NTNU – Trondheim
Norwegian University of
Science and Technology

Department of Mathematical Sciences

Examination paper for **TMA4267 Linear Statistical Models**

Academic contact during examination:

Phone:

Examination date: August 2014

Examination time (from–to):

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

Language: English

Number of pages: 7

Number pages enclosed: 0

Checked by:

Date Signature

Problem 1 The Multivariate Normal Distribution

Let $\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ be a bivariate normal random vector with mean $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{X}) = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ and covariance matrix $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{X}) = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix}$.

a) Let $\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$, where $Y_1 = 3X_1 - 2X_2$ and $Y_2 = X_1 + X_2$.

What is the distribution of $\boldsymbol{Y}$?
Let $Z = X_1 + aX_2$. How can you choose a so that Z and Y_2 are independent?

In Figure 1 you find the eigenvalues and eigenvectors of the covariance matrix $\boldsymbol{\Sigma}$.

b) Let $f(\boldsymbol{x})$ denote the probability density function (pdf) of $\boldsymbol{X}$.

Describe the graph of the equation $f(\boldsymbol{x}) = d$, where $d > 0$ is a constant?
What value of d would give a graph where the probability that $\mathbf{X}$ is inside the area enclosed by the graph equals 95%?
Make a drawing of the graph, for the value of d found above.

```
> sigma <- matrix(c(1,0.5,0.5,2),ncol=2)
> eigen(sigma)
$values
[1] 2.2071068 0.7928932

$vectors
          [,1]       [,2]
[1,] 0.3826834 -0.9238795
[2,] 0.9238795  0.3826834
```

Figure 1: Eigenvalues and vectors of the covariance matrix of Problem 1b.

Problem 2 Predicting fat content in meat

The fat content of meat can be measured using analytical chemistry. However, this is a time consuming method. In an experimental setting researchers used near infrared transmission to measure absorbances in a 100 channel spectrum (wavelength range 850–1050 nm). This was done for 215 samples of finely chopped meat. For each sample the fat content was also measured. The aim of the experiment was to develop a prediction method for the fat content of meat, based on the 100 absorbances.

a) A multiple linear regression (MLR) model was fit to the data set of 215 samples, with the 100 absorbances as covariates (named `xV1`, `xV2`, ..., `xV100`) and the logarithm of the fat content as response. An excerpt of the results of the analysis is found in Figure 2, and residual plots are found in Figure 3. In Figure 4 the estimated regression coefficients are shown graphically.

The p-value for the variable xV100 is replaced by a question mark in the print-out in Figure 2. Write down the null- and alternative hypotheses being tested. Is the missing p-value below or above 0.05?

How would you *briefly* judge the model fit?

Explain the concept of *overfitting* in MLR.
Do you think overfitting can be a problem in the regression performed here? Justify your answer.

b) A principal component analysis was performed on the 215 observations of 100 absorbances.

What is the mathematical definition of the principal component loadings and scores?

In Figure 6 the estimated principal components loadings are shown for the first three principal components. How may you interpret each of the three principal components?

Refer to the print-out from performing the principal component analysis on the meat absorbances in R in Figure 5. What is the percentage of total variance explained by the first three principal components? How can this principal component analysis be used in a regression analysis with the logarithm of the fat content of meat as reponse? What could be the reasoning behind doing this instead of using the regression in a).

```
> full <- lm(y~x)
> summary(full)
Call:
lm(formula = y ~ x)
Residuals:
     Min       1Q   Median       3Q      Max
-0.52853 -0.11046  0.00315  0.11128  0.53530

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.7503     0.3598   2.085 0.039264 *
xV1          2404.3987   576.0421   4.174 5.87e-05 ***
xV2         -3926.1439  1058.6867  -3.709 0.000324 ***

Information on xV3 to xV98 not included.

xV99         1051.4592  1517.3418   0.693 0.489743
xV100        -222.4339   688.7246  -0.323 ?
---

Residual standard error: 0.2341 on 114 degrees of freedom
Multiple R-squared:  0.9544,        Adjusted R-squared:  0.9145
F-statistic: 23.88 on 100 and 114 DF,  p-value: < 2.2e-16

> ad.test(rstudent(full))

        Anderson-Darling normality test

data:  rstudent(full)
A = 0.6021, p-value = 0.1166
```

Figure 2: Excerpt from print-out from MLR of 100 absorbances vs. the logarithm of meat fat content in Problem 2a.

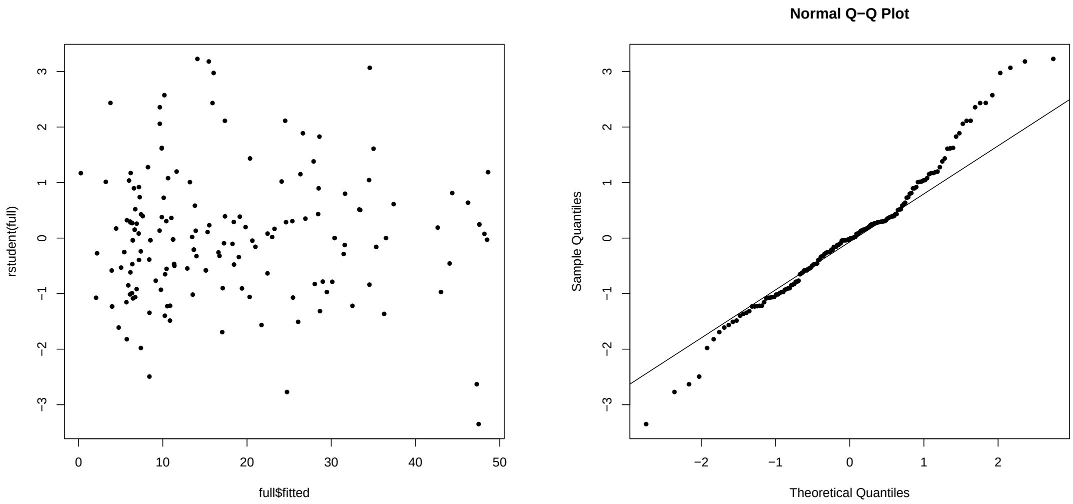

Figure 3: Residual plots (studentized residual versus fitted values in the left panel, normal plot based on studentized residuals in the right panel) for the MLR of the 100 absorbances vs. meat fat content in Problem 2a.

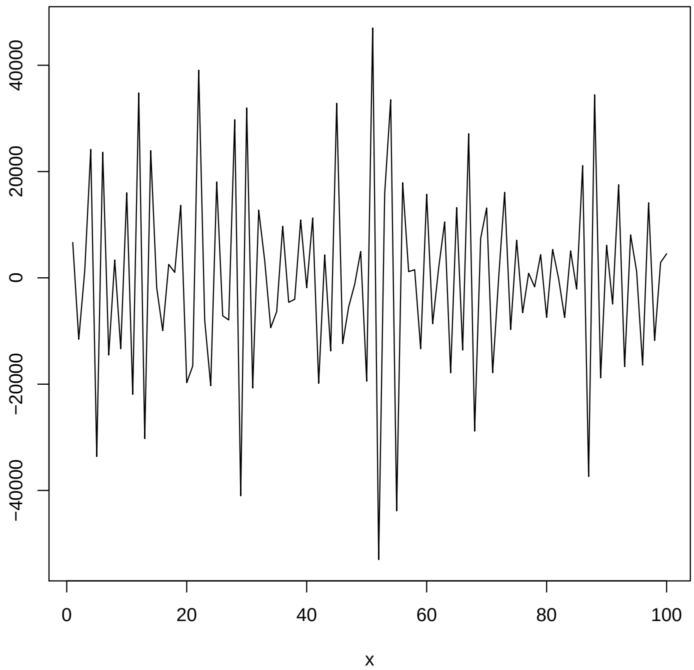

Figure 4: Estimated regression coefficients (vertical axis) for the 100 absorbances (horizontal axis) in Problem 2a.

```
> res <- prcomp(x,scale=TRUE)
> summary(res)
# only the first 6 out of 100 principal
#components are presented

Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6
Standard deviation     9.9311 0.9847 0.52851 0.33827 0.08038 0.05123
Proportion of Variance 0.9863 0.0097 0.00279 0.00114 0.00006 0.00003
Cumulative Proportion  0.9863 0.9960 0.99875 0.99990 0.99996 0.99999
```

Figure 5: Excerpt from print-out Problem 2b.

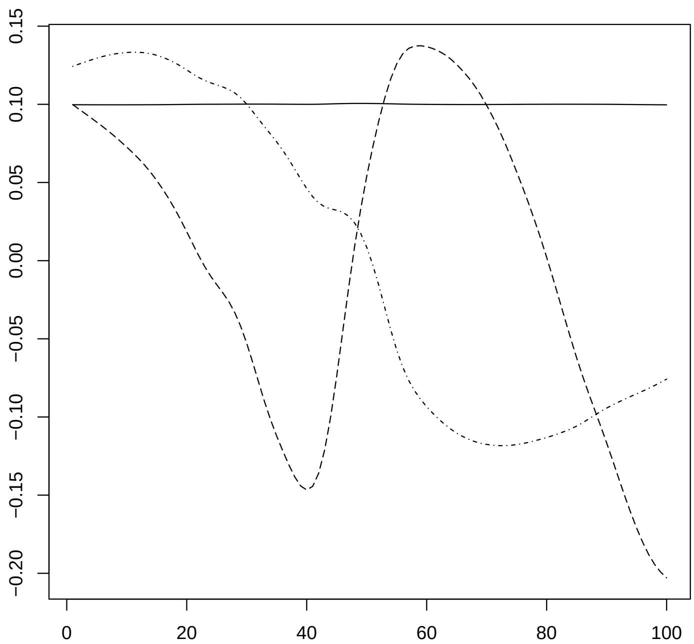

Figure 6: The estimated principal component loadings for the 100 absorbances for the meat data in Problem 2b. The horizontal axis gives the 100 absorbances and the vertical axis gives the estimated loadings for the three first principal components. The estimated loadings for the first principal component is depictured using a solid curve, the second curve is dash-dotted and the third is dashed.

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 model can be written in matrix notation as

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

where $\boldsymbol{Y}$ is a 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 a n-dimensional vector of random errors.

Assume that $n > p$ and that $\boldsymbol{X}$ has rank p.

Define the matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$.

a) What type of matrix is $\boldsymbol{H}$? Justify your answer.
Find the rank of $\boldsymbol{H}$.
How would you graphically interpret the vector $\boldsymbol{H}\boldsymbol{Y}$?

Answer the same three questions for the matrix $\boldsymbol{I} - \boldsymbol{H}$, using the findings you already have for $\boldsymbol{H}$. Here $\boldsymbol{I}$ is the $n \times n$ identity matrix.

Further, 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.

b) Let $\mathrm{SSE} = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$. Derive the distribution of SSE.
Use this to suggest an unbiased estimator for σ^2, and call the estimator $\hat{\sigma}^2$.
Find the variance of $\hat{\sigma}^2$.

Define two constant matrices $\boldsymbol{A} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ and $\boldsymbol{B} = (\boldsymbol{I} - \boldsymbol{H})$.

c) What are the dimensions of the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$?
Show that $\boldsymbol{AY}$ and $\boldsymbol{BY}$ are independent random vectors.
Use this to prove that the least squares estimator $\hat{\boldsymbol{\beta}}$ and SSE are independent random variables. What is the use of this result in multiple linear regression?