Eksamen i Statistisk Dataanalyse 2
(kursusnr.: 210006)

4. april 2009

Alle sædvanlige hjælpemidler, herunder bøger og lommeregner men *ikke* PC, er tilladt. Opgavesættet består af 8 sider med i alt 3 opgaver, der indgår med vægtningen 40 %, 35 % og 25 % i bedømmelsen. Til nogle opgaver er vedlagt 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 (6 spørgsmål)

Ved Institut for Produktionsdyr og Heste på Det Biovidenskabelige Faktultet ønsker man at udvikle måleudstyr til bestemmelse af halthed hos heste. For at afprøve en ny målemetode, har man udført et forsøg, hvor 8 heste på skift udstyres med en særlig sko på hvert af de 4 ben. Skoen er designet til at give hesten en fornemmelse af, at den er halt på det pågældende ben.

Data består af 40 målinger af halthedsgraden (`y`) svarende til, at der på hver af de 8 heste er lavet en normalmåling (uden den særlige sko) samt 4 halthedsmålinger (med den særlige sko placeret på hvert af de 4 ben). Til forsøget er knyttet faktorerne `hest` og faktoren `halt` med niveauerne `normal`, `vforben`, `hforben`, `vbagben` og `hbagben`. Formålet med den statistiske analyse er at bestemme, hvordan halthedsgraden afhænger af, hvilket ben hesten er halt på.

1. Opstil den statistiske model svarende til `modelA` i R-udskriften.
2. Opstil hypotesen om, at effekten af `hest` kan ignoreres og benyt R-udskriften til at udføre et test af hypotesen.
3. Tag udgangspunkt i slutmodellen fra spørgsmål 2. Opstil hypotesen om at effekten af `halt` kan ignoreres og benyt R-udskriften til at udføre et test for hypotesen.
4. Datasættet indholder udover `hest` og `halt` desuden faktoren `h2` med niveauerne `for`, `bag` og `norm`. Faktoren beskriver, om observationen svarer til en måleserie, hvor hesten var halt på et forben, et bagben eller om hesten ikke var halt. Tegn et faktordiagram, hvor faktorerne `hest`, `halt` og `h2` indgår med systematisk effekt.
5. Opskriv den statistiske model svarende til `modelD` og test, om `modelD` giver en rimelig beskrivelse af data. Forklar i ord hvad det er, man tester for. Formuler slutmodellen af den statistiske analyse i spørgsmål 1.-5. og angiv parameterestimater under slutmodellen.
6. Tag udgangspunkt i slutmodellen fra spørgsmål 5. og formuler hypotesen om, at der ikke er forskel på forbens- og bagbenshalthed. Benyt R-udskriften til at udføre et test af hypotesen.

Udskrift af R-kørsel (lettere redigeret):

```
> data  ### et udpluk af datasættet

   hest    halt   h2         y
1    b1  normal norm -5.670551
2    b1 vforben  for -3.088227
3    b1 vbagben  bag -3.459759
4    b1 hforben  for -2.727496
5    b1 hbagben  bag -3.414645
6    b2  normal norm -4.957237
7    b2 vforben  for -2.708009
8    b2 vbagben  bag -4.381775
9    b2 hforben  for -2.739101
.
.
39   b8 hforben  for -4.111774
40   b8 hbagben  bag -4.217532

> modelA<-lm(y~halt+hest)
> modelB<-lm(y~halt)
> modelC<-lm(y~hest)
> modelD<-lm(y~h2)
> modelE<-lm(y~1)

> anova(modelB,modelA)

  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     35 22.3963
2     28 18.1645  7    4.2318 0.9319 0.4978

> anova(modelC,modelA)

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     32 66.176
2     28 18.164  4    48.012 18.502 1.538e-07 ***

> anova(modelE,modelB)

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     39 70.408
2     35 22.396  4    48.012 18.758 2.549e-08 ***

> anova(modelD,modelB)

  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     37 22.825
2     35 22.396  2     0.429 0.3352 0.7175

> anova(modelE,modelD)

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     39 70.408
2     37 22.825  2    47.583 38.566 8.907e-10 ***
```

```
> modelB1<-lm(y~halt-1)
> summary(modelB1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
haltnormal   -5.7223     0.2828 -20.233  < 2e-16 ***
halthbagben  -3.7802     0.2828 -13.366 2.60e-15 ***
halthforben  -2.8237     0.2828  -9.984 8.84e-12 ***
haltvbagben  -3.5012     0.2828 -12.380 2.41e-14 ***
haltvforben  -2.6522     0.2828  -9.378 4.43e-11 ***

Residual standard error: 0.7999 on 35 degrees of freedom

> est1<-c(0,1,-1,0,0)
> est2<-c(0,0,0,1,-1)
> est3<-c(0,1,-1,1,-1)
> estB<-rbind(est1,est2,est3)
> estimable(modelB1,estB,conf.int=0.95)
      Estimate Std. Error   t value DF    Pr(>|t|)  Lower.CI    Upper.CI
est1 -0.956452  0.3999666 -2.391330 35 0.022296409 -1.768427 -0.14447666
est2 -0.848948  0.3999666 -2.122547 35 0.040940879 -1.660923 -0.03697252
est3 -1.805400  0.5656382 -3.191793 35 0.002983598 -2.953707 -0.65709336

> modelD1<-lm(y~h2-1)
> summary(modelD1)

Coefficients:
       Estimate Std. Error t value Pr(>|t|)
h2bag   -3.6407     0.1964  -18.54  < 2e-16 ***
h2for   -2.7380     0.1964  -13.94 2.63e-16 ***
h2norm  -5.7223     0.2777  -20.61  < 2e-16 ***

Residual standard error: 0.7854 on 37 degrees of freedom

> est4<-c(1,-1,0)
> est5<-c(1,0,-1)
> est6<-c(0,1,-1)
> estD<-rbind(est4,est5,est6)
> estimable(modelD1,estD,conf.int=0.95)
      Estimate Std. Error   t value DF     Pr(>|t|)  Lower.CI   Upper.CI
est4 -0.902700  0.2776912 -3.250733 37 2.455756e-03 -1.465356 -0.3400441
est5  2.081669  0.3401009  6.120741 37 4.309709e-07  1.392560  2.7707793
est6  2.984369  0.3401009  8.774953 37 1.425327e-10  2.295260  3.6734793

> qt(0.975,35)
[1] 2.030108
> qt(0.975,37)
[1] 2.026192
> qt(0.975,40)
[1] 2.021075
```

Opgave 2 (5 spørgsmål)

Enzymer, som spalter ATP (ATPaser), har stor betydning for ionbalancen i levende organismer. Et forsøg udført af Ian Max Møller i 1976 ved KVL havde til formål at undersøge, hvordan ATPase aktiviteten i hvederødder afhænger af dyrkningsforhold og temperatur. Ved forsøget indgår 6 forskellige dyrkningsbetingelser givet ved faktoren `treat`. For hver dyrkningsbetingelse forberedtes 3 reagensglas (`tube`), og en koncentration (`y`), som benyttes til at beregne ATPase aktiviteten, måltes for hver af de 18 reagensglas ved temperaturerne (`temp`) 5, 10, 15, 20, 25, 30 og 35 grader Celsius. I første omgang opfattes `temp` som en faktor på 7 niveauer.

1. Opskriv den statistiske model svarende til `model1` i R-udskriften.

2. Ifølge Arrhenius ligning er der en lineær sammenhæng mellem `y` og den reciprokke Kelvin temperatur. I R-udskriften er den reciprokke Kelvin temperatur beskrevet ved variablen `tempreci`. Opskriv den statistiske model som beskriver, at der for hver af de 6 behandlinger er en lineær sammenhæng mellem `y` og reciprok Kelvin temperatur.

3. Udfør et test for om Arrhenius ligning er opfyldt for datasættet, ved at undersøge om modellen fra spørgsmål 1. kan reduceres til modellen under spørgsmål 2.

4. Reducer modellen med henblik på at undersøge, hvordan `y` afhænger af dyrkningsbetingelse og temperatur. Afrapporter parameterestimater under slutmodellen, og angiv den forventede værdi af koncentrationen `y` for dyrkningsbetingelse `CaCl2:1` ved temperaturen 20 grader Celcius svarende til en reciprok Kelvin temperatur på

$$\texttt{tempreci} = 1/(273+20) = 0.00341.$$

5. Betragt en statistisk model, hvor middelværdistrukturen er givet ved følgende ligning

$$\mathrm{E}\texttt{y}_i = \alpha(\texttt{treat}_i) + \beta(\texttt{treat}_i) \cdot \texttt{tempreci}_i + \gamma \cdot \texttt{tempreci}_i^2 \qquad (1)$$

   Hvilken hypotese udtrykt ved parametrene i modellen (1) ovenfor kan benyttes til at teste, om Arrhenius ligning er opfyldt? Forklar hvordan du ved hjælp af R-udskriften kan udføre dette test.

Udskrift af R-kørsel (lettere redigeret):

```
> data  ### et udpluk af datasættet

    tube      treat temp    tempreci         y
1      1   Backgr:1    5 0.003597122 0.0000000
2      1   Backgr:1   10 0.003533569 0.3364722
3      1   Backgr:1   15 0.003472222 0.7884574
4      1   Backgr:1   20 0.003412969 1.3862944
5      1   Backgr:1   25 0.003355705 1.6094379
6      1   Backgr:1   30 0.003300330 1.7917595
7      1   Backgr:1   35 0.003246753 2.1041342
8      2   Backgr:1    5 0.003597122 1.5686159
9      2   Backgr:1   10 0.003533569 2.1747517
10     2   Backgr:1   15 0.003472222 2.3608540
11     2   Backgr:1   20 0.003412969 2.6100698
12     2   Backgr:1   25 0.003355705 2.8332133
13     2   Backgr:1   30 0.003300330 2.9444390
14     2   Backgr:1   35 0.003246753 3.0773123
15     3   Backgr:1    5 0.003597122 1.7578579
16     3   Backgr:1   10 0.003533569 2.0668628
17     3   Backgr:1   15 0.003472222 2.3321439
18     3   Backgr:1   20 0.003412969 2.5802168
19     3   Backgr:1   25 0.003355705 2.7911651
20     3   Backgr:1   30 0.003300330 2.9231616
21     3   Backgr:1   35 0.003246753 3.1135153
22     4    MgCl2:1    5 0.003597122 1.1631508
23     4    MgCl2:1   10 0.003533569 1.7227666
.
.
124   18  CaCl2:100   25 0.003355705 3.8607297
125   18  CaCl2:100   30 0.003300330 4.1478853
126   18  CaCl2:100   35 0.003246753 4.3067642

> library(nlme)
> model1<-lme(y~treat*temp,random=~1|tube,method="ML")
> model2<-lme(y~treat+temp,random=~1|tube,method="ML")
> model3<-lme(y~treat*tempreci,random=~1|tube,method="ML")
> model4<-lme(y~treat+tempreci,random=~1|tube,method="ML")

> anova(model2,model1)
       Model df       AIC      BIC   logLik   Test  L.Ratio p-value
model2     1 14  15.47402 55.18197  6.26299
model1     2 44 -59.01829 65.77811 73.50915 1 vs 2 134.4923  <.0001

> anova(model3,model1)
       Model df       AIC       BIC   logLik   Test  L.Ratio p-value
model3     1 14 -87.75283 -48.04488 57.87641
model1     2 44 -59.01829  65.77811 73.50915 1 vs 2 31.26546  0.4024

> anova(model4,model3)
       Model df       AIC       BIC   logLik   Test  L.Ratio p-value
model4     1  9  12.89128  38.41782  2.55436
model3     2 14 -87.75283 -48.04488 57.87641 1 vs 2 110.6441  <.0001
```

```
> model2ny<-lme(y~treat+temp,random=~1|tube,method="REML")
> summary(model2ny)

Random effects:
 Formula: ~1 | tube
        (Intercept)  Residual
StdDev:   0.3906985 0.1902626

Fixed effects: y ~ treat + temp
                    Value Std.Error  DF   t-value p-value
(Intercept)     0.9891440 0.2330867 102  4.243675  0.0000
treatBackgr:100 0.5979827 0.3243627  12  1.843562  0.0901
treatCaCl2:1    0.8911299 0.3243627  12  2.747325  0.0177
treatCaCl2:100  1.5895022 0.3243627  12  4.900385  0.0004
treatMgCl2:1    0.2308058 0.3243627  12  0.711567  0.4903
treatMgCl2:100  0.9094157 0.3243627  12  2.803700  0.0159
temp10          0.4199698 0.0634209 102  6.621950  0.0000
temp15          0.7695201 0.0634209 102 12.133547  0.0000
temp20          1.1636373 0.0634209 102 18.347863  0.0000
temp25          1.4455135 0.0634209 102 22.792397  0.0000
temp30          1.7041987 0.0634209 102 26.871263  0.0000
temp35          1.9567305 0.0634209 102 30.853104  0.0000

> model3ny<-lme(y~treat+treat:tempreci-1,random=~1|tube,method="REML")
> summary(model3ny)

Random effects:
 Formula: ~1 | tube
        (Intercept)  Residual
StdDev:    0.394751 0.1179786

Fixed effects: y ~ treat + treat:tempreci - 1
                             Value Std.Error  DF   t-value p-value
treatBackgr:1               18.023   0.78748  12  22.88626       0
treatBackgr:100             15.047   0.78748  12  19.10725       0
treatCaCl2:1                21.897   0.78748  12  27.80592       0
treatCaCl2:100              22.876   0.78748  12  29.04935       0
treatMgCl2:1                25.998   0.78748  12  33.01469       0
treatMgCl2:100              27.286   0.78748  12  34.64939       0
treatBackgr:1:tempreci   -4673.084 220.47139 103 -21.19587       0
treatBackgr:100:tempreci -3627.158 220.47139 103 -16.45183       0
treatCaCl2:1:tempreci    -5546.090 220.47139 103 -25.15560       0
treatCaCl2:100:tempreci  -5628.270 220.47139 103 -25.52835       0
treatMgCl2:1:tempreci    -6939.768 220.47139 103 -31.47695       0
treatMgCl2:100:tempreci  -7117.904 220.47139 103 -32.28493       0
```

```
> model4ny<-lme(y~treat+tempreci-1,random=~1|tube,method="REML")
> summary(model4ny)
Linear mixed-effects model fit by REML
 Data: NULL
       AIC      BIC   logLik
  9.022365 34.03448 4.488818

Random effects:
 Formula: ~1 | tube
        (Intercept)  Residual
StdDev:   0.3905593 0.1922538

Fixed effects: y ~ treat + tempreci - 1
                    Value Std.Error  DF   t-value p-value
treatBackgr:1      21.151   0.55116  12  38.37558       0
treatBackgr:100    21.749   0.55116  12  39.46053       0
treatCaCl2:1       22.042   0.55116  12  39.99240       0
treatCaCl2:100     22.741   0.55116  12  41.25949       0
treatMgCl2:1       21.382   0.55116  12  38.79434       0
treatMgCl2:100     22.061   0.55116  12  40.02558       0
tempreci        -5588.712 146.67238 108 -38.10337       0

> tempreci2<-tempreci*tempreci
> modelalt<-lme(y~treat+treat:tempreci+tempreci2-1,random=~1|tube,method="REML")
> summary(modelalt)

Random effects:
 Formula: ~1 | tube
        (Intercept)  Residual
StdDev:   0.3951453 0.1083446

Fixed effects: y ~ treat + treat:tempreci + tempreci2 - 1
                            Value Std.Error  DF   t-value p-value
treatBackgr:1                 -25       9.6  12 -2.572097  0.0245
treatBackgr:100               -28       9.6  12 -2.882703  0.0138
treatCaCl2:1                  -21       9.6  12 -2.167738  0.0510
treatCaCl2:100                -20       9.6  12 -2.065538  0.0612
treatMgCl2:1                  -17       9.6  12 -1.739617  0.1075
treatMgCl2:100                -15       9.6  12 -1.605258  0.1344
tempreci2                -3648945  817029.1 102 -4.466114  0.0000
treatBackgr:1:tempreci      20296    5594.5 102  3.627888  0.0004
treatBackgr:100:tempreci    21342    5594.5 102  3.814844  0.0002
treatCaCl2:1:tempreci       19423    5594.5 102  3.471841  0.0008
treatCaCl2:100:tempreci     19341    5594.5 102  3.457152  0.0008
treatMgCl2:1:tempreci       18030    5594.5 102  3.222725  0.0017
treatMgCl2:100:tempreci     17851    5594.5 102  3.190884  0.0019
```

Opgave 3 (3 spørgsmål)

Ved et dyrkningsforsøg råder man over 20 marker, som hver især kan opdeles i et antal mindre parceller (forsøgsenheder). Formålet med forsøget er at finde ud af, om en given behandling har en positiv effekt på udbyttet. I forsøget indgår ud over behandlingsfaktoren, `treat`, med niveauer `T` (behandlet) og `U` (ubehandlet) desuden tre forskellige sorter givet ved faktoren `sort` med niveauerne `1,2,3`. Ved forsøget har man således interesse i at afprøve de 6 forskellige kombinationer givet ved produktfaktoren `treat` $\times$ `sort`.

1. Hvilken forsøgsplan ville du benytte, hvis det var muligt at opdele hver af de 20 marker i helt op til 6 forskellige parceller. Forklar hvordan randomiseringen bør foretages.

2. Af praktiske årsager viser det sig kun at være muligt at opdele de enkelte marker i 3 parceller, således at det totale antal forsøgsenheder bliver 60. En mulig forsøgsplan består i at sørge for, at netop 10 marker behandles (mens 10 forbliver ubehandlede) samtidig med at det sikres, at hver af de 3 sorter optræder på en af parcellerne inden for hver mark. Opskriv et faktordiagram for denne forsøgsplan. Hvilken type forsøg er der tale om, og hvordan bør randomiseringen foretages?

3. Betragt i stedet følgende forsøgsplan, hvor `b1,b2,...,b20` refererer til de forskellige marker, og f.eks. `T3` betyder at pågældende parcel er sået med `sort=3` og skal behandles (`treat=T`). Tabellen angiver således under hver af de 20 marker, hvilken behandling der skal benyttes på de tre parceller inden for marken.

b7	b19	b13	b8	b12	b14	b4	b20	b17	b6
T2	T3	U3	T1	U1	U3	U1	T3	U1	T3
U2	U2	T2	U1	U2	T3	T2	T1	U3	U1
U3	U1	T1	U3	T2	U2	U3	U3	U2	T2
b11	b15	b2	b9	b10	b1	b5	b16	b3	b18
U3	T2	T3	T1	T2	T1	T3	T3	T1	U1
T2	T1	U1	U2	U1	U2	U2	U2	T2	T1
T3	T3	U3	U1	T1	U3	T2	T1	U2	T3

   Hvilken type forsøg er der tale om (begrund svaret)? Hvordan bør randomiseringen, som ligger til grund for forsøgsplanen, være foretaget?