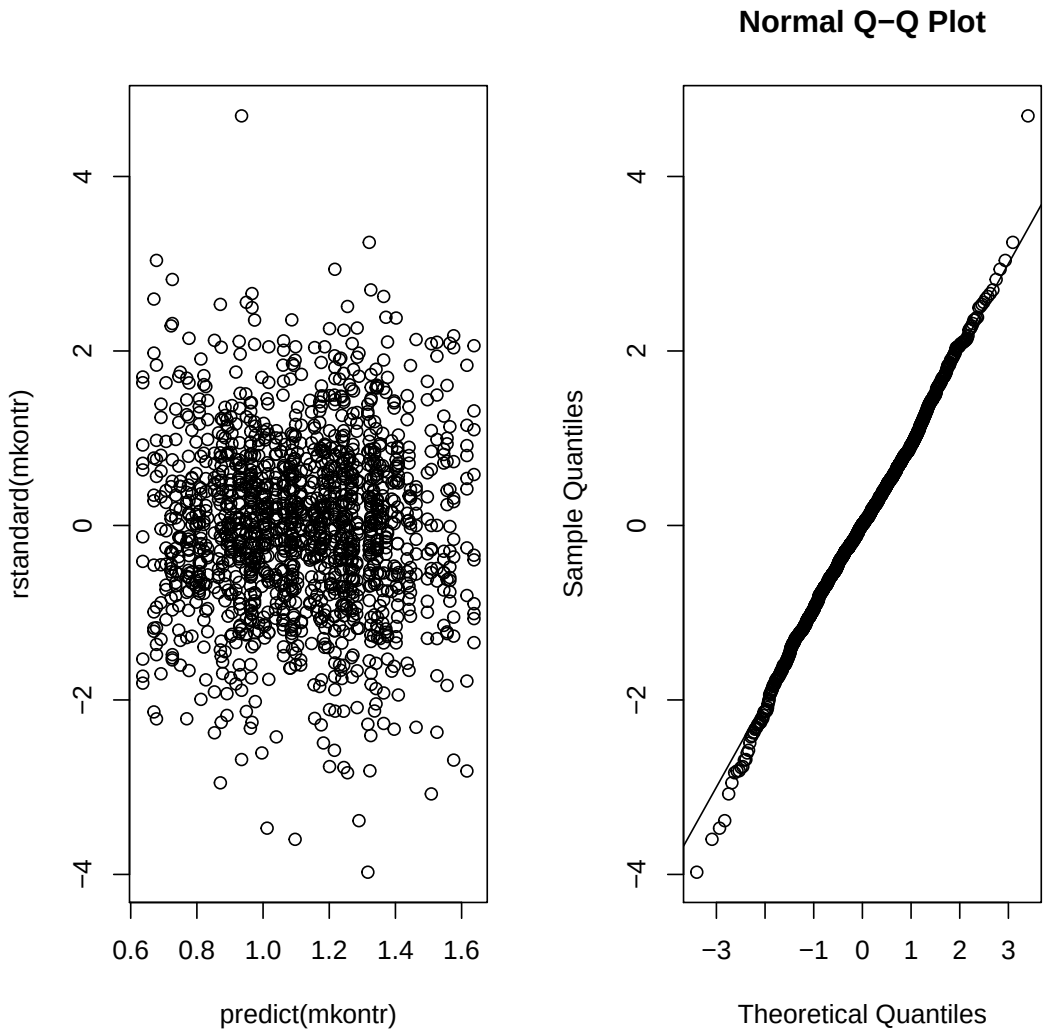
Eksempel på R-kode som kunne være brugt til løsning af opgave 1

```
### indlaesning af data
vdata <- read.table(file = "vdata.txt", header = T)
# vdata <- read.table(file = file.choose(), header = T)
### lav variable om til faktorer
vdata$subj <- factor(vdata$subj)
vdata$eksp <- factor(vdata$eksp)
vdata$fposition <- factor(vdata$position)
vdata$subjeksp <- vdata$subj:vdata$eksp
head(vdata)

##   subj eksp size position        v fposition subjeksp
## 1    1    1    S       15 1.253992        15      1:1
## 2    1    1    S       15 1.145326        15      1:1
## 3    1    1    S       15 1.146460        15      1:1
## 4    1    1    S       15 1.143079        15      1:1
## 5    1    1    S       15 1.200853        15      1:1
## 6    1    1    S       15 1.186442        15      1:1
```

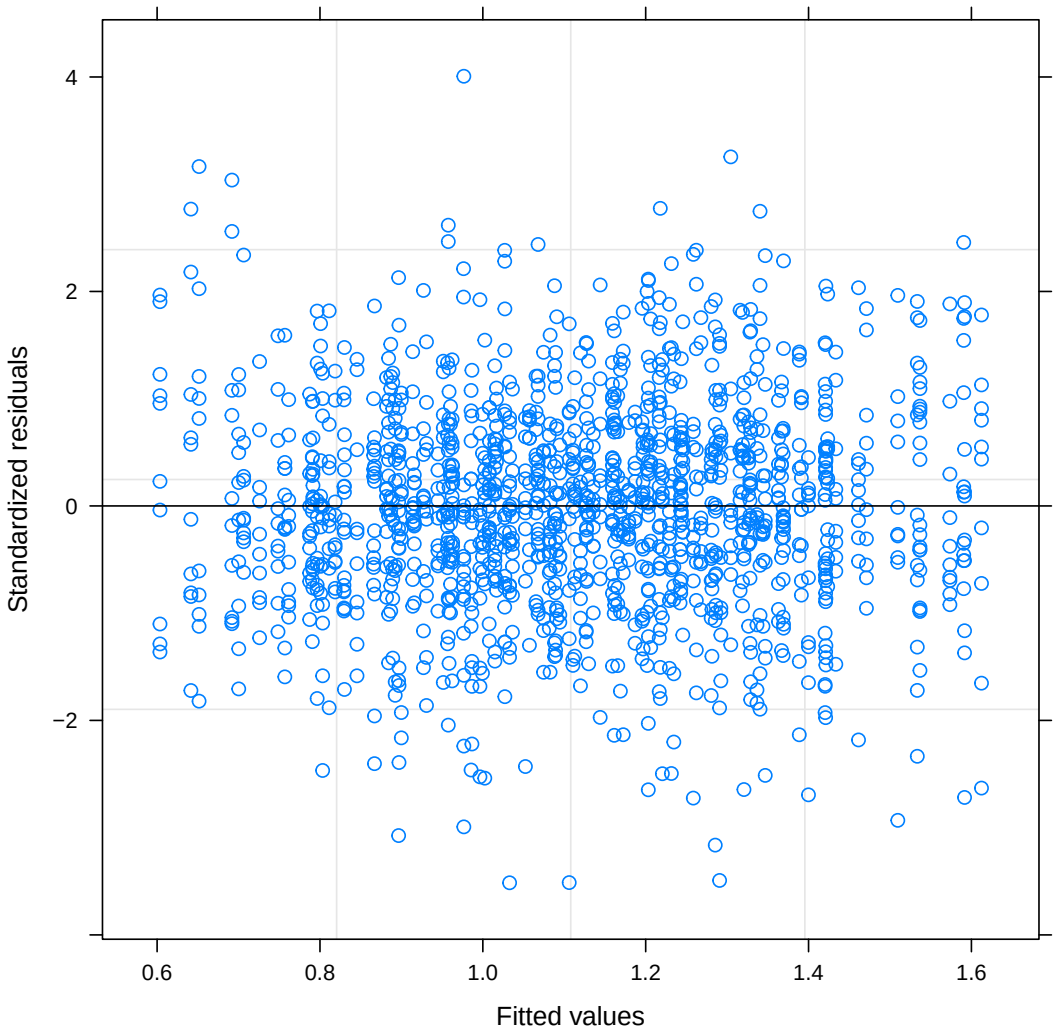
Residualplot for udgangsmodel (2 metoder)

```
### mest brugt paa kurset
mkontr <- lm(v ~ eksp + subjeksp, data = vdata)
par(mfrow = c(1,2))
plot(predict(mkontr), rstandard(mkontr))
qqnorm(rstandard(mkontr))
abline(0,1)
```




```
par(mfrow = c(1,1))

### fit af udgangsmodel og tilhoerende residualplot
library(nlme)
m0 <- lme(v ~ fposition*size, random = ~ 1|subj/subjeksp
          , data = vdata, method = "ML")
plot(m0)
```




```
### model reduktion (position som faktor)
m1 <- lme(v ~ fposition + size, random = ~ 1|subj/subjeksp
          , data = vdata, method = "ML")
anova(m1, m0)

##    Model df       AIC       BIC   logLik   Test  L.Ratio p-value
## m1     1 10 -2674.191 -2621.058 1347.095
## m0     2 18 -2664.021 -2568.383 1350.011 1 vs 2 5.830761  0.6662

m2a <- lme(v ~ fposition, random = ~ 1|subj/subjeksp
           , data = vdata, method = "ML")
anova(m2a, m1)

##     Model df       AIC       BIC   logLik   Test  L.Ratio p-value
## m2a     1  8 -2303.880 -2261.374 1159.940
## m1      2 10 -2674.191 -2621.058 1347.095 1 vs 2 374.3106  <.0001
```

```r
m2b <- lme(v ~ size, random = ~ 1|subj/subjeksp
           , data = vdata, method = "ML")
anova(m2b, m1)
```

```
##     Model df       AIC       BIC   logLik   Test  L.Ratio p-value
## m2b     1  6 -2588.724 -2556.844 1300.362
## m1      2 10 -2674.191 -2621.058 1347.095 1 vs 2 93.46693  <.0001
```

```r
### slutmodel (position som faktor)
m1final <- lme(v ~ fposition + size-1, random = ~ 1|subj/subjeksp
               , data = vdata, method = "REML")
summary(m1final)$tTable
```

```
##                    Value   Std.Error  DF   t-value      p-value
## fposition15    1.1669363 0.054698226 134  21.33408 6.366712e-45
## fposition22.5  1.1568179 0.054698226 134  21.14909 1.568637e-44
## fposition30    1.1198508 0.054698226 134  20.47326 4.386557e-43
## fposition37.5  1.0858238 0.054698226 134  19.85117 9.903854e-42
## fposition45    1.0606049 0.054698226 134  19.39011 1.029912e-40
## sizeS         -0.1918564 0.008812282 134 -21.77148 7.679890e-46
## sizeT          0.1828133 0.008812282 134  20.74529 1.139849e-43
```

```r
intervals(m1final)
```

```
## Approximate 95% confidence intervals
##
##  Fixed effects:
##                    lower       est.      upper
## fposition15    1.0587527  1.1669363  1.2751198
## fposition22.5  1.0486343  1.1568179  1.2650014
## fposition30    1.0116672  1.1198508  1.2280343
## fposition37.5  0.9776403  1.0858238  1.1940074
## fposition45    0.9524213  1.0606049  1.1687884
## sizeS         -0.2092856 -0.1918564 -0.1744272
## sizeT          0.1653842  0.1828133  0.2002425
## attr(,"label")
## [1] "Fixed effects:"
##
##  Random Effects:
##   Level: subj
##                      lower      est.     upper
## sd((Intercept)) 0.1070793 0.1703319 0.2709484
##   Level: subjeksp
##                       lower       est.      upper
## sd((Intercept)) 0.02631223 0.03274949 0.04076162
##
```

```
##  Within-group standard error:
##      lower       est.      upper
## 0.08976317 0.09321365 0.09679677

### slutmodel (position som faktor) - alternativ parametrisering
m1final2 <- lme(v ~ size + fposition - 1, random = ~ 1|subj/subjeksp
                , data = vdata, method = "REML")
summary(m1final2)$tTable

##                     Value  Std.Error  DF    t-value      p-value
## sizeM          1.16693628 0.05469825 134 21.3340715 6.366947e-45
## sizeS          0.97507988 0.05469825 134 17.8265293 3.533630e-37
## sizeT          1.34974960 0.05469825 134 24.6762869 1.104095e-51
## fposition22.5 -0.01011841 0.01137661 134 -0.8894044 3.753797e-01
## fposition30   -0.04708551 0.01137661 134 -4.1387984 6.134926e-05
## fposition37.5 -0.08111244 0.01137661 134 -7.1297525 5.643896e-11
## fposition45   -0.10633143 0.01137661 134 -9.3464918 2.688382e-16

### modelreduktion (position som numerisk variabel)
### test mod additiv model
mlinadd <- lme(v ~ position + size-1, random = ~ 1|subj/subjeksp
               , data = vdata, method = "ML")
anova(mlinadd, m1)

##         Model df       AIC       BIC   logLik   Test  L.Ratio p-value
## mlinadd     1  7 -2677.182 -2639.989 1345.591
## m1          2 10 -2674.191 -2621.058 1347.095 1 vs 2 3.009027  0.3902

### test mod vekselvirkningsmodel
mlin <- lme(v ~ size + size:position-1, random = ~ 1|subj/subjeksp
            , data = vdata, method = "ML")
anova(mlin, m0)

##      Model df       AIC       BIC   logLik   Test  L.Ratio p-value
## mlin     1  9 -2673.230 -2625.411 1345.615
## m0       2 18 -2664.021 -2568.383 1350.011 1 vs 2 8.791689  0.4567

### test for om haeldning afhaenger af 'size'
anova(mlinadd, mlin)

##         Model df       AIC       BIC   logLik   Test    L.Ratio p-value
## mlinadd     1  7 -2677.182 -2639.989 1345.591
## mlin        2  9 -2673.230 -2625.411 1345.615 1 vs 2 0.04809938  0.9762

### slutmodel (position som kovariat)
mlinaddfinal <- lme(v ~ position + size-1, random = ~ 1|subj/subjeksp
                    , data = vdata, method = "REML")
summary(mlinaddfinal)$tTable
```

```
##                 Value    Std.Error  DF   t-value      p-value
## position -0.003782092 0.0003390754 137 -11.15413 5.873549e-21
## sizeM     1.231469474 0.0551686449 137  22.32191 1.768083e-47
## sizeS     1.039613071 0.0551686449 137  18.84427 7.273448e-40
## sizeT     1.414282792 0.0551686449 137  25.63563 3.926137e-54
```

```
### slutmodel (position som kovariat) - alternativ parametrisering
mlinaddfinal2 <- lme(v ~ position + size, random = ~ 1|subj/subjeksp
                     , data = vdata, method = "REML")
summary(mlinaddfinal2)$tTable
```

```
##                   Value    Std.Error   DF   t-value      p-value
## (Intercept)  1.231469474 0.0551685209 1350  22.32196 3.381720e-94
## position    -0.003782092 0.0003390751  137 -11.15414 5.873213e-21
## sizeS       -0.191856404 0.0088094303  137 -21.77853 2.475819e-46
## sizeT        0.182813318 0.0088094303  137  20.75200 4.007218e-44
```

```
intervals(mlinaddfinal2)$fixed
```

```
##                   lower         est.        upper
## (Intercept)  1.12324413  1.231469474  1.339694818
## position    -0.00445259 -0.003782092 -0.003111594
## sizeS       -0.20927645 -0.191856404 -0.174436361
## sizeT        0.16539328  0.182813318  0.200233360
## attr(,"label")
## [1] "Fixed effects:"
```

Opgave 2

1. For at argumentere for at forsøget *ikke* er et BIBD kan man lave en co-incidensmatrix

	A	B	C	D	E	F	G
A	4	2	2	2	2	2	2
B		4	2	3	1	2	2
C			4	1	3	2	2
D				4	2	2	2
E					4	2	2
F						4	2
G							4

   Ved at ombytte behandling `E` fra mark 2 med behandling `D` fra mark 7 fås et BIBD.

   Randomiseringen foretages i to trin

   - Ved lodtrækning bestemmes hvordan de 7 blokke i forsøgsplanen, skal svare til de konkrete 7 marker, som skal indgå i forsøget.

- For hver af de 7 marker foretages en lodtrækning om, hvordan de fire sorter fra den relevante blok i forsøgsplanen skal fordeles på de 4 forsøgsenheder inden for marken.

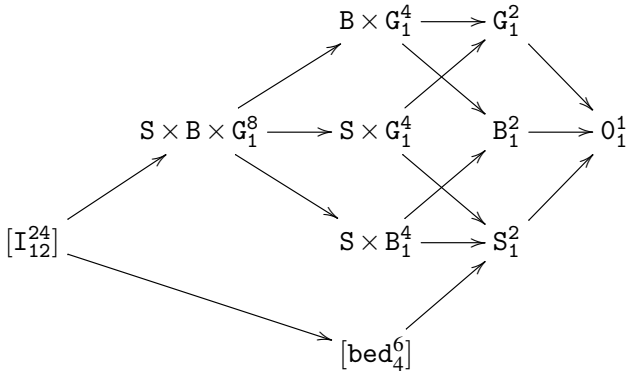
2. Det vil være oplagt at afprøve de 4 kombinationer af B og G præcis en gang i hvert bed. Forsøgsplanen kan da karakteriseres som et splitplot forsøg med bed som helplot, sort som helplotfaktor og $\mathtt{B} \times \mathtt{G}$ som delplotfaktor.

   Den naturlige statistiske model til analyse af forsøget bliver

$$Y_i = \alpha(\mathtt{S} \times \mathtt{B} \times \mathtt{G}_i) + A(\mathtt{bed}_i) + e_i,$$

   hvor $A(1), \ldots A(6)$ er uafhængige $\sim N(0, \sigma^2_{\mathtt{bed}})$ og $e_1, \ldots, e_{24}$ er uafhængige $\sim N(0, \sigma^2)$. Det forventes ikke, at man diskuterer muligheden for at inddrage vekselvirkninger mellem bed og B hhv. G i modellen, selvom dette faktisk er muligt.

   Faktordiagrammet for forsøget ser ud som følger



3. Forsøget er et 2^n-forsøg med 3 faktorer (=8 behandlingskombinationer), der skal fordeles ud på blokke (=bede) af størrelse 4. Man kan derfor benytte kompendiets Theorem 9.11 (lige-ulige reglen) til allokering af behandlinger på par af blokke, således at man for hvert par styrer, hvilken effekt der konfunderes med blok.

   I denne delopgave er der det ekstra krav, at behandlingen S (=sort) skal afprøves to gange for hver blok. Dette sker automatisk uanset, hvad man vælger at konfundere, bortset fra at man naturligvis ikke må konfundere hovedeffekten af S. En mulig forsøgsplan kunne være at benytte sig af partiel konfundering, hvor man konfunderer $\mathtt{S} \times \mathtt{B} \times \mathtt{G}$ på blok 1+2, $\mathtt{S} \times \mathtt{B}$ på blok 3+4 og $\mathtt{S} \times \mathtt{G}$ på blok 5+6. Dette svarer til følgende forsøgsplan

S	B	G	b1	b2	b3	b4	b5	b5
1	1	1	x			x		x
1	1	2		x		x	x	
1	2	1		x	x			x
1	2	2	x		x		x	
2	1	1		x	x		x	
2	1	2	x		x			x
2	2	1	x			x	x	
2	2	2		x		x		x

Opgave 3

1. Variablen `arena` bør indgå i modellen med tilfældig effekt. Som udgangspunkt vil det være naturligt også at inkludere variablen `colony` som tilfældig effekt. Det er ikke klart fra udplukket af datasættet, om faktoren `colony` er grovere end faktoren `arena`. Derfor er den sikre løsning, at fitte modellen i R ved brug af `lmer`-funktionen. Hvis `colony` var grovere end `arena` ville vi også kunne have fittet modellen med `lme`.

   Ved at betragte residualplottene i opgaveformuleringen (svarende til modeller hvor `arena` inddrages som *fixed effect*) ser vi, at responsvariablen bør log-transformeres for at opnå varianshomogenitet.

   Konklusionen er at den bedste model til beskrivelse af datasættet er

   $$\log(\texttt{tid}_i) = \alpha(\texttt{exposed}_i) + \beta \cdot \texttt{age}_i + A(\texttt{colony}_i) + B(\texttt{arena}_i) + e_i,$$

   hvor $A(\texttt{A}), \ldots, A(\texttt{F})$ er uafhængige $\sim N(0, \sigma^2_{\texttt{colony}})$, $B(\texttt{A1}), \ldots, B(\texttt{A61})$ er uafhængige $\sim N(0, \sigma^2_{\texttt{arena}})$ og $e_1, \ldots, e_{366}$ er uafhængige $\sim N(0, \sigma^2)$. Dette svarer til `modelD` i R-udskriften.

   Alternativt, hvis man ikke ønsker at inddrage `colony` i modellen (og det kræver i det mindste en kommentar), så kan datasættet analyseres med udgangspunkt i `modelE`.

2. Parameterestimaterne til beskrivelse af middelværdistrukturen estimeres til

   $$\hat{\alpha}(\texttt{yes}) = 4.929, \quad \hat{\alpha}(\texttt{no}) - \hat{\alpha}(\texttt{yes}) = -0.113, \quad \hat{\beta} = -0.089.$$

   og parameterestimaterne for variansparametrene er

   $$\hat{\sigma}_{\texttt{colony}} = 0.440, \quad \hat{\sigma}_{\texttt{arena}} = 0.454, \quad \hat{\sigma} = 1.509.$$

   Parameterestimaterne rapporteret ovenfor er alle trukket fra `modelD` i R-udskriften.

   For at diskutere hvilke variable, der bidrager væsentlig til beskrivelse af variationen i data, så er man nødt til at kigge på konfidensintervallerne for parametrene fra `modelD`. Omkring parametrene som indgår i beskrivelse af middelværdistrukturen konkluderes følgende:

   - Konfidensintervallet for β indeholder 0, hvilket betyder at der ikke er en signfikant effekt af alder (`age`) på responsvariablen `tid`.
   - Konfidensintervallet for forskellen $\alpha(\texttt{no}) - \alpha(\texttt{yes})$ indeholder 0, hvilket betyder at der *ikke* er signfikant forskel på den tid, arbejderbierne er i kontakt med dronningen afhængigt af om de er inficerede med *nosema ceranae* eller ej.

   Betragtes konfidensintervallerne for variansparametrene så ses faktisk, at konfidensintervallet for $\sigma_{\texttt{colony}}$ indeholder 0. Man kan derfor argumentere for, at denne varianskomponent kan udelades fra modellens, svarende til at vi i princippet lige så godt kunne benytte `modelE`.

   Hvis man på baggrund af delspørgsmål 1. valgte at benytte `modelE` til analyse af forsøget, så bliver konklusionerne vedr. parametrene i middelværdistrukturen uændrede i forhold til konklusionerne fra `modelD`. Dog bliver tingene lidt lettere: for `modelE` kan man af R-udskriften aflæse både p-værdier og 95 %-konfidensintervaller.

3. Som en del af R-udskriften er `estimable`-funktionen benyttet til at udtrække forskellige linearkombinationer af parametrene i et par af modellerne. Man kan argumentere for, at vi helst ville have trukket estimater ud fra `modelD`, men blandt de to modeller som er benyttet i R-udskriften er det i alle tilfælde meste oplagt at kigge på resultaterne fra `modelE`. Et andet godt argument er, at vi i delspørgsmål 2. så, at den tilfældige effekt af $\sigma_{\texttt{colony}}$ kan undværes, hvilket fører os over i `modelE`.

   Ud for den relevante linearkombination af parametrene (svarende til `est3`) aflæses, at *logaritmen(!)* til den forventede kontakttid for en 10 dage gammel arbejderbi som ikke har være eksponeret for *nosema ceranae* er 3.889 [KI: 3.522-4.257].

   Omregnes dette (ved at tage eksponentialfunktionen) fås resultatet 48.9 sekunder [tilbageregnet KI: 33.9- 70.6] på oprindelig skala. Dette harmonerer fint med, at man på baggrund af de tidligere forsøg så en kontakttid på ca. 1 % af en time svarende til $0.01 \cdot 3600 = 36$ sekunder.