Eksamen i Statistisk Dataanalyse 2, 3. april 2014

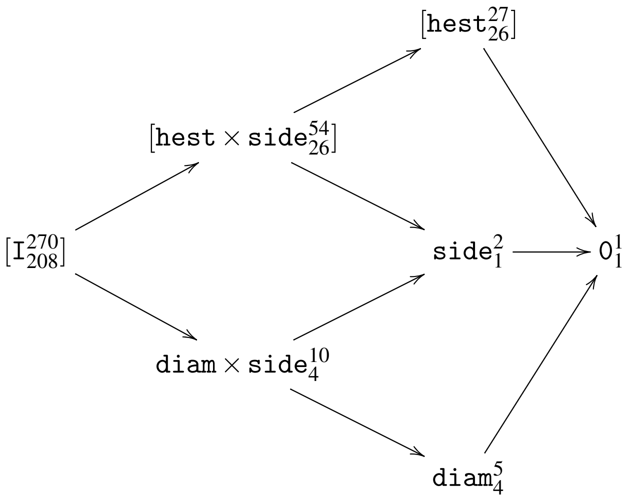
Vejledende besvarelse

Opgave 1

1. Ved besvarelsen af delspørgsmål 1.-2. bør `diam` og `side` indgå i modellen som systematiske effekter (faktorer), mens `hest` bør indgå som tilfældig effekt. Heraf følger desuden, at $\texttt{diam} \times \texttt{side}$ skal indgå som systematisk effekt, samt at $\texttt{hest} \times \texttt{side}$ skal være en tilfældig effekt. Den statistiske model bliver

$$S_i = \alpha(\texttt{diam} \times \texttt{side}_\texttt{i}) + A(\texttt{hest} \times \texttt{side}_i) + B(\texttt{hest}_i) + e_i,$$

hvor $A()'$ erne er uafhængige $\sim N(0, \sigma^2_{\texttt{hest} \times \texttt{side})}$, $B()'$ erne er uafhængige $\sim N(0, \sigma^2_{\texttt{hest})}$ og e_i'erne er uafhængige $\sim N(0, \sigma^2)$. Det tilhørende faktordiagram ser ud som følger Et faktordiagram for forsøget ser ud som følger



2. Med udgangspunkt i modellen fra 1. testes i første omgang, om vi kan fjerne vekselvirkningen $\texttt{diam} \times \texttt{side}$. Dette svarer til at modellen reduceres til

$$S_i = \beta(\texttt{diam}_i) + \gamma(\texttt{side}_i) + A(\texttt{hest} \times \texttt{side}_i) + B(\texttt{hest}_i) + e_i,$$

hvor den tilfældige del af modellen stadig ser ud som beskrevet under 1. Vi finder at vekselvirkningen kan fjernes ($L.ratio = 1.739, p = 0.7836$).

Dernæst konstateres, at hovedeffekten af `side` kan fjernes fra modellen ($L.ratio = 1.809, p = 0.1786$), svarende til at modellen reduceres til

$$S_i = \beta(\texttt{diam}_i) + A(\texttt{hest} \times \texttt{side}_i) + B(\texttt{hest}_i) + e_i, \qquad (1)$$

hvor den tilfældige del af modellen stadig ser ud som beskrevet under 1.

Endelig konstateres, at der er en signifikant effekt af `diam` ($L.ratio = 69.01, p < 0.0001$), hvorfor slutmodellen bliver (1).

Parameterestimaterne for de systematiske effekter under slutmodellen kan angives som

$$\hat{\beta}(8) = 4.740[4.538 - 4.942] \qquad \hat{\beta}(10) = 4.950[4.748 - 5.152]$$
$$\hat{\beta}(12) = 5.12[4.916 - 5.319] \qquad \hat{\beta}(14) = 5.201[4.999 - 5.403]$$
$$\hat{\beta}(16) = 5.214[5.012 - 5.415]$$

Parameterestimaterne for variansparametrene i slutmodellen bliver

$$\hat{\sigma}_{\texttt{hest}} = 0.356[0.202 - 0.627] \quad \hat{\sigma}_{\texttt{hest}\times\texttt{side}} = 0.452[0.336 - 0.607] \quad \hat{\sigma} = 0.329[0.299 - 0.362].$$

3. Med udgangspunkt i slutmodellen fra delspørgsmål 2.

$$S_i = \beta(\texttt{diam}_i) + A(\texttt{hest} \times \texttt{side}_i) + B(\texttt{hest}_i) + e_i,$$

kan man teste om modellen kan reduceres til

$$S_i = \mu + \delta \cdot \texttt{diam}_i + A(\texttt{hest} \times \texttt{side}_i) + B(\texttt{hest}_i) + e_i,$$

hvor `diam` indgår som en numerisk variabel. Man finder at likelihood ratio teststørrelsen bliver $L.Ratio = 8.266$ svarende til en *approximativ* $p - value = 0.0408$. På baggrund heraf forkastes hypotesen om, at der en lineær sammenhæng mellem diameter og symmetriscoren `S`. Man kan argumentere for, at det ville være fornuftigt at prøve at simulere p-værdien.

4. Opgaven kan besvares ved at sammenligne værdien på 5.63 fra det tidligere studie med konfidensintervallerne for den forventede symmetriscore fra slutmodellen i delspørgsmål 1.-3. Benyttes f.eks. slutmodellen (1) fra delspørgsmål 2. ses f.eks., at konfidensintervallerne *ikke* indeholder 5.63 uanset om diameteren i cirklen er $\texttt{8}, \texttt{10}, \texttt{12}, \texttt{14}$ eller `16` meter. Det ser med andre ord ud til, at heste som løber ligeud har en mere symmetrisk gang end heste som løber i cirkler, i hvert tilfælde når cirklens diameter ligger i intervallet fra $8 - 16$ meter. Kigger man nærmere på estimaterne fra slutmodellen fra delspørgsmål 1.-3. får man faktisk et indtryk, at symmetriscoren ikke ville ændre sig selvom diameteren i cirklen blev gjort endnu større. Dette kunne undersøges mere formelt i fremtidige studier.

5. På baggrund af analyserne fra delspørgsmål 1.-2. har vi set, at både effekten af $\texttt{diam} \times \texttt{side}$ og `side` kan testes væk. Der er således ikke en systematisk tendens til, at hestes gang bliver mere (eller mindre) symmetrisk, når de løber i cirkler mod højre end når de løber mod venstre. Imidlertid har vi inkluderet en tilfældig effekt af $\texttt{hest} \times \texttt{side}$ i modellen. Denne effekt giver i princippet mulighed for at beskrive, om hver enkelt hests gang ikke behøver være lige symmetrisk mod højre og venstre. På baggrund af estimaterne i delspørgsmål 2. kan vi beregne, hvor stor

en del af den totalt variation i symmetriscoren `S` som skyldes varianskomponenten fra $\mathtt{hest} \times \mathtt{side}$

$$\frac{0.452^2}{0.356^2 + 0.452^2 + 0.329^2} = 0.465$$

svarende til 46.5%. Opgaven lægger op til at teste, om varianskomponenten svarende til $\mathtt{hest} \times \mathtt{side}$ kan sættes til 0 svarende til hypotesen, $H_0 : \sigma^2_{\mathtt{hest} \times \mathtt{side}} = 0$. Hypotesen forkastes ($F = 10.419, p < 0.0001$), hvorfor vi konkluderer at variationen fra $\mathtt{hest} \times \mathtt{side}$ bidrager signifikant til beskrivelsen af variationen i datamaterialet. Konklusionen er, at enkelte heste som udgangspunkt har en foretrukken omløbsretning, men at man overordnet set ikke kan konkludere at heste som helhed hellere vil løbe mod højre end mod venstre.

Opgave 2

1. Forsøget kan opfattes som et splitplot design med smagsdommer som helplot, `konsistens` som helplot-faktor og vekselvirkningen $\mathtt{sukker} \times \mathtt{aroma}$ som delplot-faktor. Man kan ligeledes tænke på forsøget som et 2^n-forsøg med 3 faktorer på hver 2 niveauer, hvor man lader par bestående af 2 smagsdommere smage på de 8 yoghurt-kombinationer på en måde, så hovedeffekten af `konsistens` konfunderes med `smagsdommer`.

   Faktordiagrammet hørende til forsøget bliver

$$\begin{tikzcd}
 & & \mathtt{s} \times \mathtt{a}^4_1 \arrow[r] \arrow[dr] & \mathtt{a}^2_1 \arrow[ddr] & \\
 & \mathtt{s} \times \mathtt{a} \times \mathtt{k}^8_1 \arrow[ur] \arrow[r] \arrow[dr] & \mathtt{a} \times \mathtt{k}^4_1 \arrow[ur] \arrow[ddr] & \mathtt{s}^2_1 \arrow[dr] & \\
 {[\mathtt{I}^{32}_{18}]} \arrow[ur] \arrow[dr] & & \mathtt{s} \times \mathtt{k}^4_1 \arrow[ur] \arrow[dr] & & \mathtt{0}^1_1 \\
 & {[\mathtt{d}^8_6]} \arrow[rr] & & \mathtt{k}^2_1 \arrow[ur] &
\end{tikzcd}$$

   og den tilhørende statistiske model indholder vekselvirningen $\mathtt{sukker} \times \mathtt{aroma} \times \mathtt{konsistens}$ som systematisk effekt og `smagsdommer` som tilfældig effekt

$$y_i = \alpha(\mathtt{sukker} \times \mathtt{aroma} \times \mathtt{konsistens}_i) + A(\mathtt{smagsdommer}_i) + e_i,$$

   hvor $A(1), \ldots, A(8)$ er uafhængige $\sim N(0, \sigma^2_{sdommer})$ og $e_1, \ldots, e_{32}$ er uafhængige $\sim N(0, \sigma_{2)}$. Man kunne inkludere parvis vekselvirkninger mellem `smagsdommer` og `sukker` hhv `aroma`, men dette er ikke påkrævet, som beskrevet i opgaveformuleringen. I givet fald skal disse vekselvirkninger inddrages som tilfældige effekter.

   Randomiseringen foretages i to trin. Først udvælges ved randomisering hvilke 4 `smagsdommere` der skal prøve yoghurt med `vandig` konsistens og hvilke 4 der skal prøve yoghurt med `tyk` konsistens. For hver af de 8 `smagsdommere` bør der dernæst trækkes lod om, i hvilken rækkefølge yoghurt med de 4 kombinationer af `aroma` og `sukker` skal afprøves.

2. Der er tale om et 2^n-te forsøg med 3 faktorer, hvor der skal anvendes partiel konfundering af forskellige effekter på hvert par af smagsdommere. Den tilhørende forsøgsplan bliver som anført nedenfor, hvis man konfunderer `s` $\times$ `a` for `d=3,4`, `s` $\times$ `k` for `d=5,6` og `a` $\times$ `k` for `d=7,8`

`kombination nr`	`s`	`a`	`k`	d=1	d=2	d=3	d=4	d=5	d=6	d=7	d=8
1	1	1	1	x			x		x		x
2	1	1	2		x		x	x		x	
3	1	2	1		x	x			x	x	
4	1	2	2	x		x		x			x
5	2	1	1		x	x		x			x
6	2	1	2	x		x			x	x	
7	2	2	1	x			x	x		x	
8	2	2	2		x		x		x		x

3. Opgaven løses ved at sætte 4 markeringer ud for hver af dommerne `d=6` og `d=7` på en sådan måde, at hver af de 7 kombinationer ender med at blive afprøvet 4 gange i forsøgsplanen, og så hvert par af kombinationer kommer til at optræde 2 gange i forsøgsplanen under den samme dommer (-dvs i samme søjle). Man kan evt. vedlægge en koincidensmatrix til sin besvarelse, som dokumentation for, at man har gjort det rigtigt.

`kombination nr`	`sukker`	`aroma`	`konsistens`	1	2	3	4	5	6	7
2	1	1	2		x	x			o	o
3	1	2	1		x	x	x	x		
4	1	2	2	x		x		x	o	
5	2	1	1	x		x	x			o
6	2	1	2	x	x			x		o
7	2	2	1	x	x		x		o	
8	2	2	2				x	x	o	o

Opgave 3

1. Analysen bør tage udgangspunkt i den største model, der indeholder vekselvirkningen mellem `gender` og `depression` samt `vas.smerter` (-som kovariat). På baggrund af residualplot og QQ-plot for modellerne `model1` og `smodel1` kan man argumentere for at antagelserne om varianshomogenitet og normaltfordelte fejl med rimelighed er opfyldt for begge modeller. Der er måske en svag tendens til, at der opnås en lidt højere grad af varianshomogenitet, hvis man anvender `smodel1`, hvor man bruger $\sqrt{psqi}$ som responsvariabel. Begge dele vil blive anset som korrekt, under forudsætning af at der gives en fornuftig begrundelse for valget af model baseret på figurerne.

   Resultaterne i resten af den vejledende besvarelse er baseret på udgangsmodellen (`smodel1`)

$$\sqrt{\texttt{psqi}_i} = \alpha(\texttt{gender} \times \texttt{depression}_i) + \gamma \cdot \texttt{vas.smerter}_i + e_i,$$

   hvor e_i'erne er uafhængige $\sim N(0, \sigma^2)$.

2. I første omgang reduceres modellen til en model uden vekselvirkning mellem `gender` og `depression`

$$\sqrt{\texttt{psqi}_i} = \beta(\texttt{gender}_i) + \delta(\texttt{depression}_i) + \gamma \cdot \texttt{vas.smerter}_i + e_i.$$

   Den tilhørende $F-$teststørrelse bliver 2.0941 svarende til $p = 0.1487$ (-se ud for `anova(smodel2,smodel1)` i `R`-udkskriften). Vi godkender hypotesen om, at der ikke er vekselvirkning.

   Dernæst konstateres, at man kan fjerne hovedvirkningen af `gender` ($F = 3.5087, p = 0.06186$) kan fjernes (-se `anova(smodel4,smodel2)`), således at vores nye model er

$$\sqrt{\texttt{psqi}_i} = \delta(\texttt{depression}_i) + \gamma \cdot \texttt{vas.smerter}_i + e_i.. \qquad (2)$$

   Endelig kan man af `R`-udskriften efter `anova(smodel5,smodel4)` se, at `depression` ($F = 7.679, p = 0.0059$) er signifikant associeret med søvnkvalitet. Af `summary(smodel4)` ses desuden, at `vas.smerter` ($t = 8.862, p < 0.0001$) er signifikant associeret med søvnkvalitet, således at (2) bliver vores slutmodel.

   Parameterestimaterne under slutmodellen bliver

$$\begin{aligned} \hat{\delta}(0) &= 2.1758[2.0579 - 2.2938] & \hat{\delta}(1) - \hat{\delta}(0) &= 0.3305[0.0960 - 0.5650] \\ \hat{\gamma} &= 0.0126[0.0099 - 0.0153] & \hat{\sigma}^2 &= 0.692^2. \end{aligned}$$

   Slutmodellen udtrykker, at søvnkvaliteten (-målt via $\sqrt{\texttt{psqi}}$) afhænger linært af `vas.smerter` og at personer med depression sover dårligere. Derimod ser køn (`gender`) ikke ud til at hænge sammen med søvnkvaliteten, vel at mærke når man samtidig har justeret modellen for `depression` og `vas.smerter`.

3. Estimatet beregnes på baggrund af slutmodellen (2) fra delspørgsmål 2. Man skal til det formål benytte udskriften fra `smodel4` i `R`-udskriften, når man anvender `estimable` med coefficienterne (=bestillingslisten) svarende til `est2`. Den forventede værdi af $\sqrt{\texttt{psqi}}$ (-husk at vi regner på transformeret respons!) bliver 2.884 med tilhørende 95 %-konfidensinterval $[2.660 - 3.107]$. Dette estimat bør tilbagetransformeres ved at tage kvadratet på resultat, så man får et estimat for `psqi` som er $8.315[KI : 7.078 - 9.651]$.

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

```
> data1<-read.table(file="data1.txt",header=T)
> library(nlme)
> ### lav ny faktor svarende til vekselvirkningen hest x side
> data1$hestside<-data1$hest:data1$side
> ### fit udgangsmodellen
> mod0<-lme(S~factor(diam)*side,random=~1|hest/hestside,data1,method="ML")
> ### test for om vekselvirkningen mellem diam og side kan fjernes
> mod1<-lme(S~factor(diam)+side,random=~1|hest/hestside,data1,method="ML")
> anova(mod1,mod0)
```

```
     Model df      AIC      BIC    logLik   Test  L.Ratio p-value
mod1     1  9 324.2630 356.6488 -153.1315
mod0     2 13 330.5241 377.3036 -152.2620 1 vs 2 1.738948  0.7836

> ### test for om den første af hovedvirkningerne af hhv diam og side
> ### kan fjernes (mod additiv model)
> mod2a<-lme(S~factor(diam),random=~1|hest/hestside,data1,method="ML")
> mod2b<-lme(S~side,random=~1|hest/hestside,data1,method="ML")
> anova(mod2a,mod1)

      Model df      AIC      BIC    logLik   Test  L.Ratio p-value
mod2a     1  8 324.0719 352.8592 -154.0359
mod1      2  9 324.2630 356.6488 -153.1315 1 vs 2 1.808852  0.1786

> anova(mod2b,mod1)

      Model df      AIC      BIC    logLik   Test  L.Ratio p-value
mod2b     1  5 385.2727 403.2648 -187.6363
mod1      2  9 324.2630 356.6488 -153.1315 1 vs 2 69.00966  <.0001

> ### test for om den siste af hovedvirkningerne af hhv diam og side
> ### kan fjernes
> mod3<-lme(S~1,random=~1|hest/hestside,data1,method="ML")
> anova(mod3,mod2a)

      Model df      AIC      BIC    logLik   Test  L.Ratio p-value
mod3      1  4 385.0815 399.4752 -188.5408
mod2a     2  8 324.0719 352.8592 -154.0359 1 vs 2 69.00966  <.0001

> ### genfitter slutmodel med REML estimation
>
> mod2afinal<-lme(S~factor(diam)-1,random=~1|hest/hestside,data1,method="REML")
> summary(mod2afinal)$tTable

                  Value Std.Error  DF  t-value       p-value
factor(diam)8  4.739894 0.1023335 212 46.31811 7.461516e-113
factor(diam)10 4.950094 0.1023335 212 48.37218 1.665503e-116
factor(diam)12 5.118152 0.1023335 212 50.01444 2.478212e-119
factor(diam)14 5.200888 0.1023335 212 50.82293 1.073594e-120
factor(diam)16 5.213630 0.1023335 212 50.94745 6.645272e-121

> VarCorr(mod2afinal)

            Variance     StdDev
hest =      pdLogChol(1)
(Intercept) 0.1264613    0.3556139
```

```
hestside =  pdLogChol(1)
(Intercept) 0.2041804    0.4518632
Residual    0.1083929    0.3292307

> intervals(mod2afinal)

Approximate 95% confidence intervals

 Fixed effects:
                 lower     est.    upper
factor(diam)8  4.538173 4.739894 4.941616
factor(diam)10 4.748373 4.950094 5.151816
factor(diam)12 4.916431 5.118152 5.319874
factor(diam)14 4.999167 5.200888 5.402610
factor(diam)16 5.011909 5.213630 5.415352
attr(,"label")
[1] "Fixed effects:"

 Random Effects:
  Level: hest
                    lower      est.     upper
sd((Intercept)) 0.2016337 0.3556139 0.6271832
  Level: hestside
                    lower      est.     upper
sd((Intercept)) 0.3363002 0.4518632 0.6071373

 Within-group standard error:
    lower      est.     upper
0.2993382 0.3292307 0.3621084

> ### fit model, hvor diam indgår som en kovariat
> modlin<-lme(S~diam,random=~1|hest/hestside,data1,method="ML")
> anova(modlin,mod2a)

       Model df      AIC      BIC    logLik   Test  L.Ratio p-value
modlin     1  5 326.3376 344.3297 -158.1688
mod2a      2  8 324.0719 352.8592 -154.0359 1 vs 2 8.265746  0.0408

> ### i stedet for den approximative p-værdi baseret på likelihood ratio testet
> ### kan man benytte simulation
>
> set.seed(2014)
> sim<-simulate.lme(modlin,m2=mod2a,nsim=1000)
> lr.sim<-2*(sim$alt$ML-sim$null$ML)
> psim<-sum(lr.sim>8.266)/1000
> psim
```

```
[1] 0.043

> ### F-test for om den tilfældige effekt af hest x side kan fjernes fra  modellen
>
> m1<-lm(S~factor(diam)+hestside+hest,data1)
> m2<-lm(S~factor(diam)+hest,data1)
> anova(m2,m1)

Analysis of Variance Table

Model 1: S ~ factor(diam) + hest
Model 2: S ~ factor(diam) + hestside + hest
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    239 53.471
2    212 22.979 27    30.492 10.419 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```