Eksamen i Statistisk Dataanalyse 2
(kursusnr.: 210006)

7. april 2011

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 8 sider med i alt 3 opgaver, der indgår med vægtningen 40 %, 35 % og 25 % 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 2 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 (4 spørgsmål)

Ved et fodringsforsøg blev 38 lam randomiseret til 4 forskellige behandlingsgrupper. Efter 6 ugers behandling målte man glukose-indholdet i blodet en halv time før fodring samt 1 og 2.5 timer efter fodring (måletidspunktet er givet ved variablen `time`). Formålet med forsøget var at undersøge, hvordan behandlingen givet ved faktoren `treat` påvirker glukose-indholdet beskrevet ved variablen `glu` i forbindelse med fodring. Data til opgaven er venligst stillet til rådighed af Anna Hauntoft Kongsted.

Data er udleveret på vedlagte USB-stick under filnavnet `ahk.txt` og for at besvare opgaven fuldstændigt, vil det være nødvendigt at køre udvalgte R-kommandoer på din egen medbragte computer. Data kan f.eks. indlæses ved brug af kommandoen

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

hvor du vælger filen `ahk.txt`. De første linjer i datasættet er organiseret som vist nedenfor

```
   lamb treat time  glu
1  1114     A -0.5 3.93
2  1114     A  1.0 7.21
3  1114     A  2.5 5.09
4  1115     B -0.5 5.39
5  1115     B  1.0 4.70
```

```
6  1115     B  2.5 5.40
7  1116     A -0.5 4.60
8  1116     A  1.0 5.60
9  1116     A  2.5 5.70
10 1117     B -0.5 5.09
```

Bemærk at nogle af variablene skal laves om til faktorer inden den statistiske analyse.

1. Man ønsker at tage udgangspunkt i en passende *random intercept* model, som gør det muligt at undersøge, hvordan glukoseindholdet (`glu`) afhænger af faktorerne behandling (`treat`) og måletidspunkt (`time`). Opskriv dit forslag til en udgangsmodel for den statistiske analyse.

2. Reducer modellen fra delspørgsmål 1. mest muligt og angiv parameterestimater for slutmodellen. Forklar i ord hvad slutmodellen udtrykker. Husk også at angive 95 %-konfidensintervaller for parametrene i middelværdistrukturen. I forbindelse med reduktionen af modellen bedes du i din besvarelse opskrive de statistiske modeller, som du benytter dig af undervejs.

3. Angiv et estimat og et 95 %-konfidensinterval for tilvæksten i glukose fra en halv time før fodring til 2.5 timer efter fodring for et lam, som har modtaget behandling `A`.

4. Er det rimeligt at antage, at glukosekoncentrationen hen over hele perioden er den samme for lam som har modtaget behandlingerne `B` og `D`?

Opgave 2 (5 spørgsmål)

Med henblik på at undersøge formen af gulerødder blev der i sommeren 2010 udført et dyrkningsforsøg. For at gøre det simpelt opfatter vi formen af en gulerod som en keglespids (dvs. et kræmmerhus). I datasættet indgår sammenhørende værdier af omkreds (`omkreds`) i den tykke ende og længde (`length`) for 67 gulerødder (-alle mål angivet i cm). De 67 gulerødder fordeler sig på 3 forskellige sorter givet ved variablen `variety` med 3 niveauer `gul, orange` og `rød`. Omkredsen af gulerødderne i datasættet ligger alle mellem 3.9 cm og 10.9 cm. Et udsnit af datasættet ses nedenfor

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

```
> head(roots, 10)
```

```
   variety omkreds length
1      gul     8.1   12.5
2      gul    10.6   11.0
3      gul     8.0    7.5
4      gul     8.8   11.5
5      gul    10.9   11.5
6      gul     7.3    8.5
7      gul     9.6   11.5
```

```
8   orange     5.9    3.5
9   orange     5.6    6.5
10  orange     8.0    8.0
```

Man kan argumentere for, at hvis både små og store gulerødder har nogenlunde samme form, så bør der være en lineær sammenhæng mellem længde og diameter. Da omkreds og diameter for en cirkel hænger sammen via formlen

$$\texttt{omkreds} = \pi \cdot \texttt{diameter} \approx 3.14 \cdot \texttt{diameter}$$

så vil teorien også medføre en linæer sammenhæng mellem variablene længde (`length`) og `omkreds`, der findes i datasættet.

Ved besvarelsen af opgaven skal du benytte R-udskriften sidst i opgaven. Bemærk at du ikke nødvendigvis skal bruge alle dele af R-udskriften.

1. Opskriv (på papir) en statistisk model der beskriver, at der er en lineær sammenhæng mellem længde og omkreds for hver af de 3 sorter i forsøget. Benyt figur 1 (på en af de senere sider) til at argumentere for, at din model giver en rimelig beskrivelse af data.

2. Foretag en statistisk analyse med henblik på at undersøge, om de 3 gulerodssorter i forsøget kan antages at have den samme form. Angiv estimater for alle parametre i slutmodellen samt 95 %-konfidensintervaller for parametrene, der indgår i beskrivelsen af middelværdistrukturen.

3. Ud fra et praktisk synspunkt bør længden af en gulerod nærme sig 0, når omkredsen bliver tilpas lille. Argumentér for at dette kan bekræftes af den statistiske analyse ved at opskrive en relevant hypotese samt resultatet af det tilhørende test.

4. Kan man på baggrund af den statistiske analyse konkludere, at en rød gulerod er omtrent 3 gange så lang som den er bred (=diameteren)?

5. Angiv et estimat og et 95 %-konfidensinterval for længder af `orange` gulerødder med en omkreds på hhv 2, 5 og 15 cm og kommentér resultatet i lyset af tidligere delspørgsmål i opgaven.

Udskrift af R-kørsel (lettere redigeret):

```
> ### Nogle statistiske modeller og test:
>
> modelA<-lm(length~variety*omkreds,data=roots)
> modelB<-lm(length~variety+omkreds-1,data=roots)
> modelC<-lm(length~variety-1,data=roots)
> modelD<-lm(length~omkreds,data=roots)
> modelE<-lm(length~omkreds-1,data=roots)
> modelF<-lm(length~1,data=roots)

> anova(modelB, modelA)
```

```
  Res.Df      RSS Df Sum.of.Sq         F    Pr..F.
1     63 90.16979 NA        NA        NA        NA
2     61 88.52296  2  1.646835 0.5674061 0.5699579
```

```
> anova(modelC, modelA)
```

```
  Res.Df       RSS Df Sum.of.Sq        F       Pr..F.
1     64 228.18577 NA        NA       NA           NA
2     61  88.52296  3  139.6628 32.07993 1.428875e-12
```

```
> anova(modelC, modelB)
```

```
  Res.Df      RSS Df Sum.of.Sq        F       Pr..F.
1     64 228.1858 NA        NA       NA           NA
2     63  90.1698  1  138.0160 96.42926 2.533941e-14
```

```
> anova(modelD, modelB)
```

```
  Res.Df      RSS Df Sum.of.Sq        F    Pr..F.
1     65 93.59795 NA        NA       NA        NA
2     63 90.16979  2  3.428151 1.197594 0.3086989
```

```
> anova(modelF, modelC)
```

```
  Res.Df      RSS Df Sum.of.Sq        F       Pr..F.
1     66 294.5672 NA        NA       NA           NA
2     64 228.1858  2   66.3814 9.309102 0.0002827061
```

```
> anova(modelF, modelD)
```

```
  Res.Df       RSS Df Sum.of.Sq        F       Pr..F.
1     66 294.56716 NA        NA       NA           NA
2     65  93.59795  1  200.9692 139.5650 7.789958e-18
```

```
> ### Dele af summary() på udvalgte modeller:
>
> summary(modelB)
```

```
               Estimate Std. Error    t value     Pr(>|t|)
varietygul    -0.1410516  0.8445838 -0.1670073 8.678992e-01
varietyorange -0.6227173  0.7279907 -0.8553918 3.955758e-01
varietyrød    -0.7706681  0.6855899 -1.1240950 2.652377e-01
omkreds        1.1146433  0.1135093  9.8198401 2.533941e-14

Residual standard error: 1.196 on 63 degrees of freedom
```

```
> confint(modelB)

                  2.5 %    97.5 %
varietygul    -1.828818 1.5467144
varietyorange -2.077491 0.8320561
varietyrød    -2.140710 0.5993739
omkreds        0.887813 1.3414736

> summary(modelD)

             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) -0.9602306  0.6525882 -1.471419 1.460029e-01
omkreds      1.1904812  0.1007707 11.813764 7.789958e-18

Residual standard error: 1.2 on 65 degrees of freedom

> confint(modelD)

                2.5 %    97.5 %
(Intercept) -2.2635392 0.3430779
omkreds      0.9892282 1.3917342

> summary(modelE)

        Estimate Std. Error  t value     Pr(>|t|)
omkreds 1.045995 0.02283673 45.80319 9.526988e-52

Residual standard error: 1.211 on 66 degrees of freedom

> confint(modelE)

          2.5 %   97.5 %
omkreds 1.000400 1.091590

> summary(modelF)

            Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 6.552239  0.2580969 25.38674 9.008437e-36

Residual standard error: 2.113 on 66 degrees of freedom

> confint(modelF)

              2.5 %   97.5 %
(Intercept) 6.036932 7.067546
```

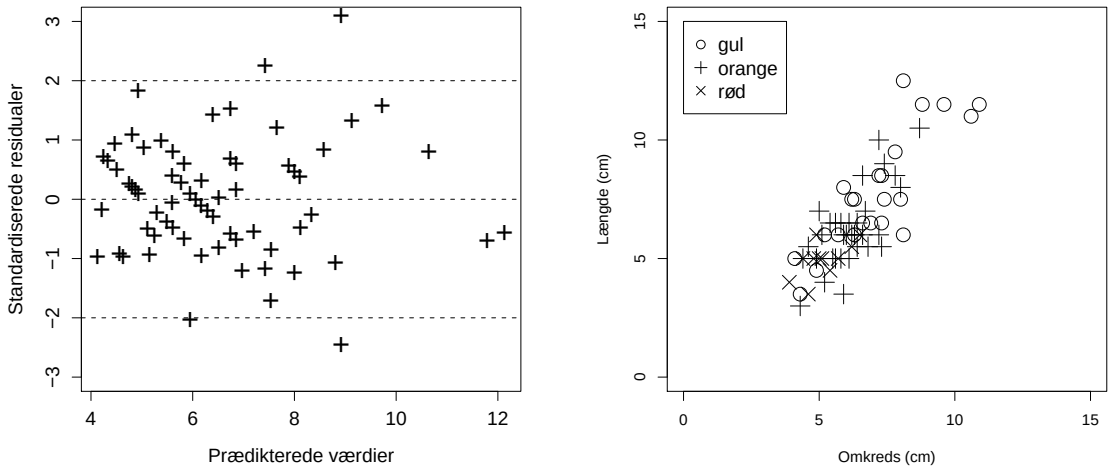
```
> library(gmodels)
> estB.1 <- c(1, 1, 0, 2)
> estB.2 <- c(0, 1, 0, 2)
> estB.3 <- c(1, 1, 0, 5)
> estB.4 <- c(0, 1, 0, 5)
> estB.5 <- c(1, 1, 0, 15)
> estB.6 <- c(0, 1, 0, 15)
> estB <- rbind(estB.1, estB.2, estB.3, estB.4, estB.5, estB.6)
> estD.1 <- c(0, 2)
> estD.2 <- c(1, 2)
> estD.3 <- c(0, 5)
> estD.4 <- c(1, 5)
> estD.5 <- c(0, 15)
> estD.6 <- c(1, 15)
> estD <- rbind(estD.1, estD.2, estD.3, estD.4, estD.5, estD.6)

> estimable(modelB, estB)

        Estimate Std. Error   t value DF     Pr(>|t|)
estB.1  1.465518  1.3181356  1.111811 63 2.704443e-01
estB.2  1.606569  0.5145745  3.122132 63 2.710799e-03
estB.3  4.809448  0.9916780  4.849808 63 8.451822e-06
estB.4  4.950499  0.2455134 20.163863 63 0.000000e+00
estB.5 15.955881  0.3806207 41.920685 63 0.000000e+00
estB.6 16.096933  1.0264230 15.682552 63 0.000000e+00

> estimable(modelD, estD)

        Estimate Std. Error   t value DF    Pr(>|t|)
estD.1  2.380962  0.2015414 11.813764 65 0.000000000
estD.2  1.420732  0.4584393  3.099062 65 0.002867745
estD.3  5.952406  0.5038535 11.813764 65 0.000000000
estD.4  4.992175  0.1973082 25.301411 65 0.000000000
estD.5 17.857218  1.5115604 11.813764 65 0.000000000
estD.6 16.896987  0.8878394 19.031580 65 0.000000000
```



Figur 1: Residualplot af standardiserede residualer tegnet op imod prædikterede værdier svarende til `modelA` (venstre figur). Længdemålinger tegnet op imod omkredsen (højre figur).

Opgave 3 (3 spørgsmål)

I forbindelse med et mejeriforsøg på Det Biovidenskabelige Fakultet ønsker man at sammenligne to behandlinger `B1` og `B2` givet ved faktoren `beh`. Forsøget udføres i en række kamre hver med plads til to prøver, og vi råder over i alt 6 kamre. I første omgang beslutter man sig for også at anvende to forskellige `doser` (`D1` og `D2`) og at benytte følgende forsøgsplan, som gentages 3 gange, således at alle 6 kamre bliver anvendt en gang.

kammer 1	kammer 2
`B1,D1`	`B1,D2`
`B2,D2`	`B2,D1`

1. Beskriv hvilken type forsøg, der er tale om og giv et forslag til en forbedring af forsøgsplanen til afprøvning af de 4 kombinationer af behandling (`beh`) og `dosis` i de 6 kamre.

I den resterende del af opgaven ønsker vi ikke længere at afprøve begge doser, og du skal således se helt bort fra denne faktor. Til gengæld ønsker man at variere temperaturen (`temp`), der i hvert kammer kan instilles på to niveauer `høj` og `lav`. Der benyttes stadig 6 kamre med plads til totalt 12 prøver ved udførelsen af forsøget.

2. Giv et forslag til en forsøgsplan, hvis man både ønsker at undersøge effekten af behandling (`beh`) og temperatur (`temp`). Opskriv en statistisk model til analyse af forsøget og forklar, hvordan randomiseringen bør foretages.

3. Da man også ønsker at kunne inddrage dag-til-dag variationen ved vurderingen af forsøgets resultater, vælger man at udføre forsøget over 3 dage. Hver dag anvendes

2 af de 6 kamre på en sådan måde, at alle 4 kombinationer af temperatur (`temp`) og behandling (`beh`) afprøves på hver af de 3 dage. Hvert af de 6 kamre anvendes således præcis en gang i forsøget. Opskriv en statistisk model og et tilhørende faktordiagram som viser, hvordan man kan inddrage faktoren `dag` ved analysen af forsøgsresultaterne.