Eksamen i Statistisk Dataanalyse 2
(LMAF10070)

11. april 2013

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 10 sider med i alt 3 opgaver, der indgår med vægtningen 30 %, 30 % og 40 % i bedømmelsen.

Til besvarelse af opgave 3 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 3 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 (3 spørgsmål)

Vi betragter i denne opgave et dyrkningsforsøg med 8 marker, som hver er opdelt i 4 forsøgsenheder (plots). De 32 forsøgsenheder i forsøget skal beplantes med et antal forskellige sorter.

1. Giv et forslag til en forsøgsplan som kan benyttes, hvis der i forsøget skal indgå 4 forskellige sorter. Opskriv et faktordiagram og en tilhørende statistisk model for analyse af forsøget og forklar, hvordan randomiseringen bør foretages.

2. Antag nu i stedet, at der i forsøget skal indgå 8 forskellige sorter, og at vi kan tænke på de 8 sorter som givet ved kombinationer af 3 forskellige faktorer (`A`,`B` og `C`) hver med 2 niveauer. Giv et forslag til en forsøgsplan. Husk at begrunde dit svar.

3. Man beslutter sig for kun at anvende 4 sorter (`s1`,`s2`,`s3`,`s4`) i forsøget. Til gengæld beslutter man sig desuden for, at halvdelen af markerne skal gødes, og at halvdelen af markerne skal overdækkes med plastic indtil risikoen for nattefrost er overstået. Giv et forslag til et forsøgsdesign, hvor hver kombination af `sort` (4 niveauer), `gødning` (2 niveauer) og `plastic` (2 niveauer) afprøves lige mange gange i forsøgsplanen. Forklar hvordan randomiseringen bør fortages og tegn et faktordiagram for forsøget.

Opgave 2 (3 spørgsmål)

I forbindelse med kirurgiske indgreb er det er veldokumenteret, at mænd og kvinder kan have en meget forskellig subjektiv opfattelse af smerter, og at forskellene varierer meget mellem forskellige typer af indgreb. I denne opgave betragter vi et datasæt, hvor har man bedt 23 kvinder (`koen=f`) og 23 mænd (`koen=m`) kvantificere den gennemsnitlige subjektive `smerte` over de 3 første døgn efter kikkertkirurgisk (laparaskopisk) operation for lyskebrok (-højere talværdi svarer til højere smerte). Desuden har man registreret patienternes `alder` samt patienternes smerte (`s0`) umiddelbart inden operationen.

Data til opgaven er venligst stillet til rådighed af Mette Astrup Tolver. Et udpluk af datasættet ses nedenfor.

```
> data2<-read.table(file="laplyske2013sd2.txt",header=T)
> data2

  koen alder s0 smerte
1    m    58  0  19.25
2    m    64  5  17.75
3    f    37  5  59.00
4    m    46  1  28.00
5    m    62 14  44.00
6    m    32  4  19.67

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

   koen alder s0 smerte
44    f    33 14   55.5
45    f    59  0   28.0
46    f    59 44   48.5
```

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. Opskriv en statistisk model som bør tages som udgangspunkt for en statistisk analyse af hvordan patienternes subjektive smertepåvirkning (`smerte`) efter operationen afhænger af de øvrige variable i datasættet.

2. Reducer modellen med henblik på at undersøge hvilke variable i datasættet, som har betydning for patienternes smerteopfattelse over de første 3 døgn efter operationen. Angiv parameterestimater for samtlige parametre i slutmodellen, og forklar i ord, hvad modellen udtrykker.

Uanset hvilken slutmodel du nåede frem til i delspørgsmål 2. bedes du benytte resultaterne fra modellen `m3` fra R-udskriften ved besvarelse af følgende delspørgsmål.

3. Angiv et estimat og et 95 %-konfidensinterval for den forventede smertepåvirkning efter operationen for en 50 årig mand, som før operationen havde en smerte på 20.

```
> ### Nogle statistiske modeller og test:
> m1<-lm(smerte~s0+koen*alder,data2)
> m2<-lm(smerte~koen*alder,data2)
> m3<-lm(smerte~s0+koen+alder,data2)
> m4<-lm(smerte~s0+koen,data2)
> m5<-lm(smerte~s0+alder,data2)
> m6<-lm(smerte~alder,data2)
> m7<-lm(smerte~s0,data2)
> m8<-lm(smerte~koen+alder,data2)
> m9<-lm(smerte~koen,data2)
> m10<-lm(smerte~1,data2)

> anova(m2,m1)

Analysis of Variance Table

Model 1: smerte ~ koen * alder
Model 2: smerte ~ s0 + koen * alder
  Res.Df   RSS Df Sum of Sq      F  Pr(>F)
1     44 17573
2     43 16295  1    1278.3 3.3732 0.07318 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> anova(m3,m1)

Analysis of Variance Table

Model 1: smerte ~ s0 + koen + alder
Model 2: smerte ~ s0 + koen * alder
  Res.Df   RSS Df Sum of Sq     F Pr(>F)
1     44 16399
2     43 16295  1    103.83 0.274 0.6034

> anova(m4,m1)

Analysis of Variance Table

Model 1: smerte ~ s0 + koen
Model 2: smerte ~ s0 + koen * alder
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     45 16703
2     43 16295  2    408.42 0.5389 0.5873
```

```
> anova(m4,m3)

Analysis of Variance Table

Model 1: smerte ~ s0 + koen
Model 2: smerte ~ s0 + koen + alder
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     45 16703
2     44 16399  1    304.59 0.8173 0.3709

> anova(m5,m3)

Analysis of Variance Table

Model 1: smerte ~ s0 + alder
Model 2: smerte ~ s0 + koen + alder
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1     45 21290
2     44 16399  1    4891.7 13.125 0.0007506 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> anova(m6,m5)

Analysis of Variance Table

Model 1: smerte ~ alder
Model 2: smerte ~ s0 + alder
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     46 22475
2     45 21290  1    1184.6 2.5037 0.1206

> anova(m7,m5)

Analysis of Variance Table

Model 1: smerte ~ s0
Model 2: smerte ~ s0 + alder
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     46 21633
2     45 21290  1    342.58 0.7241 0.3993
```

```
> anova(m8,m2)

Analysis of Variance Table

Model 1: smerte ~ koen + alder
Model 2: smerte ~ koen * alder
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     45 17642
2     44 17573  1    68.617 0.1718 0.6805

> anova(m8,m3)

Analysis of Variance Table

Model 1: smerte ~ koen + alder
Model 2: smerte ~ s0 + koen + alder
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     45 17642
2     44 16399  1    1243.1 3.3353 0.0746 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> anova(m6,m8)

Analysis of Variance Table

Model 1: smerte ~ alder
Model 2: smerte ~ koen + alder
  Res.Df   RSS Df Sum of Sq      F   Pr(>F)
1     46 22475
2     45 17642  1    4833.2 12.329 0.001027 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> anova(m9,m8)

Analysis of Variance Table

Model 1: smerte ~ koen
Model 2: smerte ~ koen + alder
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     46 17972
2     45 17642  1    330.67 0.8435 0.3633
```

```
> anova(m10,m9)

Analysis of Variance Table

Model 1: smerte ~ 1
Model 2: smerte ~ koen
  Res.Df   RSS Df Sum of Sq     F    Pr(>F)
1     47 22844
2     46 17972  1    4871.9 12.47 0.0009529 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> anova(m10,m6)

Analysis of Variance Table

Model 1: smerte ~ 1
Model 2: smerte ~ alder
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     47 22844
2     46 22475  1     369.3 0.7559 0.3891

> ### dele af summary() på udvalgte modeller:
> summary(m1)

               Estimate Std. Error    t value   Pr(>|t|)
(Intercept) 49.94995081 17.1978239  2.9044344 0.00578841
s0           0.36394499  0.1981587  1.8366335 0.07318287
koenm       -8.02852169 23.9110048 -0.3357668 0.73867957
alder       -0.08227208  0.3151942 -0.2610203 0.79532294
koenm:alder -0.23296231  0.4450566 -0.5234442 0.60335253

Residual standard error: 19.47 on 43 degrees of freedom

> summary(m2)

               Estimate Std. Error    t value    Pr(>|t|)
(Intercept)  56.2196892 17.3042138  3.2489017 0.002222337
koenm       -10.1943821 24.5175072 -0.4158001 0.679577305
alder        -0.1127082  0.3231353 -0.3487957 0.728906286
koenm:alder  -0.1891101  0.4562435 -0.4144937 0.680526498

Residual standard error: 19.98 on 44 degrees of freedom
```

```
> summary(m3)

               Estimate Std. Error   t value     Pr(>|t|)
(Intercept)  56.1098314 12.4372287  4.511442 0.0000474938
s0            0.3583803  0.1962341  1.826290 0.0745965690
koenm       -20.1939044  5.5739901 -3.622881 0.0007506051
alder        -0.1992598  0.2204128 -0.904030 0.3709027571

Residual standard error: 19.31 on 44 degrees of freedom

> ### estimable anvendt på modellen m3:
> library(gmodels)

> est1<-c(0,20,0,50)
> est2<-c(0,20,-1,50)
> est3<-c(0,20,1,50)
> est4<-c(1,20,1,50)
> est5<-c(1,20,-1,50)
> est6<-c(1,1,1,1)
> est<-rbind(est1,est2,est3,est4,est5,est6)

> estimable(m3,est,conf.int=0.95)

     Estimate Std. Error t value DF Pr(>|t|) Lower.CI Upper.CI
est1  -2.7954    11.7747 -0.2374 44   0.8134 -26.5257  20.9350
est2  17.3985    13.1184  1.3263 44   0.1916  -9.0398  43.8368
est3 -22.9893    12.9358 -1.7772 44   0.0825 -49.0596   3.0811
est4  33.1205     4.1904  7.9039 44   0.0000  24.6753  41.5658
est5  73.5084     8.9387  8.2236 44   0.0000  55.4936  91.5231
est6  36.0750    12.2810  2.9375 44   0.0052  11.3243  60.8258

> ### dele af summary() på udvalgte modeller:
> summary(m4)

               Estimate Std. Error   t value     Pr(>|t|)
(Intercept)  45.6928684  4.6712528  9.781716 1.033464e-12
s0            0.3620426  0.1957936  1.849104 7.101836e-02
koenm       -20.2698475  5.5620298 -3.644326 6.915203e-04

Residual standard error: 19.27 on 45 degrees of freedom
```

```
> summary(m5)

             Estimate Std. Error    t value    Pr(>|t|)
(Intercept) 46.7530782 13.7075509  3.4107536 0.001378087
s0           0.3498198  0.2210807  1.5823171 0.120581050
alder       -0.2112944  0.2483107 -0.8509274 0.399315044

Residual standard error: 21.75 on 45 degrees of freedom

> summary(m6)

             Estimate Std. Error    t value     Pr(>|t|)
(Intercept) 51.7353007 13.5573200  3.8160419 0.0004031086
alder       -0.2193356  0.2522838 -0.8694004 0.3891420957

Residual standard error: 22.1 on 46 degrees of freedom

> summary(m7)

            Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 35.66714  4.2494004 8.393452 7.831088e-11
s0           0.35367  0.2203705 1.604888 1.153629e-01

Residual standard error: 21.69 on 46 degrees of freedom

> summary(m8)

              Estimate Std. Error    t value     Pr(>|t|)
(Intercept)  61.1564396 12.4370436  4.9172811 1.210994e-05
koenm       -20.0713280  5.7163810 -3.5111949 1.027060e-03
alder        -0.2075697  0.2260116 -0.9184028 3.633046e-01

Residual standard error: 19.8 on 45 degrees of freedom

> summary(m9)

             Estimate Std. Error   t value     Pr(>|t|)
(Intercept)  50.35417   4.034765 12.480075 2.261338e-16
koenm       -20.14917   5.706019 -3.531213 9.529455e-04

Residual standard error: 19.77 on 46 degrees of freedom

> summary(m10)

            Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 40.27958   3.182137 12.65803 9.435106e-17

Residual standard error: 22.05 on 47 degrees of freedom
```

Opgave 3 (4 spørgsmål)

I forbindelse med rehabilitering af patienter, som har gennemgået et langt sygdomsforløb, ønsker man at undersøge effekten af et træningsprogram på patienternes kredsløbsfunktion. I denne opgave ser vi på et datasæt fra 135 patienter, som er blevet randomiseret til to forskellige træningsprogrammer givet ved faktoren `gruppe` med niveauerne `I` (intervention) og `K` (kontrol). Hver patients maksimale iltoptagelse (`vo2max`) er blevet målt tre gange (-efter 1 måned, efter 7 måneder og efter 13 måneder), og variablen `tid` i datasættet angiver tidspunktet for målingen. I datasættet indgår desuden information om køn (`sex=M` eller `sex=F`) samt alder (`age`) målt i år ved forsøgsperiodens start. Variablen `patient` er en kode, der angiver hvilken patient de enkelte målinger stammer fra.

Data til opgaven stammer fra *The Copenhagen PACT Study* og er venligst stillet til rådighed af Julie Midtgaard. Data er udleveret på vedlagte USB-nøgle under filnavnet `data3.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. starte med at indlæse data i `R` med kommandoen

```
> data3<-read.table(file.choose(),header=T)
```

hvorefter du vælger filen `data3.txt` fra USB-nøglen.

Et udpluk af datasættet er vist nedenfor

```
    patient gruppe tid sex age vo2max
1 PI1MP0206      I   1   K  49  2.223
2 PI2MM0960      I   1   K  51  2.250
3 PI4EN0588      I   1   K  44  2.543
4 PI5HK0582      I   1   K  55  1.919

[ ... flere datalinjer her ...]

        patient gruppe tid sex age vo2max
403  PK99DS0381      K  13   M  67  1.841
404 PK103RS1574      K  13   K  40  2.746
405 PK105MG2224      K  13   K  46  2.594
```

Nedenfor er angivet R-kode som fitter otte forskellige statistiske modeller til datasættet fra filen `data3.txt`.

```
> library(nlme)
> modelA<-lme(log(vo2max)~gruppe*sex*tid+age,random=~1|patient
+ ,data3,corr=corGaus(form=~tid|patient,nugget=T))
> modelB<-lme(log(vo2max)~gruppe*sex*factor(tid)+age,random=~1|patient
+ ,data3,corr=corGaus(form=~tid|patient,nugget=T))
> modelC<-lm(log(vo2max)~gruppe*sex*factor(tid)+age+factor(patient),data3)
> modelD<-lme(log(vo2max)~gruppe*sex*factor(tid)+age,random=~1|patient
+ ,data3)
```

```
> modelE<-lme(vo2max~gruppe*sex*tid+age,random=~1|patient
+ ,data3,corr=corGaus(form=~tid|patient,nugget=T))
> modelF<-lme(vo2max~gruppe*sex*factor(tid)+age,random=~1|patient
+ ,data3,corr=corGaus(form=~tid|patient,nugget=T))
> modelG<-lm(vo2max~gruppe*sex*factor(tid)+age+factor(patient),data3)
> modelH<-lme(vo2max~gruppe*sex*factor(tid)+age,random=~1|patient
+ ,data3)
```

1. Angiv en af modellerne `modelA-modelH`, som med rimelighed kan benyttes som udgangspunkt for en statistisk analyse af datasættet. Ved besvarelsen af dette delspørgsmål er det væsentligt, at du argumenterer grundigt for dit valg af model. I den forbindelse er det nødvendigt, at du fitter nogle af modellerne i R og i din besvarelse begrunder, hvordan du bruger R-kørslen til at vælge mellem modellerne.

   Bemærk: Der kan være flere korrekte og næsten lige gode svar på dette spørgsmål.

2. Med udgangspunkt i dit valg af statistisk model fra 1. bedes du foretage en statistisk analyse (modelreduktion) med henblik på at undersøge, hvordan om udviklingen i `vo2max` afhænger af `tid, gruppe, sex` og `age`. Undervejs skal du tydeligt gøre rede for, hvilke modeller du tester imod hinanden, ligesom du bedes angive teststørrelser og p-værdier svarende til de enkelte test, som du foretager. Sørg for tydeligt at angive din slutmodel.

3. Benyt din slutmodel fra 2. til at angive et estimat og et 95 %-konfidensinterval for den forventede værdi af `vo2max` 13 måneder (`tid=13`) efter forsøgsperiodens start for en 50-årig mand (`age=50,sex=M`) i interventionsgruppen (`gruppe=I`).

4. Er der på baggrund af den statistiske analyse af datasættet belæg for at hævde, at patienternes kredsløbsfunktion (`vo2max`) forbedres fra 7 til 13 måneder efter forsøgsperiodens start?

   Hint: Ved besvarelsen af dette delspørgsmål kan du forsøge at kvantificere ændringerne fra 7 til 13 måneder for passende valg af de øvrige variable i modellen. Alternativt kan du udføre et relevant test i `R` f.eks. ved at tilføje en ny tidsfaktor med kun 2 niveauer (`1` og `7:13`) til datasættet. Dette kan du gøre ved at tilføje følgende fire datalinjer i starten af dit R-program.

   ```
   > data3$tidny<-factor(data3$tid) ### laver faktor version af tid
   > levels(data3$tidny)            ### nytid har 3 niveauer: 1,7,13

   [1] "1"  "7"  "13"

   > levels(data3$tidny)<-c("1","7:13","7:13") ### slår 2 af niveauerne sammen
   > levels(data3$tidny) ### nytid har nu kun 2 niveauer: 1, 7:13

   [1] "1"    "7:13"
   ```