NTNU

Kunnskap for ei betre verd

Institutt for matematiske fag

Eksamensoppgåve i **TMA4267 Lineære statistiske modellar**

Fagleg kontakt under eksamen: Øyvind Bakke

Tlf: 73 59 81 26, 990 41 673

Eksamensdato: 22. mai 2015

Eksamenstid (frå–til): 9.00–13.00

Hjelpemiddelkode/Tillatne hjelpemiddel: Gult, stempla A4-ark med eigne handskrivne notat, bestemd enkel kalkulator, *Tabeller og formler i statistikk* (Tapir forlag), *Matematisk formelsamling* (K. Rottmann)

Annan informasjon:
I vurderinga tel kvart av dei åtte bokstavpunkta likt.

Målform/språk: nynorsk

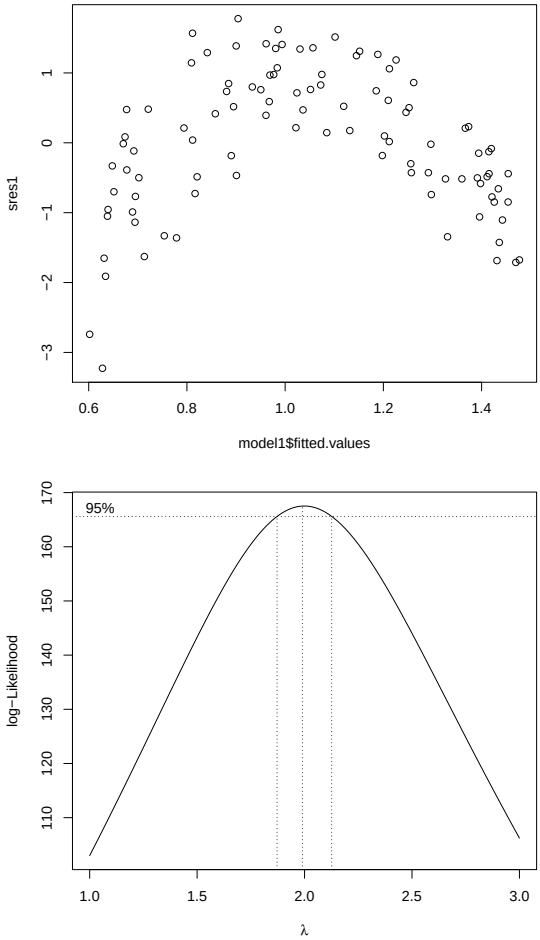
Sidetal: 4

Sidetal vedlegg: 0

Kontrollert av:

Dato Sign

```
> model1<-lm(Period~Length+Amplitude+Mass)
> summary(model1)

Call:
lm(formula = Period ~ Length + Amplitude + Mass)

Residuals:
      Min        1Q    Median        3Q       Max
-0.109411 -0.023820  0.001007  0.027937  0.063272

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.4391125  0.0138346  31.740  < 2e-16 ***
Length      0.0197488  0.0002723  72.526  < 2e-16 ***
Amplitude   0.0448392  0.0296440   1.513  0.13367
Mass        0.0232896  0.0070989   3.281  0.00144 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03644 on 96 degrees of freedom
Multiple R-squared: 0.9828,     Adjusted R-squared: 0.9823
F-statistic:  1827 on 3 and 96 DF,  p-value: < 2.2e-16

> sres1<-rstudent(model1)
> plot(model1$fitted.values,sres1)
> library(MASS)
> boxcox(model1,lambda=seq(1,3,.1))
```



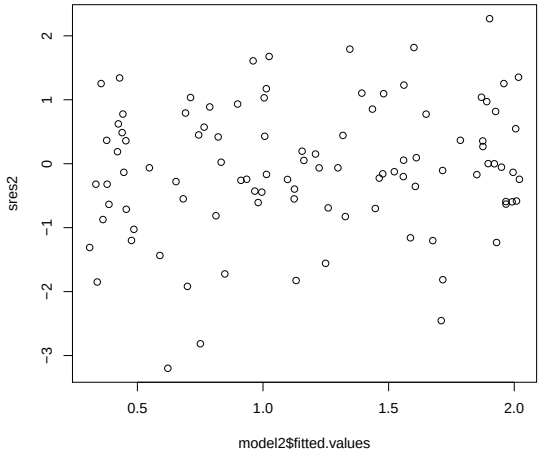
Figur 1: Modellen i oppgåve 1a: R-kode og -utskrift (venstre), residualplott (oppe til høgre) og Box–Cox-plott (nede til høgre).

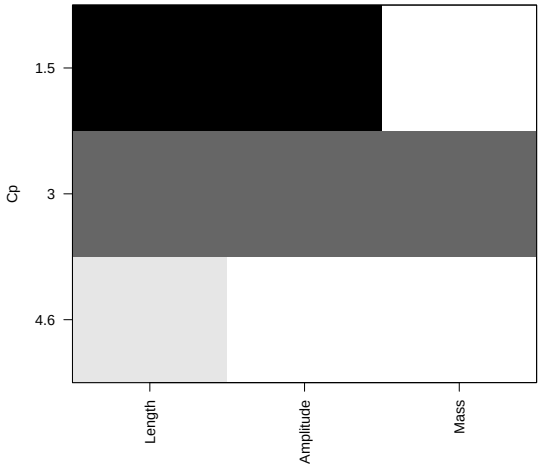
Oppgåve 1

Svingetida for ein pendel vart studert, og 100 kombinasjonar av lengda til pendelen (målt i cm), amplituden (største utslaget av svingingane frå den loddrette jamvektslinja til ei av sidene, målt i radianar) og masse (kg) vart variert. Ein multippel regresjonsmodel vart tilpassa. Figur 1 viser R-kode og -utskrift, eit residualplott og eit Box–Cox-plott.

a) Skriv opp den tilpassa regresjonsmodellen, og kommenter modelltilpassinga kort. Kva konklusjonar kan du trekke frå residualplottet? Foreslå ein transformasjon på grunnlag av Box–Cox-plottet.

```
> model2<-lm(Period^2~Length+Amplitude+Mass-1)
> summary(model2)

Call:
lm(formula = Period^2 ~ Length + Amplitude + Mass - 1)

Residuals:
      Min        1Q    Median        3Q       Max
-0.121375 -0.023555 -0.003389  0.023144  0.086937

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
Length     0.0403534  0.0002672 151.008   <2e-16 ***
Amplitude  0.0610402  0.0262051   2.329   0.0219 *
Mass      -0.0045451  0.0066159  -0.687   0.4937
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03976 on 97 degrees of freedom
Multiple R-squared: 0.9991,     Adjusted R-squared: 0.9991
F-statistic: 3.566e+04 on 3 and 97 DF,  p-value: < 2.2e-16

> sres2<-rstudent(model2)
> plot(model2$fitted.values,sres2)
> pendulum<-as.data.frame(cbind(Period,Length,Amplitude,Mass))
> library(leaps)
> best<-regsubsets(Period^2~.,data=pendulum,intercept=FALSE)
> summary(best)$which
  Length Amplitude  Mass
1   TRUE     FALSE FALSE
2   TRUE      TRUE FALSE
3   TRUE      TRUE  TRUE
> summary(best)$cp
[1] 4.569336 1.471964 3.000000
> plot(best,scale="Cp")
```





Figur 2: Modellen i oppgåve 1b: R-kode og -utskrift (venstre), residualplott (oppe til høgre) og ein grafisk tabell over beste delmengder, der Mallows' C_P blir brukt som observator for å ordne modellane (nede til høgre). Merk at opplysningane i den grafiske tabellen òg er i R-utskrifta.

Tilnærmingsformelen $T \approx 2\pi\sqrt{L/g}$ for svingetida T for ein pendel, der L er lengda og $g \approx 9.8$ m/s^2 er tyngdeakselerasjonen, viser at det kan vere rimeleg å bruke kvadratet av svingetida i staden for svingetida som responsvariabel i ein regresjonsmodell, og også at konstantleddet (skjæringspunktet) blir sløyfa. Figur 2 viser R-kode og -utskrift, eit residualplott og eit plott av beste delmengd-seleksjon basert på Mallows' C_P for slike modellar.

b) Føretrekker du den opphavlege modellen eller den nye modellen som nettopp er nemnd? Kva undermodell av den nye modellen ville du velje, dersom du skulle velje ein? Grunngje kort svara dine.

```
> model3<-lm(log(Period)~log(Length)+log(1+Amplitude^2/16+11*Amplitude^4/3072))
> summary(model3)

Call:
lm(formula = log(Period) ~ log(Length) + log(1 + Amplitude^2/16 +
    11 * Amplitude^4/3072))

Residuals:
     Min       1Q   Median       3Q      Max
-0.09906 -0.01002  0.00126  0.01266  0.08019

Coefficients:
                                             Estimate Std. Error  t value Pr(>|t|)
(Intercept)                                 -1.617849   0.015979 -101.247   <2e-16 ***
log(Length)                                  0.502433   0.004809  104.474   <2e-16 ***
log(1 + Amplitude^2/16 + 11 * Amplitude^4/3072)  1.260754   0.570785    2.209   0.0295 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02705 on 97 degrees of freedom
Multiple R-squared: 0.9912,     Adjusted R-squared: 0.9911
F-statistic:  5491 on 2 and 97 DF,  p-value: < 2.2e-16
```

Figur 3: Modellen i oppgåve 1c: R-kode og -utskrift.

Den meir eksakte formelen $T = 2\pi\sqrt{\frac{L}{g}}(1 + \frac{1}{16}\theta^2 + \frac{11}{3072}\theta^4 + \cdots)$, eller $\ln T = \ln(2\pi/\sqrt{g}) + \frac{1}{2}\ln L + \ln(1 + \frac{1}{16}\theta^2 + \frac{11}{3072}\theta^4 + \cdots)$, der θ er amplitude, viser at det kan vere rimeleg å bruke ein tredje modell, der både responsvariabelen og kovariatane er transformerte. Figur 3 viser R-kode og -utskrift.

c) Korleis stemmer estimata av koeffisientane med den fysiske modellen gitt over? Finn eit estimat av g, tyngdeakselerasjonen, og eit 95 %-konfidensintervall for g.

Oppgåve 2

Anta ein lineær regresjonsmodell $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, der $\boldsymbol{Y}$ er ein n-dimensjonal stokastisk vektor, X ei designmatrise av storleik $n \times p$, $\boldsymbol{\beta}$ ein p-dimensjonal parametervektor (koeffisientvektor) og $\boldsymbol{\epsilon}$ n-dimensjonal multinormal med forventningsverdi $\mathbf{0}$ og kovariansematrise $\sigma^2 I$, der I er identitetsmatrisa av storleik $n \times n$.

Anta vidare at søylene i X er ortogonale.

a) Vis at minstekvadratestimatoren for β_j, element j i $\boldsymbol{\beta}$, berre avheng av søyle j i X (dvs. kovariatvektor j) og $\boldsymbol{Y}$.

I eit tovegs faktorielt ikkje-replikert 2^2-forsøk er nivåa koda -1 og 1. Responsvektoren er $(6 \quad 4 \quad 10 \quad 7)^{\mathrm{T}}$, som svarer til nivåa $(-1 \quad 1 \quad -1 \quad 1)^{\mathrm{T}}$ av den første faktoren og $(-1 \quad -1 \quad 1 \quad 1)^{\mathrm{T}}$ av den andre faktoren.

b) Estimer interaksjonseffekten (to gongar koeffisienten) av dei to faktorane.

Oppgåve 3

Anta ein lineær regresjonsmodell $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, der $\boldsymbol{Y}$ er ein n-dimensjonal stokastisk vektor, X ei designmatrise av storleik $n \times p$, $\boldsymbol{\beta}$ ein p-dimensjonal parametervektor og $\boldsymbol{\epsilon}$ n-dimensjonal multinormal med forventningsverdi $\mathbf{0}$ og kovariansmatrise $\sigma^2 I$, der I er identitetsmatrisa av storleik $n \times n$.

Vi ser på ein redusert modell som berre inkluderer dei r første kovariatane, der $r < p$. La X_0 vere designmatrisa som berre består av dei første r søylene i X. La $\hat{\boldsymbol{\beta}}_{(0)} = (X_0^{\mathrm{T}} X_0)^{-1} X_0^{\mathrm{T}} \boldsymbol{Y}$ vere minstekvadratestimatoren av parametrene i undermodellen, og la $\hat{\boldsymbol{\beta}}_0$ vere $\hat{\boldsymbol{\beta}}_{(0)}$ utvida med nullar, slik at $\hat{\boldsymbol{\beta}}_0$ har lengd p, det vil seie $\hat{\boldsymbol{\beta}}_0^{\mathrm{T}} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_{(0)}^{\mathrm{T}} & \mathbf{0}^{\mathrm{T}} \end{pmatrix}$, der $\mathbf{0}$ er nullvektoren av lengd $p - r$.

Vi ønsker å måle kor god undermodellen er ved

$$J_0 = \frac{1}{\sigma^2} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_0)^{\mathrm{T}} X^{\mathrm{T}} X (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_0),$$

som vart kalla «the scaled sum of squared errors» av Mallows. Det er sjølvsagt eit problem at parametrane $\boldsymbol{\beta}$ (og σ^2) er ukjende. Vi antar at den opphavlege modellen er «sann», slik at $E\boldsymbol{Y} = X\boldsymbol{\beta}$.

a) Kva er kovariansmatrisa til $\hat{\boldsymbol{\beta}}_{(0)}$? Finn kovariansmatrisa $\mathrm{Cov}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_0)$ til $\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_0$. Vis at trasen til $\frac{1}{\sigma^2} X^{\mathrm{T}} X \,\mathrm{Cov}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_0)$ er r.

La $H_0 = X_0 (X_0^{\mathrm{T}} X_0)^{-1} X_0^{\mathrm{T}}$ vere projeksjonsmatrisa som projiserer på søylerommet til X_0 («hattematrisa» til undermodellen).

b) Vis at $E(X(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_0)) = (I - H_0) X\boldsymbol{\beta}$. Finn EJ_0. (Vink: Bruk traseformelen, $E(\boldsymbol{Z}^{\mathrm{T}} A \boldsymbol{Z}) = \mathrm{tr}(A \,\mathrm{Cov}\, \boldsymbol{Z}) + (E\boldsymbol{Z}^{\mathrm{T}}) A (E\boldsymbol{Z})$.)

La $\mathrm{SSE}_0 = \boldsymbol{Y}^{\mathrm{T}} (I - H_0) \boldsymbol{Y}$ vere feilkvadratsummen (residualkvadratsummen) til undermodellen.

c) Vis at han har forventningsverdi $E\,\mathrm{SSE}_0 = (n - r)\sigma^2 + \boldsymbol{\beta}^{\mathrm{T}} X^{\mathrm{T}} (I - H_0) X\boldsymbol{\beta}$. Kombiner uttrykka for EJ_0 and $E\,\mathrm{SSE}_0$ for å vise at $EJ_0 = \frac{1}{\sigma^2} E\,\mathrm{SSE}_0 - n + 2r$. Diskuter kort korleis dette motiverer bruk av Mallows sin C_P-observator i seleksjon av undermodellar.