NTNU – Trondheim
Norwegian University of
Science and Technology

Department of Mathematical Sciences

Examination paper for **TMA4267 Linear Statistical Models**

Academic contact during examination: Mette Langaas

Phone: 988 47 649

Examination date: 22 May 2014

Examination time (from–to): 09:00–13:00

Permitted examination support material: C: Yellow, stamped A5 sheet with your own handwritten notes, Tabeller og formler i statistikk (Tapir forlag), K. Rottmann: Matematisk formelsamling. Specified calculator.

Language: English

Number of pages: 8

Number pages enclosed: 0

Checked by:

Date Signature

Problem 1 Random vector

Let $\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$ be a random vector with mean $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{X}) = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$ and covariance matrix $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{I} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$. Further, let $\boldsymbol{A} = \begin{pmatrix} \frac{2}{3} & -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} \end{pmatrix}$ be a matrix of constants.

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

a) Find $\mathrm{E}(\boldsymbol{Y})$ and $\mathrm{Cov}(\boldsymbol{Y})$.
Are X_1 and X_2 independent?
Are Y_1 and Y_2 independent? Justify your answers.
Find the mean of $\boldsymbol{X}^T\boldsymbol{AX}$.

For the rest of this problem assume that $\boldsymbol{X}$ is trivariate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ given in the start of this problem.

b) Show that $\boldsymbol{A}$ is a symmetric projection matrix. Find the rank of $\boldsymbol{A}$.
Derive the distribution of $\boldsymbol{X}^T\boldsymbol{AX}$.
Find the probability that $\boldsymbol{X}^T\boldsymbol{AX}$ is smaller than 6.

Problem 2 Galapagos species

This data set concerns the number of species of tortoise on the various Galapagos Islands, and is taken from the book "Practical Regression and Anova using R" by Julian J. Faraway.

The data set contains measurements on 30 islands, and we study the following 6 variables:

- `Species`: The number of species of tortoise found on the island.
- `Area`: The area of the island (km^2).
- `Elevation`: The highest elevation of the island (m).
- `Nearest`: The distance from the nearest island (km).
- `Scruz`: The distance from Santa Cruz island (km).
- `Adjacent`: The area of the adjacent island (km^2).

Summary statistics are given below for the Galapagos data set.

```
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
```

A multiple linear regression model was fitted to the Galapagos data set, with `Species` as response and the remaining five variables as covariates. Call this Model A. Code and printout from R is found in Figure 1 and accompanying plots in Figures 2 and 3.

a) Write down the fitted regression model, and comment *briefly* on the model fit. What conclusions can you draw from the residual plots and the Box–Cox transformation plot?

The cube root transformation of `Species` will from now on be used as response in a new multiple linear regression model, with the same five covariates as for Model A. Call this Model B. Code and printout from R is found in Figure 4 and accompanying plots in Figure 5.

b) In the printout in Figure 4 from fitting Model B four numerical values are substituted by question marks. Calculate numerical values for each of these, and explain what each of the numbers means.

Would you prefer Model B to Model A? Justify *briefly* your answer.

c) The results from performing best subset selection is reported in Figure 6, where also R^2 and R^2_{adj} is listed numerically for the five models reported.

Write down the definition for R^2 and R^2_{adj} and explain how you can use these to compare the different models.

Choose the "best" out of these five models. Justify your choice.

Finally, the results from fitting lasso regression to the Galapagos data set is reported in Figure 7.

Explain what the lasso regression does differently from least squares regression.

Can R^2_{adj} be used to select the penalty parameter for lasso? Justify your answer.

Write down the fitted regression model for the lasso regression.

```
> 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)
```

Figure 1: Printout from statistical analyses for Model A for the Galapagos data set.

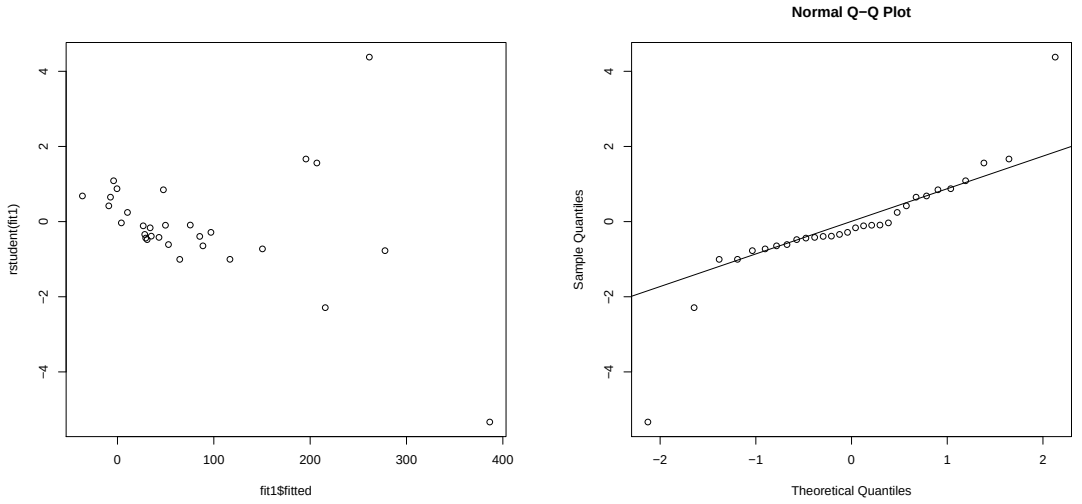

Figure 2: Residual plots (studentized residual versus fitted values in the left panel, normal plot based on studentized residuals in the right panel) for Model A for the Galapagos data set.

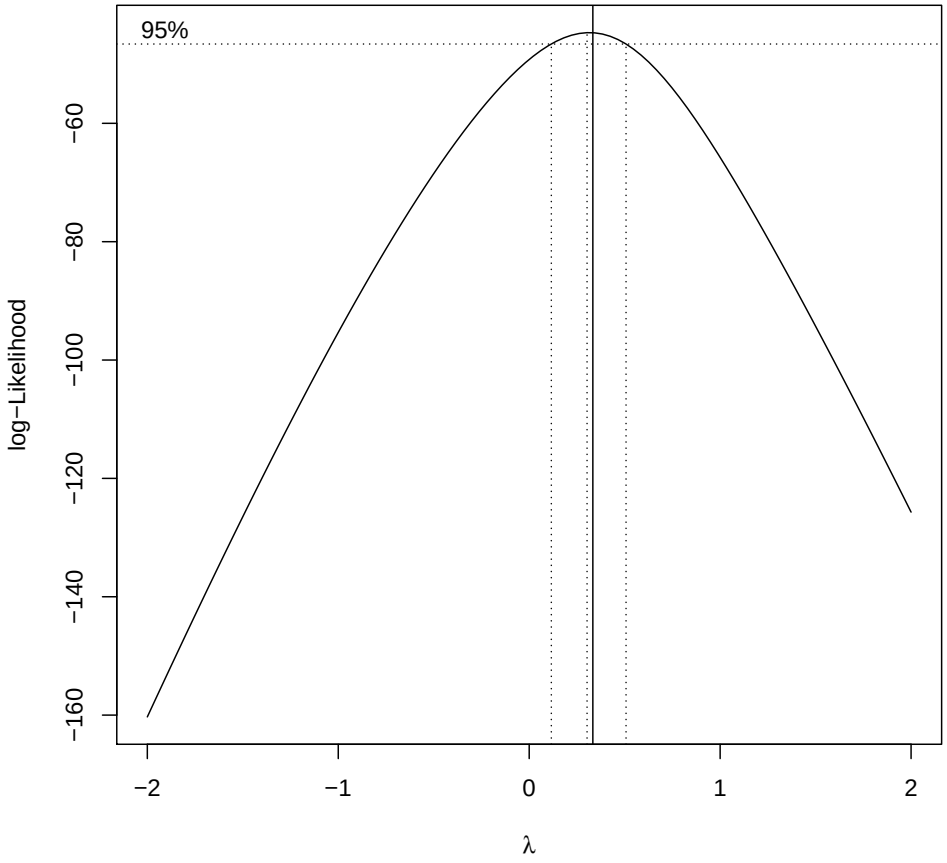

Figure 3: Box–Cox tranformation plot based on Model A for the Galapagos data set.

```
> 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
```

Figure 4: Printout from R of fitting the multiple linear regression model B for the Galapagos island data set. Response is cube root of `Species`.

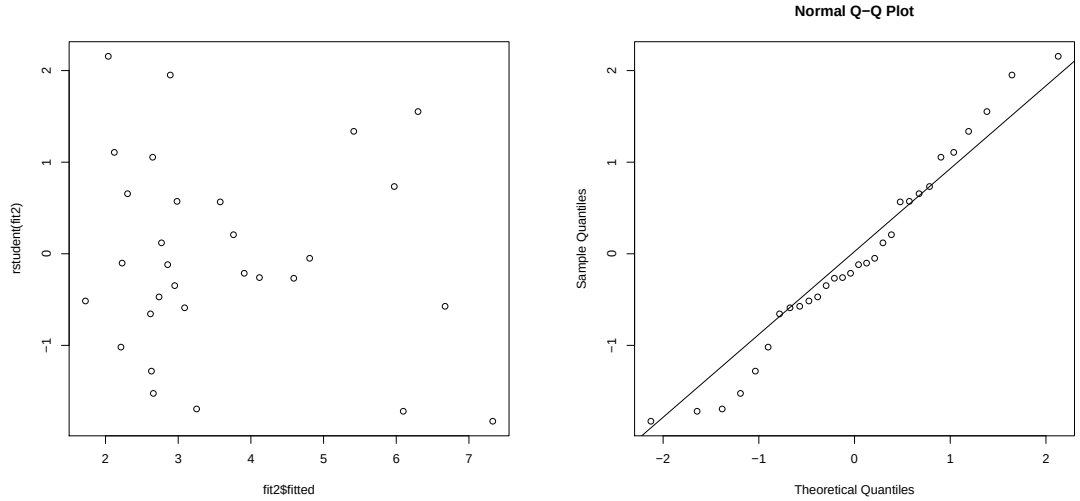

Figure 5: Residual plots (studentized residual versus fitted values in the left panel, normal plot based on studentized residuals in the right panel) for Model B (cube root of `Species`) for the Galapagos data set.

```
> 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
```

Figure 6: Printout from R of fitting best subset selection to the Galapagos island data set. Response is cube root of `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    .
```

Figure 7: Printout from R after fitting lasso regression to the Galapagos data-set. Response is cube root of `Species`.

Problem 3 Design of experiments

In a pilot study with four factors A, B, C and D, the 8 experiments listed below were run.

	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) What type of experiment is this?

What is the generator and the defining relation for the experiment?

What is the resolution of the experiment?

Write down the alias structure of the experiment.

Problem 4 Multiple linear regression

The classical multiple linear regression (MLR) model can be written in matrix notation as

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

where $\boldsymbol{Y}$ is an n-dimensional random column vector, $\boldsymbol{X}$ is a fixed design matrix with n rows and p columns, $\boldsymbol{\beta}$ is an unknown p-dimensional vector of regression coefficients and $\boldsymbol{\varepsilon}$ is an n-dimensional column vector of random errors.

Further, in the classical multiple linear model we generally assume that the vector of random errors $\boldsymbol{\varepsilon}$ is multivariate normal with mean $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ and covariance matrix $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}$, where $\boldsymbol{I}$ is the $n \times n$ identity matrix.

We will now study slightly different situation. Assume that $\boldsymbol{\varepsilon}$ is multivariate normal with mean $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ and covariance matrix $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{V}$, where $\boldsymbol{V}$ is a known positive definite $n \times n$ matrix. The unknown parameters in this model are the regression coefficients $\boldsymbol{\beta}$ and the variance parameter σ^2.

a) Write down and explain the definition of the inverse square root matrix $\boldsymbol{V}^{-\frac{1}{2}}$.

Use the inverse square root matrix to define three new quantities

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

Use these new quantities together with the method of least squares to derive an unbiased estimator for $\boldsymbol{\beta}$, in terms of $\boldsymbol{X}$, $\boldsymbol{V}$ and $\boldsymbol{Y}$.

Show that the estimator is unbiased.

Is the ordinary least square estimator $\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ unbiased in this model? Justify your answer. Comment on your findings.

We go back to the classical MLR with identically normally distributed random errors, $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}$, but now look at misspecification of $\mathrm{E}(\boldsymbol{Y})$. Suppose that the true model is

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

where we have partitioned the design matrix into two parts $\boldsymbol{X}_1$ $(n \times p_1)$ and $\boldsymbol{X}_2$ $(n \times p_2)$ and $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are unknown p_1- and p_2-dimensional vectors of regression coefficients $(p = p_1 + p_2)$.

Assume that we ignore the covariates in $\boldsymbol{X}_2$ and fit the model

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

Here $\boldsymbol{\alpha}_1$ is used in place of $\boldsymbol{\beta}_1$ to emphasize that $\boldsymbol{\alpha}_1$ (and estimates thereof) will in general be different from $\boldsymbol{\beta}_1$ in the true model.

The least squares estimator for model (2) is $\hat{\boldsymbol{\alpha}}_1 = (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{Y}$.

b) Find the mean and covariance matrix of $\hat{\boldsymbol{\alpha}}_1$ under the true model (1).

Under which conditions is $\hat{\boldsymbol{\alpha}}_1$ an unbiased estimator of $\boldsymbol{\beta}_1$? Justify your answer.