Written examination, date: 8th of December 2020 *Page 1 of 40 pages* *Enclosure: XX pages*
Course name: Multivariate Statistics
Course number: 02409
Aids allowed: All
Exam duration: 4 hours
Weighting: The questions are given equal weight

This exam is answered by:

_____ _____ _____
(name) (signature) (study no.)

There is a total of 30 questions for the 6 problems. The answers to the 30 questions must be written into the table below.

Problem	1	1	1	1	1	1	1	1	2	2
Question	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	2.1	2.2
Answer	5	2	4	1	5	5	2	3	1	5

Problem	2	2	2	3	3	3	4	4	4	4
Question	2.3	2.4	2.5	3.1	3.2	3.3	4.1	4.2	4.3	4.4
Answer	4	4	5	2	2	1	3	1	4	2

Problem	4	4	5	5	5	5	5	5	5	5
Question	4.5	4.6	5.1	5.2	5.3	5.4	5.5	5.6	5.7	5.8
Answer	2	5	1	1	1	3	3	1	1	5

The possible answers for each question are numbered from 1 to 6. If you enter a wrong number, you may correct it by crossing the wrong number in the table and writing the correct answer immediately below. If there is any doubt about the meaning of a correction then the question will be considered not answered.

Only the front page must be returned. The front page must be returned even if you do not answer any of the questions or if you leave the exam prematurely. Drafts and/or comments are not considered, only the numbers entered above are registered.

A correct answer gives 5 points, a wrong answer gives – 1 point. Unanswered questions or a 6 (corresponding to "don't know") give 0 points. The total number of points needed for a satisfactorily answered exam is determined at the final evaluation of the exam. Especially note that the grade 10 may be given even if only one answer is wrong or unanswered.
Remember to write your name, signature, and study number on the front page.

Problem 1.

Enclosure A with SAS program and SAS output belongs to this problem. We consider data from a Portuguese study on grades of students in Mathematics in High School. The data is from https://archive.ics.uci.edu/ml/datasets/Student+Performance and was originally collected by *P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.*

The following variables are included in our analysis. It is *not* important to understand the meaning of the variables in details.

Variables	Meaning
age	student's age (numeric: from 15 to 22)
traveltime	home to school traveltime (numeric: 1:<15 min., 2: 15 to 30 min., 3: 30 min. to 1 hour, or 4: >1 hour)
studytime	weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
famrel	quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
freetime	free time after school (numeric: from 1 - very low to 5 - very high)
goout	going out with friends (numeric: from 1 - very low to 5 - very high)
dalc	workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
walc	weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
health	current health status (numeric: from 1 - very bad to 5 - very good)
absences	number of school absences (numeric: from 0 to 93)

We will now look at the correlation between these variables and to investigate underlying patters we will further perform a factor analysis on the data

Insert the Data in R:

```
data=read.table("student-mat.csv",sep=";",header=TRUE)

# Data Transformation
data$G1 <- as.numeric(data$G1)
data$G2 <- as.numeric(data$G2)
data <- data[data$G3 != 0, ]  # Remove rows where G3 is 0
```

Question 1.1.

We test whether the correlation between *dalc* and *studytime* is zero against all alternatives. The p-value for this test falls in the range

We use

Theorem 1.37

Let $R = R_{ij|m+1\ldots p}$ be the empirical partial correlation coefficient between Z_i and Z_j conditioned on (or: for given) $Z_{m+1,\ldots,Z_p}$. It is assumed to be computed from the unbiased estimates of the variance-covariance matrix and from n observations. Then

$$\frac{R}{\sqrt{1-R^2}}\sqrt{n-2-(p-m)} \sim \mathrm{t}(n-2-(p-m)),$$

if $\rho_{ij|m+1,\ldots,p} = 0$.

We insert

$$\frac{-0.19982}{\sqrt{1-(-0.19982)^2}}\sqrt{357-2} \sim t(357-2)$$

$$-3.8424 \sim t(355)$$

In R:

```
data_P1 <- data[c("age", "traveltime", "studytime", "famrel", "freetime", "goout", "Dalc", "Walc", "health", "absences")]

# Calculate Pearson correlations
cor_matrix <- cor(data_P1, method = "pearson")

# Print the correlation matrix
print(cor_matrix)

# Extract the pearson correlation between Dalc and Studytime
R = cor_matrix['Dalc', 'studytime']
n = dim(data_P1)[1]

Test_statistic = R*sqrt(n-2)/sqrt(1-R^2)

# Calculate cdf1
cdf1 <- 2 * pt(Test_statistic, df = 255)

# Print the result
print(cdf1)
```

Result:

```
> print(cdf1)
[1] 0.0001538467
```

ANSWER 5

Question 1.2.

The partial correlation between *dalc* and *studytime* when conditioned on *walc* is

We use from page 34

$$\rho_{ij|k} = \frac{\rho_{ij} - \rho_{ik}\rho_{jk}}{\sqrt{(1-\rho_{ik}^2)(1-\rho_{jk}^2)}}.$$

ij: Corr(dalc, studytime) =-0.19982
ik: Corr(dalc, walc) =0.64492
jk: Corr(studytime, walc) =-0.24760
|

$$\rho_{ik|j} = \frac{-0.19982 - 0.64492 \cdot (-0.24760)}{\sqrt{(1 - 0.64492^2) \cdot (1 - 0.24760^2)}} = -0.0542$$

In R:

```
# Install the ppcor package
#install.packages("ppcor")

# Load the ppcor package
library(ppcor)

# Specify the variable names
var_names <- c("Dalc", "studytime", "Walc")

# Calculate the partial correlation matrix using the "pcor" function
partial_corr_matrix <- pcor(data[, var_names], method = "pearson")

# print the partial correlation
print("Partial Correlation Matrix")
print(partial_corr_matrix$estimate)

# Print the partial correlation matrix
print(paste0("Partial Correlation between Dalc and studytime conditioned on Walc: ",
partial_corr_matrix$estimate[1,2]))
```

Result In R:

```
> print(paste0("Partial Correlation between Dalc and studytime conditioned on Walc: ", partial_corr_matrix$estimate[1,2]))
[1] "Partial Correlation between Dalc and studytime conditioned on Walc: -0.0542081890133211"
```

ANSWER 2

Question 1.3.

The 99% confidence interval of the correlation between *freetime* and *studytime* is

We use from page 40

Theorem 1.40

Assume the situation is as in the previous theorem. We consider the hypothesis

$$H_0 : \rho_{ij|m+1,\ldots,p} = \rho_0$$

versus

$$H_1 : \rho_{ij|m+1,\ldots,p} \neq \rho_0.$$

We let

$$Z = \frac{1}{2}\log\frac{1 + R_{ij|m+1,\ldots,p}}{1 - R_{ij|m+1,\ldots,p}}$$

and

$$z_0 = \frac{1}{2}\log\frac{1+\rho_0}{1-\rho_0}.$$

Under H_0 we will have

$$(Z - z_0) \cdot \sqrt{n - (p-m) - 3} \quad \text{approx.} \sim \mathrm{N}(0,1).$$

Also shown in example 1.41.

We have n=357, we insert

$$P\left(-2.576 < (Z - z)\sqrt{357 - 0 - 3} < 2.576\right) \approx 99\%$$

$$P(-2.576 - 18.8149 < -18.8149z < 2.576 - 18.8149Z) \approx 99\%$$

$$P(\mathrm{Z} - 0.1369 < z < Z + 0.1369) \approx 99\%$$

We find *corr(freetime,studytime)* = -0.15253

$$Z = \frac{1}{2}\log\frac{1 - 0.15253}{1 + 0.15253} = -0.1537$$

And we get the z-limits

$$[-0.2906, -0.0168\,]$$

We then need to transform it

$$\left[\frac{e^{2\cdot -0.2906} - 1}{e^{2\cdot -0.2906} + 1}, \frac{e^{2\cdot -0.0168} - 1}{e^{2\cdot -0.0168} + 1}\right] = [-0.2827, -0.0168]$$

Answer 4

Question 1.4.

If we performed a principal component analysis on the standardized data, the first 3 components would describe the following amount of the variance in the data

We use

Remark 6.7

From the theorem we have that if we seek the linear combination of the original variables which explains most of the variation in these, then the first principal component is the solution. If we seek the m variables which explain most of the original variation, then the solution is the m first principal components. A measure of how well these describe the original variation is found by means of theorems 6.3 and 6.5 which show that the m first principal components describe the fraction

$$\frac{\lambda_1 + \cdots + \lambda_m}{\lambda_1 + \cdots + \lambda_m + \cdots + \lambda_k}$$

In R:

```
# Principal Component Analysis on the correlation matrix.
pca <- princomp(data_P1,cor=TRUE,scores=TRUE)

# variance for each Principal Component:
var <- pca$sdev^2

# proportion of variance:
varPC <- var/sum(var)

# cumulative variance:
cumu = c(1:10)
for (i in 1:10){
  cumu[i] = sum(varPC[1:i])
}
results_PCA <- data.frame("eigenvalues"=var,"proportion"=varPC,
                  "cumulative" = cumu)

print("Eigenvalues of the Correlation matrix:")
results_PCA
```

Result in R:

```
> results_PCA
        eigenvalues proportion cumulative
Comp.1    2.2664796 0.22664796  0.2266480
Comp.2    1.3054944 0.13054944  0.3571974
Comp.3    1.1526639 0.11526639  0.4724638
Comp.4    1.0411603 0.10411603  0.5765798
Comp.5    0.9802314 0.09802314  0.6746030
Comp.6    0.9075189 0.09075189  0.7653548
Comp.7    0.7121938 0.07121938  0.8365742
Comp.8    0.7054789 0.07054789  0.9071221
Comp.9    0.6304035 0.06304035  0.9701625
Comp.10   0.2983754 0.02983754  1.0000000
```

ANSWER 1

Question 1.5.

Unrotated factor 1 explains what fraction of the variance in the data

We use from page 404

Furthermore, we assume that the observations are standardised in such a way that $V(X_i) = 1, \forall i$ i.e. that the variance-covariance matrix for $\boldsymbol{X}$ is equal to its correlation matrix which is denoted

$$D(X) = \mathbf{R} = \begin{pmatrix} 1 & \cdots & r_{1k} \\ \vdots & & \vdots \\ r_{k1} & \cdots & 1 \end{pmatrix}.$$

The total variance is thus equal to the number of variables, i.e. 10

In R:

```
#### Factor analysis 3 Factors ####
#install.packages("psych")
# Load the psych package
library(psych)
fa3 <- principal(cor(data_P1),nfactors = 3,rotate = "none")
fa3l <- fa3$loadings[,1:3]

#Variance explained unrotated factor:
var3 <- c(sum(fa3l[,1]^2),sum(fa3l[,2]^2),sum(fa3l[,3]^2))
print("Variance explained by unrotated factor 1 is: ")
print(var3[1])
```

Result in R:

```
[1] "Variance explained by unrotated factor 1 is: "
> print(var3[1])
[1] 2.26648
```

ANSWER 5

Question 1.6.

Rotated factor 2 explains the following fraction of the variance in freetime

We have page 405

$$\text{Cov}(X_i, F_j) = \text{Cov}(\sum_v a_{iv} F_v + G_i, F_j) = a_{ij}.$$

i.e. the factor loadings are correlations. We can then use from page 29

and the *squared coefficient of correlation represents the reduction in variance. i.e. the fraction of Y's variance, which can be explained by X,* since

$$\rho^2 = \frac{\mathrm{V}(Y) - \mathrm{V}(Y|X = x)}{\mathrm{V}(Y)}.$$

In R:

Slight different results compared to SAS roughly after the third decimal number
The RC3 and RC2 factor column in R are the Factor2 and Factor3 column in SAS

```
#rotated:
rfa3 <- principal(cor(data_P1),nfactors = 3,rotate = "varimax")
rfa3l <- rfa3$loadings[,1:3]
print(rfa3l)
```

Result in R:

```
> print(rfa3l)
                RC1         RC3         RC2
age         0.04857767  0.17137066  0.7778925
traveltime  0.30342618 -0.10610087  0.1340196
studytime  -0.50246062  0.00230782  0.1991215
famrel     -0.40356782  0.67761108  0.1020048
freetime    0.19724933  0.67638186 -0.1796006
goout       0.42604732  0.49519015  0.2113759
Dalc        0.76977899  0.20252830  0.1119688
Walc        0.83107939  0.18000404  0.1177870
health      0.11691323  0.31399856 -0.3230075
absences    0.15907581 -0.14942344  0.6281552
>
```

We take the observation in row freetime and column RC3 (rotated factor 2 in R) which is equal to 0.67638186.

The answer is thus $0.6763^2=0.4575$

ANSWER 5

Question 1.7.

In the factor model the uniqueness of *freetime* is:

We find on page 404

$$V(X_j) = a_{j1}^2 + \cdots + a_{jm}^2 + \delta_j = 1.$$

Here we introduce the notation

$$h_j^2 = a_{j1}^2 + \cdots + a_{jm}^2, \qquad j = 1, \ldots, k.$$

These quantities are called ***communalities*** and h_j^2 describes how large a proportion of X_j's variance is due to the m common factors. Correspondingly δ_j gives the ***uniqueness*** in X_j's variance. i.e. the proportion of X_j's variance which is not due to the m common factors.

In R:

#Communality:
rfa3com <- rfa3$communality
print("Final Communality Estimates rotated:")
rfa3com

Result in R:

```
[1] "Final Communality Estimates rotated:"
> rfa3com
      age traveltime  studytime     famrel   freetime      goout       Dalc       Walc     health   absences
 0.6368445  0.1212861  0.2921214  0.6324287  0.5286561  0.4714094  0.6461144  0.7369682  0.2165977  0.4422114
> |
```

1-0.52865610 = 0.4713439
ANSWER 2

Question 1.8.

We now consider the loadings of the initial and rotated factors with the following possible factor interpretations:

A: Mainly an average of family relations free time and going out
B: Mainly a contrast: family relations and study time vs. alcohol consumption, and going out
C: Mainly a contrast: Family relations, health and free time vs. absences, age and travel time
D: Mainly an contrast: free time and health vs. age and absences
E: Mainly an average of study time family relations and age
F: Mainly a contrast: studytime vs. alcohol consumption, going out and free time

If the interpretations of the three factors are Factor1~P, Factor2~Q and Factor3~R, we shall write UF(P,Q,R) for the unrotated factors and RF(P,Q,R) for the rotated factors. Going from the unrotated facor model (UF) to the rotated (RF) we get the following interpretations of the three factors:

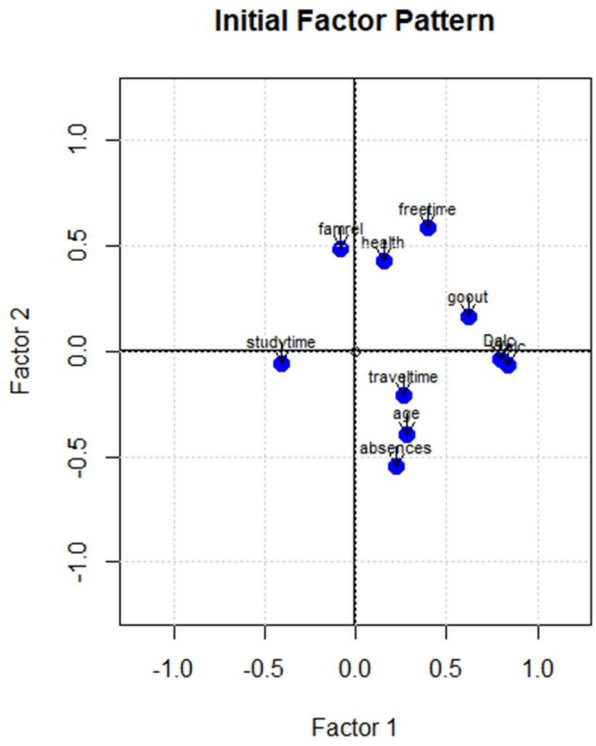
$$UF(F, C, E) \rightarrow RF(B, A, D)$$

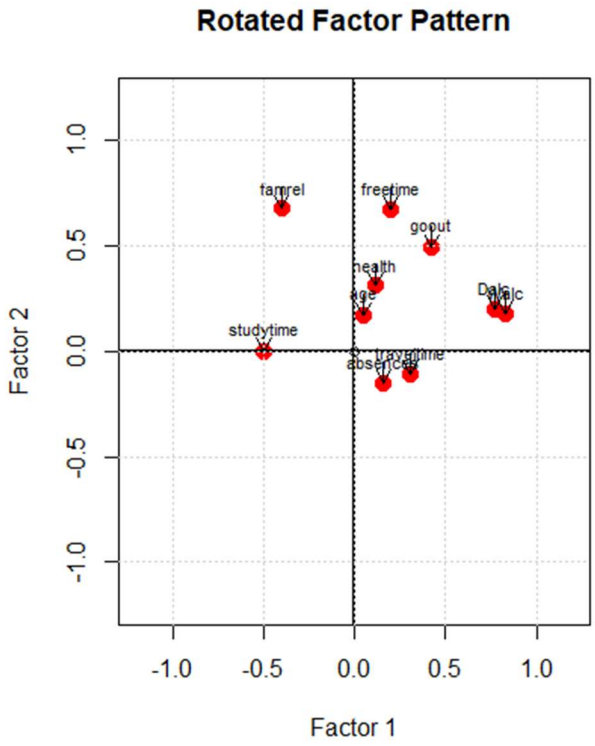
In R:

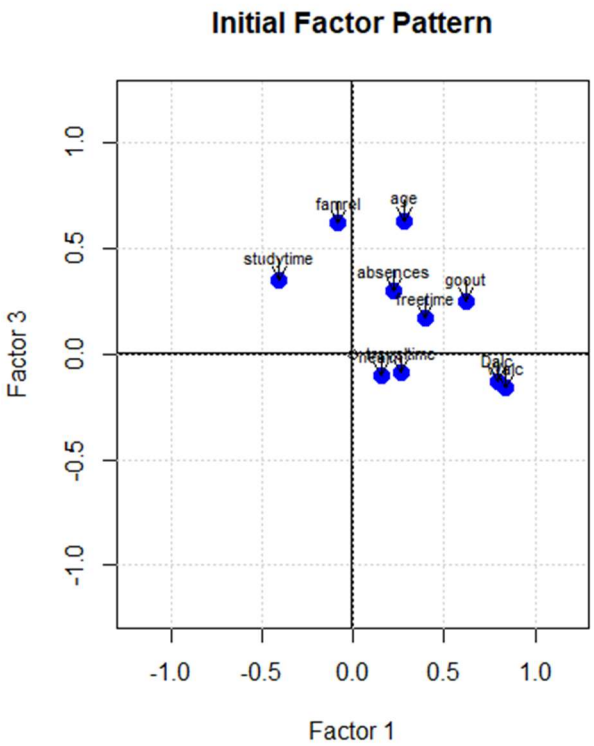
# Plots for factor analysis with 3 factors:
par(mfrow = c(1, 2))
circle <- seq(-3.2, 3.2, by = 0.1)

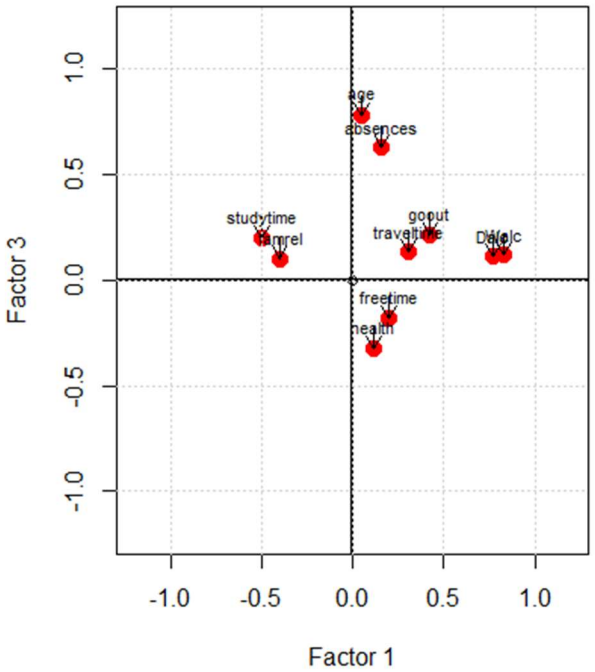
```
# Different combinations of plots
ij <- matrix(c(1, 1, 2, 2, 3, 3), ncol = 2)

Names <- c("age", "traveltime", "studytime", "famrel", "freetime", "goout", "Dalc", "Walc", "health",
"absences")
for (i in 1:3) {
  l <- ij[i, 1]
  k <- ij[i, 2]

  # Plot for the Factors
  plot(0, 0, xlim = c(-1.2, 1.2), ylim = c(-1.2, 1.2), xlab = paste0("Factor ", l),
       ylab = paste0("Factor ", k), main = "Initial Factor Pattern")
  points(fa3l[, l], fa3l[, k], col = 'blue', pch = 19, cex = 1.5)
  text(fa3l[, l], fa3l[, k] + 0.1, Names, cex = 0.7)
  arrows(fa3l[, l], fa3l[, k], fa3l[, l], fa3l[, k] + 0.1, length = 0.1, code = 1, angle = 30)
  abline(h = 0, v = 0, col = 'black', lwd = 2)  # Add thick black zero lines
  grid()

  # Plot for rotated Factors
  plot(0, 0, xlim = c(-1.2, 1.2), ylim = c(-1.2, 1.2), xlab = paste0("Factor ", l),
       ylab = paste0("Factor ", k), main = "Rotated Factor Pattern")
  points(rfa3l[, l], rfa3l[, k], col = 'red', pch = 19, cex = 1.5)
  text(rfa3l[, l], rfa3l[, k] + 0.1, Names, cex = 0.7)
  arrows(rfa3l[, l], rfa3l[, k], rfa3l[, l], rfa3l[, k] + 0.1, length = 0.1, code = 1, angle = 30)
  abline(h = 0, v = 0, col = 'black', lwd = 2)  # Add thick black zero lines
  grid()
}
```

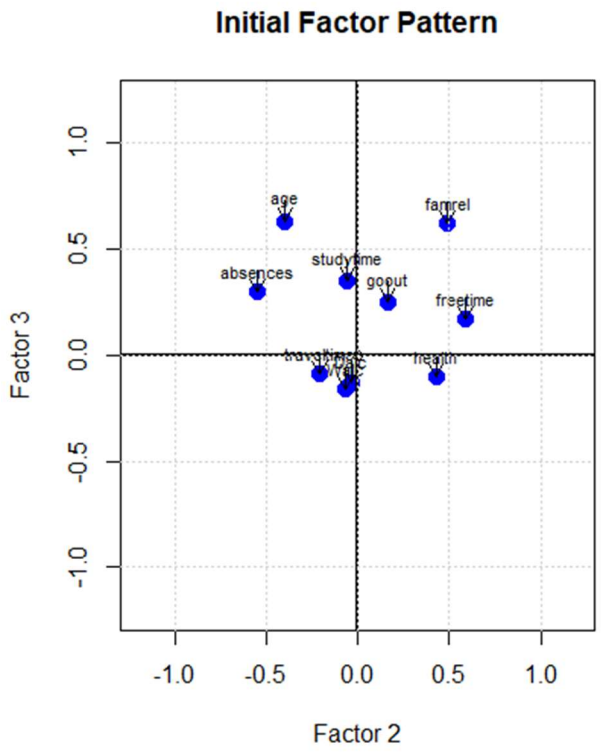
Result in R:

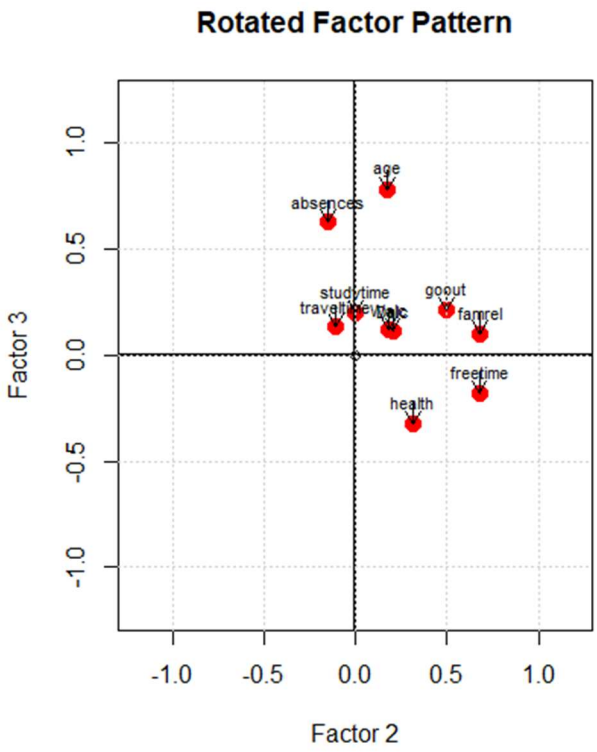

Initial Factor Pattern
Factor 2
Factor 1
famrel
freetime
health
goout
studytime
Dalc
Walc
traveltime
age
absences
-1.0
-0.5
0.0
0.5
1.0


Rotated Factor Pattern
Factor 2
Factor 1
famrel
freetime
goout
health
age
Dalc
Walc
studytime
absences
-1.0
-0.5
0.0
0.5
1.0


Initial Factor Pattern
Factor 3
Factor 1
famrel
age
studytime
absences
goout
freetime
Dalc
Walc
-1.0
-0.5
0.0
0.5
1.0


Rotated Factor Pattern
Factor 3
Factor 1
age
absences
studytime
famrel
goout
traveltime
Dalc
Walc
freetime
health
-1.0
-0.5
0.0
0.5
1.0


Initial Factor Pattern
Factor 3
Factor 2
1.0
0.5
0.0
-0.5
-1.0
age
famrel
absences
studytime
goout
freetime
health


Rotated Factor Pattern
Factor 3
Factor 2
1.0
0.5
0.0
-0.5
-1.0
age
absences
studytime
goout
famrel
freetime
health

ANSWER 3

Problem 2.

You are encouraged to use statistical software to solve this problem.

We still consider the data from problem 1, but now only a small subset of it. We consider the following variables.

Variables	Meaning
age	student's age (numeric: from 15 to 22)
traveltime	home to school traveltime (numeric: 1:<15 min., 2: 15 to 30 min., 3: 30 min. to 1 hour, or 4: >1 hour)
goout	going out with friends (numeric: from 1 - very low to 5 - very high)
health	current health status (numeric: from 1 - very bad to 5 - very good)
absences	number of school absences (numeric: from 0 to 93)
G1	Start semester grading
G3	Final grading

We have the following 20 observations. They are also given in the text file 'Problem2dataset.txt'.

Obs	age	traveltime	goout	health	absences	G1	G3
1	18	2	4	3	6	5	6
2	17	1	3	3	4	5	6
3	15	1	2	3	10	7	10
4	15	1	2	5	2	15	15
5	16	1	2	5	4	6	10
6	16	1	2	5	10	15	15
7	16	1	4	3	0	12	11
8	17	2	4	1	6	6	6
9	15	1	2	1	0	16	19
10	15	1	1	5	0	14	15
11	15	1	3	2	0	10	9
12	15	3	2	4	4	10	12
13	15	1	3	5	2	14	14
14	15	2	3	3	2	10	11
15	15	1	2	3	0	14	16
16	16	1	4	2	4	14	14
17	16	1	3	2	6	13	14
18	16	3	2	4	4	8	10
19	17	1	5	5	16	6	5
20	16	1	3	5	4	8	10

We will now try to predict G3 as a function of the other variables with the following model:

$$G3 = \mu + \beta_1 \cdot age + \beta_2 \cdot traveltime + \beta_3 \cdot goout + \beta_4 \cdot health + \beta_5 \cdot absences + \beta_6 \cdot G1 + \epsilon$$

Where μ is the intercept and ϵ is the error term.

Insert the Data in R:

```
# Read the data from the text file
examdata <- read.table("Exam2020_Problem2dataset.txt", header = TRUE)

# View the first rows of the data
head(examdata)
```

Question 2.1.

The first variable to be eliminated when performing backwards elimination is:

In R:

Running the regression in R:

```
# Perform linear regression
model <- lm(G3 ~ age + traveltime + goout + health + absences + G1, data = examdata)

# Print regression summary
summary(model)
```

Result in R:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.90255    7.51908   0.918  0.37533
age          0.01645    0.46579   0.035  0.97237
traveltime  -0.10128    0.41806  -0.242  0.81235
goout       -1.23724    0.36628  -3.378  0.00495 **
health      -0.23517    0.20714  -1.135  0.27675
absences     0.02455    0.07920   0.310  0.76150
G1           0.82181    0.09529   8.624 9.73e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.119 on 13 degrees of freedom
Multiple R-squared:  0.9421,    Adjusted R-squared:  0.9153
F-statistic: 35.23 on 6 and 13 DF,  p-value: 2.618e-07
```

It is clear that the age variable has the biggest p-value in the model, so we can drop it first from the model.

Other way with backward elimination (See the first variable needs to be eliminated)

```
# Fit the initial linear regression model
second_model <- lm(G3 ~ age + traveltime + goout + health + absences + G1, data = examdata)

# Perform backward stepwise regression
final_model <- step(second_model, direction = "backward")

# View the final model summary
summary(final_model)
```

Result in R:

```
Start:  AIC=9.87
G3 ~ age + traveltime + goout + health + absences + G1

             Df Sum of Sq     RSS    AIC
- age         1     0.002  16.271  7.872
- traveltime  1     0.073  16.342  7.961
- absences    1     0.120  16.389  8.018
- health      1     1.613  17.882  9.761
<none>                     16.269  9.871
- goout       1    14.279  30.548 20.471
- G1          1    93.079 109.348 45.976

Step:  AIC=7.87
G3 ~ traveltime + goout + health + absences + G1

             Df Sum of Sq     RSS    AIC
- traveltime  1     0.073  16.344  5.962
- absences    1     0.131  16.401  6.032
- health      1     1.615  17.885  7.765
<none>                     16.271  7.872
- goout       1    18.724  34.994 21.189
- G1          1   117.289 133.559 47.976
```

We can notice that the first variable needs to be eliminated when performing backward elimination is the age variable.

ANSWER 1

Question 2.2.

The observation with the highest leverage is:

In R:

```
Obs <- 1:length(examdata$G3)
DFFITS <- round(dffits(second_model), 4)
R_student <- round(rstudent(second_model),4)
HatDiagH <- round(hatvalues(second_model),4)
Residual <- round(residuals(second_model),4)
Covratio <- round(covratio(second_model), 4)

Stats  <- data.frame(Obs,Residual,R_student,HatDiagH,Covratio,DFFITS)
print(Stats)
```

Result in R:

```
> print(Stats)
   Obs Residual R_student HatDiagH Covratio  DFFITS
1    1   0.4021    0.4673   0.4440   2.7759  0.4176
2    2  -0.8709   -0.9570   0.3426   1.5918 -0.6908
3    3   0.1338    0.1895   0.6309   4.6462  0.2477
4    4  -0.7739   -0.7598   0.1980   1.5715 -0.3776
5    5   1.5568    1.8724   0.3411   0.4419  1.3471
6    6  -0.9867   -1.3248   0.5309   1.4360 -1.4094
7    7  -0.2717   -0.2794   0.2983   2.3850 -0.1822
8    8  -0.8736   -0.9442   0.3217   1.5632 -0.6502
9    9   1.5127    1.8660   0.3746   0.4706  1.4441
10  10  -1.1402   -1.2468   0.3032   1.0713 -0.8225
11  11  -2.0840   -2.7119   0.2975   0.0878 -1.7646
12  12   0.2535    0.3006   0.4716   3.1447  0.2840
13  13   0.2852    0.2912   0.2874   2.3394  0.1849
14  14   0.2033    0.2028   0.2561   2.2984  0.1190
15  15   0.6267    0.5886   0.1396   1.6677  0.2371
16  16   0.7514    0.7727   0.2679   1.7027  0.4674
17  17   0.2868    0.2797   0.2191   2.1428  0.1481
18  18  -0.1194   -0.1340   0.4138   2.9562 -0.1126
19  19  -0.0425   -0.0625   0.6592   5.1275 -0.0870
20  20   1.1505    1.1676   0.2025   1.0339  0.5883
> |
```

We can notice that observation 19 has the highest leverage which is equal to 0.6592.

ANSWER 5

Question 2.3.

The 99% confidence interval for the expected value of observation no. 3 is:

In R:

```
# Calculate confidence intervals for coefficients (alpha = 0.01)
conf_intervals <- confint(model, level = 0.99)

# Print the confidence intervals
print(conf_intervals)

# Calculate confidence limits for the mean (clm)
clm <- predict(model, interval = "confidence", level = 0.99)
print(clm)
```

Result in R:

```
> print(clm)
         fit       lwr       upr
1   5.597918  3.352598  7.843238
2   6.870894  4.898625  8.843164
3   9.866151  7.189567 12.542734
4  15.773878 14.274330 17.273427
5   8.443158  6.475149 10.411168
6  15.986724 13.531291 18.442157
7  11.271661  9.431077 13.112245
8   6.873612  4.962442  8.784782
9  17.487256 15.424867 19.549644
10 16.140211 14.284562 17.995859
11 11.084004  9.246126 12.921883
12 11.746538  9.432374 14.060702
13 13.714832 11.908188 15.521476
14 10.796652  9.091281 12.502023
15 15.373306 14.114223 16.632389
16 13.248642 11.504523 14.992760
17 13.713173 12.135824 15.290521
18 10.119371  7.951713 12.287029
19  5.042485  2.306417  7.778554
20  8.849534  7.333142 10.365927
```

Alternative, one could use:

Theorem 2.15

Let the situation be as above. Then the $(1-\alpha)$-*confidence interval* for the expected value of a new observation Y will be

$$[u - \mathrm{t}(n-k)_{1-\frac{\alpha}{2}} s\sqrt{c}, \quad u + \mathrm{t}(n-k)_{1-\frac{\alpha}{2}} s\sqrt{c}].$$

calculate it from the predicted value and Std Error Mean Predict or the predicted value, the RMSE and leverage.

ANSWER 4

Question 2.4.

The observation that – if deleted – will lead to the largest overall change in the parameter estimates is:

In R:

```
CookD <- round(cooks.distance(model), 5)
Stats$CookD <- CookD
print(Stats)

# Calculate the maximum Cook's Distance value
max_cook <- max(Stats$CookD)

# Create a plot for Cook's Distance with adjusted y-axis limits
plot(Stats$Obs, Stats$CookD, type = "p", pch = 19, col = "blue",
     xlab = "Observation Number", ylab = "Cook's Distance",
     main = "Cook's Distance Plot",
     ylim = c(0, 1.25 * max_cook))

# Identify and label observations with Cook's Distance > 0.05
high_cook_obs <- Stats$Obs[Stats$CookD > 0.05]
```

text(high_cook_obs, Stats$CookD[Stats$CookD > 0.05], labels = high_cook_obs, pos = 3)

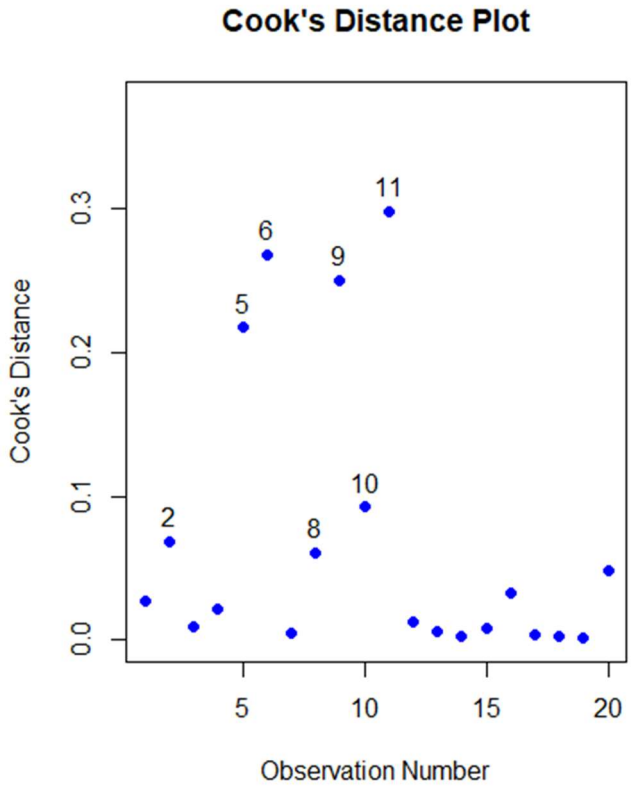
Result in R:

We look at DFFITS column:

	Obs	Residual	R_student	HatDiagH	Covratio	DFFITS	CookD
1	1	0.4021	0.4673	0.4440	2.7759	0.4176	0.02650
2	2	-0.8709	-0.9570	0.3426	1.5918	-0.6908	0.06862
3	3	0.1338	0.1895	0.6309	4.6462	0.2477	0.00947
4	4	-0.7739	-0.7598	0.1980	1.5715	-0.3776	0.02105
5	5	1.5568	1.8724	0.3411	0.4419	1.3471	0.21735
6	6	-0.9867	-1.3248	0.5309	1.4360	-1.4094	0.26821
7	7	-0.2717	-0.2794	0.2983	2.3850	-0.1822	0.00510
8	8	-0.8736	-0.9442	0.3217	1.5632	-0.6502	0.06090
9	9	1.5127	1.8660	0.3746	0.4706	1.4441	0.25015
10	10	-1.1402	-1.2468	0.3032	1.0713	-0.8225	0.09270
11	11	-2.0840	-2.7119	0.2975	0.0878	-1.7646	0.29879
12	12	0.2535	0.3006	0.4716	3.1447	0.2840	0.01239
13	13	0.2852	0.2912	0.2874	2.3394	0.1849	0.00526
14	14	0.2033	0.2028	0.2561	2.2984	0.1190	0.00218
15	15	0.6267	0.5886	0.1396	1.6677	0.2371	0.00845
16	16	0.7514	0.7727	0.2679	1.7027	0.4674	0.03221
17	17	0.2868	0.2797	0.2191	2.1428	0.1481	0.00337
18	18	-0.1194	-0.1340	0.4138	2.9562	-0.1126	0.00196
19	19	-0.0425	-0.0625	0.6592	5.1275	-0.0870	0.00117
20	20	1.1505	1.1676	0.2025	1.0339	0.5883	0.04810

Observation 11
Cooks D yields the same result

Plot:



ANSWER 4

Question 2.5.

Using only 3 of the independent variables the best model as measured by R^2 is:

In R:

```
# Create a data frame to store the results
results_Q25 <- data.frame(Model = character(0), R_Square = numeric(0), stringsAsFactors = FALSE)

# Define the combinations of independent variables
combinations <- list(
  c("health", "absences"),
  c("traveltime", "health"),
  c("goout","health", "G1"),
  c("traveltime","goout", "G1"),
  c("goout", "absences","G1"),
  c("age", "goout", "G1")
)

# Fit models and calculate R-squared for each combination
for (i in 1:length(combinations)) {
  independent_vars <- combinations[[i]]
  formula <- as.formula(paste("G3 ~", paste(independent_vars, collapse = "+")))
  model <- lm(formula, data = examdata)
```

```
  rsquare <- summary(model)$r.squared
  results_Q25 <- rbind(results_Q25, data.frame(Model = paste(independent_vars, collapse = "    "), R_Square
= rsquare))
}

# Print the results
print(results_Q25)
```

Result in R:

```
> # Print the results
> print(results_Q25)
                    Model   R_Square
1       health   absences 0.24298408
2     traveltime   health 0.06841184
3     goout   health   G1 0.94128134
4 traveltime   goout   G1 0.93628362
5   goout   absences   G1 0.93624114
6        age   goout   G1 0.93623248
```

ANSWER 5

Problem 3.

We yet again consider the data from problem 1, but will now investigate their relation to the final grading.

Variables	Meaning
age	student's age (numeric: from 15 to 22)
traveltime	home to school traveltime (numeric: 1:<15 min., 2: 15 to 30 min., 3: 30 min. to 1 hour, or 4: >1 hour)
studytime	weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
famrel	quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
freetime	free time after school (numeric: from 1 - very low to 5 - very high)
goout	going out with friends (numeric: from 1 - very low to 5 - very high)
dalc	workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
walc	weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
health	current health status (numeric: from 1 - very bad to 5 - very good)
absences	number of school absences (numeric: from 0 to 93)
G3	Final grading

We consider five models (in the notation below a parameter is implicitly fitted to each of the independent variables).

M1 $G3 = \mu + \text{age} + \text{traveltime} + \text{studytime} + \text{famrel} + \text{freetime} + \text{goout} + \text{dalc} + \text{walc} + \text{health} + \text{absences} + \epsilon$

M2 $G3 = \mu + \text{traveltime} + \text{studytime} + \text{famrel} + \text{freetime} + \text{goout} + \text{dalc} + \text{walc} + \epsilon$

M3 $G3 = \mu + \text{studytime} + \text{famrel} + \text{goout} + \text{dalc} + \text{walc} + \epsilon$

M4 $G3 = \mu + \text{studytime} + \epsilon$

M5 $G3 = \mu + \epsilon$

Where μ is the intercept and ϵ is the error term.

Question 3.1.

We test in model M1 if the parameter for *absences* is significantly different from zero against all alternatives. The p-value for this test is:

In R:

```
# Fit model M1
model_M1 <- lm(G3 ~ age + traveltime + studytime + famrel + freetime + goout + Dalc + Walc + health + absences, data = data)

# Obtain the summary of the model
summary(model_M1)
```

Result in R:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.59606    2.39491   6.930 2.07e-11 ***
age         -0.19584    0.13544  -1.446  0.14909
traveltime  -0.31985    0.24404  -1.311  0.19086
studytime    0.31950    0.20743   1.540  0.12441
famrel       0.10396    0.19170   0.542  0.58795
freetime     0.09192    0.17645   0.521  0.60275
goout       -0.39699    0.17622  -2.253  0.02490 *
Dalc        -0.02373    0.23876  -0.099  0.92090
Walc        -0.14040    0.18389  -0.763  0.44571
health      -0.19435    0.11960  -1.625  0.10509
absences    -0.06845    0.02078  -3.293  0.00109 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.093 on 346 degrees of freedom
Multiple R-squared:  0.1078,    Adjusted R-squared:  0.08205
F-statistic: 4.182 on 10 and 346 DF,  p-value: 1.744e-05
```

ANSWER 2

Question 3.2.

We test in model M1 if the parameter for *absences* is significantly different from zero against all alternatives. The usual test for this has under the nul-hypothesis the following distribution

We use

$$\frac{\hat{\theta}_{i_0} - c}{\sqrt{\hat{V}(\hat{\theta}_{i_0})}} \sim t(f), \quad f = n - \mathrm{rk}(x)$$

This is used in setting up a test for the hypothesis in

Theorem 2.23

Let the situation be as above. Then the critical region for testing H_0 against H_1 at significance level α is

$$C_\alpha = \left\{ (y_1, \cdots, y_n) \;\middle|\; \hat{\theta}_{i_0} < c - t(f)_{1-\frac{\alpha}{2}} \sqrt{\hat{V}(\hat{\theta}_{i_0})} \text{ or } \hat{\theta}_{i_0} > c + t(f)_{1-\frac{\alpha}{2}} \sqrt{\hat{V}(\hat{\theta}_{i_0})} \right\}$$

And find
N = 357
Rk(x) = 11 (since we have 11 estimated parameters)

f = 357 – 11 = 346 leading to t(346)
ANSWER 2

Question 3.3.

We sequentially test M1 through M5, starting from M1. As a result of this sequential test, the simplest model we accept is

Let us start by collating the data. We test

$$H_0: \theta \in H_{i+1} \quad \text{against} \quad \theta \in H_i \backslash H_{i+1}$$

Bmo.

$$\frac{\left(SS_{res}(H_{i+1}) - SS_{res}(H_i)\right)/\left(DF_{res}(H_{i+1}) - DF_{res}(H_i)\right)}{SS_{res}(H_i)/DF_{res}(H_i)} > F_{1-\alpha}\left(DF_{res}(H_{i+1}) - DF_{res}(H_i), DF_{res}(H_i)\right)$$

In R:

```
anova(model_M1)
summary(model_M1)
anova_M1 = anova(model_M1)
model  M2 <- lm(G3 ~ traveltime + studytime + famrel + freetime + goout + Dalc + Walc, data = data)
anova(model_M2)
summary(model_M2)
anova_M2 = anova(model_M2)
model_M3 <- lm(G3 ~ studytime + famrel + goout + Dalc + Walc, data = data)
anova(model_M3)
summary(model_M3)
anova_M3 = anova(model_M3)
model_M4 <- lm(G3 ~ studytime, data = data)
anova(model_M4)
summary(model_M4)
anova_M4 = anova(model_M4)
model  M5 <- lm(G3 ~ 1, data = data)
anova(model_M5)
summary(model_M5)
anova_M5 = anova(model_M5)

#### Compare M1 vs M2
F_observed_Model_H1 = ((anova_M2$`Sum Sq`[8]-anova_M1$`Sum Sq`[11])/(anova_M2$Df[8]-
anova_M1$Df[11]))/(anova_M1$`Sum Sq`[11]/anova_M1$Df[11])
cat("The test statistic for model H1 is:", F_observed_Model_H1, "\n")


# Calculate the critical F-value
alpha = 0.05
df11 <- anova_M2$Df[8]-anova_M1$Df[11]  # Degrees of freedom for the numerator
df21 <- anova_M1$Df[11]  # Degrees of freedom for the denominator
# Calculate p-value for the F-test
p_value_Model_H1 <- 1 - pf(F_observed_Model_H1, df11, df21)
cat("The p-value for model H1 is:", p_value_Model_H1, "\n")

################### Compare M2 vs M3 ##########################
F_observed_Model_H2 = ((anova_M3$`Sum Sq`[6]-anova_M2$`Sum Sq`[8])/(anova_M3$Df[6]-
anova_M2$Df[8]))/(anova_M2$`Sum Sq`[8]/anova_M2$Df[8])
cat("The test statistic for model H2 is:", F_observed_Model_H2, "\n")

# Calculate the critical F-value
alpha = 0.05
df12 <- anova_M3$Df[6]-anova_M2$Df[8]  # Degrees of freedom for the numerator
df22 <- anova_M2$Df[8]  # Degrees of freedom for the denominator
# Calculate p-value for the F-test
p_value_Model_H2 <- 1 - pf(F_observed_Model_H2, df12, df22)
cat("The p-value for model H2 is:", p_value_Model_H2, "\n")
```

```
################### Compare M3 vs M4 ##########################
F_observed_Model_H3 = ((anova_M4$`Sum Sq`[2]-anova_M3$`Sum Sq`[6])/(anova_M4$Df[2]-
anova_M3$Df[6]))/(anova_M3$`Sum Sq`[6]/anova_M3$Df[6])
cat("The test statistic for model H3 is:", F_observed_Model_H3, "\n")

# Calculate the critical F-value
alpha = 0.05
df13 <- anova_M4$Df[2]-anova_M3$Df[6]  # Degrees of freedom for the numerator
df23 <- anova_M3$Df[6]  # Degrees of freedom for the denominator
# Calculate p-value for the F-test
p_value_Model_H3 <- 1 - pf(F_observed_Model_H3, df13, df23)
cat("The p-value for model H3 is:", p_value_Model_H3, "\n")

################### Compare M4 vs M5 ##########################
F_observed_Model_H4 = ((anova_M5$`Sum Sq`[1]-anova_M4$`Sum Sq`[2])/(anova_M5$Df[1]-
anova_M4$Df[2]))/(anova_M4$`Sum Sq`[2]/anova_M4$Df[2])
cat("The test statistic for model H4 is:", F_observed_Model_H4, "\n")

# Calculate the critical F-value
alpha = 0.05
df14 <- anova_M5$Df[1]-anova_M4$Df[2]  # Degrees of freedom for the numerator
df24 <- anova_M4$Df[2]  # Degrees of freedom for the denominator
# Calculate p-value for the F-test
p_value_Model_H4 <- 1 - pf(F_observed_Model_H4, df14, df24)
cat("The p-value for model H4 is:", p_value_Model_H4, "\n")

# Create a data frame with the results
variation  M1  M2 = anova  M2$`Sum Sq`[8]-anova  M1$`Sum Sq`[11]
variation_M2_M3 = anova_M3$`Sum Sq`[6]-anova_M2$`Sum Sq`[8]
variation_M3_M4 = anova_M4$`Sum Sq`[2]-anova_M3$`Sum Sq`[6]
variation_M4_M5 = anova_M5$`Sum Sq`[1]-anova_M4$`Sum Sq`[2]

results_table <- data.frame(
  Model = c("M1 vs M2", "M2 vs M3", "M3 vs M4","M4 vs M5"),
  Variation = c(variation_M1_M2,variation_M2_M3,variation_M3_M4,variation_M4_M5),
  Test_Statistic = c(F_observed_Model_H1, F_observed_Model_H2,
F_observed_Model_H3,F_observed_Model_H4),
  P_Value = c(p_value_Model_H1, p_value_Model_H2, p_value_Model_H3,p_value_Model_H4)
)

# Print the results table
print(results_table)
```

Result in R:

```
> results_table
     Model Variation Test_Statistic      P_Value
1 M1 vs M2 167.73563       5.846159 0.0006635838
2 M2 vs M3  25.00208       1.254838 0.2864051989
3 M3 vs M4 147.64496       3.699730 0.0057656078
4 M4 vs M5  59.56735       5.794361 0.0165874858
```

As we can see we reject M2 and thus accept M1
ANSWER 1

Problem 4.

As we are getting close to Christmas, we will now consider the production of fermented herring. A delicacy in the Nordic countries, and a must on the table for 'juleforkost'. The data is from http://models.kvl.dk/Ripening_of_Herring and is described in this article: *Rasmus Bro, Henrik Hauch Nielsen, Guðmundur Stefánsson, Torstein Skåra, A Phenomenological Study of Ripening of Salted Herring. Assessing homogeneity of data from different countries and laboratories; J. Chemom., 16:81-88, 2002*

The data compares three countries Denmark, Norway, Iceland and 5 different treatments, e.g. if the herring is beheaded and gutted or only gutted. Note that there are missing values in the dataset, so not all observations read are necessarily used.

We will only consider a subset of variables. A detailed understanding of the variables is not necessary.

Variable	Meaning	Decription
ProteinB	Protein, brine	Solubilisation of protein fragments and salt soluble protein
AshM	Ash, muscle	Salt uptake (salt content generally 1 % lower than ashM)
TCAB	Trichloroacetic acid soluble nitrogen, brine	Level of small nitrogenous compounds and protein degradation products that is solubilised in brine. Smell of brine is a traditional quality parameter.
TCAM	Trichloroacetic acid soluble nitrogen, muscle	Level of protein degradation (caused by enzymes)
TCAIndexM	Trichloroacetic acid index, muscle	Level of protein degradation relative to total protein content
TCAIndexB	Trichloroacetic acid index, brine	Level of protein degradation relative to total protein content
Water	Water, muscle	

We will start by investigating if there is a difference between countries and treatments with a model of the form:

$$[\text{ProteinB} \quad \text{TCAIndexM} \quad \text{TCAIndexB} \quad \text{TCAM} \quad \text{TCAB}] = \mu + country_k + treatment_m$$

 # Load the dataset in R

```
data_P4_trial1=read.csv("Exam_2020_without_cleaning.csv")
selected_data_trial1 <- subset(data_P4_trial1, select = c("ProteinB", "TCAIndexM", "TCAIndexB",
"TCAM", "TCAB", "Treatment", "Country"))
```

Question 4.1.

Using only *ProteinB* and *TCAIndexM* to test for treatment effect, the usual test-statistic (Wilk's Lambda/Anderson's U) is:

We use:

Theorem 4.26

The ratio test at level α for test of H_0 against H_1 is given by the critical region

$$\{y_{11}, \ldots, y_{km} \left| \frac{\det(\mathbf{q}_1)}{\det(\mathbf{q}_1 + \mathbf{q}_2)} \leq U(p, k-1, (k-1)(m-1))_\alpha \right.\}.$$

The ratio test at level α for test of K_0 against K_1 is given by the critical region

$$\{y_{11}, \ldots, y_{km} \left| \frac{\det(\mathbf{q}_1)}{\det(\mathbf{q}_1 + \mathbf{q}_3)} \leq U(p, m-1, (k-1)(m-1))_\alpha \right.\}.$$

In R:

```
# Load the necessary libraries (if not already loaded)
library(car)

# Perform MANOVA using Anova()
manova_result_0 <- Anova(
  lm(cbind(ProteinB, TCAIndexM, TCAIndexB, TCAM, TCAB) ~ Country + Treatment, data =
selected_data_trial1),
  type = 3  # Type III sums of squares
)

# Print MANOVA results
summary(manova_result_0)
manova_summary = summary(manova_result_0)

# Extract the table with "Sum of squares and products for error"
SS_error = manova_summary$SSPE
print("Sum of squares and products for error is:")
print(SS_error)

# Extract the values for ProteinB and TCAIndexM
subset_matrix_q1 <- SS_error[c("ProteinB", "TCAIndexM"), c("ProteinB", "TCAIndexM")]

# Print the subset matrix
print(subset_matrix_q1)


print("Sum of squares and products for the hypothesis for treatment:")
SS_treatment = manova_summary$multivariate.tests$Treatment$SSPH
# Extract the values for ProteinB and TCAIndexM
subset_matrix_q3 <- SS_treatment[c("ProteinB", "TCAIndexM"), c("ProteinB", "TCAIndexM")]

# Print the subset matrix
print(subset_matrix_q3)

Test_statistic = det(subset_matrix_q1)/(det(subset_matrix_q1+subset_matrix_q3))
cat("The test statistic is:", Test_statistic, "\n")
```

Results in R:

q1:

```
Type III MANOVA Tests:

Sum of squares and products for error:
           ProteinB TCAIndexM  TCAIndexB       TCAM       TCAB
ProteinB   550.7594  1935.952  -243.0044  310.1376  392.6320
TCAIndexM 1935.9519 11609.487  9408.2291 1933.3853 1994.4879
TCAIndexB -243.0044  9408.229 58232.8075 1741.7389 2690.9405
TCAM       310.1376  1933.385  1741.7389  341.4653  332.7548
TCAB       392.6320  1994.488  2690.9405  332.7548  444.3292
```

q3:

```
Term: Treatment

Sum of squares and products for the hypothesis:
            ProteinB TCAIndexM  TCAIndexB       TCAM       TCAB
ProteinB    1.082918   19.3252   106.6121   5.657764   6.233976
TCAIndexM  19.325203   344.8679  1902.5468 100.965618 111.248404
TCAIndexB 106.612147 1902.5468 10495.8584 557.001194 613.728660
TCAM        5.657764   100.9656   557.0012  29.559310  32.569761
TCAB        6.233976   111.2484   613.7287  32.569761  35.886809
```

The test-statistic is: $\frac{\det(\boldsymbol{q1})}{\det(\boldsymbol{q1}+\boldsymbol{q3})} = \frac{\det(\begin{bmatrix} 550.75940273 & 1935.9518735 \\ 1935.9518735 & 11609.48673 \end{bmatrix})}{\det(\begin{bmatrix} 550.75940273 & 1935.9518735 \\ 1935.9518735 & 11609.48673 \end{bmatrix}+\begin{bmatrix} 1.0829176098 & 19.325203335 \\ 19.325203335 & 344.86786487 \end{bmatrix})} = 0.9540$

R result for test-statistic: The test statistic is: 0.9539673

ANSWER 3

Question 4.2.

If we only consider *country* effect on the individual variables, the variable with largest country effect as measured by F-value is:

In R:

```
print("Sum of squares and products for error is:")
print(SS_error)
print("Sum of squares and products for the hypothesis for country:")
SS  country = manova_summary$multivariate.tests$Country$SSPH
print(SS_country)
```

Results in R:

We find **q1** and **q2** in the results:

q1:

```
Type III MANOVA Tests:

Sum of squares and products for error:
           ProteinB TCAIndexM  TCAIndexB       TCAM       TCAB
ProteinB   550.7594  1935.952  -243.0044  310.1376  392.6320
TCAIndexM 1935.9519 11609.487  9408.2291 1933.3853 1994.4879
TCAIndexB -243.0044  9408.229 58232.8075 1741.7389 2690.9405
TCAM       310.1376  1933.385  1741.7389  341.4653  332.7548
TCAB       392.6320  1994.488  2690.9405  332.7548  444.3292
```

q2:

```
> print(SS_country)
           ProteinB TCAIndexM TCAIndexB       TCAM       TCAB
ProteinB  22.348872 1.5570432  49.76397  8.8822477 19.120286
TCAIndexM  1.557043 0.1084790   3.46705  0.6188251  1.332108
TCAIndexB 49.763974 3.4670500 110.80886 19.7779979 42.574918
TCAM       8.882248 0.6188251  19.77800  3.5301256  7.599091
TCAB      19.120286 1.3321080  42.57492  7.5990910 16.358111
```

We calculate the F-values, while ignoring the degrees of freedom, as they are the same across.

This is simply a SSeffect / SSerror, as known from introduction to statistics. We also use it for model selection, see e.g. page 230 in the notes. If you needed the degrees of freedom (which we do not in this case), refer to Remark 2.25 on page 133.

ProteinB	$\frac{22.348871964}{550.75940273} = 0.0406$
TCAIndexM	$\frac{0.1084790115}{11609.48673} = 9.3440e-06$
TCAIndexB	$\frac{110.80886257}{58232.80745} = 0.0019$
TCAM	$\frac{3.530125592}{341.46534636} = 0.0103$
TCAB	$\frac{16.358110533}{444.32920438} = 0.0368$

ProteinB is the largest

ANSWER 1

We will now try to use the variables to discriminate between the observations and see if we can tell the country of origin. To that end we will consider

- a full model using all variables: *water ashm ProteinB TCAIndexM TCAIndexB TCAM TCAB*
- a reduced using only *ProteinB TCAIndexM TCAIndexB TCAM TCAB*

Question 4.3.

The number of misclassifications in the full model is

 # Load the dataset in R

```
data_P4_trial2=read.csv("Exam_2020_cleaning_data_without_transf.csv")
selected_data_trial2 <- subset(data_P4_trial2, select = c("ProteinB", "TCAIndexM", "TCAIndexB",
"TCAM", "TCAB", "Treatment", "Country"))
```

In R:

```
data_P4_trial2$Country <- ifelse(data_P4_trial2$Country == 1, "A",
                  ifelse(data_P4_trial2$Country == 2, "B",
                       ifelse(data_P4_trial2$Country == 3, "C", data_P4_trial2$Country)))
library(MASS)


# Prior probabilities fo the classes
# SAS by default has equal prior probabilities, so we need equal prior in R to receive the same results
different_countries <- unique(data_P4_trial2$Country)
num_countries  <- length(different_countries)
prior  <-  rep(1/num_countries, num_countries)
###
#linear discriminant analysis
# Define the classes (mfr)
data_P4_trial2$class <- as.factor(data_P4_trial2$Country)
# Define the variables for the analysis
variables <- c("Water", "AshM", "ProteinB", "TCAIndexM", "TCAIndexB" ,"TCAM" ,"TCAB")
# Perform Linear Discriminant Analysis
z <- lda(class ~ ., data = data_P4_trial2[, c("class", variables)],prior=prior)


Class_Level_Information = data.frame("Frequency" = z$counts,"Proportion"=z$counts/z$N,"Prior"=z$prior)
print("Class Level Information:")
Class_Level_Information


n <- nrow(data  P4  trial2)
Classes <- nlevels(data_P4_trial2$Country)

paste0("DF Within Classes = ",n-Classes)
paste0("DF Between Classes = ",Classes-1)

zpred <- predict(z)

#Confusion Matrix:
print("Confusion Matrx:")
xtabs(~data_P4_trial2$Country+zpred$class)
```

Result in R:

```
[1] "Confusion Matrx:"
> xtabs(~data_P4_trial2$Country+zpred$class)
                        zpred$class
data_P4_trial2$Country  A  B  C
                     A 29 16 20
                     B  8 50  0
                     C  7  5 82
```

We count the off-diagonal elements
16+20+8+0+7+5 = 56

ANSWER 4

Question 4.4.

We now test if *water* and *ashm* contribute to the discrimination between the country 1 and 2 using Linear Discriminant Analysis. The usual test statistic is given by:

We use

Theorem 5.21

The critical region for testing the hypothesis that the last $p-q$ variables do not contribute to the discrimination between the populations π_1 and π_2, i.e. the hypothesis that $\Delta^2_{(2|1)} = 0$ against all alternatives is

$$\left\{x_{11}, \cdots, x_{2n_2} \;\middle|\; \frac{n_1+n_2-p-1}{p-q} \frac{d^2-d_1^2}{(n_1+n_2)(n_1+n_2-2)/(n_1n_2)+d_1^2} > F(p-q, n_1+n_2-p-1)_{1-\alpha}\right\}$$

Here d^2 and d_1^2 are the observed values of D^2 and D_1^2.

In R:

```
data_P4_trial2$Country = as.factor(data_P4_trial2$Country)
library(Rfast)
pcov  <- pooled.cov(as.matrix(data_P4_trial2[,variables]),data_P4_trial2$Country)
Means <- as.matrix(z$means)
invCov <- solve(pcov)
# Extract unique levels from data_P4_trial2$Country
unique_levels <- levels(data_P4_trial2$Country)
num_col <- length(unique(data_P4_trial2$Country))

# Create an empty matrix to store the Mahalanobis distances with the equal priors ################
maha <- matrix(c(rep(0, num_col^2)), ncol = num_col)

# Define the names for rows and columns
rownames(maha) <- unique_levels
colnames(maha) <- unique_levels

for (i in 1:num_col) {
  for (j in 1:num_col) {
    mu <- Means[i, ] - Means[j, ]
    maha[i, j] <- mu %*% invCov %*% mu
  }
}

# squared Mahalanobis distances between the 3 country types.
```

```
# maha    : Assuming equal priors

print("Generalized Squared Distance to countries (equal priors):")
maha

############ Reduced model ##############
variables_reduced <- c("ProteinB", "TCAIndexM", "TCAIndexB" ,"TCAM" ,"TCAB")
# Perform Linear Discriminant Analysis
z_reduced <- lda(class ~ ., data = data_P4_trial2[, c("class", variables_reduced)],prior=prior)

library(Rfast)
pcov_reduced  <- pooled.cov(as.matrix(data_P4_trial2[,variables_reduced]),data_P4_trial2$Country)
Means_reduced <- as.matrix(z_reduced$means)
invCov <- solve(pcov_reduced)

# Extract unique levels from data_P4_trial2$Country
unique_levels <- levels(data_P4_trial2$Country)
num_col <- length(unique(data_P4_trial2$Country))

# Create an empty matrix to store the Mahalanobis distances with the equal priors ###############
maha_reduced <- matrix(c(rep(0, num_col^2)), ncol = num_col)


# Define the names for rows and columns
rownames(maha_reduced) <- unique_levels
colnames(maha_reduced) <- unique_levels

for (i in 1:num  col) {
  for (j in 1:num_col) {
    mu <- Means_reduced[i, ] - Means_reduced[j, ]
    maha_reduced[i, j] <- mu %*% invCov %*% mu
  }
}

# squared Mahalanobis distances between the 3 country types.
# maha    : Assuming equal priors

print("Generalized Squared Distance to countries (equal priors):")
maha_reduced

#### Class Level Information from the previous question
print("Class Level Information:")
Class_Level_Information
```

Results in R:

Full model

```
[1] "Generalized Squared Distance to countries (equal priors):"
> maha
          A        B        C
A 0.000000 2.135084 2.637393
B 2.135084 0.000000 6.969306
C 2.637393 6.969306 0.000000
>
```

Reduced model

```
> maha_reduced
          A         B        C
A 0.0000000 0.7091284 1.518500
B 0.7091284 0.0000000 2.463813
C 1.5184996 2.4638131 0.000000
```

Class Level Information

```
[1] "Class Level Information:"
> Class_Level_Information
  Frequency Proportion     Prior
A        65  0.2995392 0.3333333
B        58  0.2672811 0.3333333
C        94  0.4331797 0.3333333
```

We have n1=65 and n2 = 58 from the class level information table and we know that the full model has p=7 variables and the reduced has q=5. Inserting these values into the equation of the theorem 5.21, we have:

$$\frac{65+58-7-1}{7-5}\cdot\frac{2.13508-0.70913}{(65+58)(65+58-2)/(65\cdot 58)+0.70913}$$

ANSWER 2

Question 4.5.

We now consider the reduced model. The class sensitivity is [country1, country2, country3]

We either use 1- error rate, where the error rate is given in the output. Alternatively we use

	Classified as		Sum
	π_i	not π_i	
Nature π_i	$tp_i = N_{ii}$ = # true positive	$fn_i = N_{i.} - N_{ii}$ = # false negative	$P_i = N_{i.}$ = # from π_i
Nature not π_i	$fp_i = N_{.i} - N_{ii}$ = # false positive	$tn_i = N_{..} - N_{i.} - N_{.i} + N_{ii}$ = # true negative	$NN_i = N_{..} - N_{i.}$ = # in all classes but π_i
Sum	$CP_i = N_{.i}$ = # clas. as π_i	$CN_i = N_{..} - N_{.i}$ = # clas. as not π_i	$TN = N_{..}$ = total # classified

Table 5.7 – The binary confusion matrix for class π_i based on the $k \times k$ confusion matrix.

Measure	Formula
Average class accuracy	$\frac{1}{k}\sum_{i=1}^{k}\frac{tp_i+tn_i}{TN} = \frac{2}{kN_{..}}\sum_{i=1}^{k}N_{ii} + \frac{(k-2)}{k} = 1 - \frac{2}{k}\frac{1}{N_{..}}\sum_{i\neq j}N_{ij}$
Average class error rate, misclassification rate	$\frac{1}{k}\sum_{i=1}^{k}\frac{fp_i+fn_i}{TN} = \frac{2}{k} - \frac{2}{k}\frac{1}{N_{..}}\sum_{i=1}^{k}N_{ii} = \frac{2}{k}\frac{1}{N_{..}}\sum_{i\neq j}N_{ij}$
Average class precision	$\frac{1}{k}\sum_{i=1}^{k}\frac{tp_i}{tp_i+fp_i} = \frac{1}{k}\sum_{i=1}^{k}\frac{N_{ii}}{N_{.i}}$
Average class sensitivity	$\frac{1}{k}\sum_{i=1}^{k}\frac{tp_i}{tp_i+fn_i} = \frac{1}{k}\sum_{i=1}^{k}\frac{N_{ii}}{N_{i.}}$
Average class specificity	$\frac{1}{k}\sum_{i=1}^{k}\frac{tn_i}{fp_i+tn_i} = \frac{1}{k}\sum_{i=1}^{k}\frac{N_{..}-N_{i.}-(N_{.i}-N_{ii})}{N_{..}-N_{i.}}$

Table 5.8 – The average uncertainties for the k−class classification problem

In R:

```
# Define the variables for the analysis
variables_reduced <- c("ProteinB", "TCAIndexM", "TCAIndexB" ,"TCAM" ,"TCAB")
# Perform Linear Discriminant Analysis
z_reduced <- lda(class ~ ., data = data_P4_trial2[, c("class", variables_reduced)],prior=prior)


Class_Level_Information_reduced = data.frame("Frequency" =
z_reduced$counts,"Proportion"=z_reduced$counts/z_reduced$N,"Prior"=z_reduced$prior)
print("Class Level Information:")
Class_Level_Information_reduced


n <- nrow(data_P4_trial2)
Classes <- nlevels(data_P4_trial2$Country)

paste0("DF Within Classes = ",n-Classes)
paste0("DF Between Classes = ",Classes-1)

zpred_reduced <- predict(z_reduced)

#Confusion Matrix:
print("Confusion Matrix:")
xtabs(~data_P4_trial2$Country+zpred_reduced$class)
```

Result in R:

```
[1] "Confusion Matrix:"
> xtabs(~data_P4_trial2$Country+zpred_reduced$class)
                      zpred_reduced$class
data_P4_trial2$Country  A  B  C
                     A 32 16 17
                     B  7 46  5
                     C 19  9 66
```

Country1: sensitivity = $\frac{32}{32+16+17} = 0.4923$

Country2: sensitivity = $\frac{46}{46+7+} = 0.7931$

Country3: sensitivity = $\frac{66}{66+19} = 0.7021$

ANSWER 2

Question 4.6.

We only consider country 1 and 2 in the reduced model. The usual test-statistic for difference in mean values is.

We use

Theorem 5.12

Using the significance level α, the critical area for a test of the hypothesis $\mu_1 = \mu_2$ against all alternatives becomes

$$C = \left\{ x_{11}, \cdots, x_{1n_1}, x_{21}, \cdots, x_{2n_2} \;\middle|\; \frac{n_1+n_2-p-1}{p(n_1+n_2-2)} \cdot \frac{n_1 n_2}{n_1+n_2} d^2 > F(p, n_1+n_2-p-1)_{1-\alpha} \right\}$$

Here d^2 is the observed value of D^2.

We have found in R from the previous questions the following tables:

```
> maha_reduced
          A         B        C
A 0.0000000 0.7091284 1.518500
B 0.7091284 0.0000000 2.463813
C 1.5184996 2.4638131 0.000000
```

```
[1] "Class Level Information:"
> Class_Level_Information
  Frequency Proportion     Prior
A        65  0.2995392 0.3333333
B        58  0.2672811 0.3333333
C        94  0.4331797 0.3333333
```

We have n1=65 and n2=58

$$\frac{65+58-5-1}{5(65+58-2)} \cdot \frac{65 \cdot 58}{65+58} 0.70913 =$$

ANSWER 5

Problem 5

We consider a random variable

$$\begin{bmatrix} \boldsymbol{Y} \\ \boldsymbol{X} \end{bmatrix} = \begin{bmatrix} Y_1 \\ Y_2 \\ X_1 \\ X_2 \end{bmatrix}$$

with expectation vector and dispersion matrix equal to

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho & 0 & \rho \\ \rho & 1 & \rho & 0 \\ 0 & \rho & 1 & \rho \\ \rho & 0 & \rho & 1 \end{bmatrix}$$

In the sequel you may find the following expressions useful

$$\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}^{-1} = \frac{1}{1-\rho^2} \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}$$

$$\det\left\{ \begin{bmatrix} 0 & \rho \\ \rho & 0 \end{bmatrix} \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}^{-1} \begin{bmatrix} 0 & \rho \\ \rho & 0 \end{bmatrix} - a \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right\} = \left\{ \frac{\rho^2}{1-\rho} - a(1-\rho) \right\} \left\{ \frac{\rho^2}{1+\rho} - a(1+\rho) \right\}$$

$$\det \begin{bmatrix} 1 & 0 & \rho \\ 0 & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix} = 1 - 2\rho^2$$

$$\det \begin{bmatrix} 1 & \rho & 0 \\ \rho & 1 & \rho \\ 0 & \rho & 1 \end{bmatrix} = 1 - 2\rho^2$$

$$\det \boldsymbol{\Sigma} = 1 - 4\rho^2$$

Question 5.1.

For which values of ρ is $\mathbf{\Sigma}$ a proper dispersion matrix?

We use

Theorem 1.5

The variance-covariance matrix $\mathbf{\Sigma}$ for a multidimensional random variable is positive semidefinite. This is a necessary and sufficient condition.

The principal minors are

$$[1], \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \begin{bmatrix} 1 & \rho & 0 \\ \rho & 1 & \rho \\ 0 & \rho & 1 \end{bmatrix}, \begin{bmatrix} 1 & \rho & 0 & \rho \\ \rho & 1 & \rho & 0 \\ 0 & \rho & 1 & \rho \\ \rho & 0 & \rho & 1 \end{bmatrix}$$

with determinants

$$1, 1-\rho^2, 1-2\rho^2, 1-4\rho^2$$

These are all positive if

$$1-4\rho^2 > 0$$

or

$$\rho^2 < \frac{1}{4} \Leftrightarrow |\rho| < \frac{1}{2}$$

ANSWER 1

Question 5.2.

The variance $\mathrm{V}(Y_1 - Y_2)$ of $Y_1 - Y_2$ is

We use

Remark 1.10 Rules for computing moments of simple functions

$\mathrm{E}(a+bX) = a + b\,\mathrm{E}(X)$ $\mathrm{E}(X+Y) = \mathrm{E}(X) + \mathrm{E}(Y)$	$\mathrm{E}(\mathbf{A}+\mathbf{X}) = \mathbf{A} + \mathrm{E}(\mathbf{X})$ $\mathrm{E}(\mathbf{AX}) = \mathbf{A}\,\mathrm{E}(\mathbf{X})$ $\mathrm{E}(\mathbf{XB}) = \mathrm{E}(\mathbf{X})\mathbf{B}$ $\mathrm{E}(\mathbf{X}+\mathbf{Y}) = \mathrm{E}(\mathbf{X}) + \mathrm{E}(\mathbf{Y})$
$\mathrm{V}(a+bX) = b^2\,\mathrm{V}(X)$ $\mathrm{V}(X+Y) = \mathrm{V}(X) + \mathrm{V}(Y) + 2\mathrm{Cov}(X,Y)$ $= \mathrm{V}(X) + \mathrm{V}(Y)$ iff X, Y independent	$\mathrm{D}(\mathbf{b}+\mathbf{X}) = \mathrm{D}(\mathbf{X})$ $\mathrm{D}(\mathbf{AX}) = \mathbf{A}\,\mathrm{D}(\mathbf{X})\mathbf{A}^T$ $\mathrm{D}(\mathbf{X}+\mathbf{Y}) = \mathrm{D}(\mathbf{X}) + \mathrm{D}(\mathbf{Y}) + \mathrm{C}(\mathbf{X},\mathbf{Y}) + \mathrm{C}(\mathbf{Y},\mathbf{X})$ $= \mathrm{D}(\mathbf{X}) + \mathrm{D}(\mathbf{Y})$ iff X, Y independent
$\mathrm{Cov}(X,X) = \mathrm{V}(X)$ $\mathrm{Cov}(aX,bY) = ab\mathrm{Cov}(X,Y)$ $\mathrm{Cov}(X+U,Y) = \mathrm{Cov}(X,Y) + \mathrm{Cov}(U,Y)$ $\mathrm{Cov}(X,Y+V) = \mathrm{Cov}(X,Y) + \mathrm{Cov}(X,V)$	$\mathrm{C}(\mathbf{X},\mathbf{X}) = \mathrm{D}(\mathbf{X})$ $\mathrm{C}(\mathbf{X},\mathbf{Y}) = \mathrm{C}(\mathbf{Y},\mathbf{X})^T$ $\mathrm{C}(\mathbf{AX},\mathbf{BY}) = \mathbf{A}\,\mathrm{C}(\mathbf{X},\mathbf{Y})\mathbf{B}^T$ $\mathrm{C}(\mathbf{X}+\mathbf{U},\mathbf{Y}) = \mathrm{C}(\mathbf{X},\mathbf{Y}) + \mathrm{C}(\mathbf{U},\mathbf{Y})$ $\mathrm{C}(\mathbf{X},\mathbf{Y}+\mathbf{V}) = \mathrm{C}(\mathbf{X},\mathbf{Y}) + \mathrm{C}(\mathbf{X},\mathbf{V})$

$$\mathrm{V}(Y_1 - Y_2) = \mathrm{V}(Y_1) + \mathrm{V}(Y_2) - 2\mathrm{Cov}(Y_1, Y_2) = 1 + 1 - 2\rho = 2 - 2\rho$$

$$[1 \quad -1] \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = 2 - 2\rho$$

ANSWER 1

Question 5.3.

The covariance $\mathrm{Cov}(Y_1 - Y_2,\ X_1 - X_2)$ is

We again use Remark 1.10 (see above)

$$\mathrm{Cov}(Y_1 - Y_2,\ X_1 - X_2) = \mathrm{Cov}(Y_1,\ X_1) - \mathrm{Cov}(Y_1, X_2) - \mathrm{Cov}(Y_2,\ X_1) + \mathrm{Cov}(Y_2,\ X_2)$$

$$= 0 - \ \rho - \ \rho + 0$$

$$= -2\,\rho$$

$$\begin{bmatrix} 1 & -1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & \rho & 0 & \rho \\ \rho & 1 & \rho & 0 \\ 0 & \rho & 1 & \rho \\ \rho & 0 & \rho & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \\ -1 \end{bmatrix} = -2\rho$$

ANSWER 1

Question 5.4.

The covariance $\mathrm{Cov}(Y_1 - Y_2,\ X_1 + X_2)$ is

We again use Remark 1.10 (see above)

$$\mathrm{Cov}(Y_1 - Y_2,\ X_1 + X_2) = \mathrm{Cov}(Y_1,\ X_1) + \mathrm{Cov}(Y_1, X_2) - \mathrm{Cov}(Y_2,\ X_1) - \mathrm{Cov}(Y_2,\ X_2)$$

$$= 0 + \ \rho - \ \rho + 0$$

$$= 0$$

$$\begin{bmatrix} 1 & -1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & \rho & 0 & \rho \\ \rho & 1 & \rho & 0 \\ 0 & \rho & 1 & \rho \\ \rho & 0 & \rho & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} = 0$$

ANSWER 3

Question 5.5.

The conditional mean $E(Y_1|X_1 = x_1)$ is

We use

Theorem 1.27

If X_2 is regularly distributed, i.e. if Σ_{22} has full rank, then the distribution of X_1 conditioned on $X_2 = x_2$ is again a normal distribution, and the following holds

$$\begin{aligned} \mathrm{E}(X_1|X_2 = x_2) &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \\ \mathrm{D}(X_1|X_2 = x_2) &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \end{aligned}$$

If Σ_{22} does not have full rank then the conditional distribution is still normal and Σ_{22}^{-1} in the above equations should be substituted by a generalised inverse Σ_{22}^{-}.

We extract the relevant parts of the dispersion matrix and get

$$D\left(\begin{bmatrix} Y_1 \\ X_1 \end{bmatrix}\right) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Since (in the notation from T1.27) Σ_{12} is 0 we simply get the mean of Y_1 which is 0

ANSWER 3

Question 5.6.

The conditional mean $E(Y_1|\boldsymbol{X} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix})$ is

We again use T1.27 (see above). We extract the relevant parts of the dispersion matrix

$$D\left(\begin{bmatrix} Y_1 \\ X_1 \\ X_2 \end{bmatrix}\right) = \begin{bmatrix} 1 & 0 & \rho \\ 0 & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}$$

We insert

$$E\left(Y_1 \left| \boldsymbol{X} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right.\right) = \mathrm{E}(Y_1) + [0 \quad \rho] \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}^{-1} (\boldsymbol{x} - \boldsymbol{0})$$

$$= \frac{1}{1-\rho^2} [0 \quad \rho] \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$= \frac{1}{1-\rho^2} [-\rho^2 \quad \rho] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$= -\frac{\rho^2}{1-\rho^2} x_1 + \frac{\rho}{1-\rho^2} x_2$$

ANSWER 1

Question 5.7.

The squared multiple correlation $\rho^2_{Y_1|X_1X_2}$ between Y_1 and $[X_1 \quad X_2]^T$ is

We use

Theorem 1.42

We consider the situation above. Let σ_i be the i'th column in Σ_{xy}, i.e. σ_i^T is the i'th row in Σ_{yx}. Further, let σ_{ii} denote the i'th diagonal element, i.e. the variance of Y_i

Then

$$\rho_{y_i|x} = \frac{\sqrt{\sigma_i^T \Sigma_{xx}^{-1} \sigma_i}}{\sqrt{\sigma_{ii}}}.$$

If we let

$$\Sigma_i = \begin{bmatrix} \sigma_{ii} & \sigma_i^T \\ \sigma_i & \Sigma_{xx} \end{bmatrix},$$

then

$$1 - \rho_{y_i|x}^2 = \frac{\det \Sigma_i}{\sigma_{ii} \det \Sigma_{xx}} = \frac{V(Y_i|X)}{V(Y_i)},$$

We again extract the relevant parts of the dispersion matrix – the same as in the previous question

$$D\left(\begin{bmatrix} Y_1 \\ X_1 \\ X_2 \end{bmatrix}\right) = \begin{bmatrix} 1 & 0 & \rho \\ 0 & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}$$

We insert

$$\rho_{Y_1|X_1X_2}^2 = 1 - \frac{\det\left(\begin{bmatrix} 1 & 0 & \rho \\ 0 & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}\right)}{1 \cdot \det\left(\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)} = 1 - \frac{1 - 2\rho^2}{1 - \rho^2} = \frac{\rho^2}{1 - \rho^2}$$

ANSWER 1

Question 5.8.

For positive ρ, the maximum squared correlation between any linear combination of Y_1 & Y_2 and any linear combination of X_1 & X_2 is

We identify the problem as canonical correlation and use

We are looking for the squared canonical correlation between Y_1 & Y_2 and X_1 & X_2. We use theorem 6.13

Theorem 6.13

Let the situation be as in the previous theorem. Then we have

$$(\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy} - \varrho_r^2\boldsymbol{\Sigma}_{yy})\boldsymbol{a}_r = \mathbf{0}$$
$$\det(\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy} - \varrho_r^2\boldsymbol{\Sigma}_{yy}) = 0$$

respectively

$$(\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx} - \varrho_r^2\boldsymbol{\Sigma}_{xx})\boldsymbol{b}_r = \mathbf{0}$$
$$\det(\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx} - \varrho_r^2\boldsymbol{\Sigma}_{xx}) = 0$$

and have

$$\boldsymbol{\Sigma}_{yx} = \begin{bmatrix} 0 & \rho \\ \rho & 0 \end{bmatrix},\ \boldsymbol{\Sigma}_{xy} = \begin{bmatrix} 0 & \rho \\ \rho & 0 \end{bmatrix},\ \boldsymbol{\Sigma}_{yy} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix},\ \boldsymbol{\Sigma}_{xx} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

The second equation in 6.13 becomes

$$\det\left\{\begin{bmatrix} 0 & \rho \\ \rho & 0 \end{bmatrix}\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}^{-1}\begin{bmatrix} 0 & \rho \\ \rho & 0 \end{bmatrix} - a\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right\} = 0$$

which is equivalent to

$$\left\{\frac{\rho^2}{1-\rho} - a(1-\rho)\right\}\left\{\frac{\rho^2}{1+\rho} - a(1+\rho)\right\}$$

the solutions for a are

$$\frac{\rho^2}{(1-\rho)^2} \text{ and } \frac{\rho^2}{(1+\rho)^2}$$

For positive ρ, the largest solution is the first, and it follows that the answer is 5.

ANSWER 5

LAST PAGE:
END OF THE EXAM SET