NTNU

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: 19 May 2017

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

Permitted examination support material: C: *Tabeller og formler i statistikk* (Tapir forlag, Fagbokforlaget), *Matematisk formelsamling* (K. Rottmann), one yellow A5 sheet with your own handwritten notes (stamped by the Department of Mathematical Sciences), specified calculator.

Other information:
All answers must be justified, and relevant calculations provided.

Language: English

Number of pages: 9

Number of pages enclosed: 0

Checked by:

Date Signature

Informasjon om trykking av eksamensoppgave
Originalen er:
1-sidig □ **2-sidig** ☒
sort/hvit ☒ **farger** □
skal ha flervalgskjema □

Problem 1 Random vector

Let $\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$ be a multivariate normal random vector with $\mathrm{E}(\boldsymbol{X}) = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$ and $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{X}) = \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}$, where ρ is a real number.

Further, let $Y_1 = \frac{1}{3}(X_1 + X_2 + X_3)$, $Y_2 = X_1 - X_2$ and $\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$.

a) Find a constant matrix $\boldsymbol{C}$ such that $\boldsymbol{Y} = \boldsymbol{C}\boldsymbol{X}$.
Find $\mathrm{E}(\boldsymbol{Y})$ and $\mathrm{Cov}(\boldsymbol{Y})$.
What is the distribution of $\boldsymbol{Y}$?
Are Y_1 and Y_2 independent?

Denote by $(\lambda_1, \boldsymbol{e}_1), (\lambda_2, \boldsymbol{e}_2), (\lambda_3, \boldsymbol{e}_3)$ eigenvalue–eigenvector pairs for the matrix $\boldsymbol{\Sigma}$, that is, $\boldsymbol{\Sigma}\boldsymbol{e}_i = \lambda_i \boldsymbol{e}_i$, for $i = 1, 2, 3$. Further, the eigenvalues are $\lambda_1 = 1 + 2\rho$ and $\lambda_2 = \lambda_3 = 1 - \rho$, and the eigenvectors $\boldsymbol{e}_1$, $\boldsymbol{e}_2$ and $\boldsymbol{e}_3$ are orthogonal and normalized column vectors.

b) For which values of ρ is the covariance matrix $\boldsymbol{\Sigma}$ positive definite?
Why do we want the covariance matrix to be positive definite?
What is the distribution of $\begin{pmatrix} \boldsymbol{e}_1^T \boldsymbol{X} \\ \boldsymbol{e}_2^T \boldsymbol{X} \\ \boldsymbol{e}_3^T \boldsymbol{X} \end{pmatrix}$?
For $\rho = 0.5$, calculate the probability $P(\boldsymbol{e}_1^T \boldsymbol{X} + \boldsymbol{e}_2^T \boldsymbol{X} + \boldsymbol{e}_3^T \boldsymbol{X} > 4)$.

Remark: The numerical values for the eigenvectors are *not needed* for any of the calculations. But, if it makes it easier for you to understand – the following are orthogonal and normalized eigenvectors for $\boldsymbol{\Sigma}$ when $\rho = 0.5$:

$$\boldsymbol{e}_1 = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \ \boldsymbol{e}_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \text{ and } \boldsymbol{e}_3 = \frac{1}{\sqrt{6}} \begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix}.$$

Problem 2 Modelling systolic blood pressure

The Framingham Heart Study is a study of the etiology (i.e. underlying causes) of cardiovascular disease, with participants from the community of Framingham in Massachusetts, USA[1] .

We will focus on modelling systolic blood pressure using data from $n = 2600$ persons. For each person in the data set we have measurements of the following seven variables:

- `SYSBP`: systolic blood pressure (mmHg),
- `SEX`: 1=male, 2=female,
- `AGE`: age (years) at examination,
- `CURSMOKE`: current cigarette smoking at examination:
  0=not current smoker, 1= current smoker,
- `BMI`: body mass index (kg/m^2),
- `TOTCHOL`: serum total cholesterol (mg/dl), and
- `BPMEDS`: use of anti-hypertensive medication at examination:
  0=not currently using, 1=currently using.

A multiple normal linear regression model was fitted to the data set with `SYSBP` as response and all the other variables as covariates. We call this `modelA`. In Figure 1 you find R commands and printout from fitting `modelA`, and in Figure 2 residual plots.

a) In the printout from `summary(modelA)` in Figure 1 *two* numerical values are replaced by question marks. Write down the mathematical formulas, calculate numerical values, and explain what each of the values means.

How would you, based on Figures 1–2, evaluate the fit of the model?

[1] For more more information about the Framingham Heart Study visit https://www.framinghamheartstudy.org/. This dataset is subset of a teaching version of the Framingham data, used with permission from the Framingham Heart Study.

```
# name of dataset: thisds
> dim(thisds)
[1] 2600    7
> colnames(thisds)
[1] "SYSBP"  "SEX"  "AGE"  "CURSMOKE"  "BMI"  "TOTCHOL"  "BPMEDS"
> modelA=lm(SYSBP~SEX+AGE+CURSMOKE+BMI+TOTCHOL+BPMEDS,data=thisds)
> summary(modelA)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 56.505170   4.668798  12.103  < 2e-16
SEX         -0.429973   0.807048  -0.533        ?
AGE          0.795810   0.048413  16.438  < 2e-16
CURSMOKE    -0.518742   0.853190  -0.608  0.54324
BMI          1.010550          ?  10.129  < 2e-16
TOTCHOL      0.028786   0.008787   3.276  0.00107
BPMEDS      19.203706   1.102547  17.418  < 2e-16
Residual standard error: 19.65 on 2593 degrees of freedom
Multiple R-squared:  0.2508,        Adjusted R-squared:  0.249
F-statistic: 144.6 on 6 and 2593 DF,  p-value: < 2.2e-16
> residA=rstudent(modelA)
> plot(modelA$fitted.values,residA,pch=20)
> qqnorm(residA,pch=20)
> qqline(residA)
> library(nortest)
> ad.test(residA)
Anderson-Darling normality test
data:  residA
A = 13.2, p-value < 2.2e-16
```

Figure 1: Printout from R commands for model A. Two numbers are replaced by question marks.

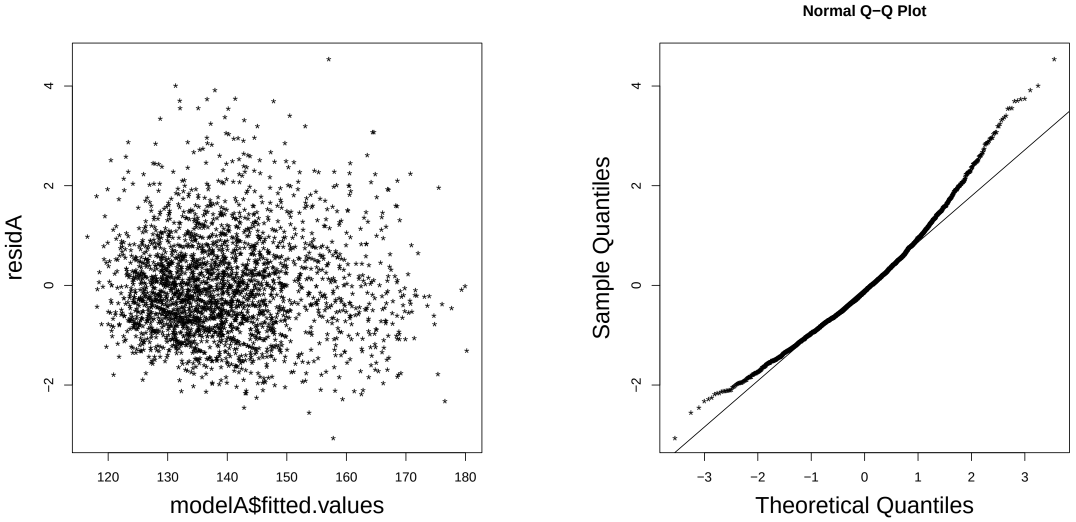

Figure 2: Residual plots for the `modelA` fit of the Framingham data.

A competing model, `modelB`, uses $-\frac{1}{\sqrt{\texttt{SYSBP}}}$ as response and the same covariates as `modelA`, see Figures 3–4.

b) Based on comparing Figures 2 and 4, would you prefer `modelA` or `modelB`? Justify your answer.

We proceed using `modelB`.

Why might a reduced model have better performance than a full model when the aim is prediction?
Explain briefly what is done in the best subset model selection, and give the reasoning behind the BIC criterion. Be sure to include an explanation of how the 6 models presented in the bottom of Figure 3 were found.
Based on the results from using the BIC criterion presented in Figure 3 choose a reduced regression model.

c) We can fit a reduced version of `modelB` by omitting the variables `SEX`, `CURSMOKE` and `TOTCHOL` from the model. We call this `modelC`, see Figure 5. To compare `modelB` and `modelC` we want to perform the following hypothesis test

$$H_0\colon \beta_{\texttt{SEX}} = \beta_{\texttt{CURSMOKE}} = \beta_{\texttt{TOTCHOL}} = 0,$$
$$H_1\colon \text{at least one of is different from } 0,$$

where $\beta_{\texttt{SEX}}$, $\beta_{\texttt{CURSMOKE}}$ and $\beta_{\texttt{TOTCHOL}}$ are the regression parameters corresponding to `SEX`, `CURSMOKE` and `TOTCHOL`, respectively.

Explain how this hypothesis test can be formulated as a linear hypothesis of the form $H_0 : \boldsymbol{C\beta} = \boldsymbol{d}$ versus $H_1 : \boldsymbol{C\beta} \neq \boldsymbol{d}$, where $\boldsymbol{\beta}$ is the complete regression parameter vector, and write out what $\boldsymbol{C}$ and $\boldsymbol{d}$ are in our case.
Calculate the appropriate test statistics and perform the hypothesis test at the significance level of your own choice.
Based on the result of the hypothesis test would you prefer `modelB` or `modelC`?

Problem 3 Design of experiments

In a pilot study with four factors A, B, C and D, the 8 experimental runs listed below are to be performed.

	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 a generator and the defining relation for the experiment?
What is the resolution of the experiment?
Write down the alias structure of the experiment.
When performing the 8 experimental runs listed above, we are instructed to do the experiments in random order. Why is that?

```
> modelB=lm(-1/sqrt(SYSBP)~SEX+AGE+CURSMOKE+BMI+TOTCHOL+BPMEDS,
data=thisds)
> summary(modelB)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.103e-01  1.383e-03 -79.745  < 2e-16 ***
SEX         -2.989e-04  2.390e-04  -1.251 0.211176
AGE          2.378e-04  1.434e-05  16.586  < 2e-16 ***
CURSMOKE    -2.504e-04  2.527e-04  -0.991 0.321723
BMI          3.087e-04  2.955e-05  10.447  < 2e-16 ***
TOTCHOL      9.288e-06  2.602e-06   3.569 0.000365 ***
BPMEDS       5.469e-03  3.265e-04  16.748  < 2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
Residual standard error: 0.005819 on 2593 degrees of freedom
Multiple R-squared:  0.2494,        Adjusted R-squared:  0.2476
F-statistic: 143.6 on 6 and 2593 DF,  p-value: < 2.2e-16
> residB=rstudent(modelB)
> par(mfrow=c(1,2))
> plot(modelB$fitted.values,residB,pch=20)
> qqnorm(residB,pch=20)
> qqline(residB)
> ad.test(residB)
Anderson-Darling normality test
data:  residB
A = 0.19209, p-value = 0.8959
> library(leaps)
> x <- model.matrix(modelB)[,-1]; dim(x)
[1] 2600    6
> colnames(x)
[1] "SEX"      "AGE"      "CURSMOKE" "BMI"      "TOTCHOL"  "BPMEDS"
> y <- -1/sqrt(thisds$SYSBP)
> allfit=regsubsets(x,y,nvmax=6)
> allsummary=summary(allfit)
> allsummary
Subset selection object
6 Variables  (and intercept)
1 subsets of each size up to 6
Selection Algorithm: exhaustive
         SEX AGE CURSMOKE BMI TOTCHOL BPMEDS
1  ( 1 ) " " " " " "      " " " "     "*"
2  ( 1 ) " " "*" " "      " " " "     "*"
3  ( 1 ) " " "*" " "      "*" " "     "*"
4  ( 1 ) " " "*" " "      "*" "*"     "*"
5  ( 1 ) "*" "*" " "      "*" "*"     "*"
6  ( 1 ) "*" "*" "*"      "*" "*"     "*"
> allsummary$bic
[1] -332.1004 -589.2250 -700.2825 -704.0729 -697.5698 -690.6913
```

Figure 3: Printout from R commands and results for `modelB`.

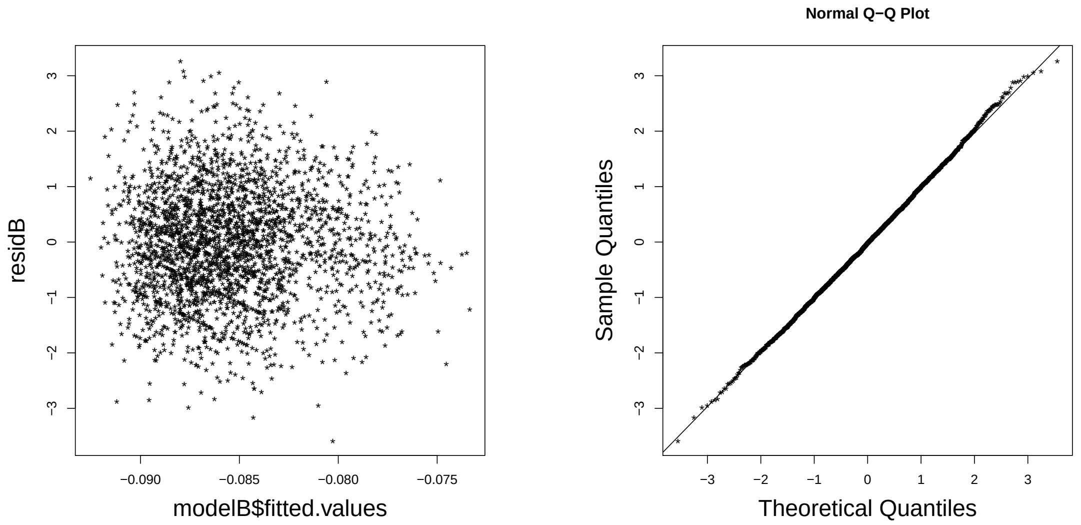

Figure 4: Residual plots for the `modelB` fit of the Framingham data.

```
> modelC=lm(-1/sqrt(SYSBP)~AGE+BMI+BPMEDS,data=thisds)
> summary(modelC)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.093e-01  1.152e-03  -94.91   <2e-16 ***
AGE          2.449e-04  1.389e-05   17.63   <2e-16 ***
BMI          3.199e-04  2.902e-05   11.02   <2e-16 ***
BPMEDS       5.490e-03  3.254e-04   16.87   <2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
Residual standard error: 0.005831 on 2596 degrees of freedom
Multiple R-squared:  0.2453,        Adjusted R-squared:  0.2444
F-statistic: 281.3 on 3 and 2596 DF,  p-value: < 2.2e-16
```

Figure 5: Printout from R commands and results for `modelC`.

Problem 4 Underfitting

The normal multiple linear regression 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 parameters and $\boldsymbol{\varepsilon}$ is an n-dimensional column vector of multivariate normal random errors with mean $\mathrm{E}(\boldsymbol{\varepsilon}) = \boldsymbol{0}$ and covariance matrix $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}$, where $\boldsymbol{I}$ is the $n \times n$ identity matrix. Assume that $n > p$ and that $\boldsymbol{X}$ has rank p.

Assume now that we partition our model into two terms:

$$\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon},$$

where $\boldsymbol{X}_1$ is an $n \times k$ matrix with rank k, $\boldsymbol{X}_2$ is an $n \times (p-k)$ matrix of rank $p-k$, $\boldsymbol{\beta}_1$ is a k-dimensional vector of regression parameters and $\boldsymbol{\beta}_2$ is a $p-k$-dimensional vector of regression parameters.

We then erroneously underfit the regression model, that is, we only fit a part of the true model,

$$\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{\alpha}_1 + \boldsymbol{\varepsilon}^*.$$

The least squares estimator for $\boldsymbol{\alpha}_1$ in the underfitted model is given as $\hat{\boldsymbol{\alpha}}_1 = (\boldsymbol{X}_1^T\boldsymbol{X}_\mathbf{1})^{-1}\boldsymbol{X}_1^T\boldsymbol{Y}$. The sums of squares of errors for the underfitted model is $\mathrm{SSE}_1 = (\boldsymbol{Y} - \boldsymbol{X}_1\hat{\boldsymbol{\alpha}}_1)^T(\boldsymbol{Y} - \boldsymbol{X}_1\hat{\boldsymbol{\alpha}}_1)$. Further, let $H_1 = \boldsymbol{X}_1(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T$.

a) First, show that $\boldsymbol{H}_1$ is idempotent and find the trace of $\boldsymbol{H}_1$.
Then, show that $\mathrm{SSE}_1 = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{Y}$.

Finally, find the expected value of the variance estimator

$$S^2 = \frac{\mathrm{SSE}_1}{n-k}.$$

The answer should be a function of n, k, σ^2, $\boldsymbol{\beta}_2$, $\boldsymbol{X}_1$ (or $\boldsymbol{H}_1$) and $\boldsymbol{X}_2$.

Hint: the so-called trace formula is given as: $\mathrm{E}(\boldsymbol{Y}^\mathrm{T}\boldsymbol{C}\boldsymbol{Y}) = \mathrm{tr}(\boldsymbol{C}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^\mathrm{T}\boldsymbol{C}\boldsymbol{\mu}$, where $\boldsymbol{Y}$ is a random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and $\boldsymbol{C}$ is a conformable constant matrix.

Problem 5 Independence of linear combinations

Let $\boldsymbol{X}$ be a p-dimensional normal random vector with mean $\boldsymbol{\mu}$ and positive definite covariance matrix $\boldsymbol{\Sigma}$. The moment generating function of $\boldsymbol{X}$ is $M_{\boldsymbol{X}}(\boldsymbol{t}) = \exp(\boldsymbol{t}^T\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t})$, where $\boldsymbol{t}$ is a p-dimensional real column vector.

Further, let $\boldsymbol{A}$ be a $q \times p$ constant matrix and $\boldsymbol{B}$ an $r \times p$ constant matrix.

a) Use the moment generating function to show that $\begin{pmatrix} \boldsymbol{AX} \\ \boldsymbol{BX} \end{pmatrix}$ is multivariately normally distributed.
Then derive a condition for when $\boldsymbol{AX}$ and $\boldsymbol{BX}$ are independent.