Eksamen i Statistisk Dataanalyse 2
(kursusnr.: 210006)

10. april 2008

Alle sædvanlige hjælpemidler, herunder bøger og lommeregner men *ikke* PC, er tilladt. Opgavesættet består af 10 sider med i alt 3 opgaver, der alle ønskes besvaret. Til nogle 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 (5 spørgsmål)

I et nordengelsk studie udvalgte man tilfældigt 50 jordstykker, hver på 1 hektar. På hvert jordstykke estimerede man den totale biomasse af vegetationen. Desuden registrerede man jordtypen og inddelte den i tre kategorier (kalkholdig, ler, muld) samt hvorvidt området var naturfredet eller ej. Endelig blev det noteret, hvor højt over havets overflade jordstykket var beliggende (målt i meter).

Data er gengivet i tabellen nedenfor. Bemærk at der ikke er lige mange jordstykker for de forskellige kombinationer af jordtype og fredningsstatus. Datasættet `biomassedata` er indlæst i R og indeholder variablene `jord`, `fredet`, `hojde` og `biomasse`.

Jordtype	Kalkholdig		Ler		Muld	
	højde	biomasse	højde	biomasse	højde	biomasse
Naturfredet område	82	2.11	130	1.70	91	1.81
	161	1.91	359	0.98	110	1.77
	153	1.72	209	1.67	248	1.53
	67	2.09	146	1.79	40	2.14
			331	1.36	338	1.05
			482	0.77	171	1.42
					107	1.65
Normalt område	178	1.68	116	2.01	21	2.06
	79	2.23	65	2.07	86	1.75
	84	2.19	117	1.88	237	1.39
	25	2.21	5	2.14	122	1.73
	146	1.96	64	2.15	277	1.33
	67	2.07	70	2.04	239	1.44
	118	2.01	161	1.74	206	1.64
	42	2.20	23	2.15	371	0.92
	57	2.24	7	2.29	236	1.37
	56	2.25	237	1.52		
	42	2.23	240	1.57		
	68	2.03	436	1.18		

I R-kørslen nedenfor er modellerne `model1`–`model5` fittet, og der er kørt diverse `anova`-, `summary`- og `estimable`-kommandoer.

1. Specificér den statistiske model svarende til `model1` i R-kørslen nedenfor.
2. Brug R-kørslen til at reducere modellen mest muligt. Husk undervejs at specificere hvilke hypoteser du tester, og hvilke modeller der indgår i analysen.
3. Hvad er konklusionen vedrørende de forklarende variables effekt på biomassen? Angiv herunder relevante estimater.
4. Angiv et estimat og et 95%-konfidensinterval for den forventede biomasse på et fredet jordstykke med muldjord beliggende 150 m over havets overflade.
5. Specificér en model hvor effekten af højden over havet tillades at være forskellig for de tre jordtyper. Angiv desuden en R-kommando der fitter modellen.

Udskrift af R-kørsel (lettere redigeret):

```
> biomassedata
   fredet jord hojde biomasse
1  normal  ler   116     2.01
2  normal muld    21     2.06
3  fredet  ler   130     1.70
4  normal  ler    65     2.07
5  normal  ler   117     1.88
6  fredet kalk    82     2.11
.
.
.
49 normal kalk    68     2.03
50 normal  ler   436     1.18

> attach(biomassedata)

### Modeller:

> model1 = lm(biomasse ~ hojde + fredet + jord + fredet:jord)
> model2 = lm(biomasse ~ hojde + jord + fredet)
> model3 = lm(biomasse ~ hojde + jord)
> model4 = lm(biomasse ~ hojde)
> model5 = lm(biomasse ~ jord)

### anova-kommandoer:

> anova(model2,model1)
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     45 0.39997
2     43 0.38664  2   0.01333 0.7415 0.4824
```

```
> anova(model3,model2)
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     46 0.42485
2     45 0.39997  1   0.02488 2.7988 0.1013

> anova(model4,model3)
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     48 0.84201
2     46 0.42485  2   0.41716 22.584 1.469e-07 ***

> anova(model5,model3)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     47 5.2254
2     46 0.4248  1    4.8005 519.77 < 2.2e-16 ***

### Summary-kommandoer:

> summary(model1)
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         2.286616   0.050143  45.602  < 2e-16 ***
hojde              -0.002843   0.000141 -20.162  < 2e-16 ***
fredetnej           0.049658   0.054976   0.903 0.371418
jordler            -0.123049   0.065255  -1.886 0.066111 .
jordmuld           -0.213490   0.059730  -3.574 0.000883 ***
fredetnej:jordler   0.046906   0.074130   0.633 0.530247
fredetnej:jordmuld -0.041253   0.073479  -0.561 0.577426

> summary(model2)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.2928629  0.0357850  64.073  < 2e-16 ***
hojde       -0.0029068  0.0001302 -22.318  < 2e-16 ***
jordler     -0.0862220  0.0342955  -2.514   0.0156 *
jordmuld    -0.2309939  0.0354480  -6.516 5.33e-08 ***
fredetnej    0.0488634  0.0292075   1.673   0.1013

> summary(model3)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.3337354  0.0266539  87.557  < 2e-16 ***
hojde       -0.0029542  0.0001296 -22.798  < 2e-16 ***
jordler     -0.0860908  0.0349595  -2.463   0.0176 *
jordmuld    -0.2357826  0.0360164  -6.547 4.37e-08 ***

> summary(model4)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.2645782  0.0311841   72.62   <2e-16 ***
hojde       -0.0032020  0.0001657  -19.32   <2e-16 ***

> summary(model5)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.07063    0.08336  24.840  < 2e-16 ***
jordler     -0.34785    0.11457  -3.036   0.0039 **
jordmuld    -0.50812    0.11789  -4.310 8.27e-05 ***
```

Estimable-kommandoer:

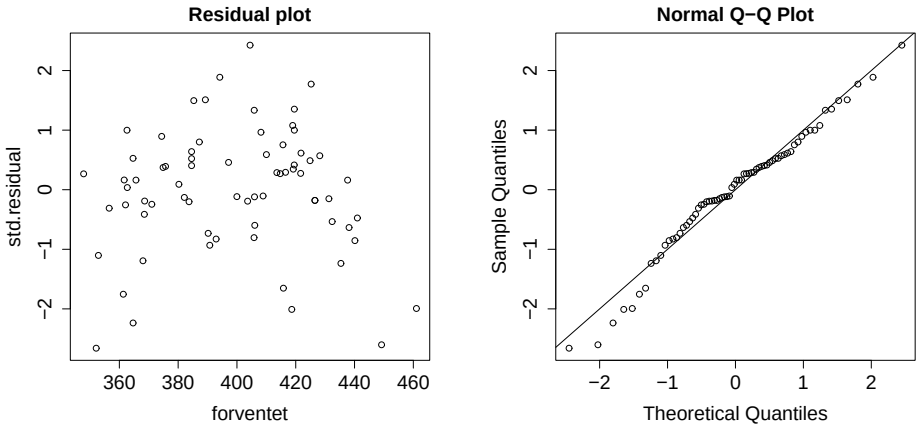
```
> y1 = c(1,1.50,0,1,1)
> y2 = c(1,150,1,1,0)
> y3 = c(1,150,0,1,1)

> forventet2 = rbind(y1, y2, y3)
> estimable(model2, forventet2, conf.int=0.95)
   Estimate Std. Error  t value DF Pr(>|t|) Lower.CI Upper.CI
y1 2.106372 0.03371385 62.47794 45        0 2.038469 2.174275
y2 1.539629 0.04625472 33.28589 45        0 1.446468 1.632791
y3 1.674715 0.02669686 62.73078 45        0 1.620945 1.728485

> x1 = c(1,1.50,0,1)
> x2 = c(1,150,1,1)
> x3 = c(1,150,0,1)

> forventet3 = rbind(x1, x2, x3)
> estimable(model3, forventet3, conf.int=0.95)
   Estimate Std. Error  t value DF Pr(>|t|) Lower.CI Upper.CI
x1 2.093521 0.03346282 62.56260 46        0 2.026164 2.160879
x2 1.568729 0.04368971 35.90614 46        0 1.480786 1.656671
x3 1.654819 0.02436471 67.91870 46        0 1.605776 1.703863
```

Opgave 2 (4 spørgsmål)

For at undersøge sammenhængen mellem puls og omgangstid på løbeture i Parc Montsouris er indsamlet data fra 14 løbeture hver bestående af 5 omgange à 1460 m. Desuden registrerede man, om turene fandt sted om morgenen eller ej. I R-kørslen nedenfor er data indlæst i R i datasættet `montsouris` med variablene `dag`, `morgen`, `puls` (målt i slag per minut) og `tid` (målt i sekunder).

dag	morgen	puls	tid	puls	tid	puls	tid	puls	tid	puls	tid
1	ja	143	445	156	431	156	428	165	383	163	401
2	ja	154	429	168	425	175	381	177	379	183	354
3	nej	160	416	168	390	168	388	168	389	179	331
4	ja	148	437	152	433	154	427	159	399	167	381
5	ja	149	433	155	428	164	384	170	369	172	369
6	ja	151	428	158	415	160	402	169	378	173	360
7	nej	161	405	165	410	175	346	179	344	168	398
8	ja	149	440	152	422	155	401	162	382	169	347
9	ja	157	422	163	399	167	394	173	358	175	363
10	nej	153	404	168	382	172	371	177	350	170	365
11	nej	154	425	157	425	157	425	161	417	164	417
12	nej	161	433	164	430	168	423	169	419	170	416
13	nej	161	403	170	402	177	367	178	367	179	363
14	ja	161	439	165	429	166	424	170	408	175	386

I `R`-udskriften nedenfor er modellerne `modelA`–`modelE` samt `model0`-`model1` fittet, og der er kørt diverse `anova`-, `summary`- og `intervals`-kommandoer. Desuden er der lavet tegninger til modelkontrol af `modelA`.



Bemærk at delspørgsmålene 3 og 4 kan besvares uafhængigt af delspørgsmål 2.

1. Opskriv den statistiske model svarende til `modelA`. Benyt residualplottet og QQ-plottet ovenfor til kort at gøre rede for, om modellen giver en god beskrivelse af data.

2. Opstil hypotesen om at effekten af `dag` kan ignoreres og benyt R-udskriften til at udføre et test for hypotesen.

3. Reducér den systematiske del af modellen fra spørgsmål 1 mest muligt og angiv parameterestimaterne for alle parametrene i slutmodellen. Husk at anføre hvilke hypoteser du tester, og hvilke modeller som indgår i analysen.

4. Angiv et estimat og et konfidensinterval for den forventede forbedring i omgangstiden (målt i sekunder), hvis personen løber om morgenen og øger sin puls med 15 slag per minut.

Udskrift af R-kørsel (lettere redigeret):

```
> montsouris
   dag morgen puls tid
1    1     ja  143 445
2    1     ja  156 431
3    1     ja  156 428
4    1     ja  165 383
5    1     ja  163 401
6    2     ja  154 429
7    2     ja  168 425
.
.
.
69  14     ja  170 408
70  14     ja  175 386

> attach(montsouris)

### Modeller:

> library(nlme)
> modelA=lme(tid~morgen*puls,random=~1|dag,method="ML")
> modelB=lme(tid~morgen+puls,random=~1|dag,method="ML")
> modelC=lme(tid~morgen,random=~1|dag,method="ML")
> modelD=lme(tid~puls,random=~1|dag,method="ML")
> modelE=lme(tid~1,random=~1|dag,method="ML")
> model0=lm(tid~puls*morgen+dag)
> model1=lm(tid~puls*morgen)

### Anova-kommandoer:

> anova(model1,model0)
Analysis of Variance Table

Model 1: tid ~ puls * morgen
Model 2: tid ~ puls * morgen + dag
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     66 18359.0
2     54  4867.8 12   13491.3 12.472 1.251e-11 ***

> anova(modelB,modelA)
       Model df      AIC      BIC    logLik   Test   L.Ratio p-value
modelB     1  5 556.6793 567.9218 -273.3397
modelA     2  6 558.5207 572.0117 -273.2604 1 vs 2 0.1585538  0.6905
```

```
> anova(modelC,modelB)
       Model df      AIC      BIC    logLik   Test  L.Ratio p-value
modelC     1  4 671.2799 680.2739 -331.6400
modelB     2  5 556.6793 567.9218 -273.3396 1 vs 2 116.6006  <.0001

> anova(modelD,modelB)
       Model df      AIC      BIC    logLik   Test   L.Ratio p-value
modelD     1  4 555.2749 564.2689 -273.6375
modelB     2  5 556.6793 567.9218 -273.3397 1 vs 2 0.5956431  0.4402

> anova(modelE,modelC)
       Model df      AIC      BIC    logLik   Test  L.Ratio p-value
modelE     1  3 670.2828 677.0283 -332.1414
modelC     2  4 671.2799 680.2739 -331.6400 1 vs 2 1.002912  0.3166

> anova(modelE,modelD)
       Model df      AIC      BIC    logLik   Test  L.Ratio p-value
modelE     1  3 670.2828 677.0283 -332.1414
modelD     2  4 555.2749 564.2689 -273.6375 1 vs 2 117.0079  <.0001

### Estimater m.m.:

> summary(modelA)

Random effects:
 Formula: ~1 | dag
        (Intercept) Residual
StdDev:    14.00388 9.340028

Fixed effects: tid ~ morgen * puls
                  Value Std.Error DF    t-value p-value
(Intercept)    908.0586  32.45027 54  27.983080  0.0000
morgennej      -16.4624  58.34431 12  -0.282160  0.7826
puls            -3.1091   0.19699 54 -15.783095  0.0000
morgennej:puls   0.1362   0.34841 54   0.391032  0.6973

> intervals(modelA)
Approximate 95% confidence intervals

 Fixed effects:
                      lower        est.       upper
(Intercept)     844.8858922 908.0586069 971.2313216
morgennej      -139.8982935 -16.4624138 106.9734660
puls             -3.4926233  -3.1091299  -2.7256365
morgennej:puls   -0.5420277   0.1362389   0.8145054
```

```
> summary(modelB)

Random effects:
 Formula: ~1 | dag
        (Intercept) Residual
StdDev:    13.86436 9.374056

Fixed effects: tid ~ morgen + puls
               Value Std.Error DF   t-value p-value
(Intercept) 900.7781 26.803641 55  33.60656  0.0000
morgennej     6.1214  8.034293 12   0.76191  0.4608
puls         -3.0643  0.161770 55 -18.94253  0.0000

> intervals(modelB)
Approximate 95% confidence intervals

 Fixed effects:
                 lower       est.      upper
(Intercept) 848.226072 900.778114 953.330156
morgennej   -11.004611   6.121391  23.247393
puls         -3.381497  -3.064327  -2.747156

> summary(modelC)

Random effects:
 Formula: ~1 | dag
        (Intercept) Residual
StdDev:    10.98187 25.91155

Fixed effects: tid ~ morgen
               Value Std.Error DF  t-value p-value
(Intercept) 402.8250  5.726907 56 70.33902  0.0000
morgennej    -8.7917  8.747994 12 -1.00499  0.3347

> intervals(modelC)
Approximate 95% confidence intervals

 Fixed effects:
                lower       est.      upper
(Intercept) 391.51771 402.825000 414.132294
morgennej   -27.57765  -8.791667   9.994313

> summary(modelD)

Random effects:
 Formula: ~1 | dag
        (Intercept) Residual
StdDev:    14.16531 9.377646

Fixed effects: tid ~ puls
               Value Std.Error DF   t-value p-value
(Intercept) 901.7988 26.639399 55  33.85207       0
puls         -3.0546  0.160017 55 -19.08918       0
```

```
> intervals(modelD)
Approximate 95% confidence intervals

 Fixed effects:
                 lower       est.      upper
(Intercept) 849.180491 901.798848 954.417204
puls         -3.370655  -3.054589  -2.738522

> summary(modelE)

Random effects:
 Formula: ~1 | dag
        (Intercept) Residual
StdDev:    11.81239 25.91153

Fixed effects: tid ~ 1
               Value Std.Error DF  t-value p-value
(Intercept) 399.0571  4.454389 56 89.58739       0

> intervals(modelE)
Approximate 95% confidence intervals

 Fixed effects:
               lower     est.    upper
(Intercept) 390.1979 399.0571 407.9164
```

Opgave 3 (3 spørgsmål)

Ved et kostforsøg ønskes et antal behandlinger afprøvet på forskellige forsøgspersoner, men ikke nødvendigvis således at hver person prøver alle behandlinger. Nedenfor diskuteres forskellige spørgsmål som har at gøre med selve forsøgsdesignet.

1. Vi antager først, at der indgår 8 behandlinger i forsøget, og at disse adskiller sig fra hinanden ved, at man kan variere 3 faktorer `A`, `B` og `C` hver på 2 niveauer. Det foreslås at lade 2 personer afprøve 4 behandlinger hver med følgende forsøgsplan

Person				
1	$a_1b_1c_1$	$a_1b_2c_1$	$a_2b_1c_2$	$a_2b_2c_2$
2	$a_1b_1c_2$	$a_2b_1c_1$	$a_1b_2c_2$	$a_2b_2c_1$

   Hvilken af faktorerne $\texttt{A} \times \texttt{B}$, $\texttt{A} \times \texttt{C}$ og $\texttt{B} \times \texttt{C}$ er konfunderet med person?

2. Forsøget udvides til at omfatte 9 behandlinger, hvoraf hver person i forsøget skal afprøve de 3. Er det muligt at udføre forsøget som et balanceret ufuldstændigt blokforsøg, hvis der inddrages præcis 15 personer i forsøget?

3. Behandlingsfaktoren er i virkeligheden produktfaktoren, $\texttt{K} \times \texttt{M}$, af de to faktorer kosttilskud (`K`) og måltid (`M`), som hver optræder på 3 niveauer. Det besluttes at lade 9 personer indgå i forsøget, således at hver person afprøver en behandling per dag over en periode på 9 dage. Af praktiske årsager kan det dog kun lade sig gøre at afprøve et kosttilskud på hver forsøgsperson, så det er ikke muligt at lave et fuldstændigt balanceret design. Af det endelige forsøgsdesign, som ses nedenfor, fremgår for eksempel, at person 3 kun afprøver behandlinger med kosttilskud 2. Tilsvarende bemærkes at der hver dag gives det samme måltid til alle personer i forsøget.

person dag	1	2	3	4	5	6	7	8	9
1	K_1M_1	K_2M_1	K_2M_1	K_2M_1	K_3M_1	K_3M_1	K_1M_1	K_3M_1	K_1M_1
2	K_1M_3	K_2M_3	K_2M_3	K_2M_3	K_3M_3	K_3M_3	K_1M_3	K_3M_3	K_1M_3
3	K_1M_1	K_2M_1	K_2M_1	K_2M_1	K_3M_1	K_3M_1	K_1M_1	K_3M_1	K_1M_1
4	K_1M_2	K_2M_2	K_2M_2	K_2M_2	K_3M_2	K_3M_2	K_1M_2	K_3M_2	K_1M_2
5	K_1M_2	K_2M_2	K_2M_2	K_2M_2	K_3M_2	K_3M_2	K_1M_2	K_3M_2	K_1M_2
6	K_1M_3	K_2M_3	K_2M_3	K_2M_3	K_3M_3	K_3M_3	K_1M_3	K_3M_3	K_1M_3
7	K_1M_3	K_2M_3	K_2M_3	K_2M_3	K_3M_3	K_3M_3	K_1M_3	K_3M_3	K_1M_3
8	K_1M_2	K_2M_2	K_2M_2	K_2M_2	K_3M_2	K_3M_2	K_1M_2	K_3M_2	K_1M_2
9	K_1M_1	K_2M_1	K_2M_1	K_2M_1	K_3M_1	K_3M_1	K_1M_1	K_3M_1	K_1M_1

   Forsøgsresultaterne tænkes analyseret ved at opfatte `person` og `dag` som tilfældige faktorer sammen med relevante systematiske faktorer. Opstil den relevante statistiske model og tegn det tilhørende faktordiagram.