NTNU

Kunnskap for en bedre verden

Institutt for matematiske fag

Eksamensoppgave i **TMA4267 Lineære statistiske modeller**

Faglig kontakt under eksamen: Mette Langaas

Tlf: 988 47 649

Eksamensdato: 22. mai 2014

Eksamenstid (fra–til): 09.00-–13.00

Hjelpemiddelkode/Tillatte hjelpemidler: C: Gult, stemplet A5-ark med dine egne håndskrevne notater, Tabeller og formler i statistikk (Tapir forlag), K. Rottmann: Matematisk formelsamling. Bestemt kalkulator.

Målform/språk: bokmål

Antall sider: 8

Antall sider vedlegg: 0

Kontrollert av:

Dato Sign

Oppgave 1 Stokastisk vektor

La $\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$ være en stokastisk vektor med forventningsverdi $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{X}) = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$ og kovariansmatrise $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{I} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$. La videre $\boldsymbol{A} = \begin{pmatrix} \frac{2}{3} & -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} \end{pmatrix}$ være en matrise av konstanter.

Definer $\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} = \boldsymbol{AX}$.

a) Finn $\mathrm{E}(\boldsymbol{Y})$ og $\mathrm{Cov}(\boldsymbol{Y})$.
Er X_1 og X_2 uavhengige?
Er Y_1 og Y_2 uavhengige? Begrunn svarene.
Finn forventningsverdien av $\boldsymbol{X}^T\boldsymbol{AX}$.

Anta i resten av denne oppgaven at $\boldsymbol{X}$ er trivariat normalfordelt med forventningsverdi $\boldsymbol{\mu}$ og kovariansmatrise $\boldsymbol{\Sigma}$ gitt i begynnelsen av denne oppgaven.

b) Vis at $\boldsymbol{A}$ er en symmetrisk projeksjonsmatrise. Finn rangen av $\boldsymbol{A}$.
Utled fordelingen til $\boldsymbol{X}^T\boldsymbol{AX}$.
Finn sannsynligheten for at $\boldsymbol{X}^T\boldsymbol{AX}$ er mindre enn 6.

Oppgave 2 Galápagos-arter

Dette datasettet omhandler antall skilpaddearter på de forskjellige Galápagos-øyene, og er tatt fra boka «Practical Regression and Anova using R» av Julian J. Faraway.

Datasettet inneholder målinger på 30 øyer, og vi studerer følgende 6 variabler:

- `Species`: Antall arter av skilpadde funnet på øya.
- `Area`: Arealet av øya (km^2).
- `Elevation`: Største høyde over havet på øya (m).
- `Nearest`: Avstanden til nærmeste øy (km).
- `Scruz`: Avstanden til øya Santa Cruz (km).
- `Adjacent`: Arealet av nærmeste øy (km^2).

Oppsummeringsobservatorer er gitt nedenfor for Galápagos-datasettet.

```
Summary statistics for the Galapagos data set
        Species      Area Elevation Nearest  Scruz Adjacent
Min.       2.00    0.0100     25.00    0.20   0.00     0.03
1st Qu.   13.00    0.2575     97.75    0.80  11.02     0.52
Median    42.00    2.5900    192.00    3.05  46.65     2.59
Mean      85.23  261.7000    368.00   10.06  56.98   261.10
3rd Qu.   96.00   59.2400    435.20   10.02  81.08    59.24
Max.     444.00 4669.0000   1707.00   47.40 290.20  4669.00
```

En multippel lineær regresjonsmodell ble tilpasset Galápagos-datasettet, med `Species` som respons og de resterende fem variablene som kovariater. Kall dette modell A. Kode og utskrift fra R finnes i figur 1 og tilhørende plott i figur 2 og 3.

a) Skriv ned den tilpassede regresjonsmodellen, og kommenter modelltilpasningen *kort*. Hvilke konklusjoner kan du trekke fra residualplottene og Box–Cox-transformasjonsplottet?

Kubikkrot-transformasjonen av `Species` vil fra nå av bli brukt som respons i en ny multippel lineær regresjonsmodell, med de samme fem kovariatene som i modell A. Kall dette modell B. Kode og utskrift fra R finnes i figur 4 og tilhørende plott i figur 5.

b) I utskriften i figur 4 for tilpasning av modell B er fire tallverdier byttet ut med spørsmålstegn. Regn ut tallverdier for hver av disse, og forklar hva hvert av tallene betyr.

Foretrekker du modell B fremfor modell A? Begrunn svaret *kort*.

c) Resultatene av en «best subset selection» er gitt i figur 6, der også tallverdiene av R^2 og R^2_{adj} er listet opp for hver av de fem modellene.

Skriv ned definisjonen av R^2 and R^2_{adj} og forklar hvordan du kan bruke disse til å sammenligne de forskjellige modellene.

Velg den «beste» av disse fem modellene. Begrunn valget ditt.

Til slutt vises resultatene av å tilpasse lasso-regresjon til Galápagos-datasettet i figur 7.

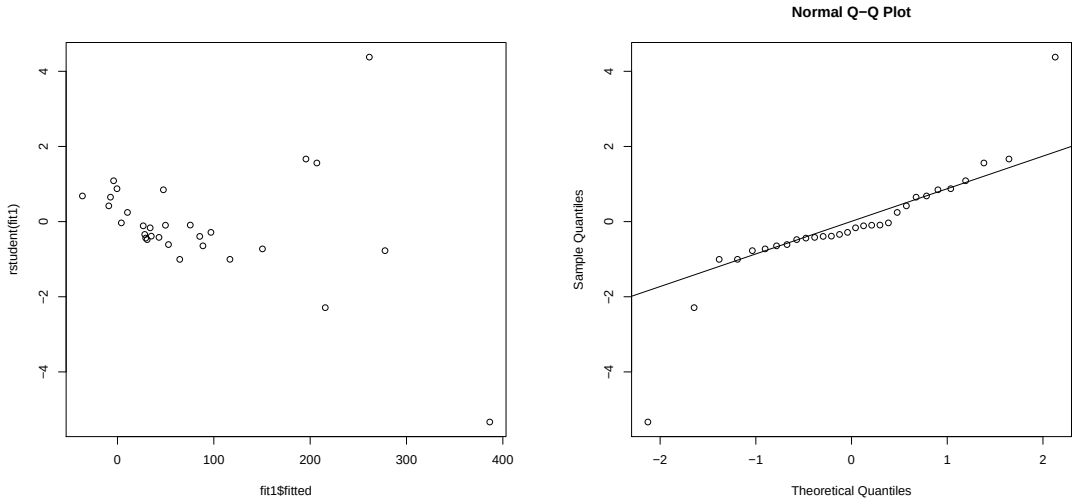
Forklar hva som skiller lasso-regresjon fra minste kvadratsum-regresjon.

Kan R^2_{adj} brukes for å velge straffeparameter for lasso? Begrunn svaret.

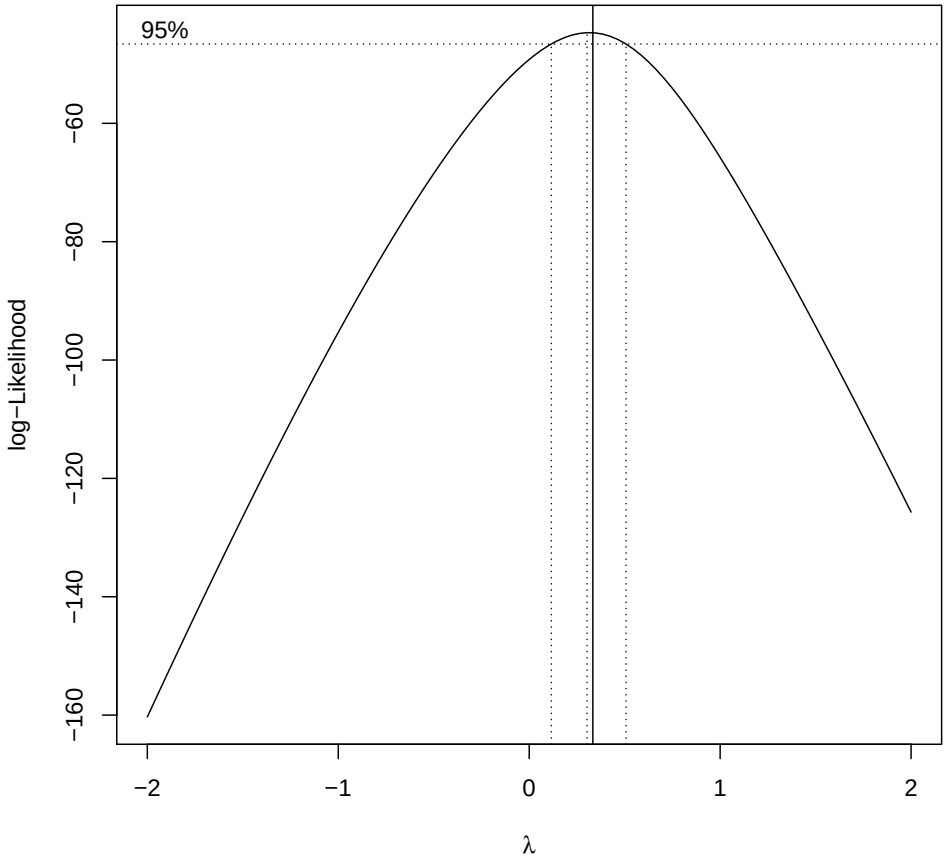
Skriv ned den tilpassede regresjonsmodellen for lassoregresjonen.

```
> fit1 <- lm(Species~Area+Elevation+Nearest+Scruz+Adjacent,
data=gala)
> summary(fit1)

Call:
lm(formula = Species ~ Area + Elevation +
    Nearest + Scruz + Adjacent,
    data = gala)

Residuals:
     Min       1Q   Median       3Q      Max
-111.679  -34.898   -7.862   33.460  182.584

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.068221  19.154198   0.369 0.715351
Area        -0.023938   0.022422  -1.068 0.296318
Elevation    0.319465   0.053663   5.953 3.82e-06 ***
Nearest      0.009144   1.054136   0.009 0.993151
Scruz       -0.240524   0.215402  -1.117 0.275208
Adjacent    -0.074805   0.017700  -4.226 0.000297 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 60.98 on 24 degrees of freedom
Multiple R-squared:  0.7658,        Adjusted R-squared:  0.7171
F-statistic:  15.7 on 5 and 24 DF,  p-value: 6.838e-07

> plot(fit1$fitted,rstudent(fit1))
> qqnorm(rstudent(fit1))
> qqline(rstudent(fit1))
> ad.test(rstudent(fit1))
        Anderson-Darling normality test
data:  rstudent(fit1)
A = 1.7071, p-value = 0.0001729
> boxcox(fit1)
> abline(v=1/3,lty=1)
```

Figur 1: Utskrift fra statistisk analyse for modell A for Galápagos-datasettet.



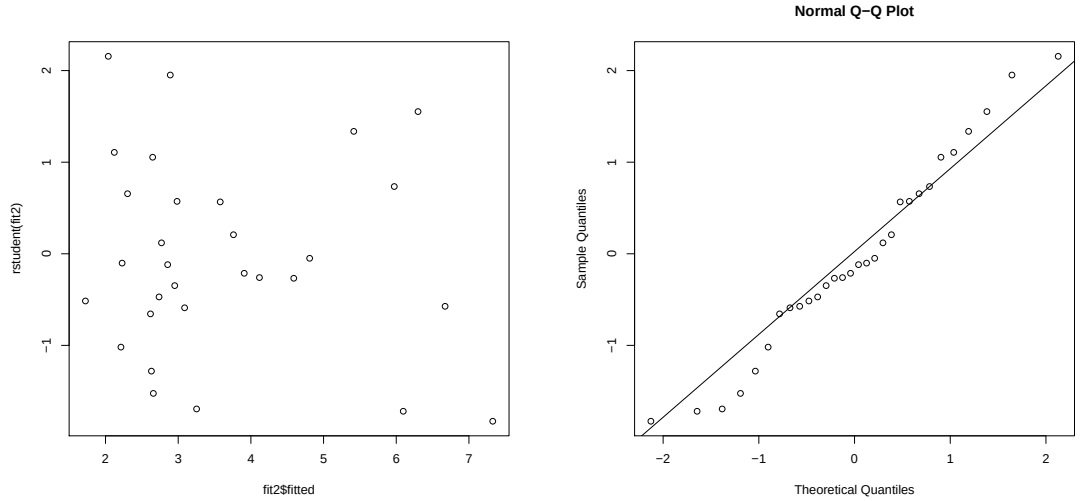
Figur 2: Residualplott (studentiserte residualer mot tilpassede verdier til venstre, normalplott basert på studentiserte residualer til høyre) for modell A for Galápagos-datasettet.



Figur 3: Box–Cox-tranformasjonsplott basert på modell A for Galápagos-datasettet.

```
> fit2 <- lm(Species^(1/3)~Area+Elevation+Nearest+Scruz+Adjacent,
x=TRUE,data=gala)
> summary(fit2)

Call:
lm.default(formula = Species^(1/3) ~ Area + Elevation + Nearest +
    Scruz + Adjacent, data = gala, x = TRUE)

Residuals:
     Min       1Q   Median       3Q      Max
-1.54306 -0.47863 -0.08499  0.56349  1.83283

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)         ?  0.3052013   7.365 1.32e-07
Area       -0.0007349  0.0003573  -2.057        ?
Elevation   0.0054510  0.0008551   6.375 1.37e-06
Nearest     0.0118152          ?   0.703  0.48855
Scruz      -0.0045951  0.0034322  -1.339  0.19317
Adjacent   -0.0010597  0.0002820  -3.757  0.00097
---
Residual standard error: 0.9716 on 24 degrees of freedom
Multiple R-squared:  0.7543,        Adjusted R-squared:  ?
F-statistic: 14.74 on 5 and 24 DF,  p-value: 1.192e-06

> plot(fit2$fitted,rstudent(fit2))
> qqnorm(rstudent(fit2))
> qqline(rstudent(fit2))
> ad.test(rstudent(fit2))
        Anderson-Darling normality test
data:  rstudent(fit2)
A = 0.2639, p-value = 0.6738
```

Figur 4: Utskrift fra R for tilpassing av den multiple lineære regresjonsmodellen B for Galápagos-datasettet. Responsen er kubikkrota av `Species`.



Figur 5: Residualplott (studentiserte residualer mot tilpassede verdier til venstre, normalplott basert på studentiserte residualer til høyre) for modell B (kubikkrot av `Species`) for Galápagos-datasettet.

```
> x <- fit2$x[,-1]
> y <- gala$Species^(1/3)
> library(leaps)
> bests <- regsubsets(x,y)
> sumbests <- summary(bests)
> sumbests
Subset selection object
5 Variables  (and intercept)
1 subsets of each size up to 5
Selection Algorithm: exhaustive
         Area Elevation Nearest Scruz Adjacent
1  ( 1 ) " "  "*"       " "     " "   " "
2  ( 1 ) " "  "*"       " "     " "   "*"
3  ( 1 ) "*"  "*"       " "     " "   "*"
4  ( 1 ) "*"  "*"       " "     "*"   "*"
5  ( 1 ) "*"  "*"       "*"     "*"   "*"
> plot(1:5, sumbests$rsq,type="l") #solid line
> lines(1:5, sumbests$adjr2,lty=2) #dashed line
> sumbests$rsq # R^2
[1] 0.5570784 0.6893784 0.7356845 0.7492704 0.7543353
> sumbests$adjr2 # R^2_adjusted
[1] 0.5412597 0.6663694 0.7051866 0.7091536 0.7031552
```

Figur 6: Utskrift fra R for tilpasning av «best subset selection» til Galápagos-datasettet. Responsen er kubikkrota av `Species`.

```
> library(glmnet)
> fit.lasso=glmnet(x,y)
> cv.lasso=cv.glmnet(x,y)
> coef(cv.lasso)
6 x 1 sparse Matrix of class "dgCMatrix"
                       1
(Intercept) 3.5388701794
Area        .
Elevation   0.0002804519
Nearest     .
Scruz       .
Adjacent    .
```

Figur 7: Utskrift fra R etter tilpasning av lasso-regresjon til Galápagos-datasettet. Responsen er kubikkrota av `Species`.

Oppgave 3 Forsøksplanlegging

I en pilotstudie med fire faktorer A, B, C og D ble de 8 forsøkene listet nedenfor utført.

	A	B	C	D
1	−1	−1	−1	1
2	1	−1	−1	1
3	−1	1	−1	−1
4	1	1	−1	−1
5	−1	−1	1	−1
6	1	−1	1	−1
7	−1	1	1	1
8	1	1	1	1

a) Hvilken type forsøk er dette?

Hva er generatoren og den definerende relasjonen for forsøket?

Hvilken resolusjon har forsøket?

Skriv ned alias-strukturen for forsøket.

Oppgave 4 Multippel lineær regresjon

I matrisenotasjon kan den klassiske multiple lineære regresjonsmodellen (MLR) skrives som

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

der $\boldsymbol{Y}$ er en n-dimensjonal stokastisk søylevektor, $\boldsymbol{X}$ er en fast designmatrise med n rader og p søyler, $\boldsymbol{\beta}$ er en ukjent p-dimensjonal vektor av regresjonskoeffisienter, og $\boldsymbol{\varepsilon}$ er en n-dimensjonal søylevektor av stokastisk støy.

Videre antar vi generelt i den klassiske multiple lineære modellen at vektoren $\boldsymbol{\varepsilon}$ av stokastisk støy er multivariat normalfordelt med forventningsverdi $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ og kovariansmatrise $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}$, der $\boldsymbol{I}$ er $n \times n$-identitetsmatrisen.

Vi vil nå se å en noe endret situasjon. Anta at $\boldsymbol{\varepsilon}$ er multivariat normalfordelt med forventningsverdi $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ og kovariansmatrise $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{V}$, der $\boldsymbol{V}$ er en kjent postivt definitt $n \times n$-matrise. De ukjente paramtrene i denne modellen er regresjonskoeffisientene $\boldsymbol{\beta}$ og variansparameteren σ^2.

a) Skriv ned og forklar definisjonen av den inverse kvadratrotmatrisen $\boldsymbol{V}^{-\frac{1}{2}}$.

Bruk den inverse kvadratrotmatrisen for å definere tre nye størrelser

$$\begin{aligned}\boldsymbol{Y}^* &= \boldsymbol{V}^{-\frac{1}{2}}\boldsymbol{Y},\\ \boldsymbol{X}^* &= \boldsymbol{V}^{-\frac{1}{2}}\boldsymbol{X},\\ \boldsymbol{\varepsilon}^* &= \boldsymbol{V}^{-\frac{1}{2}}\boldsymbol{\varepsilon}.\end{aligned}$$

Bruk disse nye størrelsene sammen med minste kvadraters metode for å utlede en forventningsrett estimator for $\boldsymbol{\beta}$, uttrykt ved $\boldsymbol{X}$, $\boldsymbol{V}$ og $\boldsymbol{Y}$.

Vis at estimatoren er forventningsrett.

Er den vanlige minste kvadratsum-estimatoren $\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ forventningsrett i denne modellen? Kommenter det du kommer fram til.

Vi går tilbake til den klassiske MLR med identisk normalfordelte stokastiske støyledd, $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}$, men ser nå på feilspesifikasjon av $\mathrm{E}(\boldsymbol{Y})$. Anta at den sanne modellen er

$$\begin{aligned}\boldsymbol{Y} &= \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon},\\ \boldsymbol{\varepsilon} &\sim N_n(\mathbf{0}, \sigma^2\boldsymbol{I}),\end{aligned} \tag{1}$$

der vi har partisjonert designmatrisen i to deler $\boldsymbol{X}_1$ ($n \times p_1$) og $\boldsymbol{X}_2$ ($n \times p_2$), og $\boldsymbol{\beta}_1$ og $\boldsymbol{\beta}_2$ er ukjente p_1- og p_2-dimensjonale vektorer av regresjonskoeffisienter ($p = p_1 + p_2$).

Anta at vi ser bort fra kovariatene i $\boldsymbol{X}_2$ og tilpasser modellen

$$\begin{aligned}\boldsymbol{Y} &= \boldsymbol{X}_1\boldsymbol{\alpha}_1 + \boldsymbol{\delta},\\ \boldsymbol{\delta} &\sim N_n(\mathbf{0}, \tau^2\boldsymbol{I}).\end{aligned} \tag{2}$$

Her er $\boldsymbol{\alpha}_1$ brukt istedenfor $\boldsymbol{\beta}_1$ for å understreke at $\boldsymbol{\alpha}_1$ (og estimater av denne) generelt er forskjellig fra $\boldsymbol{\beta}_1$ i den sanne modellen.

Minste kvadratsum-estimatoren for modell (2) er $\hat{\boldsymbol{\alpha}}_1 = (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{Y}$.

b) Finn forventningsverdien og kovariansmatrisen til $\hat{\boldsymbol{\alpha}}_1$ under den sanne modellen (1).

Under hvilke forutsetninger er $\hat{\boldsymbol{\alpha}}_1$ en forventningsrett estimator av $\boldsymbol{\beta}_1$? Begrunn svaret.