Opgaver til kursusuge 8

Formålet med denne uges øvelser er at få repeteret kursets pensum, hvorfor vi i høj grad regner tidligere eksamensopgaver fra Statistisk Dataanalyse 2. Desuden findes nedenfor en række opgaver 8.1-8.6, som er inspireret af gamle eksamensopgaver fra kurset Statistisk Forsøgsplanlægning, nu omformuleret så de passer til Statistisk Dataanalyse 2.

Opgave 8.1.

Nedenstående tabel viser den ugentlige vækst (`vaekst`) af 30 kalkuner fodret med standardfoder tilsat forskellige mængder af A-vitamin. I tabellen er angivet logaritmen til mængden af A-vitamin (`logA`) i passende enheder samt hver af kalkunernes vækst (beregnet som en gennemsnitlig vækst gennem 8 uger).

	`logA`					
	1	2	3	4	5	6
`vaekst`	0.137	0.198	0.217	0.167	0.363	0.285
	0.160	0.155	0.330	0.373	0.368	0.280
	0.197	0.163	0.365	0.200	0.307	0.348
	0.133	0.305	0.320	0.355	0.343	0.287
	0.223	0.225	0.298	0.323	0.350	0.222

a) Opskriv de statistiske modeller svarende til de tre kald af funktionen `lm()` i R-programmet hørende til denne opgave.

b) Analysér forsøget med henblik på at finde frem til en model der beskriver data godt. Angiv estimater og konfidensintervaller for parametrene i den systematiske del af denne model.

Man er interesseret i hvilken mængde af A-vitamin der giver den største vækst. For en parabel med ligningen $y = a + bx + cx^2$ gælder at maksimum findes i den værdi x som opfylder $b + 2cx = 0$, forudsat at c er negativ.

c) Angiv et estimat for den værdi af `logA` som ifølge modellen giver størst vækst. Afgræns også denne værdi så godt du kan ud fra `R`-udskriften.

R-program og udskrift (noget beskåret) til Opgave 8.1.

```
> kalkun = read.table(file.choose(), header=T) ## fil: kalkun.txt
> kalkun
   logA     y
1     1 0.137
2     1 0.160
  . . .   (i alt 30 datalinier)
30    6 0.222

> attach(kalkun)

> logA2= logA*logA
> logA3= logA*logA*logA

> modelA= lm(y ~ logA + logA2 + logA3)
> modelB= lm(y ~ logA + logA2)
> modelC= lm(y ~ logA)
>

> summary(modelA)

Call:
lm(formula = y ~ logA + logA2 + logA3)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.132933   0.095492   1.392    0.176
logA         0.020378   0.108804   0.187    0.853
logA2        0.017453   0.034817   0.501    0.620
logA3       -0.002754   0.003290  -0.837    0.410

Residual standard error: 0.05922 on 26 degrees of freedom
Multiple R-Squared: 0.5046,     Adjusted R-squared: 0.4475
F-statistic: 8.828 on 3 and 26 DF,  p-value: 0.0003326


> anova(modelB, modelA)
Analysis of Variance Table

Model 1: y ~ logA + logA2
Model 2: y ~ logA + logA2 + logA3
  Res.Df      RSS Df Sum of Sq      F Pr(>F)
1     27 0.093644
2     26 0.091187  1  0.002457 0.7005 0.4102
```

```
> summary(modelB)

Call:
lm(formula = y ~ logA + logA2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.06354    0.04711   1.349  0.18865
logA         0.10767    0.03082   3.493  0.00166 **
logA2       -0.01146    0.00431  -2.659  0.01302 *

Residual standard error: 0.05889 on 27 degrees of freedom
Multiple R-Squared: 0.4913,     Adjusted R-squared: 0.4536
F-statistic: 13.04 on 2 and 27 DF,  p-value: 0.0001090

> anova(modelC, modelB)
Analysis of Variance Table

Model 1: y ~ logA
Model 2: y ~ logA + logA2
  Res.Df      RSS Df Sum of Sq      F  Pr(>F)
1     28 0.118162
2     27 0.093644  1  0.024518 7.0692 0.01302 *

> x.eq.4 = c(0, 1, 8)
> x.eq.5 = c(0, 1, 10)
> x.eq.6 = c(0, 1, 12)
> x.eq.7 = c(0, 1, 14)
> x.eq.8 = c(0, 1, 16)
> x.eq.9 = c(0, 1, 18)
>
> x.all = rbind(x.eq.4, x.eq.5, x.eq.6, x.eq.7, x.eq.8, x.eq.9)
>
> library(gmodels)
> estimable(modelB, x.all, conf.int= 0.95)

       Estimate Std.Err.   t value DF Pr(>|t|) Lower.CI  Upper.CI
x.eq.4  0.01598 0.00763  2.0950007 27 0.04568   0.00032  0.031640
x.eq.5 -0.00693 0.01438 -0.4822788 27 0.63349  -0.03644  0.022574
x.eq.6 -0.02985 0.02245 -1.3297864 27 0.19471  -0.07592  0.016212
x.eq.7 -0.05277 0.03082 -1.7123254 27 0.09830  -0.11602  0.010464
x.eq.8 -0.07570 0.03930 -1.9261372 27 0.06467  -0.15634  0.004939
x.eq.9 -0.09862 0.04783 -2.0618708 27 0.04896  -0.19676 -0.000480

> qt(0.975,df=27)
[1] 2.051831
```

Opgave 8.2.

I et forsøg vedrørende biologisk bekæmpelse undersøgtes tilstedeværelse af svampen *Arbuscular Mycorrhizal fungi* (AM) på hyfelængder. Roer dyrkedes i en speciel forsøgsopstilling hvor to sider af roen voksede ud i hvert sit rør med henholdsvis jord og sand. I begge sider var desuden en ud af tre mulige bakteriekulturer som var den samme i de to sider, men som varierede fra roe til roe. I forsøget var i alt 24 roer, 4 med hver kombination af bakteriekultur (1,2 eller 3) og AM (til stede eller ikke til stede). Som responsvarabel benyttedes kvadratroden af hyfelængden i hver af de to sider for hver af de 24 roer (i alt 48 resultater).

a) Angiv alle faktorer af relevans for forsøget og deres tilhørende niveauer. Hvilke af faktorerne bør indgå i modellen med tilfældig virkning og hvilke med systematisk virkning? [Vejledning: "Siden" af roen skal ikke indgå som faktor i modellen. Dette kan ikke ses ud fra ovenstående beskrivelse.]

b) Opskriv et faktordiagram for modellen som indeholder alle mulige vekselvirkninger bortset fra dem hvori roe indgår. Beregn frihedsgraderne hørende til faktordiagrammet.

I følgende tabel er angivet `MS` (Mean Square) for hver af en række effekter som indgår i den model der nu skal tages udgangspunkt i. Her står `M` for materiale (jord eller sand) og B for bakteriekultur. For to af effekterne er også frihedsgradstallet angivet.

Effekt	MS	DF
AM	52.736	
B	0.519	
M	73.075	
AM × B	1.408	
AM × M	17.605	
M × B	0.464	
AM × M × B	0.0512	
roe	0.313	18
residual	0.251	18

c) Analysér forsøget med henblik på at undersøge hvilke effekter af AM, materialet og bakteriekulturen der kan påvises ud fra data.

Som beregningshjælp kan benyttes følgende udskrift fra R:

```
> pf(0.204,df1=2,df2=18)
[1] 0.1826782
> 1-pf(76.2,df1=1,df2=19)
[1] 4.479968e-08
> pf(0.204,df1=2,df2=18)
[1] 0.1826782
> 1-pf(76.2,df1=1,df2=19)
[1] 4.479968e-08
> pf(2.01,2,20)
[1] 0.8398341
> pf(4.50,2,18)
[1] 0.9739877
> pf(1.23,2,20)
[1] 0.6865253
```

d) Præsenter grafisk estimaterne (men ikke konfidensintervaller eller anden usikkerhedsangivelse) på passende form for de effekter der ud fra ovenstående analyse er af interesse. Som hjælp hertil angiver nedstående tabel gennemsnit for de 12 kombinationer af `AM`, `M` og `B`. [Som yderligere hjælp kan oplyses at "adjusted means for den benyttede model er identiske med simple gennemsnit for alle modellens effekter].

e) Undersøg om der er en effekt af roe.

	AM ikke til stede		AM til stede	
Bakteriekultur	Sand	Jord	Sand	Jord
1	0.85	2.47	1.38	5.32
2	1.00	2.17	1.49	5.34
3	0.94	1.91	2.58	5.82

Opgave 8.3.

Ved et lagringsforsøg lagres kødstykker i 0, 1, 2, 3 eller 4 dage ved enten $-5°$C eller $+5°$C. Efter lagring måles mørheden af kødet ved at måle den kraft der skal til at skære det over. Umiddelbart ser der ud til at være to faktorer: `tid` med niveauerne (0, 1, 2, 3, 4) samt `temp` med niveauerne $(-5, +5)$. Der er dog det problem at ved lagring i 0 dage sker der slet ingen lagring, og den tiltænkte lagringstemperatur kan derfor ikke have nogen effekt. Vi ønsker at benytte modeller der afspejler dette.
To stykker kød lagres og måles for hver kombination af `temp` og `tid`. Der måles således i alt mørheden af 20 stykker kød, hvoraf de 4 stykker ikke har været lagret og derfor har en meningsløs værdi af faktoren `temp`.
Nedenfor er specificeret en række modeller ved hjælp af functionen `lm()` i R. De to variable `temp` og `tid` er i forvejen oprettet som numeriske variable i R og er ikke *på forhånd* omdannet til faktorer.

Model 1: `model1= lm(y ~ factor(temp):factor(tid))`
Model 2: `model2= lm(y ~ factor(temp) + factor(tid))`
Model 3: `model3= lm(y ~ factor(temp))`
Model 4: `model4= lm(y ~ factor(tid))`
Model 5: `model5= lm(y ~ factor(temp) + factor(temp)*tid)`
Model 6: `model6= lm(y ~ factor(temp) + tid)`
Model 7: `model7= lm(y ~ tid)`

a) Opskriv de statistiske modeller der svarer til ovenstående modeller.

b) Hvilke af modellerne giver mening ud fra den indledende beskrivelse af forsøget? (Begrund svaret).

c) Angiv hvilke af ovenstående 7 modeller der er delmodeller af hvilke andre. Svaret ønskes angivet i form af et piludiagram eller på anden entydig måde.

Opgave 8.4

I et forsøg sammenlignedes udbyttet af 8 jordbærsorter. De 8 sorter var plantet i 4 rækker som angivet i nedenstående skitse. I skitsen er angivet rækkenummer, sorten (G, V, R1, F, Re, M, E, P) samt udbyttetal i en passende enhed.

Række									
	I	G	V	R1	F	Re	M	E	P
		5.8	6.3	4.9	6.5	4.5	5.2	6.5	3.8
									H
	II	E	P	M	Re	G	V	F	R1
		6.9	7.6	7.9	5.6	7.0	5.5	4.0	2.7
									Æ
	III	V	F	R1	G	P	E	Re	M
		7.6	6.4	5.0	6.9	7.4	5.3	5.2	3.2
									K
	IV	E	Re	M	P	G	F	V	R1
		7.5	7.0	6.1	7.2	6.5	5.6	5.8	1.4

(a) Hvilken type forsøg er der tale om og hvordan udføres randomisering i et sådant forsøg. Opstil en statistisk model og analyser forsøget (Estimater for parametrene ønskes ikke angivet).

Det viser sig, at der var en hæk i højre side af forsøgsmarken som angivet i skitsen. Dette kunne tænkes at have indflydelse på forsøgsresultaterne, hvilket skal undersøges i det følgende. En rimelig antagelse er at en eventuel indflydelse af hækken på udbyttetallene aftager som den reciprokke afstand til hækken. Vi indfører derfor variablen $x = 1/\text{afstand}$ som antager værdierne

$$0.125, \quad 0.143, \quad 0.167, \quad 0.200, \quad 0.250, \quad 0.333, \quad 0.500, \quad 1.000.$$

(b) Opskriv relevante modeller og hypoteser, og udfør test for hypoteserne. Drag konklusioner vedrørende udbytte for de forskellige jordbærsorter. Sørg for herunder at få estimater for estimerede middeludbytter og/eller forskelle på middeludbytter for sorterne, og vis ved nogle eksempler hvordan du kan konstruere konfidensintervaller for forskelle i middeludbytte mellem to sorter.

Opgave 8.5

I et oxidationsforsøg med lipider var man interesseret i at undersøge den antioxidante effekt af druer. Til det formål blev 3 linolsyreopløsninger tilsat drueekstrakt med en koncentration på henholdsvis 0.0%, 0.1% og 0.2% af en given koncentration. Efter 10 minutter (tid 0) blev den relative oxygenkoncentration (%) målt 14 gange for hver opløsning med 20 sekunders mellemrum.

	Koncentration		
Tid	0.0%	0.1%	0.2%
0	42.8	81.9	93.8
20	39.8	80.6	93.5
40	36.7	79.4	93.2
60	33.7	78.2	92.8
80	30.7	77.0	92.5
100	27.8	75.7	92.4
120	24.9	74.2	92.0
140	22.1	72.8	91.8
160	19.4	71.3	91.4
180	16.8	69.9	91.2
200	14.2	68.2	90.8
220	11.6	66.6	90.6
240	9.3	64.9	90.4
260	7.1	63.0	90.1

Denne opgave kan med fordel løses ved hjælp af R-programmet med udskrift sidst i opgavesættet.

1. Tegn den relative oxygenkoncentration som funktion af tiden med et symbol for hver koncentration af antioxidanten. Opstil en statistisk model for data.

2. Tegn hældningsestimaterne som funktion af koncentrationen af drueekstraktet og opstil en statistisk model for de relative oxygenkoncentrationsdata hvor denne koncentration indgår eksplicit i hældningen. Test modellen mod den opstillet i spørgsmål 1.

3. Angiv estimater og 95%-konfidensintervaller for parametrene i den systematiske del af slutmodellen.

R-program og udskrift (lidt beskåret) til Opgave 8.5.

```
>oxygen= read.table("oxygen.dat",header=T)
>attach(oxygen)

>model1= lm(O2 ~ factor(Konc)+factor(Konc):Tid-1)
>anova(model1)

Analysis of Variance Table

Response: O2
                 Df Sum Sq Mean Sq  F value    Pr(>F)
factor(Konc)      3 201182   67061 436125.9 < 2.2e-16 ***
factor(Konc):Tid  3   2231     744   4836.8 < 2.2e-16 ***
Residuals        36      6  0.1538
```

```
>summary(model1)

Call:
lm(formula = O2 ~ factor(Konc) + factor(Konc):Tid - 1)

Residuals:
      Min        1Q    Median        3Q       Max
-0.782857 -0.183901  0.004945  0.161099  1.028571

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
factor(Konc)0       42.05714    0.19884  211.51  < 2e-16 ***
factor(Konc)0.1     82.46000    0.19884  414.69  < 2e-16 ***
factor(Konc)0.2     93.74000    0.19884  471.42  < 2e-16 ***
factor(Konc)0:Tid   -0.13841    0.00130 -106.47  < 2e-16 ***
factor(Konc)0.1:Tid -0.07183    0.00130  -55.26  < 2e-16 ***
factor(Konc)0.2:Tid -0.01421    0.00130  -10.93 5.47e-13 ***

Residual standard error: 0.3921 on 36 degrees of freedom
Multiple R-Squared:     1,      Adjusted R-squared:     1
F-statistic: 2.205e+05 on 6 and 36 DF,  p-value: < 2.2e-16

>model2= lm(O2 ~ factor(Konc)+Tid+Konc:Tid,data=oxygen)
>anova(model2)


Analysis of Variance Table

Response: O2
             Df Sum Sq Mean Sq F value    Pr(>F)
factor(Konc)  2  34345   17173 94144.4 < 2.2e-16 ***
Tid           1   1528    1528  8377.6 < 2.2e-16 ***
Tid:Konc      1    702     702  3847.7 < 2.2e-16 ***
Residuals    37      7  0.1824

> pf(7.89,df1=1,df2=36)
[1] 0.992017
> qt(0.975,df=36)
[1] 2.028094
```

Opgave 8.6

I et spiringsforsøg undersøgte man 3 græssorter af arten *rødsvingel*, nemlig Napoli, Smirna og Symphony. Fire partier blev tilfældigt udvalgt for hver sort (forskellige partier for hver sort), og fra hvert parti blev der taget 4 prøver af 100 frø. De 100 frø i hver prøve blev sået i et spiringskammer og antal spirede frø blev dernæst observeret en til to gange dagligt i omkring 20 dage. På baggrund af disse observationer blev en middelspiretid for hver prøve beregnet, og denne er angivet i nedenstående tabel i dage.

	Parti			
Sort	1	2	3	4
Napoli	5.77	5.19	5.52	5.13
	5.48	5.73	5.32	5.16
	5.08	5.24	5.14	4.76
	4.89	4.84	5.00	5.06
Smirna	4.99	4.70	4.83	5.22
	4.95	4.66	4.61	5.32
	4.79	4.81	4.56	5.52
	4.60	4.86	4.81	5.56
Symphony	4.52	4.12	4.26	4.41
	4.58	4.13	4.18	4.33
	4.21	4.83	4.47	4.93
	4.21	4.44	4.62	4.73

1. Opstil et faktordiagram samt en statistisk model for forsøget.
2. Test om eventuelle tilfældige effekter kan fjernes (dvs. sættes til 0).
3. Analyser data og angiv estimater for alle parametre i slutmodellen.
4. Angiv LSD-værdien for sammeligning mellem sorterne, idet følgende formel kan benyttes

$$LSD_{0.95} = t_{0.975,9}\sqrt{\frac{2(\hat{\sigma}^2 + 4\hat{\sigma}_p^2)}{4 \cdot 4}}.$$

Som beregningshjælp kan benyttes følgende udskrift fra R:

```
> model1=lm(x~p+s)
> model2=lm(x~s)
> anova(model2,model1)
Analysis of Variance Table

Model 1: x ~ s
Model 2: x ~ p + s
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     45 3.7799
2     36 2.1893  9    1.5907 2.9063 0.01083 *

> library(nlme)
> mod1=lme(x~s,random=~1|p,method="ML")
> mod2=lme(x~1,random=~1|p,method="ML")
```

```
> anova(mod2,mod1)
     Model df      AIC      BIC     logLik   Test  L.Ratio p-value
mod2     1  3 33.99365 39.60725 -13.996826
mod1     2  5 21.17083 30.52684  -5.585417 1 vs 2 16.82282   2e-04

> mod1final=lme(x~s-1,random=~1|p,method="REML")
> summary(mod1final)
Linear mixed-effects model fit by REML

Random effects:
 Formula: ~1 | p
        (Intercept)  Residual
StdDev:   0.1702398 0.2466033

Fixed effects: x ~ s - 1
      Value Std.Error DF  t-value p-value
s1 5.206875 0.1051010  9 49.54162       0
s2 4.924375 0.1051010  9 46.85373       0
s3 4.435625 0.1051010  9 42.20344       0

> intervals(mod1final)
Approximate 95% confidence intervals

 Fixed effects:
     lower     est.   upper
s1 4.96912 5.206875 5.44463
s2 4.68662 4.924375 5.16213
s3 4.19787 4.435625 4.67338

### Alternatively: R code to construct F-test

> model1=lm(x~p+s)
> model2=lm(x~s)
> model3=lm(x~1)
> MSe=deviance(model1)/model1$df
> MSe
[1] 0.0608132
> MSp=(deviance(model2)-deviance(model1))/(model2$df-model1$df)
> MSp
[1] 0.1767396
> MSs=(deviance(model3)-deviance(model2))/(model3$df-model2$df)
> MSs
[1] 2.436025
```