Eksamen i Statistisk Dataanalyse 2
(NMAB14002U)

9. april 2015

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 9 sider med i alt 3 opgaver, der indgår med vægtningen 40 %, 25 % og 35 % 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 3 er der vedlagt noget R-udskrift, 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 (5 spørgsmål)

Ved brug af såkaldte *in-growth cores* er det muligt at måle tørvægten af rødder under jorden. Over en periode fra januar 2009 til juni 2010 har man ved hjælp af 96 in-growth cores (`incore`) målt tørvægten (`DW`) af nytilkomne rødder. Fra hver af de 96 in-growth cores har man ved at placere et lille stykke stof i en *mesh bag* været i stand til at måle tørvægten både i overfladen (`depth=0hor`) og i 0-5 cm dybde (`depth=0-5cm`), således at der totalt er foretaget 192 målinger af tørvægten af rødder (2 målinger per in-growth core).

De 96 in-growth cores var placeret i 12 forskellige klimakamre (`octagon`), således at der var præcis 8 in-growth cores i hvert kammer. I halvdelen af klimakamrene sørgede man for at koncentrationen af CO_2 fra solnedgang til solopgang blev holdt på et kunstigt højt niveau (510 ppm). Niveauet af CO_2 er givet ved faktoren `co2` med niveauerne `0=normal` og `1=forhøjet`. Desuden sørgede man for, at de 8 in-growth cores inden for hvert klimakammer (`octagon`) modtog en af 4 klimabehandlinger svarende til kombinationer af de to faktorer `temp` (temperatur: med niveauerne `0=normal` og `1=forhøjet`) og `drought` (tørke: med niveauerne `0=normal` og `1=tørke`). Forsøget var balanceret i den forstand, at der var lige mange in-growth cores, som modtog hver af de 8 behandlingskombinationer af faktorerne `co2, temp` og `drought`.

Data til denne opgave stammer fra CLIMAITE forsøgsopstillingen og er venligst stillet til rådighed af Marie Frost Arndal. Datasættet til opgaven findes på den udleverede USB-nøgle under filnavnet `data1.txt`. For at besvare opgaven vil det være nødvendigt at køre diverse R-kommandoer på din egen medbragte computer. Du kan f.eks. indlæse

data i `R` med kommandoen

```
data1<-read.table(file.choose(),header=T,sep="\t")
```

hvorefter du vælger filen `data1.txt` fra USB-nøglen. De første linjer i datasættet bør se således ud

```
head(data1,18)
```

```
##    octagon incore co2 temp drought depth       DW
## 1        1      1   0    0       0   0-5 15.34994
## 2        1      1   0    0       0  Ohor 14.76958
## 3        1      2   0    0       0   0-5 34.86580
## 4        1      2   0    0       0  Ohor 33.61352
## 5        1      3   0    0       1   0-5 20.18645
## 6        1      3   0    0       1  Ohor 21.39042
## 7        1      4   0    0       1   0-5 26.37573
## 8        1      4   0    0       1  Ohor 10.69521
## 9        1      5   0    1       1   0-5 21.25855
## 10       1      5   0    1       1  Ohor 14.26028
## 11       1      6   0    1       1   0-5 28.51257
## 12       1      6   0    1       1  Ohor 11.20451
## 13       1      7   0    1       0   0-5 42.91128
## 14       1      7   0    1       0  Ohor 32.08564
## 15       1      8   0    1       0   0-5 33.85831
## 16       1      8   0    1       0  Ohor 29.02986
## 17       2      9   1    0       0   0-5 35.28254
## 18       2      9   1    0       0  Ohor 17.82535
```

Formålet med forsøget er at undersøge, hvordan tørvægten af rødder (`DW`) i forskellige jorddybder afhænger af klimafaktorerne.

1. Opskriv en statistisk model du vil benytte som udgangspunkt for en statistisk analyse af tørvægt (`DW`). Vekselvirkninger med eventuelle tilfældige effekter ønskes ikke inddraget i modellen. Det er *ikke* en del af opgaven at lave modelkontrol (f.eks. skal du ikke transformere `DW`).

2. Reducer den statistiske model fra 1. med henblik på at undersøge, hvordan tørvægten af rodmateriale (`DW`) i forskellig dybde afhænger af klimafaktorerne. Undervejs bedes 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. På baggrund af din besvarelse skal det være klart, hvilken slutmodel du når frem til.

   Hint: Der er flere mulige løsninger på dette delspørgsmål. For at spare tid, kan du f.eks. starte med at udføre et test, der fjerner effekten af tørke (`drought`) helt fra modellen.

3. Benyt din slutmodel fra delspørgsmål 2. til at give et estimat for den forventede tørvægt (`DW`) i dybden 0-5 cm for en kontrolprøve, hvor `temp=co2=drought=0`. Angiv desuden variansestimaterne fra din slutmodel.

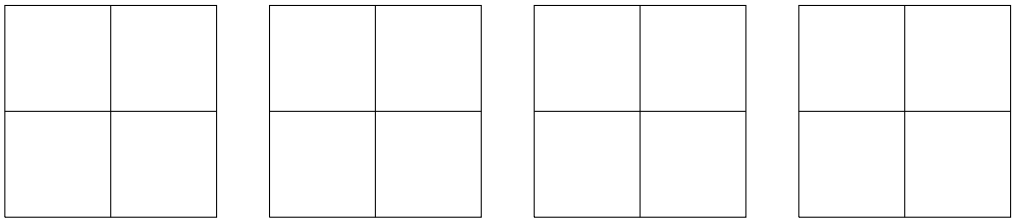
4. Benyt din slutmodel fra delspørgsmål 2. til at give et estimat og et 95 %-konfidensinterval for forskellen i den forventede tørvægt (`DW`) i overfladen (`depth=0hor`) for en kontrolprøve (`temp=co2=drought=0`) og en prøve, hvor alle klimafaktorer er modificerede (dvs. `temp=co2=drought=1`).

Lad os nu forestille os, at man fra hver in-growth core havde mulighed for at bestemme tørvægten af rødder i hhv. 5, 15, 25 og 35 cm dybde.

5. Opskriv (på papir) en statistisk model for det nye datasæt, der beskriver at der er en lineær sammenhæng mellem tørvægt af rødder (`DW`) og dybde (`depth`). Angiv desuden en linjes `R`-kode, som du ville bruge til at fitte modellen (-du kan naturligvis ikke fitte modellen i praksis!). Ved besvarelsen af delspørgsmål 5. behøver du ikke inkludere faktorerne `co2, temp` og `drought` i din statistiske model.

Opgave 2 (3 spørgsmål)

Ved et dyrkningsforsøg ønsker man at afprøve 4 behandlinger givet som kombinationer af 2 faktorer (`A` og `B`) med hver to niveauer (`1` og `2`). Forsøget udføres på 16 jordlodder, der er organiseret på 4 marker som vist på tegningen nedenfor.



1. Giv et forslag til en forsøgsplan og forklar, hvordan randomiseringen bør foretages. Opstil en statistisk model til analyse af forsøget og tegn et tilhørende faktordiagram.

Forsøget udvides nu til at omfatte 8 behandlinger givet som kombinationer af 3 behandlinger (`A, B` og `C`) med hver 2 niveauer. Den randomiserede forsøgsplan er vist nedenfor

A2B2C2	A1B1C1
A1B1C2	A2B2C1

A1B2C1	A2B1C2
A1B2C2	A2B1C1

A1B2C2	A2B1C1
A1B2C1	A2B1C2

A2B2C1	A1B1C2
A2B2C2	A1B1C1

2. Forklar hvilken type forsøg der er tale om og giv et forslag til en alternativ forsøgsplan. Argumentér kort for hvorfor der kunne være en fordel ved at anvende den alternative forsøgsplan.

Man beslutter sig i sidste ende for at udføre forsøget med to behandlingsfaktorer (`A` og `B`) med hhv. 2 og 5 niveauer (dvs. i alt 10 behandlinger). Samtidig udvides forsøget til at omfatte 10 marker.

3. Afgør om forsøget kan udføres som et balanceret ufuldstændigt blokforsøg.

Opgave 3 (4 spørgsmål)

For at sammenligne effekten af to forskellige træningsprogrammer er 147 patienter ved lodtrækning blevet inddelt i to grupper givet ved faktoren `gruppe` med niveauerne `I` (=intervention) og `K` (=kontrol). Patienternes muskelstyrke i benene er blevet målt ved forsøgets start (`benpres.0`) og efter 6 måneder (`benpres.6`). I datasættet findes desuden informationer om patienternes køn (`sex=K` eller `sex=M`) samt alder i år (`age`) ved forsøgets start.

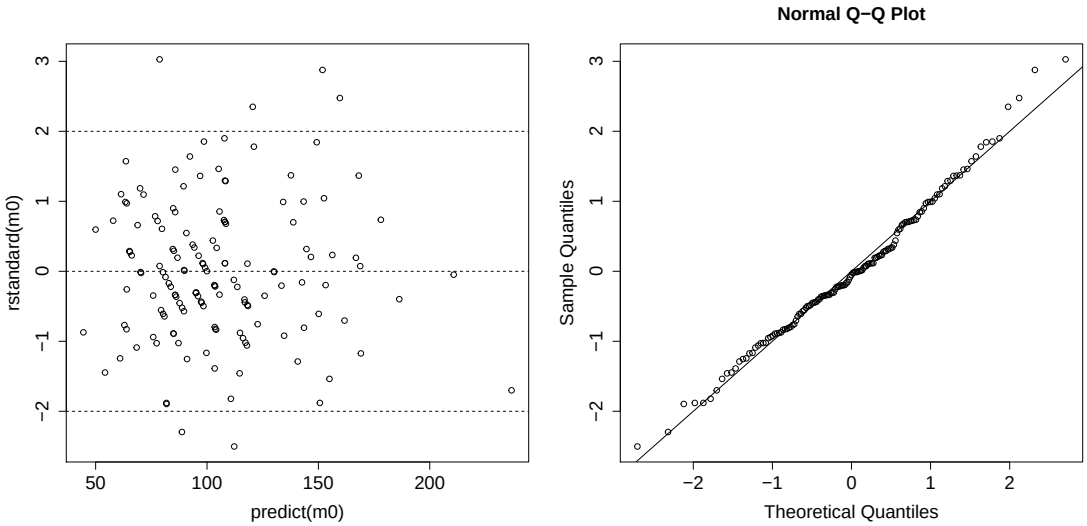
Data til opgaven stammer fra *The Copenhagen PACT Study* og er venligst stillet til rådighed af Julie Midtgaard. De første linjer af datasættet ses nedenfor.

```
dim(fulldata3)

## [1] 147   5

head(fulldata3,6)

##   gruppe sex age benpres.0 benpres.6
## 1      I   K  43       100       110
## 2      I   K  44        60        80
## 4      I   K  31        60       110
## 5      I   K  35        60        70
## 6      I   M  64       120       130
## 7      I   K  34        80       110
```

De følgende linjers `R`-kode fitter en statistisk model til datasættet og optegner nogle figurer.

```
m0<-lm(benpres.6~benpres.0+age+gruppe*sex-1,data=fulldata3)
```

```
plot(predict(m0),rstandard(m0))
abline(h=c(-2,0,2),lty=2)
qqnorm(rstandard(m0))
abline(0,1)
```



Besvar følgende 4 delspørgsmål ved brug af R-udskriften ovenfor samt sidst i opgavesættet. Bemærk, at der kan være dele af R-udskriften, som ikke skal benyttes.

1. Opskriv den statistiske model svarende til modellen `m0` som er fittet ved hjælp af R-programkoden ovenfor. Diskuter om modellen giver en god beskrivelse af variationen i data.

2. Tag udgangspunkt i parameterestimaterne fra modellen `m0`. Angiv den forventede muskelstyrke efter 6 måneder for en 50 årig kvinde i kontrolgruppen, der ved forsøgets start havde en muskelstyrke på 100 (dvs `benpres.0=100`). Angiv det tilsvarende estimat for en 50 årig mand.

3. Foretag modelreduktion med henblik på at undersøge, hvordan muskelstyrken i benene efter 6 måneder afhænger af de øvrige variable i datasættet. Angiv samtlige parameterestimater i slutmodellen.

4. Tag udgangspunkt i slutmodellen fra din analyse i delspørgsmål 3. Find den forventede muskelstyrke i benene for 50-årige mænd i interventionsgruppen, der ved forsøgets start kunne tage 100 kg i benpres. Angiv både et estimat og et 95 % - konfidensinterval.

```r
### nogle statistiske modeller og test:
m1<-lm(benpres.6~benpres.0+gruppe*sex-1,data=fulldata3)
m2<-lm(benpres.6~benpres.0+age+gruppe+sex-1,data=fulldata3)
m3<-lm(benpres.6~benpres.0+gruppe+sex,data=fulldata3)
```

```r
anova(m1,m0)
```

```
## Analysis of Variance Table
##
## Model 1: benpres.6 ~ benpres.0 + gruppe * sex - 1
## Model 2: benpres.6 ~ benpres.0 + age + gruppe * sex - 1
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1    142 41822
## 2    141 41135  1    687.37 2.3561  0.127
```

```r
anova(m2,m0)
```

```
## Analysis of Variance Table
##
## Model 1: benpres.6 ~ benpres.0 + age + gruppe + sex - 1
## Model 2: benpres.6 ~ benpres.0 + age + gruppe * sex - 1
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1    142 42311
## 2    141 41135  1    1176.2 4.0319 0.04656 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(m3,m1)
```

```
## Analysis of Variance Table
##
## Model 1: benpres.6 ~ benpres.0 + gruppe + sex
## Model 2: benpres.6 ~ benpres.0 + gruppe * sex - 1
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1    143 42781
## 2    142 41822  1    959.07 3.2564 0.07327 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(m3,m2)
```

```
## Analysis of Variance Table
##
## Model 1: benpres.6 ~ benpres.0 + gruppe + sex
## Model 2: benpres.6 ~ benpres.0 + age + gruppe + sex - 1
##   Res.Df   RSS Df Sum of Sq     F Pr(>F)
## 1    143 42781
## 2    142 42311  1     470.2 1.578 0.2111
```

```r
### dele af summary() paa udvalgte modeller:

summary(m0)
```

```
##
## Call:
## lm(formula = benpres.6 ~ benpres.0 + age + gruppe * sex - 1,
##     data = fulldata3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.217 -10.583  -0.755  11.379  51.275
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## benpres.0      0.8761     0.0476  18.408  < 2e-16 ***
## age           -0.2221     0.1447  -1.535  0.12703
## gruppeI       40.0870     9.0062   4.451 1.72e-05 ***
## gruppeK       27.6090     8.7426   3.158  0.00194 **
## sexM          23.9697     5.6327   4.255 3.78e-05 ***
## gruppeK:sexM -16.5259     8.2302  -2.008  0.04656 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.08 on 141 degrees of freedom
## Multiple R-squared:  0.9773,Adjusted R-squared:  0.9763
## F-statistic:  1012 on 6 and 141 DF,  p-value: < 2.2e-16
```

```r
summary(m1)
```

```
##
## Call:
## lm(formula = benpres.6 ~ benpres.0 + gruppe * sex - 1, data = fulldata3)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -44.11 -10.59  -0.54  10.44  51.53
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## benpres.0      0.89106    0.04682  19.033  < 2e-16 ***
## gruppeI       28.16587    4.58197   6.147 7.55e-09 ***
## gruppeK       16.09460    4.51169   3.567 0.000492 ***
## sexM          22.60479    5.58861   4.045 8.56e-05 ***
## gruppeK:sexM -14.77926    8.19004  -1.805 0.073266 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.16 on 142 degrees of freedom
## Multiple R-squared:  0.9769,Adjusted R-squared:  0.9761
## F-statistic:  1202 on 5 and 142 DF,  p-value: < 2.2e-16
```

```r
summary(m2)
```

```
##
## Call:
## lm(formula = benpres.6 ~ benpres.0 + age + gruppe + sex - 1,
##     data = fulldata3)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -41.30 -12.46  -1.10  10.40  52.29
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## benpres.0  0.87137    0.04804  18.138  < 2e-16 ***
## age       -0.18190    0.14480  -1.256 0.211105
## gruppeI   39.75816    9.10036   4.369 2.39e-05 ***
## gruppeK   25.08521    8.74368   2.869 0.004748 **
## sexM      16.81858    4.41030   3.813 0.000204 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.26 on 142 degrees of freedom
## Multiple R-squared:  0.9766,Adjusted R-squared:  0.9758
## F-statistic:  1188 on 5 and 142 DF,  p-value: < 2.2e-16
```

```r
summary(m3)
```

```
##
## Call:
## lm(formula = benpres.6 ~ benpres.0 + gruppe + sex, data = fulldata3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.958 -11.447  -0.571  10.341  52.411
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  29.83186    4.52327   6.595 7.65e-10 ***
## benpres.0     0.88424    0.04703  18.801  < 2e-16 ***
## gruppeK     -14.13957    2.86098  -4.942 2.13e-06 ***
## sexM         16.30991    4.40054   3.706    3e-04 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.3 on 143 degrees of freedom
## Multiple R-squared:  0.7823,Adjusted R-squared:  0.7777
## F-statistic: 171.2 on 3 and 143 DF,  p-value: < 2.2e-16
```

```r
### estimable anvendt paa udvalgte modeller:

library(gmodels)
est1<-c(100,50,1,0,0,0)
est2<-c(100,50,1,0,1,0)
est3<-c(100,50,1,0,1,1)
estA<-rbind(est1,est2,est3)
estimable(m0,estA,conf.int=0.95)
```

```
##      Estimate Std. Error  t value  DF Pr(>|t|) Lower.CI Upper.CI
## est1 116.5993   2.345277 49.71663 141        0 111.9628 121.2357
## est2 140.5690   4.995047 28.14167 141        0 130.6941 150.4438
## est3 124.0431   6.626266 18.71990 141        0 110.9434 137.1427
```

```r
est4<-c(0,100,0,1)
est5<-c(1,100,0,1)
est6<-c(100,100,0,1)
estB<-rbind(est4,est5,est6)
estimable(m3,estB,conf.int=0.95)
```

```
##       Estimate Std. Error   t value  DF     Pr(>|t|)   Lower.CI Upper.CI
## est4  104.7337   5.110204 20.495007 143 0.000000e+00   94.63238  114.835
## est5  134.5655   4.071676 33.049173 143 0.000000e+00  126.51709  142.614
## est6 3087.9199 449.028139  6.876896 143 1.758753e-10 2200.32951 3975.510
```