Eksamen i Statistisk Dataanalyse 2
(NMAB14002U)

6. april 2017

Alle sædvanlige hjælpemidler, herunder bøger, noter, R-programmer og lommeregner samt brug af programmet R på egen PC, er tilladt. Det er *ikke* tilladt at benytte PC til nogle former for aktivitet, som involverer opkobling til et netværk eller kommunikation med andre. Det er tilladt at skrive med blyant. Opgavesættet består af 8 sider med i alt 3 opgaver, der indgår med vægtningen 40 %, 30 % og 30 % i bedømmelsen.

Til besvarelse af opgave 1 har du fået udleveret en USB-nøgle med et datasæt, som du skal indlæse og anvende i R på din egen PC for at kunne besvare opgaven. Til opgave 3 er der vedlagt noget R-udskrift, som kan benyttes i besvarelsen (det er ikke sikkert at alle dele af udskriften skal benyttes). Husk at det er vigtigt at specificere de statistiske modeller og hypoteser du bruger, og at komme med konklusioner på analyserne.

Opgave 1 (4 spørgsmål)

I forbindelse med minkavl er der stor fokus på at etablere en effektiv produktion af dyr under hensyntagen til dyrenes velfærd. Sammenhængen mellem en minks vægt og længde har betydning for værdien af skindet, samtidig med at det kan være en væsentlig indikator for, om man har en sund population af mink. Data til denne opgave består af vægtmålinger (`vgt`: enhed g) og længdemålinger (`lgt`: enhed cm) for 1105 mink fra Københavns Universitets forsøgsfarme. Desuden har man registreret minkens køn via faktoren `koen` med niveauerne `H=hanmink` og `T=tæve`. Fra avldatabaser har man desuden indsamlet oplysninger om, hvilke mink der har samme far (givet ved faktoren `sire`), samt hvilke mink der stammer fra samme kuld (`litter`).

Data til opgaven er venligst stillet til rådighed af Anne Sofie Vedsted Hammer fra Institut for Veterinær Sygdomsbiologi, KU. Data er blevet gjort tilgængelige på vedlagte USB-nøgle. Data kan indlæses med kommandoen

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

hvorefter filen `opgave1mink.txt` vælges fra USB-nøglen. De første datalinjer ses her

```
##   sire litter koen  lgt  vgt
## 1  S11   L296    H 50.0 3180
## 2  S11   L296    H 53.0 3350
## 3  S11   L296    H 51.0 3350
## 4  S29    L34    T 45.5 2150
## 5  S29    L34    H 50.5 2990
```

```
library(nlme)
m1 <- lme(log(vgt) ~ koen*log(lgt), random = ~ 1 | sire/litter
          , data = mink)
m2 <- lme(log(vgt) ~ koen*log(lgt), random = ~ 1 | litter
          , data = mink)
m3 <- lme(vgt ~ koen*lgt, random = ~ 1 | sire/litter
          , data = mink)
m4 <- lme(vgt ~ koen*lgt, random = ~ 1 | litter
          , data = mink)
```

1. Vælg en af modellerne `m1-m4` som udgangspunkt for en statistisk analyse af data, og opskriv den tilhørende statistiske model. Du bedes argumentere for dit valg af model.

2. Foretag en statistisk analyse med udgangspunkt i din model fra delspørgsmål 1. med henblik på at undersøge sammenhængen mellem vægt og længde, herunder om sammenhængen er forskellig for de to køn. Angiv estimater og 95 % konfidensintervaller for samtlige parametre i slutmodellen og forklar i ord hvad modellen udtrykker. Du bedes tydeligt anføre de modeller du tester imod hinanden samt relevante p-værdier og teststørrelser.

3. Med udgangspunkt i din slutmodel fra delspørgsmål 2. bedes du give et bud på vægten af en hanmink på 55 cm og en hunmink på 45 cm.

4. Det planlægges at udføre et nyt forsøg, hvor man ønsker at afgøre om en ny fodertype (`treat = B`) giver længere mink end en klassisk foderblanding (`treat = A`). Til forsøget udvælges 50 minkhvalpe (alle hanmink) fra 25 kuld (2 mink per kuld) umiddelbart efter fravænning hos moderen. Ved lodtrækning allokeres en mink fra hvert kuld til hver af de to behandlinger. Minkhvalpenes længde (`lgt`) måles ved fravænning (=forsøgsstart) og løbende en gang hver uge efter behandlingens start. Giv et forslag til en statistik model du vil benytte til at afgøre, om fodertype `B` giver længere hvalpe end fodertype `A`. Husk at argumentere for dit valg af model.

   (Du kan naturligvis ikke lave den statistiske analyse, da du ikke har adgang til data.)

Opgave 2 (3 spørgsmål)

Den stolte ejer af et kolonihavehus har indkøbt 4 minidrivhuse med henblik på dyrkning af peberfrugter. Det besluttes at udføre et videnskabeligt forsøg med to forskellige sorter af peberfrugt (`sort=I` eller `sort=II`) med det formål at maksimere vægten af de indsamlede peberfrugter fra hver plante. Det er muligt at fjerne siderne på minidrivhusene, og det påtænkes at fjerne siderne på to af de fire drivhuse i forbindelse med dyrkningsforsøget. Der er plads til 4 planter i hvert drivhus.

1. Giv et forslag til en forsøgsplan, hvor der totalt udplantes 16 peberfrugtplanter (lige mange af hver af de to `sorter`). Hvilken type forsøg er der tale om, og hvordan bør randomiseringen foretages? Tegn et faktordiagram for forsøget.

Efter et besøg i plantecentret bliver kolonihavehusejeren anbefalet at dyrke sine peberfrugter i plantesække. I hvert drivhus er der plads til to plantesække, hver med plads til to planter. Den ivrige kolonihavehusejer beslutter sig for at nippe halvdelen af blomsterne af en af de to planter i hver sæk, fordi det skulle give større frugter fra de tilbageværende blomster. Niveauet `beh=+` referer til en plante, hvor en del af blomsterne er fjernet (`beh=-` betyder at ingen blomster fjernes). Et eksempel på en forsøgsplan er anført på figuren nedenfor. Opdelingen i de to plantesække inden for hvert drivhus er markeret med en lodret streg.

II,+	I,+
I,-	II,-

II,-	II,+
I,+	I,-

I,-	I,+
II,+	II,-

II,+	II,-
I,-	I,+

2. Opskriv en statistisk model til analyse af forsøget. Hvis man isoleret betragter den del af forsøget der vedrører drivhuset længst til venstre, hvilken effekt er så konfunderet med `plantesæk`? Foreslå en alternativ forsøgsplan, hvor det stadig kræves at alle kombinationer af `sort` og `beh` er repræsenteret inden for hvert drivhus. Begrund dit forslag.

På baggrund af dyrkningsforsøget beslutter ejeren af kolonihavehuset sig for at udføre et nyt forsøg det følgende år, hvor følgende forhold inddrages:

- Siderne på alle drivhusene skal være på.
- Det er for dyrt at benytte plantesække.
- Uden plantesække er der plads til 6 peberfrugtplanter i hvert drivhus.
- Der skal anvendes planter fra 7 forskellige sorter af peberfrugt i forsøget.
- Der indkøbes flere drivhuse.

3. Hvor mange drivhuse skal man totalt have for at kunne udføre forsøget som et balanceret ufuldstændigt blokforsøg? Husk at begrunde dit svar.

Opgave 3 (3 spørgsmål)

Bladlus lever af planter og er derfor et problem for økologisk landbrug, hvor der ikke må sprøjtes. Det er blevet foreslået at behandle såsæden (frøene der sås) med en bestemt svampetype inden såning. Håbet er at svampebehandlingen gør det mindre attraktivt for bladlusene at leve på planterne, dog uden at påvirke plantevæksten negativt. I denne opgave undersøges effekten på vækst af majsplanter, mere præcist højden af skuddene (`ShootHeight`) 10 dage efter såning.

Der er foretaget to eksperimenter under sammenlignelige, men ikke helt identiske omstændigheder. I hvert eksperiment indgik 36 potter. Halvdelen af potterne blev tilsået med et svampebehandlet frø (`Treatment = G`), den anden halvdel var kontroller tilsået med et ubehandlet frø (`Treatment = X`). Der er således i alt fire grupper af observationer givet ved kombinationer af faktorerne `Experiment` og `Treatment`. Potterne inden for hvert eksperiment var placeret i par

(givet ved faktoren `Pos`), således at der for hvert par var en behandlet og en ubehandlet plante. I første eksperiment (`Experiment = 1`) var der to svampebehandlede frø der ikke spirede, så der er kun er 16 planter med svampebehandling.

Data til opgaven er stillet til rådighed af Susanna Saari fra Institut for Plantevidenskab, KU.

```
head(bladlus)

##   Experiment Treatment Pos ShootHeight
## 1          1         G 1:1        56.3
## 2          1         G 2:1        62.1
## 3          1         G 3:1        77.7
## 4          1         G 4:1        67.5
## 5          1         G 5:1        77.0
## 6          1         G 6:1        62.3

bladlus$Experiment <- factor(bladlus$Experiment)
```

Besvar følgende 3 delspørgsmål ved brug af R-udskriften sidst i opgavesættet. Bemærk at der kan være dele af R-udskriften, som ikke skal benyttes.

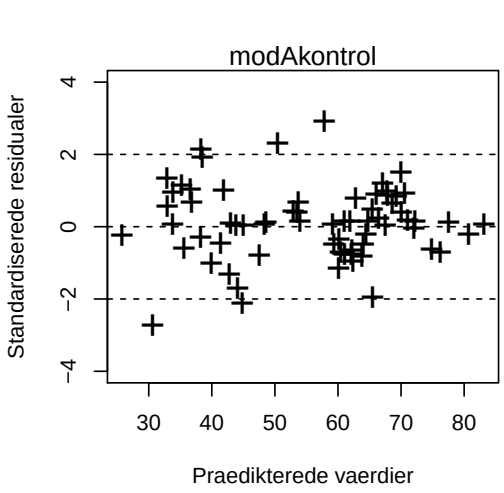
1. Argumenter for hvilken af modellerne `modelA1` og `modelB1`, som du finder mest velegnet til at analysere data. Opskriv den tilhørende statistiske model og angiv parameterestimater for samtlige parametre i modellen.

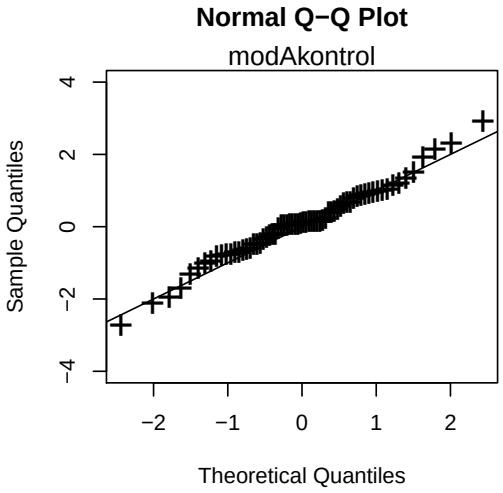
2. Foretag reduktion af modellen fra delspørgsmål 1. med henblik på at undersøge om højden af skuddene påvirkes af behandlingen, og om der er forskel mellem de to eksperimenter. Angiv et estimat og et 95 % konfidensinterval for den forventede skudhøjde i eksperiment 2, hvis der benyttes behandling X (kontrol).

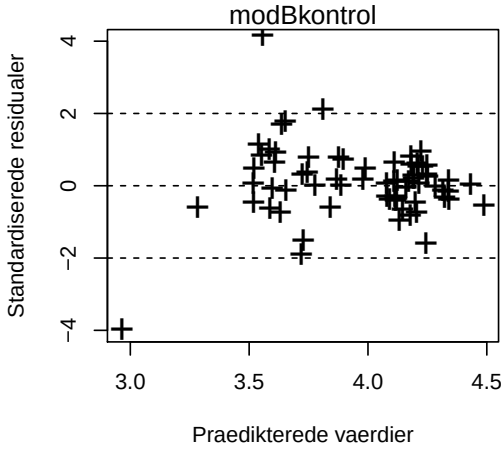
3. Diskuter om der ser ud til at være en effekt af svampebehandlingen for hver enkelt af de to eksperimenter. Kommenter desuden på, om potternes placering lader til at påvirke skudhøjden (`ShootHeight`).

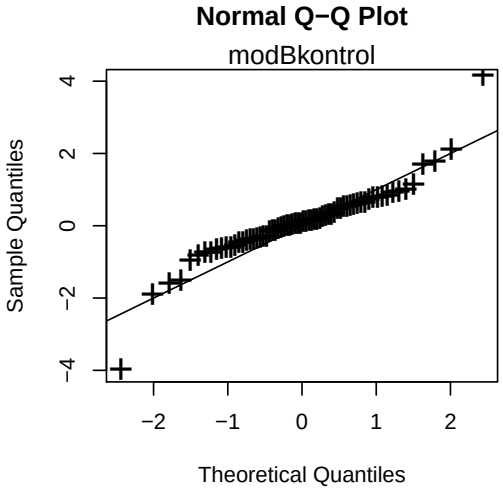
```
### indlaes R pakker
library(nlme)
library(gmodels)

### nogle statistiske modeller
modAkontrol <- lm(ShootHeight ~ Experiment * Treatment + Pos
                  , data = bladlus)
modBkontrol <- lm(log(ShootHeight) ~ Experiment * Treatment + Pos
                   , data = bladlus)
```










```r
### nogle statistiske modeller
modelA1 <- lme(ShootHeight ~ Experiment * Treatment, random = ~ 1|Pos
               , method = "ML", data = bladlus)
modelA2 <- lme(ShootHeight ~ Experiment + Treatment, random = ~ 1|Pos
               , method = "ML", data = bladlus)
modelA3 <- lme(ShootHeight ~ Experiment, random = ~ 1|Pos
               , method = "ML", data = bladlus)
modelA4 <- lme(ShootHeight ~ Treatment, random = ~ 1|Pos
               , method = "ML", data = bladlus)
modelB1 <- lme(log(ShootHeight) ~ Experiment * Treatment
               , random = ~ 1|Pos, method = "ML", data = bladlus)
modelB2 <- lme(log(ShootHeight) ~ Experiment + Treatment
               , random = ~ 1|Pos, method = "ML", data = bladlus)
modelB3 <- lme(log(ShootHeight) ~ Experiment
               , random = ~ 1|Pos, method = "ML", data = bladlus)
modelB4 <- lme(log(ShootHeight) ~ Treatment
               , random = ~ 1|Pos, method = "ML", data = bladlus)
```

```
### nogle statistiske test
anova(modelA2, modelA1)

##         Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## modelA2     1  5 566.3645 577.6070 -278.1823
## modelA1     2  6 553.8295 567.3205 -270.9148 1 vs 2 14.53504   1e-04

anova(modelA3, modelA2)

##         Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## modelA3     1  4 588.5312 597.5252 -290.2656
## modelA2     2  5 566.3645 577.6070 -278.1823 1 vs 2 24.16665  <.0001

anova(modelA4, modelA2)

##         Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## modelA4     1  4 572.0619 581.0559 -282.0310
## modelA2     2  5 566.3645 577.6070 -278.1823 1 vs 2 7.697371  0.0055

anova(modelB2, modelB1)

##         Model df      AIC      BIC     logLik   Test L.Ratio p-value
## modelB2     1  5 42.68147 53.92395 -16.340736
## modelB1     2  6 31.65597 45.14695  -9.827987 1 vs 2 13.0255   3e-04

anova(modelB3, modelB2)

##         Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## modelB3     1  4 60.32653 69.32051 -26.16326
## modelB2     2  5 42.68147 53.92395 -16.34074 1 vs 2 19.64506  <.0001

anova(modelB4, modelB2)

##         Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## modelB4     1  4 47.98942 56.98341 -19.99471
## modelB2     2  5 42.68147 53.92395 -16.34074 1 vs 2 7.307952  0.0069
```

```
### (udpluk af) summary fra udvalgte modeller samt konfidensintervaller
summary(modelA1)
```

```
##                              Value Std.Error DF   t-value      p-value
## (Intercept)             57.728593  2.995236 34 19.273473 7.092994e-20
## Experiment2            -20.111926  4.118055 34 -4.883842 2.428391e-05
## TreatmentX               5.271407  3.882746 32  1.357649 1.840726e-01
## Experiment2:TreatmentX  21.928593  5.400641 32  4.060368 2.955020e-04
```

```
intervals(modelA1)

## Approximate 95% confidence intervals
##
##  Fixed effects:
##                              lower       est.     upper
## (Intercept)             51.818016  57.728593  63.63917
## Experiment2            -28.238192 -20.111926 -11.98566
## TreatmentX              -2.408195   5.271407  12.95101
## Experiment2:TreatmentX  11.246777  21.928593  32.61041
## attr(,"label")
## [1] "Fixed effects:"
##
##  Random Effects:
##   Level: Pos
##                      lower     est.    upper
## sd((Intercept)) 0.9874535 3.996962 16.17869
##
##  Within-group standard error:
##     lower      est.     upper
##  8.659794 10.934991 13.807952
```

```
summary(modelB1)
```

```
##                              Value  Std.Error DF    t-value      p-value
## (Intercept)             4.0287268 0.07168937 34 56.1969912 3.990620e-35
## Experiment2            -0.4627322 0.09852761 34 -4.6964727 4.236347e-05
## TreatmentX              0.0883630 0.09852761 32  0.8968349 3.765048e-01
## Experiment2:TreatmentX  0.5043442 0.13727469 32  3.6739778 8.667772e-04
```

```r
intervals(modelB1)
```

```
## Approximate 95% confidence intervals
##
##  Fixed effects:
##                              lower       est.      upper
## (Intercept)             3.8872603  4.0287268  4.1701933
## Experiment2            -0.6571594 -0.4627322 -0.2683051
## TreatmentX             -0.1065127  0.0883630  0.2832387
## Experiment2:TreatmentX  0.2328314  0.5043442  0.7758569
## attr(,"label")
## [1] "Fixed effects:"
##
##  Random Effects:
##   Level: Pos
##                           lower         est.         upper
## sd((Intercept)) 4.236988e-163 1.779919e-05 7.477273e+152
##
##  Within-group standard error:
##     lower      est.     upper
## 0.2359394 0.2784439 0.3286055
```

```r
### nogle kald til estimable()
est1 <- c(0, 0, 0, 1)
est2 <- c(1, 0, 0, 1)
est3 <- c(1, 1, 1, 1)
est4 <- c(1, 1, 0, 0)
est5 <- est3 - est4
est <- rbind(est1, est2, est3, est4, est5)
estimable(modelA1, est, conf.int = 0.95)
```

```
##      Estimate Std. Error   t value DF     Pr(>|t|) Lower.CI Upper.CI
## est1 21.92859   5.400641  4.060368 32 2.955020e-04 10.92785 32.92934
## est2 79.65719   7.361968 10.820094 32 3.183898e-12 64.66135 94.65303
## est3 64.81667   2.826117 22.934885 32 0.000000e+00 59.06005 70.57328
## est4 37.61667   2.826117 13.310372 34 4.884981e-15 31.87331 43.36003
## est5 27.20000   3.753827  7.245939 32 3.120517e-08 19.55371 34.84629
```

```r
estimable(modelB1, est, conf.int = 0.95)
```

```
##       Estimate Std. Error   t value DF     Pr(>|t|)  Lower.CI  Upper.CI
## est1 0.5043442 0.13727469  3.673978 32 8.667772e-04 0.2247248 0.7839635
## est2 4.5330709 0.18510115 24.489696 32 0.000000e+00 4.1560322 4.9101096
## est3 4.1587017 0.06758938 61.528918 32 0.000000e+00 4.0210266 4.2963768
## est4 3.5659945 0.06758938 52.759684 34 0.000000e+00 3.4286364 3.7033527
## est5 0.5927072 0.09558582  6.200785 32 6.077362e-07 0.3980052 0.7874091
```