NTNU

Kunnskap for en bedre verden

Institutt for matematiske fag

Eksamensoppgave i **TMA4267 Lineære statistiske modeller**

Faglig kontakt under eksamen: Øyvind Bakke

Tlf: 73 59 81 26, 990 41 673

Eksamensdato: 22. mai 2015

Eksamenstid (fra–til): 9.00–13.00

Hjelpemiddelkode/Tillatte hjelpemidler: Gult, stemplet A4-ark med egne håndskrevne notater, bestemt enkel kalkulator, *Tabeller og formler i statistikk* (Tapir forlag), *Matematisk formelsamling* (K. Rottmann)

Annen informasjon:
I vurderingen teller hvert av de åtte bokstavpunktene likt.

Målform/språk: bokmål

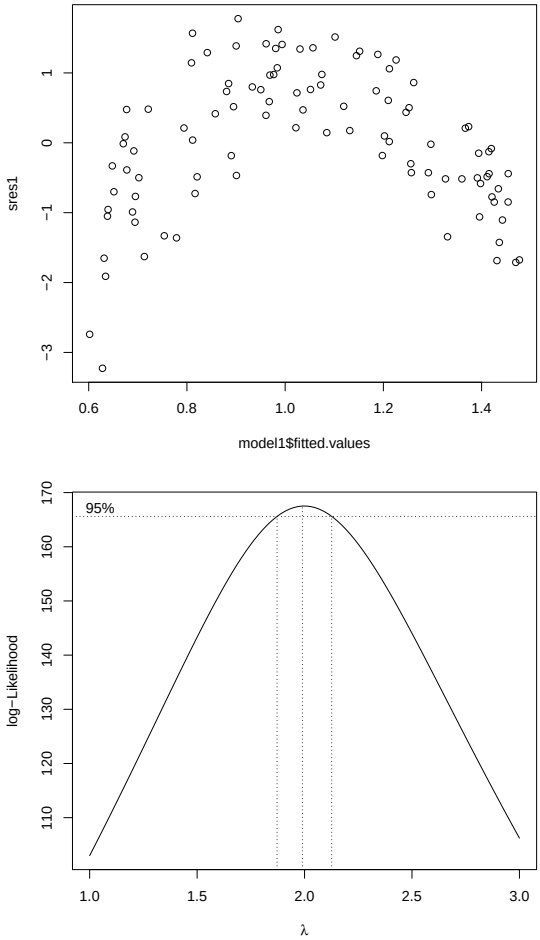
Antall sider: 4

Antall sider 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))
```



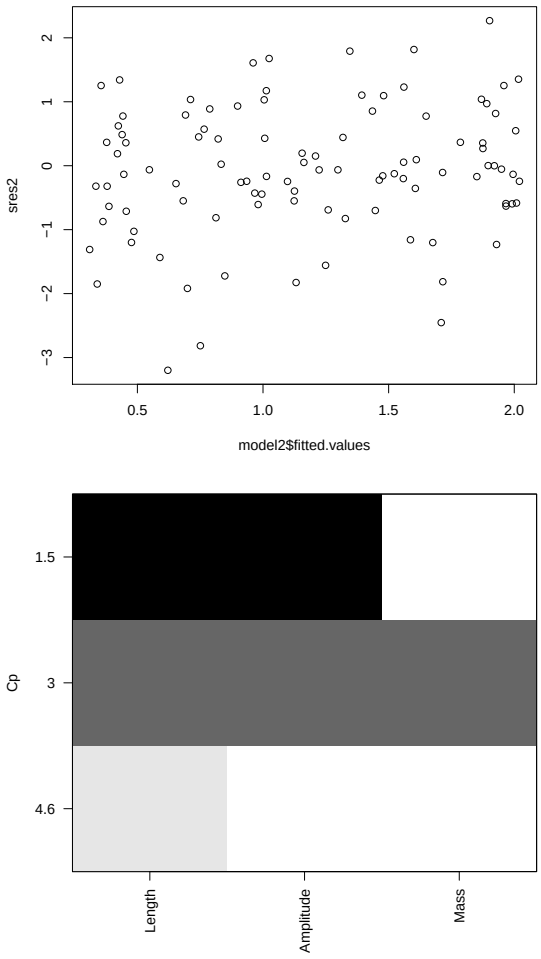
Figur 1: Modellen i oppgave 1a: R-kode og -utskrift (venstre), residualplott (oppe til høyre) og Box–Cox-plott (nede til høyre).

Oppgave 1

Svingetida for en pendel ble studert, og 100 kombinasjoner av pendelens lengde (målt i cm), amplitude (svingningenes største utslag fra den loddrette likevektslinja til en av sidene, målt i radianer) og masse (kg) ble variert. En multippel regresjonsmodell ble tilpasset. Figur 1 viser R-kode og -utskrift, et residualplott og et Box–Cox-plott.

a) Skriv opp den tilpassede regresjonsmodellen, og kommenter modelltilpasningen kort. Hvilke konklusjoner kan du trekke fra residualplottet? Foreslå en 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 oppgave 1b: R-kode og -utskrift (venstre), residualplott (oppe til høyre) og en grafisk tabell over beste delmengder, der Mallows' C_P brukes som observator for å ordne modellene (nede til høyre). Merk at opplysningene i den grafiske tabellen også er i R-utskriften.

Tilnærmingsformelen $T \approx 2\pi\sqrt{L/g}$ for svingetida T for en pendel, der L er lengden og $g \approx 9.8$ m/s^2 er tyngdeakselerasjonen, viser at det kan være rimelig å bruke kvadratet av svingetida istedenfor svingetida som responsvariabel i en regresjonsmodell, og også at konstantleddet (skjæringspunktet) sløyfes. Figur 2 viser R-kode og -utskrift, et residualplott og et plott av beste delmengde-seleksjon basert på Mallows' C_P for slike modeller.

b) Foretrekker du den opprinnelige modellen eller den nye modellen som nettopp er beskrevet? Hvilken undermodell av den nye modellen ville du velge, hvis du skulle velge en? Begrunn kort svarene 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 oppgave 1c: R-kode og -utskrift.

Den mer 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 være rimelig å bruke en tredje modell, der både responsvariabelen og kovariatene er transformert. Figur 3 viser R-kode og -utskrift.

c) Hvordan stemmer estimatene av koeffisientene med den fysiske modellen gitt over? Finn et estimat av g, tyngdeakselerasjonen, og et 95 %-konfidensintervall for g.

Oppgave 2

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

Anta videre at søylene i X er ortogonale.

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

I et toveis faktorielt ikke-replikert 2^2-forsøk er nivåene kodet -1 og 1. Responsvektoren er $(6 \;\; 4 \;\; 10 \;\; 7)^{\mathrm{T}}$, som svarer til nivåene $(-1 \;\; 1 \;\; -1 \;\; 1)^{\mathrm{T}}$ av første faktor og $(-1 \;\; -1 \;\; 1 \;\; 1)^{\mathrm{T}}$ av andre faktor.

b) Estimer interaksjonseffekten (to ganger koeffisienten) av de to faktorene.

Oppgave 3

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

Vi betrakter en redusert modell som bare inkluderer de r første kovariatene, der $r < p$. La X_0 være designmatrisen som bare består av de første r søylene i X. La $\hat{\boldsymbol{\beta}}_{(0)} = (X_0^{\mathrm{T}} X_0)^{-1} X_0^{\mathrm{T}} \boldsymbol{Y}$ være minstekvadratestimatoren av parametrene i undermodellen, og la $\hat{\boldsymbol{\beta}}_0$ være $\hat{\boldsymbol{\beta}}_{(0)}$ utvidet med nuller, slik at $\hat{\boldsymbol{\beta}}_0$ har lengde p, det vil si $\hat{\boldsymbol{\beta}}_0^{\mathrm{T}} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_{(0)}^{\mathrm{T}} & \mathbf{0}^{\mathrm{T}} \end{pmatrix}$, der $\mathbf{0}$ er nullvektoren av lengde $p - r$.

Vi ønsker å måle hvor 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 ble kalt «the scaled sum of squared errors» av Mallows. Det er selvfølgelig et problem at parametrene $\boldsymbol{\beta}$ (og σ^2) er ukjente. Vi antar at den opprinnelige modellen er «sann», slik at $E\boldsymbol{Y} = X\boldsymbol{\beta}$.

a) Hva er kovariansmatrisen til $\hat{\boldsymbol{\beta}}_{(0)}$? Finn kovariansmatrisen $\mathrm{Cov}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_0)$ 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 $H_0 = X_0(X_0^{\mathrm{T}} X_0)^{-1} X_0^{\mathrm{T}}$ være projeksjonsmatrisen som projiserer på søylerommet til X_0 («hattematrisen» 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 $\mathrm{SSE}_0 = \boldsymbol{Y}^{\mathrm{T}}(I - H_0)\boldsymbol{Y}$ være feilkvadratsummen (residualkvadratsummen) til undermodellen.

c) Vis at den har forventningsverdi $E\,\mathrm{SSE}_0 = (n - r)\sigma^2 + \boldsymbol{\beta}^{\mathrm{T}} X^{\mathrm{T}} (I - H_0) X \boldsymbol{\beta}$. Kombiner uttrykkene 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 hvordan dette motiverer bruk av Mallows' C_P-observator i seleksjon av undermodeller.