Eksamen i Statistisk Dataanalyse 2
(kursusnr.: 210006)

14. april 2007

Alle sædvanlige hjælpemidler, herunder bøger og lommeregner men *ikke* PC, er tilladt. Der er 10 sider med i alt 3 opgaver der alle ønskes besvaret og som indgår med samme vægt i bedømmelsen. Til alle opgaver er vedlagt udskrift af R-kørsler 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)

Bioprotein, dvs. protein fremstillet ud fra naturgas, anses for at være en potentielt god proteinkilde for dyr og mennesker. For at undersøge hvordan væksten påvirkes af indtagelse af Bioprotein har man udført et eksperiment hvor 40 rotter fik en del af deres proteinindtag erstattet med Bioprotein i 14 uger. Rotterne blev tilfældigt allokeret til en af fire grupper med 10 rotter i hver gruppe. For rotterne i gruppe 2, 3 og 4 bestod hhv. 6%, 12% og 24% af proteinindtaget af Bioprotein, mens gruppe 1 var en kontrolgruppe der således indtog 0% Bioprotein.

Rotterne blev vejet hver uge under forsøget. I spørgsmål 1 og 2 skal vi dog kun bruge målingerne fra uge 7 og 13, i spørgsmål 3 og 4 desuden en måling fra forsøgets start.

1. Betragt kun målingerne fra uge 7 og 13 (ikke startmålingerne). Opskriv en statistisk model for forsøget, og angiv det tilhørende faktordiagram.
2. Analysér data med særligt henblik på at undersøge om indtagelse af Bioprotein påvirker væksten af rotter. Husk i den forbindelse at angive relevante estimater og at konkludere hvilken effekt behandlingen har.

Rotterne blev som nævnt også vejet ved forsøgets start.

3. Modificér slutmodellen fra spørgsmål 2 således at startmålingerne inddrages. Giver den modificerede model en signifikant bedre beskrivelse af data end modellen fra spørgsmål 2? Ændres konklusionen vedrørende behandlingerne? (Svarene skal begrundes.)
4. Angiv et estimat og et konfidensinterval for den forventede vægt efter 13 uger for en rotte der ved forsøgets start vejede 100 g og som fik erstattet 6% af protein i kosten med Bioprotein.

Udskrift af R-kørsel (lettere redigeret):

```
##### Datasæt mm. (Tallene yderst til venstre er observationsnumrene i
##### det oprindelige datasæt og skal IKKE bruges til noget. Der er i alt
##### 80 observationer i datasættet.

> bioprotein
    rat week weight treat start
7     1    7    274     1    94
13    1   13    324     1    94
21    2    7    283     1    87
27    2   13    346     1    87
.
.
553  40    7    313     4    90
559  40   13    401     4    90

> attach(bioprotein)
> week = factor(week)
> treat = factor(treat)
> rat = factor(rat)

##### Fit af diverse modeller:

> modelA = lme(weight ~ week + treat + week:treat, random =~ 1|rat)
> modelB = lme(weight ~ week + treat, random =~ 1|rat)
> modelC = lme(weight ~ start + week + treat, random =~ 1|rat)

##### Diverse anova-kommandoer:

> anova(modelA)
            numDF denDF  F-value p-value
(Intercept)     1    36 8543.503  <.0001
week            1    36  871.427  <.0001
treat           3    36    3.975  0.0152
week:treat      3    36    1.468  0.2395

> anova(modelB)
            numDF denDF  F-value p-value
(Intercept)     1    39 8543.496  <.0001
week            1    39  841.146  <.0001
treat           3    36    3.975  0.0152

> anova(modelC)
            numDF denDF   F-value p-value
(Intercept)     1    39 10551.748  <.0001
start           1    35     8.970  0.0050
week            1    39   841.147  <.0001
treat           3    35     5.074  0.0051
```

```
#### Diverse summary-kommandoer (redigeret)

> summary(modelA)
Random effects:
 Formula: ~1 | rat
        (Intercept) Residual
StdDev:    22.51707 10.10096

Fixed effects: weight ~ week + treat + week:treat
              Value Std.Error DF  t-value p-value
(Intercept)   299.7  7.804151 36 38.40264  0.0000
week13         62.2  4.517285 36 13.76933  0.0000
treat2         24.4 11.036737 36  2.21080  0.0335
treat3         21.1 11.036737 36  1.91180  0.0639
treat4          3.3 11.036737 36  0.29900  0.7667
week13:treat2   7.3  6.388406 36  1.14270  0.2607
week13:treat3  10.8  6.388406 36  1.69056  0.0996
week13:treat4  -0.2  6.388406 36 -0.03131  0.9752

> summary(modelB)
Random effects:
 Formula: ~1 | rat
        (Intercept) Residual
StdDev:    22.47626 10.28116

Fixed effects: weight ~ week + treat
               Value Std.Error DF  t-value p-value
(Intercept) 297.4625  7.558084 39 39.35687  0.0000
week13       66.6750  2.298938 39 29.00252  0.0000
treat2       28.0500 10.564407 36  2.65514  0.0117
treat3       26.5000 10.564407 36  2.50842  0.0168
treat4        3.2000 10.564407 36  0.30290  0.7637

> summary(modelC)
Random effects:
 Formula: ~1 | rat
        (Intercept) Residual
StdDev:    19.97436 10.28116

Fixed effects: weight ~ start + week + treat
                Value Std.Error DF   t-value p-value
(Intercept) 163.52822  44.07139 39  3.710530  0.0006
start         1.40392   0.45640 35  3.076078  0.0041
week13       66.67500   2.29894 39 29.002527  0.0000
treat2       29.03275   9.51143 35  3.052406  0.0043
treat3       24.67490   9.52456 35  2.590660  0.0139
treat4        1.65568   9.51931 35  0.173929  0.8629
```

```
##### Diverse estimable-kommandoer

> forv1 = c(1,1,1,1,0,0)
> forv2 = c(1,100,0,1,0,0)
> forv3 = c(1,100,1,1,0,0)
> forv4 = c(1,100,1,0,1,0)
> forv = rbind(forv1, forv2, forv3, forv4)

> estimable(modelC, forv, conf.int=0.95)
      Estimate Std. Error  t value DF     Pr(>|t|) Lower.CI Upper.CI
forv1 260.6399  43.305018  6.01870 35 7.295451e-07 172.7260 348.5537
forv2 332.9533   7.235682 46.01547 35 0.000000e+00 318.2641 347.6425
forv3 399.6283   7.235682 55.23022 35 0.000000e+00 384.9391 414.3175
forv4 395.2704   6.983718 56.59886 35 0.000000e+00 381.0927 409.4481
```

Opgave 2 (4 spørgsmål)

Ved et dyrkningsforsøg indgår 3 faktorer med følgende niveauer:

$$\begin{aligned} \texttt{G}\ (\text{Gødning}) &: \text{g1,g2} \\ \texttt{V}\ (\text{Vanding}) &: \text{v1,v2} \\ \texttt{S}\ (\text{Sort}) &: 1,2,3,4. \end{aligned}$$

Forsøget udføres med henblik på at undersøge effekten af faktorerne:

$$\texttt{G},\ \texttt{V},\ \texttt{S},\ \texttt{G}\times\texttt{V},\ \texttt{G}\times\texttt{S},\ \texttt{V}\times\texttt{S},\ \texttt{G}\times\texttt{V}\times\texttt{S}$$

Man råder over 16 forsøgsenheder (plots) fordelt på et 4×4 kvadrat.

1. I første omgang betragtes følgende forsøgsplan, hvor tallene angiver hvilken af de 4 sorter, som skal anvendes på det pågældende plot. Planen læses således, at v1 anvendes på de første to søjler, mens niveau v2 anvendes på de sidste to søjler. Tilsvarende anvendes de to niveauer af G på plots i 1. og 3. hhv. 2. og 4. række.

	v1	v1	v2	v2
g1	1	1	1	1
g2	2	2	2	2
g1	3	3	3	3
g2	4	4	4	4

   Hvilke af faktorerne $\texttt{G}\times\texttt{V}$, $\texttt{G}\times\texttt{S}$, $\texttt{V}\times\texttt{S}$ er balancerede. Hvilke af faktorerne G, V, S er grovere/finere end hinanden?

2. Forsøgsplanen ændres nu til:

	v1	v1	v2	v2
g1	1	2	3	4
g2	2	1	4	3
g1	3	4	1	2
g2	4	3	2	1

   Tegn faktordiagrammet for modellen med systematisk effekt af faktorerne:

$$\texttt{G},\ \texttt{V},\ \texttt{S},\ \texttt{G}\times\texttt{V},\ \texttt{G}\times\texttt{S},\ \texttt{V}\times\texttt{S}$$

   og afgør, om det er muligt at udvide modellen og teste for trefaktorvekselvirkningen $\texttt{G}\times\texttt{V}\times\texttt{S}$?

3. Forsøget ændres nu til kun at omfatte to sorter. I det følgende antages således, at forsøget omfatter 3 faktorer `G, V` og `S` hver på 2 niveauer. I nedenstående forsøgsplan tænkes forsøget udført på kun 8 forsøgsenheder fordelt på 2 lige store blokke.

Blok 1	g1 v1 s1	g1 v2 s2	g2 v1 s2	g2 v2 s1
Blok 2	g1 v1 s2	g1 v2 s1	g2 v1 s1	g2 v2 s2

Afgør hvilken af faktorerne `G`, `V`, `S`, $\texttt{G} \times \texttt{V}$, $\texttt{G} \times \texttt{S}$, $\texttt{V} \times \texttt{S}$, $\texttt{G} \times \texttt{V} \times \texttt{S}$ som er konfunderet med `B` (Blok).

4. Betragt følgende to forslag til forsøgsplaner, som udvider forsøget til at omfatte 12 forsøgsenheder fordelt på 2 lige store blokke. Hver forsøgsenhed er angivet med et "x".

			Forsøgsplan 1		Forsøgsplan 2	
G	V	S	Blok 1	Blok 2	Blok 1	Blok 2
1	1	1	x x		x x	
1	1	2		x		x
1	2	1		x		x
1	2	2	x	x	x x	
2	1	1		x		x x
2	1	2	x	x	x	
2	2	1	x	x	x	
2	2	2	x			x x

Argumenter for hvilken forsøgsplan man bør foretrække. Du kan evt. benytte nedenstående R-program ved besvarelsen.

Udskrift af `R`-kørsel til brug ved besvarelse af spørgsmål 4:

```
##### Faktorerne G,V,S,B defineret svarende til forsøgsplan 1
> y=rnorm(12)

> model1=lm(y ~ G + V + S + G:V + G:S + V:S + G:V:S)
> anova(model1)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value  Pr(>F)
G          1 3.3549  3.3549  3.8382 0.12168
V          1 0.0252  0.0252  0.0288 0.87354
S          1 3.7792  3.7792  4.3236 0.10610
G:V        1 0.0126  0.0126  0.0145 0.91007
G:S        1 4.3424  4.3424  4.9679 0.08972 .
V:S        1 0.0041  0.0041  0.0047 0.94861
G:V:S      1 0.4752  0.4752  0.5437 0.50183
Residuals  4 3.4963  0.8741
```

```
> model1blok=lm(y ~ B + G + V + S + G:V + G:S + V:S + G:V:S)
> anova(model1blok)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value Pr(>F)
B          1 0.6202  0.6202  0.5417 0.5150
G          1 3.3549  3.3549  2.9305 0.1854
V          1 0.0252  0.0252  0.0220 0.8916
S          1 3.7792  3.7792  3.3011 0.1668
G:V        1 0.1746  0.1746  0.1525 0.7222
G:S        1 3.5711  3.5711  3.1194 0.1755
V:S        1 0.0006  0.0006  0.0005 0.9839
G:V:S      1 0.5297  0.5297  0.4627 0.5452
Residuals  3 3.4345  1.1448

##### Faktorer G,V,S,B defineret svarende til forsøgsplan 2
> y=rnorm(12)

> model2=lm(y ~ G + V + S + G:V + G:S + V:S + G:V:S)
> anova(model2)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value  Pr(>F)
G          1 1.50872 1.50872  5.5137 0.07868 .
V          1 0.00340 0.00340  0.0124 0.91666
S          1 0.26117 0.26117  0.9545 0.38390
G:V        1 2.71468 2.71468  9.9210 0.03452 *
G:S        1 0.03995 0.03995  0.1460 0.72182
V:S        1 0.07843 0.07843  0.2866 0.62078
G:V:S      1 0.04756 0.04756  0.1738 0.69815
Residuals  4 1.09452 0.27363

> model2blok=lm(y ~ B + G + V + S + G:V + G:S + V:S + G:V:S)
> anova(model2blok)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value  Pr(>F)
B          1 0.04154 0.04154  0.1518 0.71666
G          1 1.51474 1.51474  5.5357 0.07827 .
V          1 0.00340 0.00340  0.0124 0.91666
S          1 0.26117 0.26117  0.9545 0.38390
G:V        1 2.71468 2.71468  9.9210 0.03452 *
G:S        1 0.03995 0.03995  0.1460 0.72182
V:S        1 0.07843 0.07843  0.2866 0.62078
Residuals  4 1.09452 0.27363
```

Opgave 3 (4 spørgsmål)

DHL-stafetten er et stort motionsløb hvor hold bestående af fem personer gennemfører 5 gange 5 kilometer. Holdenes kan sammensættes mht. køn efter ønske, således at der er seks forskellige grupper (5 mænd og 0 kvinder, 4 mænd og 1 kvinde osv.). Løbet afvikles over fire dage. I tabellen nedenfor er mediantiderne for hver kombination af løbsdag og gruppe angivet i sekunder.

Mænd/Kvinder	5/0	4/1	3/2	2/3	1/4	0/5
Mandag	7930	8019	8253	8517	8793	9035
Tirsdag	7838	8021	8313	8552	8857	9061
Onsdag	7630	7858	8093	8160	8790	8785
Torsdag	7580	7766	8069	8349	8620	8672

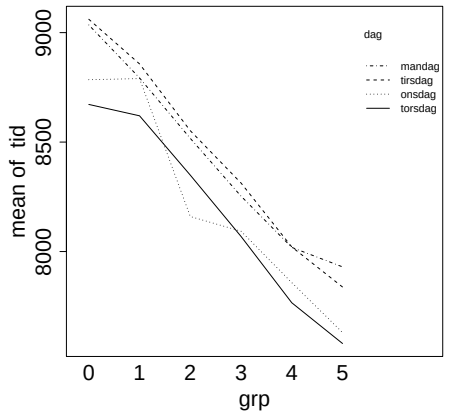
I besvarelsen kan du benytte udskriften fra R-kørslen sidst i opgaven. Figurerne nedenfor er også produceret i R-kørslen.

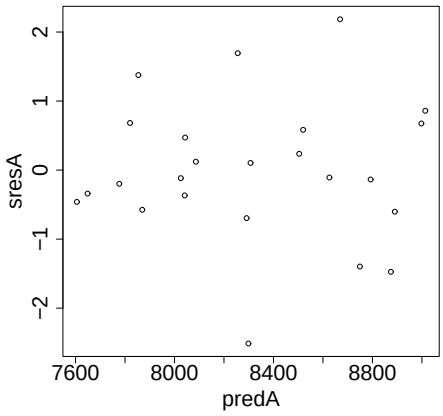
1. Opskriv modellerne svarende til `modelA` og `modelB` i R-udskriften nedenfor. Hvad er dimensionerne af de to modeller?

2. Giver figurerne nedenfor anledning til at betvivle at `modelA` giver en rimelig beskrivelse af data (begrund svaret *kort*)? Undersøg om `modelA` kan reduceres til `modelB`.

3. Angiv et estimat og et konfidensinterval for forskellen mellem mænds og kvinders femkilometertid.

4. Pga. mudder blev ruten ændret efter tirsdag således at ruten mandag og tirsdag var en anden end ruten onsdag og torsdag. Tyder data på at de to ruter var lige hurtige (begrund svaret)?





Udskrift af R-kørsel (lettere redigeret):

```
##### Datasæt mm. #####

> dhldata
       dag m k  tid
1   mandag 5 0 7930
2   mandag 4 1 8019
3   mandag 3 2 8253
.
.
23 torsdag 1 4 8620
24 torsdag 0 5 8672

> attach(dhldata)
> grp = factor(m)

##### Modeller, estimater og test #####

> modelA = lm(tid ~ dag + grp)
> modelB = lm(tid ~ dag + m)

> summary(modelA)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8997.71      42.79 210.252  < 2e-16 ***
dagonsdag    -205.17      40.35  -5.085 0.000134 ***
dagtirsdag     15.83      40.35   0.392 0.700264
dagtorsdag   -248.50      40.35  -6.159 1.83e-05 ***
grp1         -123.25      49.42  -2.494 0.024789 *
grp2         -493.75      49.42  -9.992 5.05e-08 ***
grp3         -706.25      49.42 -14.292 3.83e-10 ***
grp4         -972.25      49.42 -19.675 4.00e-12 ***
grp5        -1143.75      49.42 -23.146 3.76e-13 ***

Residual standard error: 69.88 on 15 degrees of freedom
Multiple R-Squared: 0.984,      Adjusted R-squared: 0.9754
F-statistic: 115.1 on 8 and 15 DF,  p-value: 4.666e-12

> confint(modelA)
                  2.5 %      97.5 %
(Intercept)  8906.49328  9088.92339
dagonsdag    -291.16505  -119.16829
dagtirsdag    -70.16505   101.83171
dagtorsdag   -334.49838  -162.50162
grp1         -228.57608   -17.92392
grp2         -599.07608  -388.42392
grp3         -811.57608  -600.92392
grp4        -1077.57608  -866.92392
grp5        -1249.07608 -1038.42392
```

```
> summary(modelB)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 9030.089     41.030 220.083  < 2e-16 ***
dagonsdag   -205.167     46.824  -4.382 0.000321 ***
dagtirsdag    15.833     46.824   0.338 0.738960
dagtorsdag  -248.500     46.824  -5.307 4.02e-05 ***
m           -242.236      9.693 -24.990 5.37e-16 ***

Residual standard error: 81.1 on 19 degrees of freedom
Multiple R-Squared: 0.9727,     Adjusted R-squared: 0.9669
F-statistic: 168.9 on 4 and 19 DF,  p-value: 1.450e-14

> confint(modelB)
                 2.5 %    97.5 %
(Intercept) 8944.21159 9115.9670
dagonsdag   -303.16984 -107.1635
dagtirsdag   -82.16984  113.8365
dagtorsdag  -346.50318 -150.4968
m           -262.52430 -221.9471

> anova(modelB, modelA)
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     19 124970
2     15  73256  4     51714 2.6473 0.07472 .

##### Plots #####

> interaction.plot(grp,dag,tid)

> predA = predict(modelA)
> sresA = rstandard(modelA)
> plot(predA, sresA)
```