Eksamen i Statistisk Dataanalyse 2, 7. april 2011

Vejledende besvarelse

Opgave 1

1. Den statistiske analyse kunne tage udgangspunkt i følgende model

$$\texttt{glu}_i = \gamma(\texttt{treat} \times \texttt{time}_i) + A(\texttt{lamb}_i) + e_i,$$

hvor $A(1), \ldots, A(38)$ er uafhængige $\sim N(0, \sigma_A^2)$ og $e_1, \ldots, e_{114}$ er uafhængige $\sim N(0, \sigma^2)$. Bemærk at variablen `time` indgår i udgangsmodellen som en faktor.

Der er forskellige andre muligheder for udgangsmodellen. Hvis det ikke eksplicit stod anført i opgaven, at der skulle benyttes en random intercept model, så kunne man også have forsøgt at modellere korrelationsstrukturen inden for måleserie vha. en Diggle-model. Man kunne også vælge kun at analysere målingerne `time`=1 og 2.5 timer, mens målingen til `time`=-1 kunne indgå som en kovariat (=baselinemåling).

2. Selvom variablen `time` kan opfattes som en numerisk variabel, så er det ikke påkrævet, at man ved reduktionen undersøger, om det f.eks. er rimeligt at antage, at glukoseindholdet ændrer sig lineært over tid. Der vil således ikke blive trukket ned (eller givet ekstra point) hvis man gennem hele analysen opfatter `time` som en faktor.

I første omgang undersøges om man kan fjerne vekselvirkningen $\texttt{treat} \times \texttt{time}$ svarende til modellen

$$\texttt{glu}_i = \alpha(\texttt{treat}_i) + \beta(\texttt{time}_i) + A(\texttt{lamb}_i) + e_i.$$

Gennem den resterende del af opgaven er den tilfældige del af modellen givet, som beskrevet under besvarelsen af delspørgsmål 1. ovenfor. Vi godkender hypotesen om, at der ikke er vekselvirkning mellem `treat` og `time` ($LR = 8.809, p = 0.185$).

Dernæst kan man forsøge at fjerne hovedeffekten af `time` svarende til modellen

$$\texttt{glu}_i = \alpha(\texttt{treat}_i) + A(\texttt{lamb}_i) + e_i.$$

Vi forkaster hypotesen om, at glukoseindholdet *ikke* ændrer sig over tid ($LR = 54.38, p < 0.0001$).

Alternativt kan man forsøge at fjerne hovedeffekten af `treat` svarende til modellen

$$\texttt{glu}_i = \beta(\texttt{time}_i) + A(\texttt{lamb}_i) + e_i.$$

Vi kaster hypotesen om, at der *ikke* er nogen effekt af behandlingen ($LR = 9.541, p = 0.023$). Da p-værdien for det approksimative likelihood-ratio test er test på signfikansniveuaet på 5 %, så kan man overveje at fortage et simulationsstudie med henblik på at få en mere eksakt bestemmelse af p-værdien. På baggrund af udskriften nedenfor fås en simuleret p-værdi på 0.04 , hvorfor vi holder fast i, at der er en signifikant effekt af behandlingen (`treat`).

Slutmodellen bliver den additive model med hovedeffekt af behandling (`treat`) og `time`, hvori `lamb` desuden indgår med tilfældig effekt. Den additive model udtrykker, at forskellen mellem behandlingerne med rimelighed kan antages at være den samme til hvert af de tre måletidspunkter.

Når man angiver parameterestimaterne for slutmodellen bør modeller refittes i Rved brug af `method='REML'` (=default opførslen i R). Modellen indeholder 6 parametre til beskrivelse af middelværdistrukturen, og disse kunne f.eks. angives som

$$\hat{\alpha}(\texttt{A}) + \hat{\beta}(-0.5) = 4.170 \quad [3.345, 4.994] \qquad \hat{\alpha}(\texttt{B}) + \hat{\beta}(-0.5) = 5.088 \quad [4.298, 5.878]$$
$$\hat{\alpha}(\texttt{C}) + \hat{\beta}(-0.5) = 4.159 \quad [3.369, 4.948] \qquad \hat{\alpha}(\texttt{D}) + \hat{\beta}(-0.5) = 5.450 \quad [4.625, 6.274]$$
$$\hat{\beta}(1) - \hat{\beta}(-0.5) = 2.421 \quad [1.853, 2.989] \qquad \hat{\beta}(2.5) - \hat{\beta}(-0.5) = 1.778 \quad [1.210, 2.346].$$

Parameterestimaterne til beskrivelse af variansstrukturen bliver

$$\hat{\sigma}^2 = 1.243^2 \quad \hat{\sigma}_A^2 = 0.851^2.$$

3. Da slutmodellen fra delspørgsmål 2. er den additive model, vil ændringen i glukoseindholdet over tid være uafhængig af behandling. Af `summary(m2refit)` nedenfor ses, at tilvæksten i glukose fra en halv time før fodring til 2.5 timer efter fodring estimeres til

   $$1.778 \quad [1.210, 2.346]$$

   uanset behandlingen.

4. Da faktorerne behandling (`treat`) og `time` indgår additivt i slutmodellen, så vil forskellen mellem behandling `B` og `D` være den samme hen over hele forsøgsperioden. Ved passende omparametrisering af slutmodellen (f.eks. som i `m2refitny` nedenfor) kan man direkte aflæse, at glukose-indholdet er 0.362 højere for behandling `D` end for behandling `B`, samt at der ikke er tale om en signifikant forskel mellem de to behandlinger ($p = 0.484$).

Eksempel på R-kode som kunne være benyttet ved løsning af opgave 1

Ved løsning af opgaven har jeg benyttet dele af følgende R-kørsel. Det er naturligvis ikke tanken, at dette skal skrives med ind i en besvarelse af eksamensopgaven. Det er derimod et forsøg på at vise, hvilke R-kommandoer man selv kunne have brugt under selve eksamen, for at generere det nødvendige output for at besvare eksamensspørgsmålene.

```
> ahk <- read.table(file = "ahk.txt", header = T)

> library(nlme)
> m0 <- lme(glu ~ treat * factor(time), random = ~1 | lamb, data = ahk,
+     method = "ML")
> m1 <- lme(glu ~ treat * time, random = ~1 | lamb, data = ahk,
+     method = "ML")
> m2 <- lme(glu ~ treat + factor(time), random = ~1 | lamb, data = ahk,
+     method = "ML")
> m3 <- lme(glu ~ treat + time, random = ~1 | lamb, data = ahk,
+     method = "ML")

> anova(m1, m0)

   Model df      AIC      BIC    logLik   Test  L.Ratio p-value
m1     1 10 449.3696 476.7316 -214.6848
m0     2 14 419.4090 457.7158 -195.7045 1 vs 2 37.96063  <.0001

> anova(m2, m0)

   Model df     AIC      BIC    logLik   Test  L.Ratio p-value
m2     1  8 416.218 438.1076 -200.1090
m0     2 14 419.409 457.7158 -195.7045 1 vs 2 8.809014  0.1846

> anova(m3, m1)

   Model df      AIC      BIC    logLik   Test  L.Ratio p-value
m3     1  7 446.0287 465.1821 -216.0143
m1     2 10 449.3696 476.7316 -214.6848 1 vs 2 2.659079  0.4472

> anova(m3, m2)

   Model df      AIC      BIC    logLik   Test  L.Ratio p-value
m3     1  7 446.0287 465.1821 -216.0143
m2     2  8 416.2180 438.1076 -200.1090 1 vs 2 31.81069  <.0001

> m4 <- lme(glu ~ treat, random = ~1 | lamb, data = ahk, method = "ML")
> m5 <- lme(glu ~ factor(time), random = ~1 | lamb, data = ahk,
+     method = "ML")
> anova(m4, m2)

   Model df      AIC      BIC    logLik   Test  L.Ratio p-value
m4     1  6 466.5935 483.0107 -227.2967
m2     2  8 416.2180 438.1076 -200.1090 1 vs 2 54.37549  <.0001

> anova(m5, m2)
```

```
   Model df      AIC      BIC    logLik   Test  L.Ratio p-value
m5     1  5 419.7592 433.4401 -204.8796
m2     2  8 416.2180 438.1076 -200.1090 1 vs 2 9.541154  0.0229

> ### Simulation af p-værdi:
> set.seed(1)
> sim<-simulate.lme(m5,m2=m2,nsim=100)
> lr.sim<-2*(sim$alt$ML-sim$null$ML)
> psim<-sum(lr.sim>9.541)/100
> psim

[1] 0.04

> m2refit <- lme(glu ~ treat + factor(time) - 1, random = ~1 |
+     lamb, data = ahk, method = "REML")
> summary(m2refit)

                   Value Std.Error DF   t-value      p-value
treatA          4.169610 0.4058744 34 10.273154 5.824073e-12
treatB          5.088018 0.3885506 34 13.094864 7.729419e-15
treatC          4.158684 0.3885506 34 10.703069 1.990461e-12
treatD          5.449610 0.4058744 34 13.426839 3.773376e-15
factor(time)1   2.420789 0.2851825 75  8.488562 1.400901e-12
factor(time)2.5 1.778158 0.2851825 75  6.235158 2.423625e-08

> intervals(m2refit)

Approximate 95% confidence intervals

 Fixed effects:
                   lower     est.    upper
treatA          3.344774 4.169610 4.994446
treatB          4.298388 5.088018 5.877647
treatC          3.369054 4.158684 4.948314
treatD          4.624774 5.449610 6.274446
factor(time)1   1.852677 2.420789 2.988902
factor(time)2.5 1.210045 1.778158 2.346271
attr(,"label")
[1] "Fixed effects:"

 Random Effects:
  Level: lamb
                    lower      est.    upper
sd((Intercept)) 0.5566832 0.8506086 1.299725

 Within-group standard error:
   lower     est.    upper
1.058021 1.243082 1.460511
```

```
> ahk$treat <- relevel(ahk$treat, ref = "B")
> m2refitny <- lme(glu ~ factor(time) + treat - 1, random = ~1 |
+     lamb, data = ahk, method = "REML")
> summary(m2refitny)

                     Value Std.Error DF    t-value      p-value
factor(time)-0.5  5.0880175 0.3885506 74 13.0948642 5.923811e-21
factor(time)1     7.5088070 0.3885506 74 19.3251708 1.223416e-30
factor(time)2.5   6.8661754 0.3885506 74 17.6712509 2.865306e-28
treatA           -0.9184074 0.5113572 35 -1.7960192 8.112843e-02
treatC           -0.9293333 0.4977186 35 -1.8671863 7.026933e-02
treatD            0.3615926 0.5113572 35  0.7071233 4.841717e-01
```

Opgave 2

1. Analysen bør tage udgangspunkt i en model, hvor længden (`length`) beskrives som en lineær funktion af `omkreds`, og hvor både skæring og hældning kan afhænge af sorten (`variety`)

   $$\texttt{length}_i = \alpha(\texttt{variety}_i) + \beta(\texttt{variety}_i) \cdot \texttt{omkreds}_i + e_i,$$

   hvor $e_1, \ldots, _{67}$ er uafhængige $\sim N(0, \sigma^2)$. Figur 1, der indeholder residualplot baseret på ovenstående model, giver ikke anledning til at betvivle antagelsen om varianshomogenitet, ligesom der heller ikke er noget som tyder på, at sammenhængen mellem længde og omkreds ikke kan beskrives ved en lineær funktion.

2. Først undersøges hypotesen, $H_0 : \beta(\texttt{gul}) = \beta(\texttt{orange}) = \beta(\texttt{roed})$ om at hældningen kan antages at være uafhængig af sorten. Dette svarer til den statitiske model

   $$\texttt{length}_i = \alpha(\texttt{variety}_i) + \beta \cdot \texttt{omkreds}_i + e_i.$$

   Af `R`-udskriften fremgår at hypotesen godkendes ($F = 0.567, p = 0.570$).

   Dernæst testes hypotesen, $H_0 : \alpha(\texttt{gul}) = \alpha(\texttt{orange}) = \alpha(\texttt{roed})$ om at linjernes skæring kan antages at være ens for alle sorter. Dette svarer til modellen

   $$\texttt{length}_i = \alpha + \beta \cdot \texttt{omkreds}_i + e_i.$$

   Man finder, at hypotesen godkendes ($F = 1.198, p = 0.309$) og estimater med tilhørende 95 %-konfidensintervaller for slutmodellen (hvor `variety` slet ikke indgår) bliver

   $$\hat{\alpha} = -0.960 \quad [-2.264, 0.343], \quad \hat{\beta} = 1.190 \quad [0.989, 1.392], \quad \hat{\sigma}^2 = 1.200^2 = 1.440.$$

3. Situationen beskrevet i opgaveformulering svarer til hypotesen om, at parameteren som beskriver skæringen i slutmodellen bør være $= 0$. Mere formelt kan vi formulere hypotesen som $H_0 : \alpha = 0$, og man kan direkte af `R`-udskriften aflæse, at denne hypotese accepteres ($t = -1.471, p = 0.146$).

4. Slutmodellen fra spørgsmål 2.+3. udtrykker, at

$$\texttt{length}_i = \beta\,\texttt{omkreds}_i + e_i,$$

hvor $e_1, \ldots, e_{67}$ er uafhængige $\sim N(0, \sigma^2)$. Idet vi har givet fra opgaveformuleringen at $\texttt{omkreds} = \pi \cdot \texttt{diameter}$, så vil størrelsen $\delta = \pi \cdot \beta \approx 3.14 \cdot \beta$ udtrykke forholdet mellem længde (`length`) og bredde (=`diameter`) for en gulerod. Et estimat med tilhørende 95 %-konfidensinterval for dette forhold bliver

$$3.14 \cdot 1.046 = 3.284 \quad [3.141, 3.428].$$

Da konfidensintervallet for forholdet *ikke* indeholder tallet 3 konkluderes, at det *ikke* er rimeligt at antage, at længden (`length`) af en gulerod er 3 gange større end diameteren.

5. Det vil være mest korrekt at beregne estimaterne på baggrund af slutmodellen (`modelE`)

$$\texttt{length}_i = \beta_i \cdot \texttt{omkreds}_i + e_i,$$

hvor skæringsparameteren er sat lig med 0 og hvor effekten af sort (`variety`) ikke indgår. Til dette formål benyttes blot estimat og konfidensinterval for hældningsparameteren ($\hat{\beta} = 1.046[1.000, 1.092]$), der blot ganges med den relevante `omkreds` (dvs 2,5 eller 15 cm).

Estimater med tilhørende 95 %-konfidensintervaller for længden bliver således

længde af gulerod med omkreds på 2 cm	:	2.092	[2.001, 2.183]
længde af gulerod med omkreds på 5 cm	:	5.230	[5.002, 5.456]
længde af gulerod med omkreds på 15 cm	:	15.690	[15.006, 16.374]

Det vil også være ok at beregne estimaterne for længden af gulerødderne ved brug af `modelD` i `R`-udskriften. Man bør da se på linjerne `estD.2,estD.4` og `estD.6` i `R`-udskriften efter kommandoen `estimable(modelD,estD)`. Konfidensintervallerne skal i givet fald konstrueres som estimat $\pm 1.997 \cdot$ Std.Error, idet 0.975-fraktilen i en t-fordeling med 65 frihedsgrader er

```
> qt(0.975, 65)

[1] 1.997138
```

Gulerødderne som indgår i forsøget har en `omkreds` på mellem 3.9 cm og 10.9 cm. Man må derfor gå ud fra, at estimaterne ovenfor er mest pålidelige for en gulerod med længde 5 cm, mens det er mere tvivlsomt, hvorvidt man kan ekstrapolere resultaterne til en gulerod med omkreds 15 cm. Fra spørgsmål 3. har vi set, at man på baggrund af den statistiske analyse får godkendt den meget rimelige hypotese om, at længden af en gulerod nærmer sig 0 når omkredsen nærmer sig 0. Dette tyder på, at det er rimeligt at ekstrapolere modeller til f.eks. gulerødder med en omkreds på kun 2 cm. Man kan således forvente at dette estimat er rimelig pålideligt på trods af, at der i datasættet ikke indgår gulerødder med en omkreds under 3.9 cm.

Opgave 3

1. Den skitserede består af 3 gentagelser af et 2^n-te forsøg med to faktorer (`beh, dosis`) på to blokke (`kammer`), hvor man har valgt at konfundere vekselvirkningen `beh` $\times$ `dosis` med `kammer`. Ulempen med forsøgsplanen er, at man vil have svært ved at skelne vekselvirkningen `beh` $\times$ `dosis` fra en eventuelt effekt af `kammer`.

   Som et alternativ kunne man vælge at benytte sig af partiel kunfundering, hvor man på 3 par hver bestående af 2 kamre konfunderer henholdvis `beh`, `dosis` og `beh` $\times$ `dosis`. Dette er skitseret nedenfor.

kammer 1	kammer 2	kammer 3	kammer 4	kammer 5	kammer 6
B1,D1	B1,D2	B1,D1	B1,D2	B1,D1	B2,D1
B2,D2	B2,D1	B2,D1	B2,D2	B1,D2	B2,D2

   Bemærk at dette faktisk også bliver et balanceret ufuldstændigt blokforsøg (BIBD) med 4 behandlinger (kombinationer af `beh` og `dosis`) på blokke af størrelse 2, hvor hvert par af behandlinger forekommer præcis en gang.

2. Forsøget bør udføres som et splitplot forsøg med `kammer` som helplot, temperatur (`temp`) som helplot-faktor og behandling (`beh`) som delplot-faktor. Dette begrundes med at de 12 forsøgsenheder er organiseret i blokke af størrelse 2, og at det ikke er muligt at variere temperaturen (`temp`) inden for en blok.

   Da opgaveformuleringen ikke var fuldstændig præcis på dette punkt vil der i delopgave 3.2 og 3.3 ikke blive trukket ned, hvis man har opfattet det som om, at der inden for hvert kammer skal være en forsøgsenhed med `høj` og en forsøgsenhed med `lav` temperatur. Det er dog væsentligt, at man er konsekvent ved løsningen af de to delspørgsmål.

   Randomiseringen foretages i to trin. Først udvælges ved lodtrækning tre kammernumre, som indstilles på høj temperatur, mens de øvrige tre indstilles på lav temperatur. Dernæst foretages for hvert kammer en lodtrækning af hvor de to prøver med behandling `B1` og `B2` skal placeres i forhold til hinanden.

   Ved den statistiske analyse bør `kammer` indgå med tilfældig effekt mens hovedvirkningerne `beh` og `temp` samt vekselvirkningen `beh` $\times$ `temp` indgår som systematiske faktorer således at den statistiske model bliver:

$$Y_i = \gamma(\texttt{beh} \times \texttt{temp}_i) + A(\texttt{kammer}_i) + e_i$$

   hvor $A(1),\ldots,A(6)$ er uafhængige $\sim N(0,\sigma_A^2)$ og $e_1,\ldots,e_{12}$ er uafhængige $\sim N(0,\sigma^2)$.

   Det vil ikke kunne lade sig gøre at inddrage vekselvirkningen `kammer` $\times$ `beh` i modellen som en tilfældig effekt, da der ikke er gentagelser af produktfaktoren i forsøgsdesignet. Produktfaktoren `kammer` $\times$ `temp` er lig med `kammer` og er således allerede med i modellen.

3. Faktoren `dag` bør inddrages som en tilfældig effekt i den statistiske model der således bliver

$$Y_i = \gamma(\texttt{beh} \times \texttt{temp}_i) + A(\texttt{kammer}_i) + B(\texttt{dag}_i) + e_i$$

hvor $A(1),\ldots,A(6)$ er uafhængige $\sim N(0,\sigma_A^2)$, $B(1),B(2),B(3)$ er uafhængige $\sim N(0,\sigma_B)$ og $e_1,\ldots,e_{12}$ er uafhængige $\sim N(0,\sigma^2)$. Når man skal tegne det tilhørende faktordiagram, er det vigtigt at bemærke, at faktoren `dag` er grovere end blokfaktoren `kammer`.