Eksamen i Statistisk Dataanalyse 2, 6. april 2017

Vejledende besvarelse

Opgave 1

1. Der skal tages stilling til to ting i forbindelse med valget af statistisk model: i) hvor mange tilfældige effekter bør inddrages i modellen, ii) skal responsen (`vgt`) og kovariaten `lgt` transformeres med logaritmen før analysen.

   Der er en god grund til at tro, at længden af mink fra samme kuld vil være mere ens end mink som blot har samme far. Derfor bør man som udgangspunkt inddrage både en tilfældig effekt af `sire` og en tilfældig effekt af `litter`. For at vurdere om analysen bør tage udgangspunkt i modellen `m1` eller `m3` kan man med fordel se på residualplot og QQ-plot svarende til de to modeller fittet med R-koden

```
mod2control <- lm(log(lgt) ~ log(vgt)*koen + litter, data = mink)
mod4control <- lm(lgt ~ vgt*koen + litter, data = mink)
```

   Disse plots giver ikke væsentlig anledning til at foretrække den ene model fremfor den anden (dvs. begge kan benyttes!): der lader til at være varianshomogenitet og residualerne er tilnærmelsesvist normalfordelte. Hvis man på forhånd har en ide om, at sammenhængen mellem vægt og længde bør beskrives ved en potensfunktion, så kan det være et argument for at foretrække modellen `m1`. Dette synspunkt er valgt i den vejledende besvarelse.

   Den statistiske model kan skrives som

$$\log(\texttt{vgt}_i) = \alpha(\texttt{koen}_i) + \beta(\texttt{koen}_i) \cdot \log(\texttt{lgt}_i) + A(\texttt{sire}_i) + B(\texttt{litter}_i) + e_i,$$

   hvor

   - $A(\texttt{S11}), A(\texttt{S29}), \ldots$ er uafhængige $\sim N(0, \sigma_A^2)$
   - $B(\texttt{L296}), B(\texttt{L34}), \ldots$ er uafhængige $\sim N(0, \sigma_B^2)$
   - $e_1, e_2, \ldots$ er uafhængige $\sim N(0, \sigma^2)$.

2. Udgangsmodellen fra 1. beskriver, at der er en lineær sammenhæng mellem `log(vgt)` og `log(lgt)`, hvor både skæring og hældning kan afhænge af `koen`.

   Først testes hypotesen om, at hældningen ikke afhænger af `koen` svarende til modellen

$$\log(\texttt{vgt}_i) = \alpha(\texttt{koen}_i) + \beta \cdot \log(\texttt{lgt}_i) + A(\texttt{sire}_i) + B(\texttt{litter}_i) + e_i.$$

   Hypotesen godkendes ($L.Ratio = 2.625$, $p = 0.105$). Dernæst testes hypotesen om, at skæringen ikke afhænger af `koen` svarende til modellen

$$\log(\texttt{vgt}_i) = \alpha + \beta \cdot \log(\texttt{lgt}_i) + A(\texttt{sire}_i) + B(\texttt{litter}_i) + e_i.$$

Hypotesen forkastes ($L.Ratio = 510.71, p < 0.0001$).

Man kan diskutere om det er relevant at test hypotesen om, at der ingen sammenhæng er mellem vægt og længde (svarende til $\beta = 0$), men denne hypotese afvises også klart ($L.Ratio = 366.248, p < 0.0001$).

Slutmodellen

$$\log(\texttt{vgt}_i) = \alpha(\texttt{koen}_i) + \beta \cdot \log(\texttt{lgt}_i) + A(\texttt{sire}_i) + B(\texttt{litter}_i) + e_i.$$

beskriver, at der er en lineær sammenhæng ml. `log(vgt)` og `log(lgt)`. Skæringen afhænger af `koen` og estimeres til $\hat{\alpha}(\texttt{H}) = 2.435[1.895 - 2.976]$ samt $\hat{\alpha}(\texttt{T}) = 2.079[1.564 - 2.594]$. Hældningen afhænger ikke af `koen` og estimeres til $\hat{\beta} = 1.439[1.303 - 1.574]$. Varianskomponenterne estimeres til

$$\hat{\sigma}_A = 0.0345 \quad \hat{\sigma}_B = 0.0489 \quad \hat{\sigma} = 0.0862.$$

Andelen af variansen der kan tilskrives de forskellige varianskomponenter kan findes ved at sammenligne med den totale varians $\hat{\sigma}_A^2 + \hat{\sigma}_B^2 + \hat{\sigma}^2$.

Ved bedømmelsen af besvarelserne er der **ikke slået hårdt ned på, hvordan det påvirker fortolkningen af estimaterne, at analysen er foretaget på logaritmetransformerede variable**. I princippet vil det være mest korrekt at tilbagetransformere resultater ne med eksponentialfunktionen og fortolke resultaterne via medianen (=centrum) i fordelingen. Således vil f.x. en øgning af længden (`lgt`) på 10 % føre (median-) vægten øges med en faktor $1.1^{\hat{\beta}} = 1.1^{1.439} = 1.147$ svarende til en stigning på 14.7%. Tilsvarende vil (median-) vægten for hanmink (`koen = H`) være til at en faktor $\exp(2.435 - 2.079) = 1.428$ højere (svarende til 42.8 %) end (median-) vægten for hunmink.

3. Tanken er, at man benytte estimaterne fra slutmodellen fra 2. til at udtrække et estimat svarende til de to ønskede typer af mink. Dette kan gøres enten ved håndkraft eller ved brug af `estimable()`-funktionen. Hvis der regnes på logtransformerede data, så bliver den tilbagetransformerede værdi for vægten af hanmink på 55 cm lig med 3641 g, mens vægten for hunmink på 45 cm bliver estimeres til 1911 g. (I begge tilfælde er det i princippet et estimat for medianvægten for en mink på pågældende statur.)

4. Da der foretages gentagne længdemålinger på hver minkhvalp henover forsøgsperioden, så bør den statistiske model kunne tage højde for den serielle afhængighedsstruktur mellem målinger taget på samme individ. Her kunne man tage udgangspunkt i en Diggle-model

$$\texttt{lgt}_i = \gamma(\texttt{treat} \times \texttt{uge}_i) + A(\texttt{litter}_i) + B(\texttt{subject}_i) + D_i + e_i,$$

hvor der inkluderes (nestede) tilfældige effekter af kuld (`litter`) og individ `subject`, og hvor D_i beskriver korrelationsstrukturen for Diggle-modellen.

Det betragtes som valgfrit, om man vælger at lave en model der inkluderer startlængden (ved fravænning) som en kovariat (baselinemåling) i modellen, men dette vil også være en god strategi.

Hvis man har en god grund til at tro, at længden vil udvikle sig lineært over tid, så kan man argumentere for, at modellen kunne have formen

$$\texttt{lgt}_i = \alpha(\texttt{treat}_i) + \beta(\texttt{treat}_i) \cdot \texttt{uge}_i + A(\texttt{litter}_i) + B(\texttt{subject}_i) + D_i + e_i.$$

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

```
mink <- read.table(file = file.choose(), header = T)
```

Hvis analysen tager udgangspunkt i modellen `m1`

```
library(nlme)
model1a <- lme(log(vgt) ~ koen*log(lgt), random = ~ 1 | sire/litter
          , data = mink, method = "ML")
model1b <- lme(log(vgt) ~ koen + log(lgt), random = ~ 1 | sire/litter
          , data = mink, method = "ML")
anova(model1b, model1a)

##         Model df       AIC       BIC   logLik   Test  L.Ratio p-value
## model1b     1  6 -2032.122 -2002.076 1022.061
## model1a     2  7 -2032.747 -1997.694 1023.374 1 vs 2 2.625342  0.1052

model1c <- lme(log(vgt) ~ log(lgt), random = ~ 1 | sire/litter
          , data = mink, method = "ML")
anova(model1c, model1b)

##         Model df       AIC       BIC    logLik   Test  L.Ratio p-value
## model1c     1  5 -1523.415 -1498.377  766.7072
## model1b     2  6 -2032.122 -2002.076 1022.0608 1 vs 2 510.7072  <.0001

model1d <- lme(log(vgt) ~ koen , random = ~ 1 | sire/litter
          , data = mink, method = "ML")
anova(model1d, model1b)

##         Model df       AIC       BIC   logLik   Test  L.Ratio p-value
## model1d     1  5 -1667.874 -1642.836  838.937
## model1b     2  6 -2032.122 -2002.076 1022.061 1 vs 2 366.2476  <.0001

model1brefit <- lme(log(vgt) ~ koen + log(lgt), random = ~ 1 | sire/litter
          , data = mink, method = "REML")
summary(model1brefit)

## Linear mixed-effects model fit by REML
##  Data: mink
```

```
##         AIC       BIC   logLik
##   -2011.755 -1981.725 1011.877
##
## Random effects:
##  Formula: ~1 | sire
##         (Intercept)
## StdDev:  0.03449317
##
##  Formula: ~1 | litter %in% sire
##         (Intercept)   Residual
## StdDev:  0.04887225 0.08617864
##
## Fixed effects: log(vgt) ~ koen + log(lgt)
##                  Value  Std.Error  DF    t-value p-value
## (Intercept)  2.4351007 0.27540809 910   8.841791       0
## koenT       -0.3560887 0.01393202 910 -25.559023       0
## log(lgt)     1.4385898 0.06910363 910  20.817863       0
##  Correlation:
##          (Intr) koenT
## koenT    -0.921
## log(lgt) -1.000  0.916
##
## Standardized Within-Group Residuals:
##          Min           Q1          Med           Q3          Max
## -7.302054365 -0.589773090  0.009811254  0.582085727  3.299208938
##
## Number of Observations: 1105
## Number of Groups:
##             sire litter %in% sire
##               73              193

intervals(model1brefit)

## Approximate 95% confidence intervals
##
##  Fixed effects:
##                  lower       est.      upper
## (Intercept)  1.8945919  2.4351007  2.9756095
## koenT       -0.3834313 -0.3560887 -0.3287461
## log(lgt)     1.3029688  1.4385898  1.5742108
## attr(,"label")
## [1] "Fixed effects:"
##
##  Random Effects:
##   Level: sire
##                      lower       est.      upper
## sd((Intercept)) 0.0223455 0.03449317 0.05324467
```

```
##   Level: litter
##                       lower       est.      upper
## sd((Intercept)) 0.03993574 0.04887225 0.05980851
##
##  Within-group standard error:
##      lower       est.      upper
## 0.08229483 0.08617864 0.09024575
```

```
library(gmodels)
han.55 <- c(1, 0, log(55))
hun.45 <- c(1, 1, log(45))
est <- rbind(han.55, hun.45)
est.table <- estimable(model1brefit, est, conf.int = 0.95)
exp(est.table[, c(1, 6, 7)])

##        Estimate Lower.CI Upper.CI
## han.55 3640.985 3590.152 3692.537
## hun.45 1910.725 1885.535 1936.251
```

Hvis analysen tager udgangspunkt i modellen `m3`

```
library(nlme)
model3a <- lme(vgt ~ koen*lgt, random = ~ 1 | sire/litter
          , data = mink, method = "ML")
model3b <- lme(vgt ~ koen + lgt, random = ~ 1 | sire/litter
          , data = mink, method = "ML")
anova(model3b, model3a)

##         Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## model3b     1  6 15566.94 15596.99 -7777.470
## model3a     2  7 15558.69 15593.75 -7772.347 1 vs 2 10.24655  0.0014

model3c <- lme(vgt ~ lgt, random = ~ 1 | sire/litter
          , data = mink, method = "ML")
anova(model3c, model3b)

##         Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## model3c     1  5 16004.96 16030.00 -7997.481
## model3b     2  6 15566.94 15596.99 -7777.470 1 vs 2 440.0227  <.0001

model3d <- lme(vgt ~ koen , random = ~ 1 | sire/litter
          , data = mink, method = "ML")
anova(model3d, model3b)

##         Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## model3d     1  5 15885.45 15910.49 -7937.724
## model3b     2  6 15566.94 15596.99 -7777.470 1 vs 2 320.5085  <.0001
```

```r
model3arefit <- lme(vgt ~ koen * lgt, random = ~ 1 | sire/litter
          , data = mink, method = "REML")
summary(model3arefit)

## Linear mixed-effects model fit by REML
##  Data: mink
##        AIC      BIC   logLik
##   15533.34 15568.37 -7759.67
##
## Random effects:
##  Formula: ~1 | sire
##         (Intercept)
## StdDev:    94.68898
##
##  Formula: ~1 | litter %in% sire
##         (Intercept) Residual
## StdDev:    124.1425  249.938
##
## Fixed effects: vgt ~ koen * lgt
##                  Value Std.Error  DF   t-value p-value
## (Intercept) -1208.3578  277.1648 909 -4.359708  0.0000
## koenT         215.1086  361.4730 909  0.595089  0.5519
## lgt            88.5018    5.1452 909 17.200900  0.0000
## koenT:lgt     -23.7037    7.3955 909 -3.205132  0.0014
##  Correlation:
##           (Intr) koenT  lgt
## koenT     -0.693
## lgt       -0.998  0.694
## koenT:lgt  0.616 -0.994 -0.618
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -6.96021289 -0.50648593 -0.02068344  0.51844355  4.46965376
##
## Number of Observations: 1105
## Number of Groups:
##             sire litter %in% sire
##               73              193

intervals(model3arefit)

## Approximate 95% confidence intervals
##
##  Fixed effects:
##                    lower        est.       upper
## (Intercept) -1752.31523 -1208.35785 -664.400466
## koenT        -494.31003   215.10864  924.527300
```

```
## lgt            78.40398    88.50180   98.599621
## koenT:lgt     -38.21804   -23.70371   -9.189371
## attr(,"label")
## [1] "Fixed effects:"
##
##  Random Effects:
##   Level: sire
##                    lower     est.    upper
## sd((Intercept)) 63.03735 94.68898 142.2332
##   Level: litter
##                    lower     est.    upper
## sd((Intercept)) 99.58819 124.1425 154.7509
##
##  Within-group standard error:
##    lower     est.    upper
## 238.6845 249.9380 261.7221
```

```
library(gmodels)
han.55 <- c(1, 0, 55, 0)
hun.45 <- c(1, 1, 45, 45)
est <- rbind(han.55, hun.45)
est.table <- estimable(model3arefit, est, conf.int = 0.95)
est.table
```

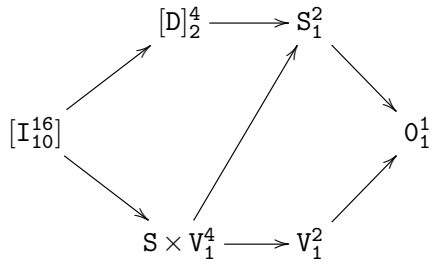
```
##        Estimate Std. Error  t value  DF Pr(>|t|) Lower.CI Upper.CI
## han.55 3659.241   20.22602 180.9175 909        0 3619.546 3698.936
## hun.45 1922.665   18.50567 103.8960 909        0 1886.346 1958.984
```

Opgave 2

1. Der skal totalt anvendes 8 planter af hver sort, så det er oplagt at lave et blokforsøg, hvor `drivhus` betragtes som blokke, og hver `sort` afprøves to gange for hvert `drivhus`. Desuden bør man være opmærksom på faktoren `side` med to niveauer (`side = nej` eller `side = ja`), der angiver om planten placeres i et drivhus med eller uden siderne på. Faktoren `side` er grovere end faktoren `drivhus`, så man kan tænke på forsøget som et splitplot-forsøg med `drivhus` som helplot, `side` som helplotfaktor og `sort` som delplotfaktor. Bemærk dog at der er to gentagelser (=planter af samme `sort`) inden for hvert `drivhus`. Hvis man alene vælger at betragte faktoren `sort` (og fejlagtigt glemmer `side`!), så kan forsøget opfattes som et fuldstændig randomiseret blokforsøg.

   Randomiseringen foretages i 2 trin: først afgøres ved lodtrækning hvilke to af de fire drivhuse, hvor siderne skal pilles af, dernæst randomiseres placeringen af de fire planter (to af hver sort) tilfældigt ud på de fire pladser inden for hvert drivhus.

   Faktordiagrammet for forsøget ser ud som følger



På diagrammet er benyttet forkortelserne: `D = drivhus`, `V = sort`, `S = side`.

Man kunne godt inddrage vekselvirkning mellem sort og drivhus (som en tilfældig effekt), men dette forventes ikke for en fuldstændig besvarelse.

2. Den oplagte mulighed er at lade begge faktorerne `drivhus` og `plantesaek` indgå i modellen med tilfældig effekt. Desuden bør vi have trefaktorvekselvirningen mellem `side`, `sort` og `beh` med som systematisk effekt, hvilket *er* muligt da vi har to gentagelser for hver kombination af de tre faktorer. Den tilsvarende statistiske model bliver således

$$Y_i = \alpha(\mathtt{side}\times\mathtt{sort}\times\mathtt{beh}_i) + A(\mathtt{drivhus}_i) + B(\mathtt{plantesaek}_i) + e_i,$$

hvor $A(1),\ldots A(4)$ er uafhængige $\sim N(0,\sigma^2_{\mathtt{A}})$, $B(1),\ldots B(8)$ er uafhængige $\sim N(0,\sigma^2_{\mathtt{B}})$ og $e_1,\ldots,e_{24}$ er uafhængige $\sim N(0,\sigma^2)$.

Det er ikke påkrævet, at man diskuterer muligheden for at inddrage øvrige vekselvirkninger. Hvis man kaster sig ud i dette, kan konkluderes at vekselvirkningerne mellem `plantesaek` og alle andre faktorer allerede er med i modellen. Derfor er det kun til diskussion, om man også bør inkludere vekselvirkningerne `drivhus` $\times$ `beh` og `drivhus` $\times$ `sort`. Hvis man ønsker at inddrage disse vekselvirkninger i modellen, så bør de indgå med tilfældig effekt.

Anskuer man det første drivhus som et 2^2-forsøg på to blokke (=`plantesaek`ke) af størrelse 2, så ses at det er vekselvirkningen mellem `sort` og `beh` som er konfunderet med blokeffekten. Da denne effekt er konfunderet med `plantesaek` inden for hvert `drivhus`, så kunne man argumentere for, at en form for partiel konfundering kunne være en bedre løsning. Da man samtidig skal fjerne siderne fra to af de fire drivhuse, så kan man argumentere for, at det kunne være fornuftigt at konfundere vekselvirkningen i to af de fire drivhuse (et med og et uden sider), og at konfundere en af hovedeffekterne i de to øvrige drivhuse. Der er flere fornuftig svar på dette delspørgsmål, men det anses for centralt at man diskuterer partiel konfundering.

3. Der lægges op til at lave et ufuldstændigt blokforsøg med syv behandlinger ($v_T = 7$) og blokstørrelse $r_B = 6$. Ifølge kompendiets Theorem 9.6 skal der være mindst $v_B = 7$ blokke (=drivhuse). *Hvis* der anvendes præcis 7 drivhuse, så skal hver behandling forekomme

$$r_T = \frac{r_B \cdot v_B}{v_T} = \frac{6\cdot 7}{7} = 6$$

gange, og hvert par af behandlinger skal mødes

$$\lambda = \frac{r_T\cdot(r_B-1)}{v_T-1} = \frac{6\cdot(6-1)}{7-1} = 5$$

gange i forsøgsplanen.

Hvis man overhovedet prøver at realisere denne forsøgsplan, så skal hver behandling forekomme i 6 af de 7 blokke, dvs man skal blot fjerne hver af de 7 behandlinger, fra netop en af de 7 blokke. Der er med andre ord kun een mulig forsøgsplan af denne type (dvs med 7 blokke / drivhuse). Det viser sig (f.x. ved opskrivning af en koincidensmatrix), at dette faktisk er et balanceret ufuldstændigt blokforsøg (BIBD).

Opgave 3

1. Der bør benyttes en statistisk model med en systematisk (fixed) effekt af vekselvirkningen mellem de to faktorer `Experiment` og `Treatment`. Endelig bør placeringen af planterne (givet ved faktoren `Pos`) inddrages som en tilfældig effekt i modellen. Forskellen på de to modeller `modelA1` og `modelB1` ligger i om `ShootHeight` eller den logaritmetransformerede skudhøjde benyttes som respons i modellen. Ved at betragte residual plot for de to modeller ses, at antagelsen om varianshomogenitet virker rimelig for begge modeller. Derimod er der en tendens til, at de standardiserede residualer i højere grad kan beskrives ved en standard normalfordeling for modellen `modelA1`. Man kan også inddrage den observation, at der er et par ret store residualer (numerisk værdi omkring 4), hvilket ikke bør optræde i et datasæt af denne størrelse.

   Modellen kan opskrives som

   $$\texttt{ShootHeight}_i = \gamma(\texttt{Experiment} \times \texttt{Treatment}_i) + A(\texttt{Pos}_i) + e_i,$$

   hvor $A(\texttt{1:1}), A(\texttt{2:1}), \ldots$ er uafhængige $\sim N(0, \sigma_A^2)$ og $e_1, \ldots, e_{70}$ er uafhængige $\sim N(0, \sigma^2)$.

   Parameterestimaterne fra modellen (aflæses ud fra `summary(modelA1)`)

   $$\begin{aligned}
   \hat{\gamma}(\texttt{1,G}) &= 57.729 \\
   \hat{\gamma}(\texttt{2,G}) &= 57.729 + (-20.112) = 37.617 \\
   \hat{\gamma}(\texttt{1,X}) &= 57.729 + 5.271 = 63.000 \\
   \hat{\gamma}(\texttt{2,G}) &= 57.729 + (-20.112) + 5.271 + 21.929 = 64.816 \\
   \hat{\sigma}_A &= 4.000 \\
   \hat{\sigma} &= 10.935
   \end{aligned}$$

   (Der er naturligvis andre måder, hvorpå man kan vælge at afrapportere estimaterne.)

2. Først undersøges om der kan ses bort fra vekselvirkningen mellem `Experiment` og `Treatment` svarende til, at vi reducerer modellen til

   $$\texttt{ShootHeight}_i = \alpha(\texttt{Experiment}_i) + \beta(\texttt{Treatment}_i) + A(\texttt{Pos}_i) + e_i.$$

   Vi forkaster hypotesen ($L.Ratio = 14.54, p < 0.0001$) og konkludere, at der *er* en signifant vekselvirkning. Modellen fra 1. benyttes derfor som slutmodel for analysen.

   Estimatet for kombinationen `Experiment = 2, Treatment = X` fås ved at lægge de fire estimater fra `modelA1` sammen. Et 95 % - konfidensinterval kan aflæses ved at anvende `estimable()`-funktionen med linearkombinationen svarende til vektoren `est3`.

Vi finder således, at estimatet med konfidensinterval bliver $64.817[95\%KI : 59.060 - 70.573]$.

3. Effekterne af behandlingen i hhv. eksperiment 1 og 2 kan findes ved at udtrække relevante kontraster (=kombinationer af parametre) fra slutmodellen (`modelA1`). Effekten af behandlingen i eksperiment 1 kan aflæses direkte (kaldet `TreatmentX` i `R`-udskriften)

$$5.271[95\%KI : (-2.408) - 12.951].$$

Heraf ses bl.a. at der *i*kke er en signifikant effekt af behandlingen i `Experiment = 1`.

Ønsker man at estimere behandlingseffekten i `Experiment = 2`, så kan man benytte kaldet til `estimable()`-funktionen svarende til `est5`. Heraf ses, at behandlingseffekten i `Experiment = 2` estimeres til $27.200[95\%KI : 19.554 - 34.846]$. Der er således en signfikant (positiv) effekt af behandlingen i `Experiment = 2` på skudhøjden.

Betydningen af potternes placering er modelleret vha. størrelsen af den tilfældige effekt af `Pos`. Estimat samt 95 % - konfidensinterval for σ_A er $4.000[95\%KI : 0.987 - 16.179]$. Da 0 ikke er indeholdt i konfidensintervallet kan man konkludere, at potternes placering bidrager til variationen af skudhøjden i forsøget.

Hvis vi selv havde haft adgang til data så ville det være naturligt formelt at teste hypotesen om at $\sigma_A = 0$. Det er også positivt, hvis man som en del af besvarelse angiver, hvor stor en andel af den totale variation, der skyldes potternes placering.