# Cover page

**Department of Mathematical Sciences**

**Examination paper for TMA4267 Linear statistical models**

**Examination date:** 14 May 2021

**Examination time (from–to):** 9:00–13:00

**Permitted examination support material:** A / All support material is allowed

**Academic contact during examination:** Øyvind Bakke
**Phone:** 990 41 673

**Technical support during examination:** Orakel support services
**Phone:** 73 59 16 00

If you experience technical problems during the exam, contact Orakel support services as soon as possible *before the examination time expires*. If you don't get through immediately, hold the line until your call is answered.

## OTHER INFORMATION

**Make your own assumptions:** If a question is unclear/vague, make your own assumptions and specify them in your answer. Only contact academic contact in case of errors or insufficiencies in the question set.

**Cheating/Plagiarism:** The exam is an individual, independent work. During the exam it is not permitted to communicate with others about the exam questions, or distribute drafts for solutions. Such communication is regarded as cheating. All submitted answers will be subject to plagiarism control. Read more about cheating and plagiarism

**Notifications:** If there is a need to send a message to the candidates during the exam (e.g. if there is an error in the question set), this will be done by sending a notification in Inspera. A dialogue box will appear. You can re-read the notification by clicking the bell icon in the top right-hand corner of the screen. All candidates will also receive an SMS to ensure that nobody misses out on important information. Please keep your phone available during the exam.

**Weighting of the problems:** There are eight multiple choice questions and three points (1, 2a and 2b) requiring a written answer and a file upload.

Each multiple choice question has one correct answer. A correct answer will score 5 points, and a missing or wrong answer will score 0 points.

Each of the written-answer points can score 0–10 points. Thus, the maximum score on the home exam is 70.

## ABOUT SUBMISSION

**How to answer questions:** All question types other than Upload assignment must be answered directly in Inspera. In Inspera, your answers are saved automatically every 15 seconds. **NB!** We advise against pasting content from other programs, as this may cause loss of formatting and/or entire elements (e.g. images, tables).

**File upload:** All files must be uploaded *before* the examination time expires. Upload handwritten answers in PDF format.

Det er lagt til **30 minutter** til ordinær eksamenstid for digitalisering av håndskrevne svar og opplasting av filer. Tilleggstida er forbehold innlevering og inngår i gjenstående eksamenstid som vises øverst til venstre på skjermen.

**30 minutes** are added to the examination time for digitising of handwritten answers and uploading of files. The additional time is included in the remaining examination time shown in the top left-hand corner.

NB! You are responsible to ensure that the file(s) are correct and not corrupt/damaged. Check the file(s) you have uploaded by clicking “Download” when viewing the question. All files can be removed or replaced as long as the test is open.

How to digitize your sketches/calculations

How to create PDF documents

**Automatic submission:** Your answer will be submitted automatically when the examination time expires and the test closes, if you have answered at least one question. This will happen even if you do not click “Submit and return to dashboard” on the last page of the question set. You can reopen and edit your answer as long as the test is open. If no questions are answered by the time the examination time expires, your answer will not be submitted. This is considered as “did not attend the exam”.

**Withdrawing from the exam:** If you become ill, or wish to submit a blank test/withdraw from the exam for another reason, go to the menu in the top right-hand corner and click “Submit blank”. This cannot be undone, even if the test is still open.

**Accessing your answer post-submission:** You will find your answer in Archive when the examination time has expired.

# 1 Multiple choice 1

Let $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$ and $\mathbf{X}_4$ be independent and identically distributed random vectors with mean vector

$$\begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix}$$

and covariance matrix

$$\begin{pmatrix} 3 & -1 & 1 \\ -1 & 1 & 0 \\ 1 & 0 & 2 \end{pmatrix}.$$

Find the covariance matrix between $\frac{1}{2}\mathbf{X}_1 + \frac{1}{2}\mathbf{X}_2 + \frac{1}{2}\mathbf{X}_3 + \frac{1}{2}\mathbf{X}_4$ and $\mathbf{X}_1 + \mathbf{X}_2 + \mathbf{X}_3 - 3\mathbf{X}_4$.

**Select one alternative:**

- $\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$
- $\frac{39}{4}\begin{pmatrix} 3 & -1 & 1 \\ -1 & 1 & 0 \\ 1 & 0 & 2 \end{pmatrix}$
- $\begin{pmatrix} 3 & -1 & 1 \\ -1 & 1 & 0 \\ 1 & 0 & 2 \end{pmatrix}$
- $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

---

Maximum marks: 5

## 2 Multiple choice 2

The random vector

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$$

has covariance matrix

$$\begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$

What is the second principal component?

**Select one alternative:**

- $X_2$
- $0.924X_1 + 0.383X_2$
- $X_3$
- $0.383X_1 - 0.924X_2$

Maximum marks: 5

# 3 Multiple choice 3

The random vector

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$$

has covariance matrix

$$\begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{pmatrix}.$$

What is the covariance matrix of the random vector

$$\begin{pmatrix} X_1 - X_2 \\ X_2 - X_3 \end{pmatrix}?$$

**Select one alternative:**

- $\begin{pmatrix} \sigma_{11} - \sigma_{12} & \sigma_{12} - \sigma_{22} & \sigma_{13} - \sigma_{23} \\ \sigma_{12} - \sigma_{13} & \sigma_{22} - \sigma_{23} & \sigma_{23} - \sigma_{33} \end{pmatrix}$
- $\begin{pmatrix} \sigma_{11} - \sigma_{12} & \sigma_{12} - \sigma_{13} \\ \sigma_{12} - \sigma_{22} & \sigma_{22} - \sigma_{23} \\ \sigma_{13} - \sigma_{23} & \sigma_{23} - \sigma_{33} \end{pmatrix}$
- $\begin{pmatrix} \sigma_{11} - \sigma_{12} + \sigma_{21} - \sigma_{22} & \sigma_{12} - \sigma_{13} - \sigma_{22} + \sigma_{23} \\ \sigma_{12} - \sigma_{13} - \sigma_{22} + \sigma_{23} & -\sigma_{12} + \sigma_{22} - \sigma_{23} + \sigma_{33} \end{pmatrix}$
- $\begin{pmatrix} \sigma_{11} - 2\sigma_{12} + \sigma_{22} & \sigma_{12} - \sigma_{13} - \sigma_{22} + \sigma_{23} \\ \sigma_{12} - \sigma_{13} - \sigma_{22} + \sigma_{23} & \sigma_{22} - 2\sigma_{23} + \sigma_{33} \end{pmatrix}$

---

Maximum marks: 5

## 4 Multiple choice 4

Let $X_1$, $X_2$ and $X_3$ be independent random variables with mean zero and variance one. Find the mean of

$$\begin{pmatrix} X_1 & X_2 & X_3 \end{pmatrix} \begin{pmatrix} -1 & 0 & 2 \\ 0 & 1 & 1 \\ 2 & 1 & 2 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}.$$

**Select one alternative:**

- $0$
- $-1$
- $2$
- $1$

---

Maximum marks: 5

## 5 Multiple choice 5

A fitted linear model with a $10 \times 2$ design matrix has $R^2_{\text{adj}} = 0.6777$ ($R^2_{\text{adj}}$ is the adjusted $R^2$). Find $R^2$, the coefficient of determination.

**Select one alternative:**

- $0.7135$
- $0.9076$
- $0.7422$
- $0.5185$

---

Maximum marks: 5

## 6 Multiple choice 6

We have $10$ parameters $\theta_1$, …, $\theta_{10}$. We wish to test all null hypotheses of the form $\theta_i = \theta_j$, where $i \neq j$, against two-sided alternatives. What should the significance level of each test be if we want to ensure $\mathrm{FWER} \leq 0.05$ by the Bonferroni method?

**Select one alternative**

- $0.0050$
- $0.0011$
- $0.0006$
- $0.0500$

Maximum marks: 5

## 7 Multiple choice 7

We have four observations. Responses are $-1, 0, 3$ and $5$. Covariates are an intercept and a two-level factor, with the two first observations having one level and the two last observations having the other level. Find the error sum of squares (SSE) of the fitted linear model.

**Select one alternative:**

- $2.5$
- $3.75$
- $33.75$
- $32.5$

Maximum marks: 5

## 8 Multiple choice 8

In a $2^{6-2}$fractional factorial design, $E = ABC$ and $F = BCD$ are chosen as generators. What is the resolution of the design?

**Select one alternative:**

- I
- IV
- III
- II

Maximum marks: 5

## 9 Problems (written answers)

**Upload your file here. Maximum one file.**

The following file types are allowed: **.jpg,.pdf** Maximum file size is **50 GB**

Select file to upload

Maximum marks: 30

**Question 9**
Attached



**Problem 1**

Let $X$ be a $k$-dimensional random vector and $Y$ a univariate random variable such that the $k+1$-dimensional vector $(\boldsymbol{X}^T\ Y)^T$ has a multivariate normal distribution. You can use the notation $\boldsymbol{\mu_X} = E\boldsymbol{X}$, $\mu_Y = EY$, $\Sigma_{\boldsymbol{X}} = \mathrm{Cov}\,\boldsymbol{X}$, $\sigma_Y^2 = \mathrm{Var}\,Y$, $\boldsymbol{\sigma}_{\boldsymbol{X}Y} = \mathrm{Cov}(\boldsymbol{X}, Y)$ (a $k$-dimensional vector) when giving your answers.

Show that the expected value in the conditional distribution of $Y$ given $\boldsymbol{X} = \boldsymbol{x}$ is the same as the expected value of the response variable $Y$ in a multiple linear regression model having $\boldsymbol{x}$ as covariates in addition to an intercept.

What is the intercept of the regression model?

**Problem 2**

Red blood cell counts of 105 athletes were studied. A multiple regression model was fitted, with red blood cell count as response variable and sex, height, weight and sport as covariates, which are coded as follows in R:

**bloodcells** The response variable; the athlete's blood cell count divided by $10^{12}$.

**sex** A factor with two levels, coded as a dummy variable **sexmale** having the value 0 if the athlete is female and 1 if the athlete is male.

**height** The athlete's height in cm.

**weight** The athlete's weight in kg.

**sport** The athlete's sport – basketball, gymnastics, sculling, swimming or tennis. This is a factor with five levels. The factor is coded using four dummy variables, **sportgymnastics** having the value 1 if the athlete's sport is gymnastics and 0 otherwise, **sportsculling** having the value 1 if the sport is sculling and 0 otherwise, **sportswimming** having the value 1 if the sport is swimming and 0 otherwise, and **sporttennis** having the value 1 if the sport is tennis and 0 otherwise.

R output of the analysis is given below. If you want, you can download the data as shown at the start of the output.

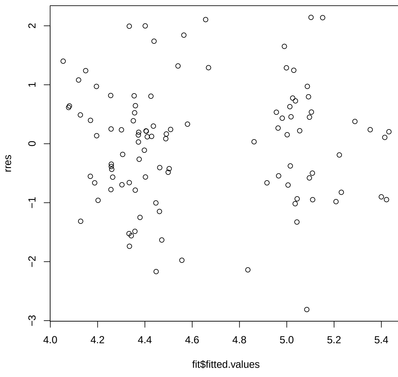
**(a)** Comment on the model fit of the model. Calculate the error sum of squares (SSE, also called residual sum of squares). How would you explain **sporttennis**'s coefficient (which is estimated at 0.272121) to a layperson?

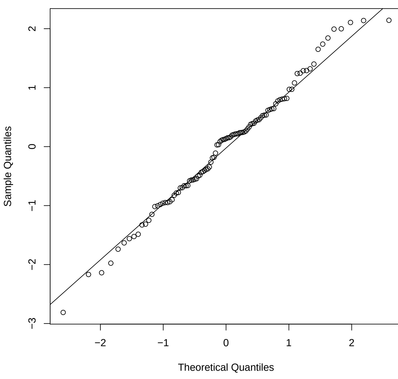
Some levels of the **sport** factor are significant and some not. Consider a model in which this factor is removed (see bottom part of R output).

**(b)** Formulate a hypothesis test where the null hypothesis corresponds to the reduced model, and perform the test at the 0.05 significance level. It is given that $P(F > 2.46548) = 0.05$ if $F$ is $F$-distributed with 4 and 97 degrees of freedom. Also comment on Mallows's $C_P$ for the two models.

*R output:*

```
> bloodcelldsta<-read.table(
+   "https://www.math.ntnu.no/emner/TMA4267/2021v/blodcells.csv")
> fit<-lm(bloodcells~.,data=bloodcelldata)
> summary(fit)

Call:
lm(formula = bloodcells ~ ., data = bloodcelldata)

Residuals:
     Min       1Q   Median       3Q      Max
-0.68471 -0.16760  0.03403  0.15263  0.50826

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.943915   0.661092   4.453 2.27e-05 ***
sexmale         0.713085   0.080672   8.839 4.29e-14 ***
height          0.012567   0.003816   3.294  0.00138 **
weight         -0.013543   0.003191  -4.244 5.03e-05 ***
sportgymnastics -0.168127   0.148127  -1.135  0.25916
sportsculling   0.216542   0.071543   3.027  0.00317 **
sportswimming   0.098028   0.082627   1.186  0.23837
sporttennis     0.272121   0.100388   2.711  0.00794 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2579 on 97 degrees of freedom
Multiple R-squared:  0.7107,    Adjusted R-squared:  0.6898
F-statistic: 34.04 on 7 and 97 DF,  p-value: < 2.2e-16

> rres<-rstudent(fit)
> plot(fit$fitted.values,rres)
> qqnorm(rres)
> qqline(rres)
```




```
> fitred<-lm(bloodcells~.-sport,data=bloodcelldata)
> summary(fitred)

Call:
lm(formula = bloodcells ~ . - sport, data = bloodcelldata)

Residuals:
     Min       1Q   Median       3Q      Max
-0.63164 -0.21403  0.00946  0.16426  0.67482

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.568475   0.505906   7.054 2.22e-10 ***
sexmale      0.763153   0.067257  11.347  < 2e-16 ***
height       0.008897   0.003285   2.709  0.00793 **
weight      -0.011434   0.002840  -4.027  0.00011 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2696 on 101 degrees of freedom
Multiple R-squared:  0.6708,    Adjusted R-squared:  0.661
F-statistic: 68.59 on 3 and 101 DF,  p-value: < 2.2e-16

> library(leaps)
> best<-regsubsets(bloodcells~.,data=bloodcelldata)
> plot(best,scale="Cp")
```

