Eksamen i Statistisk Dataanalyse 2
(kursusnr.: 210006)

15. april 2010

Alle sædvanlige hjælpemidler, herunder bøger 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 30 %, 40 % og 30 % i bedømmelsen.

Til opgave 1 er vedlagt `R`-udskrifter, som kan benyttes i besvarelsen (det er ikke sikkert at alle dele af udskriften skal benyttes). Til besvarelse af opgave 2 har du fået udleveret en USB-stick med et datasæt, som du skal indlæse og anvende i `R` på din egen PC for at kunne besvare opgaven. 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)

Ved kikkertkirurgisk (laparaskopisk) operation for lyskebrok er man interesseret i at følge patienternes ubehag i form af smerter hen over operationsforløbet. I forbindelse med et klinisk forsøg har man bedt 78 mænd graduere deres smerter umiddelbart før operationen og på forskellige tidspunkter efter operationen. I nedenstående datasæt beskriver variablen `pain` smerterne til tidspunkterne 3, 24,48 og 72 timer efter operationen set i forhold til smerterne umiddelbart før operationen. Således svarer værdien `pain=0` til, at smerterne har samme niveau som før operationen. Variablen `id` er et nummer, som identificerer patienten.

```
> data1$tidfac <- factor(data1$tid)
> head(data1)
```

```
    id tid pain tidfac
88   1   3   15      3
175  1  24   27     24
262  1  48   25     48
349  1  72   10     72
89   2   3   14      3
176  2  24   22     24
```

Data til opgaven er venligst stillet til rådighed af Mette Astrup Madsen. Til besvarelse af opgaven skal du benytte `R`-udskriften sidst i opgaven. Bemærk, at variablen `tid` kan

indgå i modellerne både som en faktor og som en kovariat. Ved besvarelsen af spørgsmål 1 nedenfor, bedes du inddrage `tid` som en faktor.

1. Opskriv en statistisk model til analyse af smerterne og udfør et test for, om smertepåvirkningen ændres over tid. Angiv parameterestimater og konfidensintervaller for parametrene i slutmodellen.

2. Undersøg om det er rimeligt at antage, at smertepåvirkningen ændrer sig lineært over tiden i den betragtede forsøgsperiode. Angiv et estimat og et 95 %-konfidensinterval for den forventede smertepåvirkning efter 24 timer.

3. Det er af særlig interesse at kende rekonvalescenstiden ved kikkertkirurgisk operation for lyskebrok. Rekonvalescenstiden er den tid der går efter operationen, før smerterne er tilbage til niveauet før operationen. Forklar, hvordan man kan benytte resultaterne af den statistiske analyse til at udtale sig om længden af rekonvalescenstiden. Er det rimeligt at hævde, at rekonvalescenstiden er 72 timer?

Udskrift af R-kørsel (lettere redigeret):

```
> ### Større udpluk af datasættet:
>
> head(data1,15)

    id tid pain tidfac
88   1   3   15      3
175  1  24   27     24
262  1  48   25     48
349  1  72   10     72
89   2   3   14      3
176  2  24   22     24
263  2  48   10     48
350  2  72    5     72
90   4   3   26      3
177  4  24    4     24
264  4  48  -12     48
351  4  72  -17     72
91   5   3   70      3
178  5  24    6     24
265  5  48   23     48

> ### Nogle statistiske modeller og test:
>
> model0<-lm(pain~tidfac+factor(id),data=data1)
> model1<-lm(pain~tidfac,data=data1)
> model2<-lm(pain~factor(id),data=data1)
> model3<-lm(pain~tid,data=data1)
> model4<-lm(pain~1,data=data1)
```

```
> anova(model1,model0)
Analysis of Variance Table

Model 1: pain ~ tidfac
Model 2: pain ~ tidfac + factor(id)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    308 302837
2    231  44009 77    258828 17.644 < 2.2e-16 ***

> anova(model2,model0)
Analysis of Variance Table

Model 1: pain ~ factor(id)
Model 2: pain ~ tidfac + factor(id)
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1    234 63034
2    231 44009  3     19025 33.287 < 2.2e-16 ***

> anova(model3,model1)
Analysis of Variance Table

Model 1: pain ~ tid
Model 2: pain ~ tidfac
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    310 302878
2    308 302837  2    41.128 0.0209 0.9793
> anova(model4,model1)
Analysis of Variance Table

Model 1: pain ~ 1
Model 2: pain ~ tidfac
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    311 321862
2    308 302837  3     19025 6.4497 0.0003013 ***

> anova(model4,model3)
Analysis of Variance Table

Model 1: pain ~ 1
Model 2: pain ~ tid
  Res.Df    RSS Df Sum of Sq     F   Pr(>F)
1    311 321862
2    310 302878  1     18984 19.43 1.44e-05 ***

> library(nlme)

> m0 <- lme(pain ~ factor(tid) - 1, random = ~1 | id, data = data1,
+     method = "ML")
```

```
> m1 <- lme(pain ~ 1, random = ~1 | id, data = data1, method = "ML")
> anova(m1, m0)

   Model df      AIC      BIC    logLik   Test  L.Ratio p-value
m1     1  3 2833.268 2844.497 -1413.634
m0     2  6 2755.196 2777.654 -1371.598 1 vs 2 84.07126  <.0001

> m2 <- lme(pain ~ tid, random = ~1 | id, data = data1, method = "ML")
> anova(m2, m0)

   Model df      AIC      BIC    logLik   Test   L.Ratio p-value
m2     1  4 2751.415 2766.387 -1371.707
m0     2  6 2755.196 2777.654 -1371.598 1 vs 2 0.2185811  0.8965

> ### dele af summary() på udvalgte modeller:
>
> summary(model1)

              Estimate Std. Error   t value     Pr(>|t|)
(Intercept)  23.705128   3.550437  6.676679 1.139422e-10
tidfac24     -6.730769   5.021076 -1.340503 1.810697e-01
tidfac48    -13.051282   5.021076 -2.599300 9.791176e-03
tidfac72    -21.141026   5.021076 -4.210457 3.349000e-05

> summary(model3)

              Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 24.5682393 3.07663991  7.985413 2.762718e-14
tid         -0.3018743 0.06848401 -4.407953 1.440132e-05

> m0refit <- lme(pain ~ tidfac - 1, random = ~1 | id, data = data1,
+     method = "REML")
> summary(m0refit)

             Value Std.Error  DF   t-value      p-value
tidfac3  23.705128  3.550437 231 6.6766791 1.798788e-10
tidfac24 16.974359  3.550437 231 4.7809211 3.107405e-06
tidfac48 10.653846  3.550437 231 3.0007141 2.988975e-03
tidfac72  2.564103  3.550437 231 0.7221935 4.709060e-01

> intervals(m0refit)

Approximate 95% confidence intervals

 Fixed effects:
             lower      est.     upper
```

```
tidfac3  16.709750 23.705128 30.700507
tidfac24  9.978980 16.974359 23.969738
tidfac48  3.658468 10.653846 17.649225
tidfac72 -4.431276  2.564103  9.559481
attr(,"label")
[1] "Fixed effects:"

 Random Effects:
  Level: id
                   lower     est.    upper
sd((Intercept)) 23.81270 28.15534 33.28994

 Within-group standard error:
   lower     est.    upper
12.59975 13.80268 15.12045

> m2refit <- lme(pain ~ tid, random = ~1 | id, data = data1,
+     method = "REML")
> summary(m2refit)

                Value  Std.Error  DF    t-value      p-value
(Intercept) 24.5682393 3.46401691 233   7.092413 1.565932e-11
tid         -0.3018743 0.03012522 233 -10.020650 6.976541e-20

> intervals(m2refit)

Approximate 95% confidence intervals

 Fixed effects:
                lower       est.      upper
(Intercept) 17.7434415 24.5682393 31.3930371
tid         -0.3612269 -0.3018743 -0.2425217
attr(,"label")
[1] "Fixed effects:"

 Random Effects:
  Level: id
                   lower     est.    upper
sd((Intercept)) 23.82010 28.16182 33.29489

 Within-group standard error:
   lower     est.    upper
12.55634 13.74973 15.05654

> library(gmodels)

> est24 <- c(1, 24)
> est36 <- c(1, 36)
```

```
> est48 <- c(1, 48)
> est60 <- c(1, 60)
> est72 <- c(1, 72)
> est84 <- c(1, 84)
> est96 <- c(1, 96)
> est <- rbind(est24, est36, est48, est60, est72, est84, est96)
> estimable(m2refit, est, conf.int = 0.95)

        Estimate Std. Error    t value  DF     Pr(>|t|)    Lower.CI
est24 17.3232562   3.304734  5.2419513 233 3.559157e-07  10.8122767
est36 13.7007647   3.282415  4.1739890 233 4.228913e-05   7.2337583
est48 10.0782732   3.299788  3.0542188 233 2.518893e-03   3.5770397
est60  6.4557817   3.356235  1.9235188 233 5.563347e-02  -0.1566649
est72  2.8332901   3.449841  0.8212815 233 4.123255e-01  -3.9635775
est84 -0.7892014   3.577688 -0.2205898 233 8.256050e-01  -7.8379543
est96 -4.4116929   3.736265 -1.1807762 233 2.388956e-01 -11.7728731
       Upper.CI
est24 23.834236
est36 20.167771
est48 16.579507
est60 13.068228
est72  9.630158
est84  6.259552
est96  2.949487
```

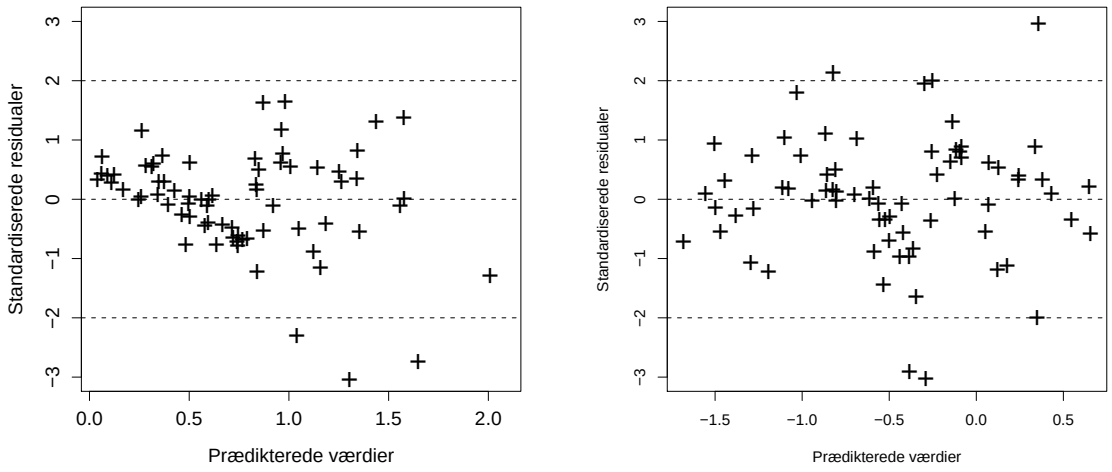
Opgave 2 (5 spørgsmål)

Insulin er et hormon, der regulerer foderindtaget hos langt de fleste pattedyr. Ved et forsøg blev 18 forskellige får randomiseret til 3 forskellige fodertyper (`CS,LS,MS`) givet ved faktoren `feed` med henblik på at undersøge, hvordan foderbehandlingerne påvirker insulin-koncentrationen i blodet. For hvert får blev foretaget målinger af insulin-koncentrationen 1 time før og 2.5 timer efter fodring. Målingerne blev foretaget i to forskellige perioder angivet ved faktoren `status`. Niveauet `status=preg` angiver at målingen er taget, mens fåret var drægtigt, mens `status=lac` angiver at målingen er foretaget efter læmning (nedkomst). Der er således for hvert får foretaget 4 insulin-målinger givet ved kombinationen af variablene `time` og `status`. Data til denne opgave er venligst stillet til rådighed af Maria Brun-Rasmussen.

Data er udleveret på vedlagte USB-stick under filnavnet `mb.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. indlæse data i `R` med kommandoen

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

hvor du vælger filen `mb.txt`. De første 6 linjer i datasættet er organiseret som vist nedenfor



Figur 1: Residualplot for modellerne `model1` og `lmodel1` beskrivet under opgaveformuleringen til spørgsmål 1.

```
     id status time feed insulin
1  4242   preg   1h   LS  0.5833
2  4242   preg 2.5h   LS  1.8922
3  5204   preg   1h   LS  0.4742
4  5204   preg 2.5h   LS  1.5054
5 60217   preg   1h   LS  0.7104
6 60217   preg 2.5h   LS  1.4463
```

Formålet med opgaven er at undersøge, hvordan insulin-koncentrationen (`insulin`) afhænger af faktorerne `feed`, `time` og `status`.

1. Opskriv en statistisk model du vil benytte som udgangspunkt for en statistisk analyse af insulin-målingerne. Du kan argumentere for dit valg af model ud fra figur 1, hvor modellen til venstre er fittet med kommandoen

   ```
   > model1 <- lm(insulin ~ time * status * feed + factor(id),
   +     data = data3)
   ```

   mens modellen til højre er fittet med kommandoen

   ```
   > lmodel1 <- lm(log(insulin) ~ time * status * feed + factor(id),
   +     data = data3)
   ```

2. Reducer den statistiske model fra spørgsmål 1 med henblik på at undersøge, hvordan insulin-koncentrationen afhænger af `feed`, `time` og `status`. Undervejs skal du tydeligt gøre rede for, hvilke modeller du tester mod hinanden, ligesom du bedes udtrække teststørrelser og p-værdier fra `R`-udskriften hørende til de enkelte test, som du foretager.

3. Angiv samtlige parameterestimater der indgår i beskrivelsen af middelværdi- og variansstruktur for din slutmodel fra analysen i spørgsmål 2. Sørg for i ord at forklare, hvad de enkelte parameterestimater beskriver.

4. Angiv et 95 %-konfidensinterval for forskellen mellem grupperne givet ved `feed=CS,time=1h,status=preg` og `feed=CS,time=2.5h,status=preg`.

5. Angiv et estimat og et 95 %-konfidensinterval for forskellen mellem grupperne givet ved `feed=CS,time=1h,status=lac` og `feed=MS,time=2.5h,status=preg`.

Opgave 3 (3 spørgsmål)

I forbindelse med et speciale på Det Biovidenskabelige Fakultet skal der udføres et forsøg, hvor tre forskellige plantesorter skal udsættes for tre forskellige behandlinger. Faktoren `sort` optræder på tre niveauer: `normal` samt to genmodificerede versioner givet ved niveauerne `X1` og `X2`. Behandlingsfaktoren antager niveauerne `A, B` eller `ingen`. Der måles en kontinuert responsvariabel, `y`, og formålet er at beskrive sammenhængen mellem respons og de to forskellige faktorer `behandling` og `sort`, som hver optræder på 3 niveauer. I forbindelse med udførelsen af forsøget råder man totalt over 36 ensartede forsøgsenheder.

I første omgang forestiller man sig kun at afprøve følgende 5 kombinationer af de to faktorer i forsøget:

`sort`	`behandling`	`antal forsøgsenhder`
normal	ingen	12
normal	A	6
normal	B	6
X1	ingen	6
X2	ingen	6

1. Opskriv en statistisk model til analyse af data fra forsøget. Diskuter kort om du synes, at der er tale om et godt forsøgsdesign. Du kan f.eks. støtte dig til et faktordiagram.

Man beslutter sig nu for, at alle 9 kombinationer af `sort` og `behandling` skal afprøves i forsøget. Af praktiske årsager samles planterne i grupper af 3, som placeres på hvert sit bord. Planter på samme bord modtager samme behandling og der sørges for, at hver `sort` er repræsenteret på alle borde. Der indgår i alt 12 borde i forsøget, og det oplyses at `behandling`sfaktoren er balanceret.

2. Hvilken type forsøg er der tale om, og hvordan bør randomiseringen foretages? Opstil et faktordiagram og opskriv en statistisk model til analyse af forsøget.

I den resterende del af opgaven kræves ikke længere, at alle 3 planter på samme bord skal modtage den samme behandling.

Langt om længe bliver det besluttet at benytte forsøgsplanen angivet i tabellen nedenfor. Bemærk, at for overskuelighedens skyld er de 9 kombinationer af `sort` og `behandling` blot angivet med cifrene $1, 2, \ldots, 9$.

Bord nr.											
1	2	3	4	5	6	7	8	9	10	11	12
1	4	6	8	7	9	2	3	8	9	*	*
3	5	7	1	4	6	5	5	3	4	*	*
2	1	1	9	2	2	8	6	7	3	*	*

3. Opskriv en koincidens-matrix der viser, hvor mange gange hvert par af de 9 kombinationer af `sort` og `behandling` optræder sammen inden for samme bord på de 10 borde, som er angivet i tabellen. Find ud af hvilke kombinationer, der skal afprøves på bord 11 og 12, for at forsøgsplanen bliver et BIBD (balanceret ufuldstændigt blokforsøg). Forklar hvordan randomiseringen skal foretages, når den færdige forsøgplan skal overføres til den konkrete forsøgsopstilling.