Opgaver til kursusuge 3

Formålet med denne uges øvelser er at opnå en vis fortrolighed med lineære modeller, herunder fortrolighed med

- hvordan modellerne skrives op, altså $Y_i = \cdots$ (husk at der ofte er flere måder at skrive samme model op på, så fortvivl ikke hvis du ikke er enig med naboen)
- fortolkning af parametre og parameterestimater
- hvordan man tester lineære modeller mod hinanden (F-testet, både forståelse af formlen og hvordan testet kan udføres i `R`)
- hvad og hvordan man konkluderer på baggrund af en statistisk analyse
- hvordan man udfører analyserne i `R`

I alle opgaverne, både dem med og uden computer, er det vigtigt at være omhyggelig med at specificere modellerne og konklusionerne.

Data til de enkelte opgaver kan findes på ugeplanen for uge 3 under Absalon.

Opgaver uden brug af computer

Opgave 3.1

Løs kompendiets opgave 4.1 (de dele vi ikke har snakket om ved forelæsningerne).

Opgave 3.2

For at undersøge om nogle sorter af vårbyg er mere påvirkelige af manganmangel end andre blev et dyrkningsforsøg med vårbygsorterne *Barke*, *Ferment*, *Linus*, *Lux* og *Paloma* foretaget. Mangan blev tilført i tre forskellige mængder (0.025, 0.05 og 0.1 ppm pr uge) og for hver kombination af sort og mangan var der 6 bygplanter. Planterne blev dyrket i spande i væksthus og 46 dage efter udplantning blev højden af hver bygplante målt i cm fra spandens låg. Data er angivet i følgende tabel.

Mangan	Sort									
ppm/uge	Barke		Ferment		Linus		Lux		Paloma	
0.025	76	72	77	79	69	72	66	68	70	73
	78	73	75	76	71	67	64	67	70	72
	79	73	76	80	68	70	65	68	70	73
0.05	81	76	83	81	74	70	66	68	76	83
	79	80	84	81	71	68	68	74	73	77
	82	83	86	82	70	73	69	70	74	76
0.1	86	84	93	86	73	77	69	70	85	83
	87	82	92	85	74	79	69	73	81	84
	87	82	93	83	76	76	75	70	82	85

(Disse data er også brugt til eksamen i Statistisk Forsøgsplanlægning, august 2004, opgave 1.)

Sidst i opgaven finder du en R-kørsel (lettere redigeret) som du kan bruge til at løse dele af opgaven.

I de første spørgsmål skal variablen der beskriver manganindholdet, bruges som *faktor*.

1. Hvilken type eksperiment er der tale om? Opskriv faktordiagrammet for forsøget.
2. Opstil en statistisk model og angiv dimensionen af modellen.
3. Reducer modellen mest muligt.
4. Angiv estimater for parametrene i slutmodellen og beregn LSD-værdien.

   Vink: Modellen med vekselvirkning svarer til en ensidet variansanalysemodel med produktfaktoren som forklarende variabel. Hvordan beregner man estimater og LSD-værdi i denne model? Du kan med fordel benytte den linje is R-udskriften, der begynder med `tapply`.

I de næste spørgsmål skal manganindholdet (`mn`) bruges som *kovariat* (numerisk variabel).

5. Opskriv modellen hvor manganmængden bruges som kovariat og hvor virkningen af manganmængden tillades at være forskellig for de forskellige sorter. Hvor stor er dimensionen af denne model?

6. Undersøg om modellen hvor manganindholdet benyttes som faktor (fra spørgsmål 1), kan reduceres til modellen fra spørgsmål 5.

   Vink: Find først ud af, hvilke to af modellerne fra R-udskriften du gerne vil sammenligne (-dvs. teste mod hinanden). Beregn dernæst F-teststørrelsen ud fra kompendiets sætning 4.14. SS_e-størrelsen for en statistisk model fås med kommandoen `deviance(..)`.

   Hvilken kommando skulle du have brugt for at få `R` til at udføre testet?

7. Opstil og test hypotesen om at effekten af mangan er den samme for alle sorter.

8. Beregn et estimat for forskellen i den forventede højde af en Ferment-bygplante og en Linus-bygplante, når de begge har fået tilført 0.075 ppm/uge. Kunne du have estimeret denne forskel i modellen fra spørgsmål 1?

Uddrag af en R-kørsel:

```
mangan <- read.table("../data/aug04_1.txt",header=T)
mangan$mnfac<-factor(mangan$mn)
head(mangan)

##       srt    mn hjd mnfac
## 1   Barke 0.025  76 0.025
## 2   Barke 0.025  72 0.025
## 3 Ferment 0.025  77 0.025
## 4 Ferment 0.025  79 0.025
## 5   Linus 0.025  69 0.025
## 6   Linus 0.025  72 0.025

tapply(mangan$hjd,mangan$mnfac:mangan$srt,mean)

##   0.025:Barke 0.025:Ferment   0.025:Linus     0.025:Lux  0.025:Paloma
##      75.16667      77.16667      69.50000      66.33333      71.33333
##    0.05:Barke  0.05:Ferment    0.05:Linus      0.05:Lux   0.05:Paloma
##      80.16667      82.83333      71.00000      69.16667      76.50000
##     0.1:Barke   0.1:Ferment     0.1:Linus       0.1:Lux    0.1:Paloma
##      84.66667      88.66667      75.83333      71.00000      83.33333

modelA<-lm(hjd~mnfac+srt+mnfac:srt,data=mangan)
deviance(modelA)

## [1] 470.6667

modelB<-lm(hjd~mnfac+srt,data=mangan)
deviance(modelB)

## [1] 606.7111
```

```
anova(modelB,modelA)
```

```
Analysis of Variance Table
```

```
Model 1: hjd ~ mnfac + srt
Model 2: hjd ~ mnfac + srt + mnfac:srt
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     83 606.71
2     75 470.67  8    136.04 2.7098 0.0112 *
```

```r
modelC<-lm(hjd~mn+srt+mn:srt,data=mangan)
summary(modelC)
```

```
##                  Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)     72.916667   1.261922 57.7822153 5.281821e-67
## mn             121.428571  19.078472  6.3646906 1.144570e-08
## srtFerment       1.333333   1.784628  0.7471213 4.571806e-01
## srtLinus        -5.833333   1.784628 -3.2686557 1.594043e-03
## srtLux          -7.500000   1.784628 -4.2025573 6.818054e-05
## srtPaloma       -5.000000   1.784628 -2.8017049 6.373188e-03
## mn:srtFerment   26.666667  26.981034  0.9883486 3.259619e-01
## mn:srtLinus    -35.238095  26.981034 -1.3060321 1.952845e-01
## mn:srtLux      -62.857143  26.981034 -2.3296788 2.234543e-02
## mn:srtPaloma    35.238095  26.981034  1.3060321 1.952845e-01
```

```r
deviance(modelC)

## [1] 509.5833

modelD<-lm(hjd~mn+srt,data=mangan)
deviance(modelD)

## [1] 630.0468
```

```r
anova(modelD,modelC)
```

```
Analysis of Variance Table

Model 1: hjd ~ mn + srt
Model 2: hjd ~ mn + srt + mn:srt
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     84 630.05
2     80 509.58  4    120.46 4.7279 0.001781 **
```

```r
1-pf(1.24,5,75)

## [1] 0.2990391

qt(0.975,75)

## [1] 1.992102
```

Opgave 3.3

I et forsøg med vinterraps undersøgtes betydningen af CO_2 på to niveauer (360 ppm og 700 ppm) for udbyttet. I forsøget indgik fire sorter, to af en gammel type (1) og to af en nye type (2). For hver kombination af CO_2 og sort blev der dyrket 50 planter der efter et stykke tid blev høstet, og udbyttet blev opgjort. I tabellen nedenfor er angivet det samlede udbytte af de 50 planter for hver kombination af CO_2 og sort.

Type	Sort	CO_2	Udbytte
1	Capitol	360	384
1	Capitol	700	369
1	Express	360	302
1	Express	700	312
2	Matador	360	277
2	Matador	700	304
2	Vestal	360	243
2	Vestal	700	265

(Data er tidligere benyttet i til eksamen i Statistisk Forsøgsplanlægning, maj 2002, opgave 1.)

Sidst i opgaven finder du en R-kørsel (lettere redigeret), som du kan bruge til at løse dele af opgaven.

1. Hvilken slags forsøg er der tale om og hvordan udføres det med hensyn til randomisering?
2. Angiv et faktordiagram, og opskriv (med papir og blyant) den statistiske model svarende til `model1` i R-kørslen nedenfor.
3. Opskriv modellerne svarende til `model2`–`model5` i R-kørslen nedenfor. Overvej hvilke test der er relevante (dvs. hvilke modeller der er delmodeller af andre modeller).
4. Reducér modellen, og angiv slutmodellen. Hvad er konklusionen mht. effekten af CO_2? Er der forskel på sorterne udover hvad der kan foklares af typen?
5. Betragt til sidst den additive model, `model2` (uanset at dette ikke er slutmodellen). På nær parameteriseringen er `model2a` identisk med `model2` — hvorfor?
6. I `model2a` benyttes sorten Capitol som reference. Lad os i stedet estimere *det gennemsnitlige udbytte* (over sort) for de to CO_2-niveauer, dvs. de såkaldte *adjusted means* for CO_2-faktoren. Opskriv først de ønskede parameterfunktioner. Find dernæst estimaterne vha. outputtet fra `estimable` nedenfor (nogle af dem er forkerte).

Uddrag af en R-kørsel:

```
data <- read.table(file="../data/co2.txt",header=T)
data$co2<-factor(data$co2)
data$type<-factor(data$type)
attach(data)


## The following object is masked from package:datasets:
##
##     co2
```

```r
names(data)
```

```
## [1] "type"    "variety" "co2"     "yield"
```

```r
model1 = lm(yield ~ co2 + type + variety + type:co2)
model2 = lm(yield ~ co2 + type + variety)
model3 = lm(yield ~ co2 + type)
model4 = lm(yield ~ type + variety)
model5 = lm(yield ~ type)
```

```
> anova(model2,model1)
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1      3 527.0
2      2 162.5  1     364.5 4.4862 0.1683

> anova(model3,model2)
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1      5 6689.5
2      3  527.0  2    6162.5 17.540 0.02211 *

> anova(model4,model2)
  Res.Df RSS Df Sum of Sq      F Pr(>F)
1      4 769
2      3 527  1       242 1.3776 0.3252

> anova(model5,model4)
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1      6 6931.5
2      4  769.0  2    6162.5 16.027 0.01231 *
```

```r
model2a = lm(yield ~ co2 + variety)
summary(model2a)
```

```
##                 Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)        371.0  10.478152 35.407007 4.954016e-05
## co2700              11.0   9.371944  1.173716 3.252093e-01
## varietyExpress     -69.5  13.253930 -5.243728 1.350279e-02
## varietyMatador     -86.0  13.253930 -6.488641 7.431346e-03
## varietyVestal     -122.5  13.253930 -9.242541 2.679727e-03
```

```r
library(gmodels)
adj1 = c(1,1,1/3,1/3,1/3)
adj2 = c(1,0,0.25,0.25,0.25)
adj3 = c(0,0,0.25,0.25,0.25)
adj4 = c(1,1,0.25,0.25,0.25)
adj = rbind(adj1,adj2, adj3, adj4)
estimable(model2a, adj, conf.int=0.95)
```

```
##      Estimate Std. Error   t value DF     Pr(>|t|)  Lower.CI  Upper.CI
## adj1 289.3333   7.157940 40.421311  3 3.331835e-05 266.55357 312.11309
```

```
## adj2 301.5000   6.626965 45.495939  3 2.337750e-05 280.41004 322.58996
## adj3 -69.5000   8.116342 -8.562971  3 3.347152e-03 -95.32982 -43.67018
## adj4 312.5000   6.626965 47.155824  3 2.099723e-05 291.41004 333.58996
```

Opgave 3.4

Løs opgave 4.3 i kompendiet.

Opgaver med brug af computer

Opgave 3.5

Løs opgave 4.4 i kompendiet.

Kommentarer, uddybninger og vink til de enkelte spørgsmål:

1. Tegn også faktordiagrammet for eksperimentet.
2. Vedr. figuren: prøv først kommandoen

   ```
   interaction.plot(barleyfac,sinapisfac,weight)
   ```

   Hvad kan denne figur bruges til? Hvorfor er den ikke egnet til af afgøre om sammenhængen mellem antal antallet af bygfrø og friskvægten er lineær?

   Prøv derefter følgende kommandoer, en ad gangen. Forsøg for hver kommando at forså hvad `R` gør.

   ```
   ave = tapply(weight, list(barley, sinapis), mean)
   ave
   ave[,1]
   ave[,2]
   ave[,3]
   plot(barley, weight, type="n")
   points(c(0,3,7,15,34,77),ave[,1], type="l")
   points(c(0,3,7,15,34,77),ave[,2], type="l", lty=2)
   points(c(0,3,7,15,34,77),ave[,3], type="l", lty=3)
   ```

   Ser der ud til at være en lineær sammenhæng mellem antal antallet af bygfrø og friskvægten?

   Opskriv derefter en lineær model og test den mod faktormodellen som beskrevet i spørgsmål 2 i kompendiet.

4. Beregn først estimatet "i hånden" vha. estimaterne fra spørgsmål 3. Brug derefter funktionen `estimable` til også at få beregnet konfidensintervallet. Alternativt kan `predict` bruges, se opgave 1.5(d).

Opgave 3.6

I et forsøg "opdrættede" man bananfluer ved syv forskellige temperaturer. For hver temperatur udtog man en stikprøve på 30 bananfluer og målte den gennemsnitlige tid til udklækning for disse 30 fluer. Forsøget blev gentaget således at man har to observationer for hver temperatur:

Temperatur	27	28	29	30	31	32	33
Tid til	1.07	1.07	0.36	0.00	0.20	0.24	0.86
udklækning	1.23	0.77	0.41	0.12	0.28	0.52	0.74

Data stammer *Statistics in Biology II* af C. I. Bliss (1970), side 50–51.

1. Indlæs data og tegn tid til udklækning mod temperaturen vha. `plot`.

Sammenhængen mellem temperatur og tid til udklækning er tydeligvis ikke lineær! Betragt modellen

$$y_i = \beta_0 + \beta_1 \cdot \texttt{temp}_i + \beta_2 \cdot \texttt{temp}_i^2 + e_i \tag{1}$$

hvor y_i er tid til udklækning for den i'te observation, $\texttt{temp}_i$ er den tilhørende temperatur og $e_1, \ldots, e_{14}$ er uafhængige $N(0, \sigma^2)$-fordelte.

2. Overvej, på grundlag af figuren, hvorfor dette kunne være en rimelig model. Er modellen en lineær model (sammenlign med definition 4.3 i kompendiet)?

3. Estimér modellens parametre. Indtegn den estimerede funktion i figuren fra spørgsmål 1.

   Vink til estimation: Lav en ny variabel, `temp2`, med de kvadrerede værdier af `temp`. Brug både `temp` og `temp2` som kovariater i kaldet til `lm`.

   Vink til graf: En nem måde: `points(temp,fitted(model), type="l")`. Forklar hvad `R` gør! Dette giver ikke en glat kurve, men det går nok endda.

4. Beregn et estimat for den temperatur hvor udklækning sker hurtigst.

   Vink: For hvilken værdi af x har andengradsplynomiet $a + bx + cx^2$ sit minimum (når $c > 0$)?

5. Fit også den ensidede variansanalysemodel hvor `temp` bruges som faktor.

6. Overbevis dig selv om at den kvadratiske regressionsmodel (1) er en delmodel af den ensidede variansanalysemodel, således at vi kan teste (1) mod den ensidede variansanalysemodel. Udfør testet.

Fra grafen synes det ret klart at sammenhængen *ikke* er lineær. Udfør, for øvelsens skyld, alligevel følgende:

7. Udtryk hypotesen om linearitet vha. parametrene i model (1) og test hypotesen.

Opgave 3.7

Eksamensopgave fra Statistisk Forsøgsplanlægning: august 2002, opgave 1.