Eksamen i Statistisk Dataanalyse 2
(LMAF10070)

12. april 2012

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. Opgavesættet består af 10 sider med i alt 3 opgaver, der indgår med vægtningen 35 %, 40 % og 25 % i bedømmelsen.

Til besvarelse af opgave 2 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 både opgave 1 og opgave 2 er der vedlagt udvalgte R-udskrifter, 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)

En forstørrelse af venstre forkammer i hjertet kan være et symptom på en lang række hjerkekarsygdomme. I forbindelse med et phd-projekt på KU-LIFE, har man indsamlet data om størrelsen (volumen) af venstre forkammer for 82 hunde, der ikke lider af hjertesygdomme. Formålet er at give en beskrivelse af normalområdet for hjertevolumen for raske hunde. Datasættet er venligst stillet til rådighed af Miriam Höllmer. Et udpluk af datasættet er vist nedenfor

```
> data1 <- read.table(file = "data1.txt", header = T)
> data1

       race vgt maxLA
1 Chihuahua 2.1  1.66
2 Chihuahua 3.3  2.21
3 Chihuahua 2.0  1.03
4 Chihuahua 2.4  0.96
5 Chihuahua 2.3  1.66
6 Chihuahua 3.0  2.18

[ ... flere datalinjer her .... ]
```

```
           race vgt maxLA
80 Grand_Danois  63 34.11
81 Grand_Danois  69 34.48
82 Grand_Danois  74 40.20
```

Datasættet indeholder ud over volumen af venstre forkammer (`maxLA`) målt i mL også variablene `race` samt hundens vægt (`vgt`) målt i kg. Der indgår 82 hunde i datasættet fordelt på følgende fire racer

```
> levels(data1$race)

[1] "Chihuahua"    "Dalmatiner"   "Grand_Danois" "Gravhund"
```

Formålet med den statistiske analyse er at undersøge, hvordan hjertevolumen af venstre forkammer afhænger af `race` og vægt (`vgt`). Ved besvarelse af opgaven skal du benytte relevante dele af `R`-udskriften nedenfor.

1. Opskriv en statistisk model som udtrykker, at der er en lineær sammenhæng mellem `log(maxLA)` og `log(vgt)`, hvor både hældning og skæring tillades at afhænge af racen (`race`). Her angiver `log` den naturlige logaritme. Angiv estimater for samtlige parametre i modellen.

2. Benyt `R`-udskriften til at reducere modellen fra spørgsmål 1. mest muligt og angiv estimater for samtlige parametre i slutmodellen. Angiv også 95 %-konfidensintervaller for parametrene i beskrivelsen af middelværdistrukturen.

3. Tag udgangspunkt i slutmodellen fra spørgsmål 2. og angiv et estimat og et 95 %-konfidensinterval for det forventede volumen af venstre forkammer (målt i mL) for en hund af racen `Dalmatiner` som vejer `35` kg.

4. Undersøg, om det på baggrund af resultaterne fra den statistiske analyse er rimeligt at antage, at det forventede hjertevolumen `maxLA` (*ikke* `log(maxLA)`) for hunde af racen `Gravhund` har følgende sammenhæng med vægten

   $$\text{forventet volumen} \rightarrow \mathbb{E}\texttt{maxLA} = \gamma \cdot \texttt{vgt}.$$

   Giv et bud på, hvad en rimelig værdi for proportionalitetskonstanten γ kunne være.

   [**Hint:** Benyt evt. følgende regneregel for logaritmer $\exp(\alpha + \beta \log(z)) = \exp(\alpha) \cdot z^{\beta}$]

Udskrift af R-kørsel (lettere redigeret):

```
> ### Nogle statistiske modeller og test:
> modA<-lm(log(maxLA)~race+race:log(vgt)-1,data1)
> modB<-lm(log(maxLA)~race+log(vgt)-1,data1)
> modC<-lm(log(maxLA)~log(vgt),data1)
> modD<-lm(log(maxLA)~race-1,data1)
> modE<-lm(log(maxLA)~1,data1)
```

```
> anova(modB, modA)

  Res.Df      RSS Df Sum.of.Sq        F    Pr..F.
1     77 5.725866 NA         NA       NA        NA
2     74 5.300248  3 0.4256180 1.980771 0.1242154

> anova(modC, modA)

  Res.Df      RSS Df Sum.of.Sq        F    Pr..F.
1     80 6.062402 NA         NA       NA        NA
2     74 5.300248  6 0.7621545 1.773484 0.1162632

> anova(modC, modB)

  Res.Df      RSS Df Sum.of.Sq        F    Pr..F.
1     80 6.062402 NA         NA       NA        NA
2     77 5.725866  3 0.3365365 1.508553 0.2189941

> anova(modD, modA)

  Res.Df      RSS Df Sum.of.Sq        F       Pr..F.
1     78 9.537980 NA         NA       NA           NA
2     74 5.300248  4  4.237732 14.79139 6.318237e-09

> anova(modD, modB)

  Res.Df      RSS Df Sum.of.Sq        F       Pr..F.
1     78 9.537980 NA         NA       NA           NA
2     77 5.725866  1  3.812114 51.26435 4.133677e-10

> anova(modE, modC)

  Res.Df        RSS Df Sum.of.Sq        F       Pr..F.
1     81 137.644869 NA         NA       NA           NA
2     80   6.062402  1  131.5825 1736.374 5.175639e-56

> anova(modE, modD)

  Res.Df       RSS Df Sum.of.Sq        F       Pr..F.
1     81 137.64487 NA         NA       NA           NA
2     78   9.53798  3  128.1069 349.2122 4.208202e-45
```

```
> ### dele af summary() på udvalgte modeller:
> summary(modA)

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
raceChihuahua              -0.3803     0.2440  -1.559  0.12336
raceDalmatiner              0.9508     1.5961   0.596  0.55321
raceGrand_Danois            0.7438     2.3630   0.315  0.75383
raceGravhund               -1.3021     0.4229  -3.079  0.00292 **
raceChihuahua:log(vgt)      0.7220     0.2350   3.073  0.00297 **
raceDalmatiner:log(vgt)     0.5702     0.4689   1.216  0.22789
raceGrand_Danois:log(vgt)   0.6688     0.5645   1.185  0.23992
raceGravhund:log(vgt)       1.3677     0.1998   6.844 1.92e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2676 on 74 degrees of freedom
Multiple R-squared: 0.9888,Adjusted R-squared: 0.9876
F-statistic: 815.4 on 8 and 74 DF,  p-value: < 2.2e-16

> confint(modA)

                               2.5 %     97.5 %
raceChihuahua             -0.8663739  0.1058718
raceDalmatiner            -2.2296124  4.1311724
raceGrand_Danois          -3.9646151  5.4521378
raceGravhund              -2.1448101 -0.4593944
raceChihuahua:log(vgt)     0.2537916  1.1902530
raceDalmatiner:log(vgt)   -0.3642098  1.5045787
raceGrand_Danois:log(vgt) -0.4560148  1.7935841
raceGravhund:log(vgt)      0.9695414  1.7659101

> ### dele af summary() på udvalgte modeller:
> summary(modB)

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
raceChihuahua     -0.6847     0.1552  -4.411 3.30e-05 ***
raceDalmatiner    -0.5902     0.4897  -1.205   0.2318
raceGrand_Danois  -0.7394     0.6013  -1.230   0.2226
raceGravhund      -0.5807     0.3057  -1.900   0.0612 .
log(vgt)           1.0232     0.1429   7.160 4.13e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2727 on 77 degrees of freedom
Multiple R-squared: 0.9879,Adjusted R-squared: 0.9871
F-statistic:  1255 on 5 and 77 DF,  p-value: < 2.2e-16
```

```
> confint(modB)

                      2.5 %     97.5 %
raceChihuahua    -0.9937611 -0.3755721
raceDalmatiner   -1.5653399  0.3849857
raceGrand_Danois -1.9366804  0.4579054
raceGravhund     -1.1894765  0.0280377
log(vgt)          0.7386497  1.3077875

> ### dele af summary() på udvalgte modeller:
> summary(modC)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.62693    0.07044   -8.90 1.38e-13 ***
log(vgt)     1.01474    0.02435   41.67  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2753 on 80 degrees of freedom
Multiple R-squared: 0.956,Adjusted R-squared: 0.9554
F-statistic:  1736 on 1 and 80 DF,  p-value: < 2.2e-16

> confint(modC)

                 2.5 %     97.5 %
(Intercept) -0.7671081 -0.4867476
log(vgt)     0.9662743  1.0631976

> ### dele af summary() på udvalgte modeller:
> summary(modD)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
raceChihuahua    0.34949    0.07292   4.793  7.7e-06 ***
raceDalmatiner   2.89021    0.07631  37.876  < 2e-16 ***
raceGrand_Danois 3.54233    0.08022  44.155  < 2e-16 ***
raceGravhund     1.56186    0.08022  19.469  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3497 on 78 degrees of freedom
Multiple R-squared: 0.9798,Adjusted R-squared: 0.9788
F-statistic: 946.6 on 4 and 78 DF,  p-value: < 2.2e-16
```

```
> confint(modD)

                     2.5 %   97.5 %
raceChihuahua    0.2043257 0.494651
raceDalmatiner   2.7382960 3.042132
raceGrand_Danois 3.3826147 3.702042
raceGravhund     1.4021415 1.721569

> ### Nogle kald til estimable()
> library(gmodels)
> estA1<-c(0,0,0,0,0,35,0,0)
> estA2<-c(0,0,0,0,0,log(35),0,0)
> estA3<-c(0,1,0,0,0,35,0,0)
> estA4<-c(0,1,0,0,0,log(35),0,0)
> estA<-rbind(estA1,estA2,estA3,estA4)
> estimable(modA,estA,conf.int=0.95)

       Estimate  Std. Error   t value DF  Pr(>|t|)   Lower.CI  Upper.CI
estA1 19.956455 16.41309218  1.215886 74 0.2278928 -12.747345 52.660255
estA2  2.027204  1.66726444  1.215886 74 0.2278928  -1.294893  5.349301
estA3 20.907235 14.81812899  1.410923 74 0.1624577  -8.618531 50.433002
estA4  2.977984  0.09285233 32.072260 74 0.0000000   2.792972  3.162996

> estB1<-c(0,0,0,0,35)
> estB2<-c(0,0,0,0,log(35))
> estB3<-c(0,1,0,0,35)
> estB4<-c(0,1,0,0,log(35))
> estB<-rbind(estB1,estB2,estB3,estB4)
> estimable(modB,estB,conf.int=0.95)

       Estimate Std. Error   t value DF     Pr(>|t|)  Lower.CI  Upper.CI
estB1 35.812650 5.00182769  7.159913 77 4.133676e-10 25.852739 45.772562
estB2  3.637898 0.50809253  7.159913 77 4.133676e-10  2.626157  4.649640
estB3 35.222473 4.51612569  7.799268 77 2.474709e-11 26.229718 44.215228
estB4  3.047721 0.06344269 48.038964 77 0.000000e+00  2.921391  3.174052

> estC1<-c(0,35)
> estC2<-c(0,log(35))
> estC3<-c(1,35)
> estC4<-c(1,log(35))
> estC<-rbind(estC1,estC2,estC3,estC4)
> estimable(modC,estC,conf.int=0.95)

       Estimate Std. Error  t value DF Pr(>|t|)  Lower.CI  Upper.CI
estC1 35.515759 0.85231370 41.66982 80        0 33.819601 37.211917
estC2  3.607740 0.08657920 41.66982 80        0  3.435441  3.780038
estC3 34.888831 0.78935670 44.19907 80        0 33.317961 36.459701
estC4  2.980812 0.03814226 78.14985 80        0  2.904906  3.056717
```

Opgave 2 (4 spørgsmål)

Data til denne opgave stammer fra CLIMAITE forsøgsopstilingen, som har til formål at undersøge de langsigtede effekter af ændringer i det danske klima frem mod år 2075.

Forsøgsopstillingen består af 12 cylindre (`octagon`), som hver er opdelt i 4 lige store plots. For hver af de 12 cylindre afprøves 4 kombinationer af faktorerne `temp` (temperatur) og `drought` (tørke), som hver har 2 niveauer (`0,1`), hvor niveauet `1` beskriver en passende ændring af klimaet i forhold til i dag. Ved lodtrækning har man desuden udvalgt 6 cylindre som modtager en øget tilførsel af CO_2 (`co2`=1), mens de resterende 6 cylindre har en normal koncentration af CO_2 (`co2`=0).

I hver af de 48 ($= 12 \cdot 4$) plots indstalleres et rør som gør det muligt at tage billeder under jorden og bestemme rodvæksten. Data til denne opgave er venligst stillet til rådighed af Marie Frost Arndal. Det skal bemærkes, at der kun er målinger fra 45 af de 48 plots til rådighed i denne opgave.

Data er udleveret på vedlagte USB-nøgle under filnavnet `roots.txt` og for at besvare opgaven fuldstændigt, vil det være nødvendigt at køre udvalgte `R`-kommandoer på din egen medbragte computer. Du kan f.eks. indlæse data i `R` med kommandoen

```
data2<-read.table(file.choose(),header=T)
```

hvorefter du vælger filen `roots.txt`. De første 6 linjer i datasættet er organiseret som vist nedenfor

```
  octagon co2 drought temp  length
1       1   0       0    0 194.386
2       1   0       1    0 156.266
3       1   0       1    1  48.881
4       1   0       0    1  64.156
5      10   1       1    0 137.201
6      10   1       1    1 172.646
```

Formålet med opgaven er at undersøge, hvordan rodlængden (`length`) afhænger af faktorerne `temp`, `drought` og `co2`. Responsvariablen, `length`, angiver den totale længde af rødderne baseret på billeder taget i det øverste jordlag (8-15 cm dybde) ca. 2 år efter forsøgets start. Du bedes foretage analysen på de utransformerede værdier af responsvariablen, `length`, og det er ikke en del af opgaven, at du skal bruge tid på at lave modelkontrol.

1. Opskriv en statistisk model du vil benytte som udgangspunkt for en statistisk analyse af rodlængden (`length`).

2. Reducer den statistiske model fra 1. med henblik på at undersøge, hvordan rodlængden (`length`) afhænger af temperatur (`temp`), tørke (`drought`) og CO_2 (`co2`). Undervejs skal du tydeligt gøre rede for, hvilke modeller du tester mod hinanden, ligesom du bedes udtrække p-værdier og teststørrelser fra `R`-udskriften hørende til de enkelte test, som du foretager.

3. Angiv samtlige parameterestimater, der indgår i beskrivelsen af middelværdi- og variansstruktur for din slutmodel fra analysen i spørgsmål 2. Sørg for i ord at forklare, hvad de enkelte parameterestimater beskriver.

I det fulde datasæt (som du *ikke* har adgang til i denne opgave) har man for hvert plot løbende registreret rodlængden (`length`) på 13 tidspunkter (`session`) hen over det andet år i forsøgsperioden. Desuden angiver variablen `tidaar` tidspunktet for de enkelte målinger målt i antal år siden forsøgets start. Første måling (`session=1`) er taget `tidaar=1.068` år efter forsøgets start.

4. Benyt R-udskriften nedenfor til at opskrive en statistisk model, som kunne danne udgangspunkt for en statistisk analyse af det fulde datasæt fra alle 13 måletidspunkter. Husk at begrunde dit svar.

Udskrift af R-kørsel:

```
> data2full <- read.table("data2full.txt", header = T)

> head(data2full, 10)

    plot octagon co2 drought temp tidaar session  length
1    1:1       1   0       0    0  1.068       1  43.210
46   1:1       1   0       0    0  1.106       2  84.118
91   1:1       1   0       0    0  1.136       3  88.683
136  1:1       1   0       0    0  1.218       4 102.491
181  1:1       1   0       0    0  1.257       5 117.279
226  1:1       1   0       0    0  1.369       6 145.286
271  1:1       1   0       0    0  1.467       7 149.577
316  1:1       1   0       0    0  1.580       8 159.481
361  1:1       1   0       0    0  1.654       9 166.748
406  1:1       1   0       0    0  1.733      10 179.277
451  1:1       1   0       0    0  1.791      11 184.151
496  1:1       1   0       0    0  1.815      12 187.398
541  1:1       1   0       0    0  1.881      13 194.386
2    1:2       1   0       1    0  1.068       1  84.205
47   1:2       1   0       1    0  1.106       2  97.242
92   1:2       1   0       1    0  1.136       3  97.525
137  1:2       1   0       1    0  1.218       4 108.479
182  1:2       1   0       1    0  1.257       5 127.715

[ ... flere datalinjer her ... ]

    plot octagon co2 drought temp tidaar session length
465 12:3      12   1       1    1  1.791      11 51.202
510 12:3      12   1       1    1  1.815      12 51.202
555 12:3      12   1       1    1  1.881      13 50.515
```

```
> library(nlme)

> model1<-lme(length~factor(session)*temp*drought*co2,random=~1|plot
+ ,cor=corGaus(form=~tidaar|plot,nugget=T)
+ ,data2full)
> model2<-lme(length~factor(session)*temp*drought*co2,random=~1|octagon/plot
+ ,cor=corGaus(form=~tidaar|octagon/plot,nugget=T)
+ ,data2full)
> model3<-lme(length~factor(session)*temp*drought*co2,random=~1|plot
+ ,data2full)
> model4<-lme(length~factor(session)*temp*drought*co2,random=~1|octagon/plot
+ ,data2full)
> anova(model1,model2,model3,model4)

       Model  df      AIC      BIC    logLik
model1     1 108 3797.554 4248.547 -1790.777
model2     2 109 3799.246 4254.416 -1790.623
model3     3 106 4535.322 4977.964 -2161.661
model4     4 107 4537.115 4983.933 -2161.558
```

Opgave 3 (3 spørgsmål)

Ved et eksperimentielt forsøg ønsker man at afprøve alle 8 kombinationer af tre faktorer temperatur (`temp`), tørke (`drought`) og CO_2 indhold (`co2`). Hver af de 3 faktorer har 2 niveauer.

Ved udførelsen af forsøget tænker man sig i første omgang at benytte 8 plots (=forsøgsenheder) fordelt på 2 blokke med hver 4 forsøgsenheder.

1. Angiv en forsøgsplan, hvor vekselvirkningen `temp` $\times$ `drought` $\times$ `co2` er konfunderet med blok.

Efter at have modtaget en større økonomisk bevilling bliver det muligt at udvide forsøgsdesignet, så man kommer til at råde over 12 blokke med hver 4 forsøgsenheder.

2. Er det muligt at udføre forsøget som et balanceret ufuldstændigt blokforsøg, hvor de 8 behandlinger afprøves på 48 forsøgsenheder fordelt på 12 lige store blokke?

Det besluttes at udføre forsøget ved at vælge 6 blokke, hvorpå alle forsøgsenheder gives niveauet `co2`=1, mens der på alle øvrige forsøgsenheder på de resterende 6 blokke benyttes behandlingen `co2`=0. På hver blok afprøves alle 4 kombinationer af faktorerne `temp` og `drought`.

3. Forklar hvilken type forsøg der er tale om. Tegn et faktordiagram og opskriv en statistisk model til analyse af et datasæt, hvor man foretager en måling, `y`, på hver af de 48 forsøgsenheder. Forklar desuden hvordan randomiseringen af behandlinger til forsøgsenheder bør foretages.