NTNU

Kunnskap for ei betre verd

Institutt for matematiske fag

# Eksamensoppgåve i **TMA4267 Lineære statistiske modellar**

**Fagleg kontakt under eksamen:** Mette Langaas

**Tlf:** 988 47 649

**Eksamensdato:** 22. mai 2014

**Eksamenstid (frå–til):** 09.00-–13.00

**Hjelpemiddelkode/Tillatne hjelpemiddel:** C: Gult, stempla A5-ark med dine eigne handskrivne notat, Tabeller og formler i statistikk (Tapir forlag), K. Rottmann: Matematisk formelsamling. Bestemd kalkulator.

**Målform/språk:** nynorsk

**Sidetal:** 8

**Sidetal vedlegg:** 0

**Kontrollert av:**

Dato Sign

## Oppgåve 1 Stokastisk vektor

La $\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$ vere ein 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 vidare $\boldsymbol{A} = \begin{pmatrix} \frac{2}{3} & -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} \end{pmatrix}$ vere ei matrise av konstantar.

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? Grunngje svara.
Finn forventningsverdien av $\boldsymbol{X}^T\boldsymbol{AX}$.

Anta i resten av denne oppgåva at $\boldsymbol{X}$ er trivariat normalfordelt med forventningsverdi $\boldsymbol{\mu}$ og kovariansmatrise $\boldsymbol{\Sigma}$ gjevne i byrjinga av denne oppgåva.

**b)** Vis at $\boldsymbol{A}$ er ei symmetrisk projeksjonsmatrise. Finn rangen av $\boldsymbol{A}$.
Utlei fordelinga til $\boldsymbol{X}^T\boldsymbol{AX}$.
Finn sannsynet for at $\boldsymbol{X}^T\boldsymbol{AX}$ er mindre enn 6.

## Oppgåve 2 Galápagos-artar

Dette datasettet handlar om talet på skjelpaddeartar på dei ulike Galápagosøyane, og er teke frå boka «Practical Regression and Anova using R» av Julian J. Faraway.

Datasettet inneheld målingar på 30 øyar, og vi studerer desse 6 variablane:

- `Species`: Talet på artar av skjelpadde funne på øya.
- `Area`: Arealet av øya ($\mathrm{km}^2$).
- `Elevation`: Største høgd over havet på øya (m).
- `Nearest`: Avstanden til næraste øy (km).
- `Scruz`: Avstanden til øya Santa Cruz (km).
- `Adjacent`: Arealet av næraste øy ($\mathrm{km}^2$).

Oppsummeringsobservatorar er gjevne nedanfor 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
```

Ein multippel lineær regresjonsmodell vart tilpassa Galápagos-datasettet, med `Species` som respons og dei resterande fem variablane som kovariatar. Kall dette modell A. Kode og utskrift frå R finst i figur 1 og tilhøyrande plott i figur 2 og 3.

**a)** Skriv ned den tilpassa regresjonsmodellen, og kommenter modelltilpassinga *kort*. Kva konklusjonar kan du trekkje frå residualplotta og Box–Cox-transformasjonsplottet?

Kubikkrot-transformasjonen av `Species` vil frå no av bli brukt som respons i ein ny multippel lineær regresjonsmodell, med dei same fem kovariatane som i modell A. Kall dette modell B. Kode og utskrift frå R finst i figur 4 og tilhøyrande plott i figur 5.

**b)** I utskrifta i figur 4 for tilpassing av modell B er fire talverdiar bytte ut med spørjeteikn. Rekn ut talverdiar for kvar av desse, og forklar kva kvart av tala tyder.

Føretrekkjer du modell B framfor modell A? Grunngje svaret *kort*.

**c)** Resultata av ein «best subset selection» er gjeve i figur 6, der òg talverdiane av $R^2$ og $R^2_{\text{adj}}$ er lista opp for kvar av dei fem modellane.

Skriv ned definisjonen av $R^2$ and $R^2_{\text{adj}}$ og forklar korleis du kan bruke desse til å samanlikne dei ulike modellane.

Vel den «beste» av desse fem modellane. Grunngje valet ditt.

Til slutt blir resultata av å tilpasse lasso-regresjon til Galápagos-datasettet vist i figur 7.

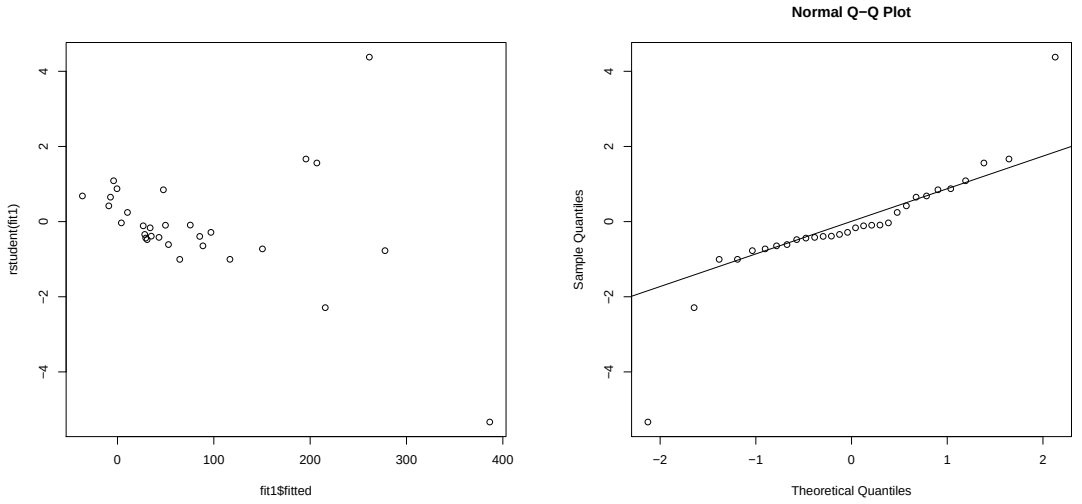
Forklar kva som skil lasso-regresjon frå minste kvadratsum-regresjon.

Kan $R^2_{\text{adj}}$ brukast for å velje straffeparameter for lasso? Grunngje svaret.

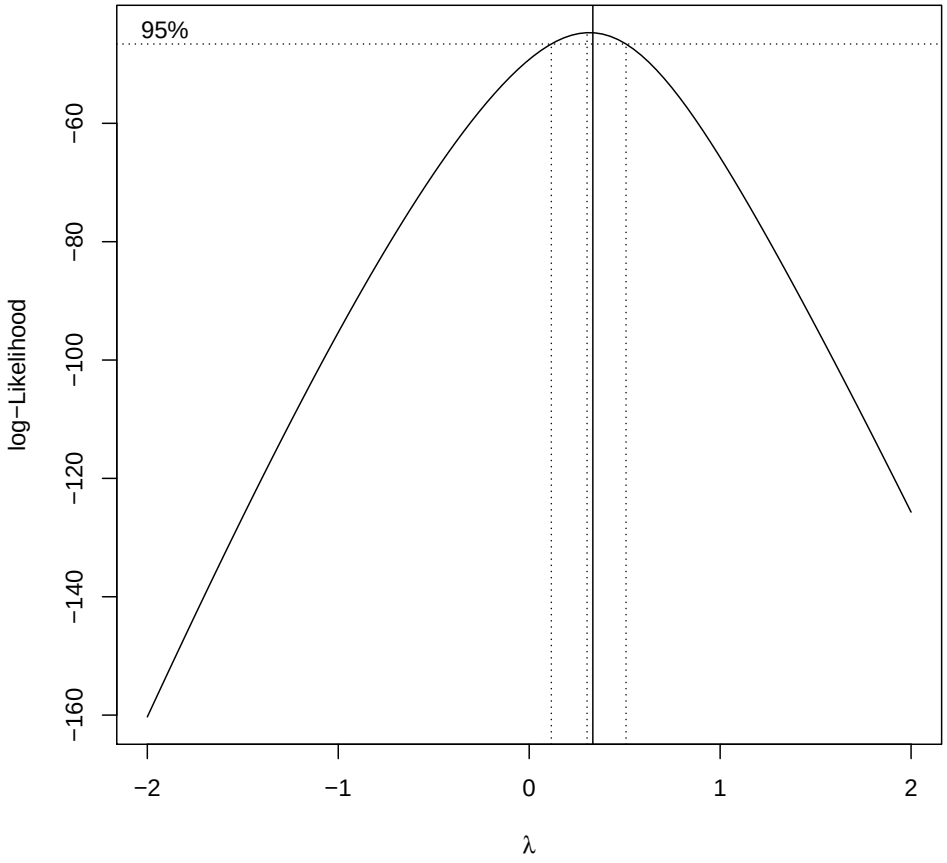
Skriv ned den tilpassa regresjonsmodellen for lassoregresjonen.

```
> 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 frå statistisk analyse for modell A for Galápagos-datasettet.



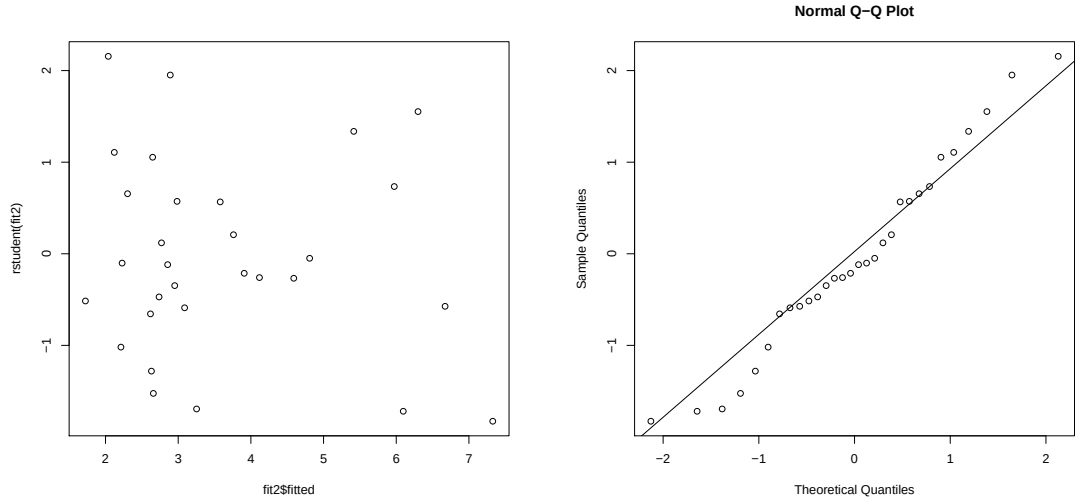

Figur 2: Residualplott (studentiserte residualar mot tilpassa verdiar til venstre, normalplott basert på studentiserte residualar til høgre) for modell A for Galápagos-datasettet.




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 frå R for tilpassing av den multiple lineære regresjonsmodellen B for Galápagos-datasettet. Responsen er kubikkrota av `Species`.




Figur 5: Residualplott (studentiserte residualar mot tilpassa verdiar til venstre, normalplott basert på studentiserte residualar til høgre) 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 frå R for tilpassing 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 frå R etter tilpassing av lasso-regresjon til Galápagos-datasettet. Responsen er kubikkrota av `Species`.

## Oppgåve 3 Forsøksplanlegging

I ein pilotstudie med fire faktorar A, B, C og D vart dei 8 forsøka lista nedanfor utførte.

| | 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)** Kva type forsøk er dette?

Kva er generatoren og den definerande relasjonen for forsøket?

Kva resolusjon har forsøket?

Skriv ned alias-strukturen for forsøket.

## Oppgåve 4 Multippel lineær regresjon

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

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

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

Vidare 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 $n \times n$-identitetsmatrisa.

Vi vil no sjå på ein litt endra 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 ei kjend postivt definitt $n \times n$-matrise. Dei ukjente parametrane i denne modellen er regresjonskoeffisientane $\boldsymbol{\beta}$ og variansparameteren $\sigma^2$.

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

Bruk den inverse kvadratrotmatrisa for å definere tre nye storleikar

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

Bruk desse nye storleikane saman med minste kvadrats metode for å utleie ein forventningsrett estimator for $\boldsymbol{\beta}$, uttrykt ved $\boldsymbol{X}$, $\boldsymbol{V}$ og $\boldsymbol{Y}$.

Vis at estimatoren er forventningsrett.

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

Vi går tilbake til den klassiske MLR med identisk normalfordelte stokastiske støyledd, $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}$, men ser no 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 designmatrisa i to delar $\boldsymbol{X}_1$ $(n \times p_1)$ og $\boldsymbol{X}_2$ $(n \times p_2)$, og $\boldsymbol{\beta}_1$ og $\boldsymbol{\beta}_2$ er ukjende $p_1$- og $p_2$-dimensjonale vektorar av regresjonskoeffisientar $(p = p_1 + p_2)$.

Anta at vi ser bort frå kovariatane i $\boldsymbol{X}_2$ og tilpassar 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 i staden for $\boldsymbol{\beta}_1$ for å understreke at $\boldsymbol{\alpha}_1$ (og estimat av denne) generelt er ulik $\boldsymbol{\beta}_1$ 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 kovariansmatrisa til $\hat{\boldsymbol{\alpha}}_1$ under den sanne modellen (1).

Under kva føresetnader er $\hat{\boldsymbol{\alpha}}_1$ ein forventningsrett estimator av $\boldsymbol{\beta}_1$? Grunngje svaret.