02409 Multivariate Statistics - 2023

4 hours written exam - All aids allowed, including internet access.

There are 7 problems with a total of 30 questions.
The questions in a specific problem should be read in order.

The questions are weighted equally

A correct answer gives 5 points, a wrong answer gives –1 point. Unanswered questions or a “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.

You can find the data needed to solve the problems here in SAS-format (https://designer.mcq.eksamen.dtu.dk/api/images/e7fa560c-0dea-4fd7-b133-cfb65d7d1edf) and in R-format (https://designer.mcq.eksamen.dtu.dk/api/images/2c1a770b-bab7-4256-b7e2-dc2001f93b7f) . The problem will specifically say, if you need to use the dataset.

For this problem we will use data from Statistikbanken. The details of the data and the individial variables, are only important to the degree that are covered in each problem and questions. A brief overview is given below.

We have people and their highest finished level of education by municipality. These are the variables *H10 H1001 H1010 ... H9099.* Each variable coves a type of educations and denote how many people in a municipality have that specific type of education as their highest level of completed education.
Added to this dataset we have the number of people in each municipality that are members of the Church of Denmark (Folkekirken), variable c*hurch.*
We have whether or not each municipality is a net recipient or donor of funding, the variable *gives.*
Finally, we have how many people in each municipality are participating in specific sports, variables such as *HorseRiding* and *Sailing.*

PROBLEM 1 - Education and church membership

We use the dataset given in the description of the exam.

We will investigate a model of the form:

$$Church = \alpha + \beta_1 \cdot H3020 + \beta_2 \cdot H4058 + \beta_3 \cdot H5020 + \beta_4 \cdot H8090 + \epsilon,$$

where $Church$ is the church membership per municipality, α is an intercept, the β's are estimated parameters, and ϵ is the residual.

H3020 is a vocational education related to food
H4058 is a short technical education, requiring a high school diploma
H5020 is a medium length pedagodigal education
H8090 is a PhD in a health related area

Question 1.1

The R^2 of the model is

Vælg en svarmulighed

- 93/97
- $1 - \frac{877734461}{1.792437 \cdot 11^{11}}$
- 0.9949
- $\frac{3072.13}{42077}$
- 0.9951
- Don't know

Question 1.2

The following number of observations have an absolute Rstudent value larger than 2

Vælg en svarmulighed

- ◯ Don't know
- ◯ 1
- ◯ 2
- ◯ 4
- ◯ 5
- ◯ 3

Question 1.3

The number of influential observations in the model, i.e., observations with both an absolute RSTUDENT larger than 2 and a leverage larger than 0.102 are:

Vælg en svarmulighed

- 1
- 2
- 3
- 4
- Don't know
- 5

Question 1.4

What observation – if removed – would change the estimated parameter vector the most (measured by an appropriate confidence region)

Vælg en svarmulighed

- ◯ Don't know
- ◯ Observation 5
- ◯ Observation 4
- ◯ Observation 3
- ◯ Observation 2
- ◯ Observation 1

Question 1.5

We now consider if there is multicollinearity present in the model:

Vælg en svarmulighed

- ◯ Yes, there are several independent variables with a VIF larger than 10 and a Tolerance less than 0.1
- ◯ No, all independent variables have a VIF smaller than 10 and a Tolerance larger than 0.1
- ◯ Yes, there are several independent variables with a VIF smaller than 10 and a Tolerance larger than 0.1
- ◯ No, since at least one variable has a Tolerance larger than 0.1 and a VIF smaller than 10
- ◯ No, all the parameters are significantly different from zero, as shown by the t-tests. We thus have certain estimates and no signs of multicollinearity.
- ◯ Don't know

Question 1.6

When performing a backwards elimination, the first variable to be eliminated has an F-value of:

Vælg en svarmulighed

- 17.9
- 3.6
- 4.2
- 4747.9
- Don't know
- 9.9

Question 1.7

Considering only the four independent variables in the model. When performing a forward selection, the first variable to be included is:

Vælg en svarmulighed

- ◯ H3020
- ◯ H5020
- ◯ H8090
- ◯ H3020 and H4085 are tied
- ◯ Don't know
- ◯ H4058

Question 1.8

Irregardless of the previous question we now consider two models:
M1: $Church = \alpha + \beta_1 \cdot H3020 + \beta_2 \cdot H4058 + \beta_3 \cdot H5020 + \beta_4 \cdot H8090 + \epsilon$,

M2: $Church = \tilde{\alpha} + \tilde{\beta}_1 \cdot H3020 + \tilde{\epsilon}$,

The usual test statistic for testing for a simpler model is:

Vælg en svarmulighed

- Don't know
- $\frac{(15460700518 - 877734461) / 96}{877734461 / 93}$
- 526.7
- $\frac{(15460700518 - 877734461) / (96 - 93)}{877734461 / 93}$
- 4747.92
- 1022.43

PROBLEM 2 - Education and municipal compensation

We still use the dataset from the exam description.

We will now look at the relationship between the number of people with a long tertiary (i.e. Masters or equivalent) education in each municipality and whether a given municipality is a net donor or recipient in the municipal tax equilisation scheme (kommunal udligning).

We will investigate this by means of a Linear Discriminant Analysis.
We have the class variable:
gives - R for Recieves and G for Gives -
and the educational variables:
H70 H7020 H7025 H7030 H7035 H7039 H7059 H7075 H7080 H7090 H7095

H70 Total (*note that the total does not equal the sum of the educations below, as some categories have been omitted*)
H7020 Educational
H7025 Humanities and Theological
H7030 Artistic
H7035 Science
H7039 Social Science
H7059 Technical Science
H7075 Food, Bio- and Laboratory Technology
H7080 Agriculture, Nature and Environment
H7090 Health Science
H7095 Police and Defence, etc.

We will use equal priors and losses.

Question 2.1

The number of resubstitution misclassifications in the model are:

Vælg en svarmulighed

- 0
- 3
- 1
- 2
- Don't know
- 7

Question 2.2

We now test whether the variables *H7020 H7025 H7030* can be omitted from the model. Using the usual test statistic we get a p-value in the range of:

Vælg en svarmulighed

- [0.01; 0.05[
- [0.005; 0.01[
- [0.001; 0.005[
- [0.05;1]
- Don't know
- [0; 0.001[

Question 2.3

We now consider the full model and change the priors to be 0.1 for the 'G' class and 0.9 for the 'R' class. The squared generalized distance from the mean of group G to itself is :

Vælg en svarmulighed

- *4.605*
- 44.333
- Don't know
- 0.211
- 32.133
- $-2 \cdot \log(0.2)$

PROBLEM 3 - Education and correlation

We still consider the dataset from the exam description.

We will now investigate the correlation between the variables: *H7035 H7095*

Question 3.1

The fraction of variance of *H7035* that can be explained by *H7095* is:

Vælg en svarmulighed

- 0.4855
- $\frac{25.50}{366.74}$
- Don't know
- 0.0001
- 0.69681
- $\frac{366.74}{730.90}$

Question 3.2

The usual test statistic for the correlation between *H7035* and *H7095* being different from zero is:

Vælg en svarmulighed

- 9.5187
- $\frac{0.69681}{\sqrt{1-0.69681^2}}\sqrt{98-2-2}$
- 90.60
- $\frac{0.69681^2}{1-0.69681^2}\frac{98-2-1}{2}$
- 12.3992
- Don't know

Question 3.3

The 90% confidence interval for the correlation between H7035 and H7095 is:

Vælg en svarmulighed

- $\hat{\rho} \pm 0.86107$
- [0.58; 0.79]
- [0.53; 0.81]
- [0.60; 0.77]
- [0.62; 0.76]
- Don't know

PROBLEM 4 - Education and sport

We will now investigate the relation between long tertiary education and locomotive sports, by means of a Canonical Correlation Analysis, and use the dataset from the exam description.

We study the relation between the education variables
H7020 H7025 H7030 H7035 H7039 H7059 H7075 H7080 H7090 H7095

and the sports variables
Cycling CanoeKayak Gliding Hanggliding HorseRiding Sailing Skiing SurfRafting

Question 4.1

The number of significant canonical correlations (at level 5%) are:

Vælg en svarmulighed

- ◯ Don't know
- ◯ 5
- ◯ 1
- ◯ 2
- ◯ 3
- ◯ 4

Question 4.2

Based on the first set of canonical variables, the least liked sport by people with a long education is:

Vælg en svarmulighed

- HorseRiding
- Cycling
- Don't know
- Sailing
- SurfRafting
- Skiing

Question 4.3

Based on the Canonical Correlation Analysis, the two educations that favor Gliding the most are

Vælg en svarmulighed

- ◯ H7059 and H7075
- ◯ H7035 and H7090
- ◯ Don't know
- ◯ H7075 and H7080
- ◯ H7030 and H7025
- ◯ H7020 and H7095

PROBLEM 5 - Sports across municipalities

For this problem, you must rely on the embedded tables and figures when answering the questions!

We will continue our investigations of the sports variables by means of a factor analysis.

Cycling CanoeKayak Gliding Hanggliding HorseRiding Sailing Skiing SurfRafting

Question 5.1

How many components/factors must we include to account for 90% of the variance

	Eigenvalue	Difference	Proportion	Cumulative
1	3.63522464	2.40142303	0.4544	0.4544
2	1.23380161	0.15264929	0.1542	0.6086
3	1.08115232	0.35973429	0.1351	0.7438
4	0.72141803	0.12339588	0.0902	0.8339
5	0.59802215	0.20522564	0.0748	0.9087
6	0.39279651	0.18852167	0.0491	0.9578
7	0.20427484	0.07096492	0.0255	0.9833
8	0.13330992		0.0167	1.0000

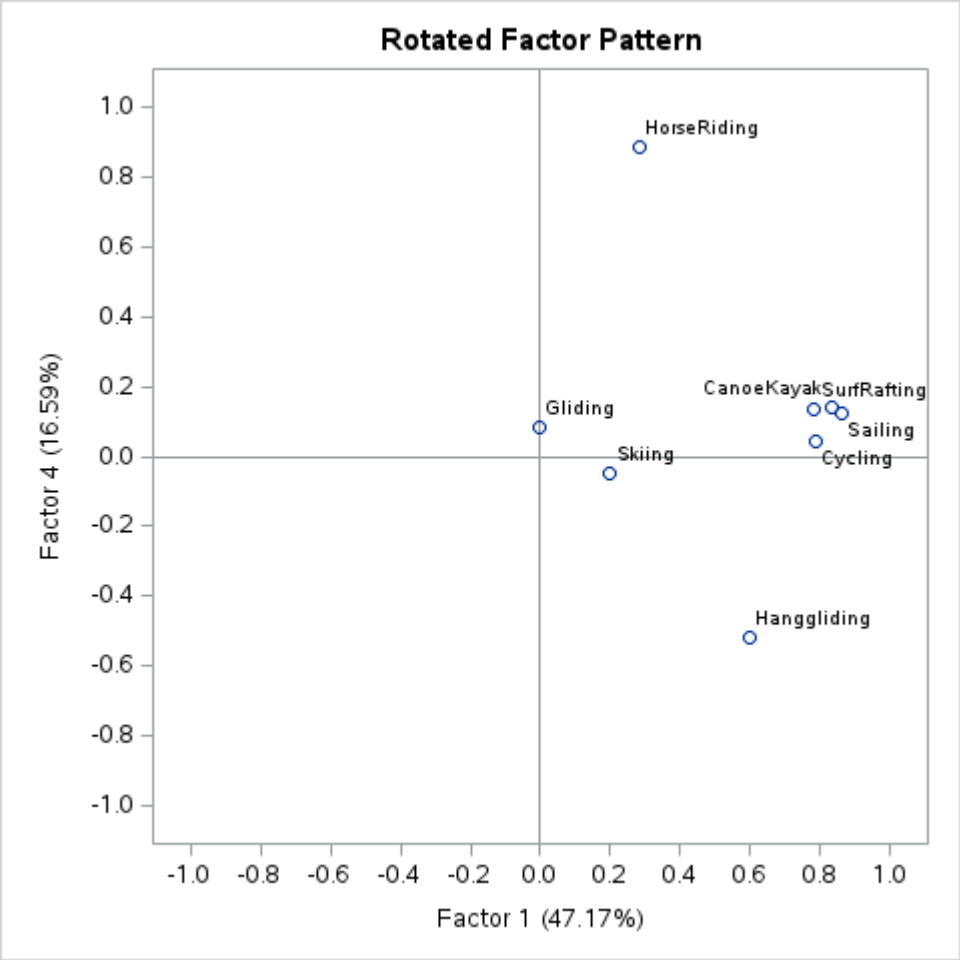
Vælg en svarmulighed

- ◯ 5
- ◯ 2
- ◯ Don't know
- ◯ 4
- ◯ 1
- ◯ 3

Question 5.2

We now perform a factor analysis with 4 factors.

The 4th VARIMAX rotated factor, can be interpreted as



Vælg en svarmulighed

- ◯ Don't know
- ◯ Contrast between *Cycling CanoeKayak Sailing SurfRafting* and *Gliding Skiing*
- ◯ Mean of *Cycling CanoeKayak Sailing SurfRafting*
- ◯ Mean of *Cycling CanoeKayak Sailing SurfRafting,* and a contrast between *HorseRiding* and *Hanggliding*
- ◯ Mean of all variables
- ◯ Contrast between *HorseRiding* and *Hanggliding*

PROBLEM 6

We now leave the exciting world of education and sports and consider something else entirely.

We consider independent random variables $Y_i \sim N(\mu_i, \sigma^2)$, organized in a two-way layout with expected values as presented in the table below.

	columns	
rows	$E\left(Y_1\right) = \mu + \alpha + \beta$	$E\left(Y_2\right) = \mu - \alpha + \beta$
	$\mathrm{E}\left(Y_3\right) = \mu - \alpha - \beta$	$E\left(Y_4\right) = \mu - \alpha + \beta$

In the sequel, you may find the following expressions useful

$$\begin{bmatrix} 4 & -2 & 2 \\ -2 & 4 & 0 \\ 2 & 0 & 4 \end{bmatrix}^{-1} = \frac{1}{8}\begin{bmatrix} 4 & 2 & -2 \\ 2 & 3 & -1 \\ -2 & -1 & 3 \end{bmatrix}$$

$$\frac{1}{8}\begin{bmatrix} 4 & 2 & -2 \\ 2 & 3 & -1 \\ -2 & -1 & 3 \end{bmatrix}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & 1 \end{bmatrix} = \frac{1}{4}\begin{bmatrix} 2 & 0 & 2 & 0 \\ 2 & -1 & 0 & -1 \\ 0 & 1 & -2 & 1 \end{bmatrix}$$

Question 6.1

The ordinary least squares estimator $\hat{\beta}$ of β is of the form $aY_1 + bY_2 + cY_3 + dY_4$, where $\begin{bmatrix} a & b & c & d \end{bmatrix}$ is:

Vælg en svarmulighed

- $\frac{1}{4}\begin{bmatrix} 0 & 1 & -2 & 1 \end{bmatrix}$
- $\frac{1}{4}\begin{bmatrix} 2 & -1 & 0 & -1 \end{bmatrix}$
- $\frac{1}{4}\begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}$
- $\frac{1}{4}\begin{bmatrix} 2 & 0 & 2 & 0 \end{bmatrix}$
- Don't know
- $\frac{1}{4}\begin{bmatrix} 1 & -1 & -1 & -1 \end{bmatrix}$

Question 6.2

The variance of $\hat{\beta}$ is:

Vælg en svarmulighed

- $\frac{1}{4}\sigma^2$
- $\frac{1}{8}\sigma^2$
- Don't know
- $\frac{1}{2}\sigma^2$
- $\frac{3}{8}\sigma^2$
- $\frac{1}{16}\sigma^2$

Question 6.3

The correlation between $\hat{\mu}$ and $\hat{\beta}$ is:

Vælg en svarmulighed

- 0
- Don't know
- $-\frac{1}{4}$
- $\frac{-1}{\sqrt{3}}$
- $-\frac{\sqrt{2}}{16}$
- $\frac{-3}{\sqrt{48}}$

Question 6.4

We want to make a change to the experimental design, such that both α and β can be uniquely estimated *and* such that their estimates becomes uncorrelated. This can be achieved with the following layout

Vælg en svarmulighed

- Don't know
- columns

  rows $E(Y_1) = \mu + \alpha + \beta$ $E(Y_2) = \mu - \alpha + \beta$

  $E(Y_3) = \mu - \alpha - \beta$ $E(Y_4) = \mu + \alpha - \beta$
- columns

  rows $E(Y_1) = \mu - \alpha - \beta$ $E(Y_2) = \mu - \alpha + \beta$

  $E(Y_3) = \mu - \alpha - \beta$ $E(Y_4) = \mu - \alpha + \beta$
- columns

  rows $E(Y_1) = \mu + \alpha + \beta$ $E(Y_2) = \mu + \alpha + \beta$

  $E(Y_3) = \mu - \alpha - \beta$ $E(Y_4) = \mu - \alpha + \beta$
- columns

  rows $E(Y_1) = \mu + \alpha + \beta$ $E(Y_2) = \mu - \alpha + \beta$

  $E(Y_3) = \mu + \alpha + \beta$ $E(Y_4) = \mu - \alpha + \beta$
- columns

  rows $E(Y_1) = \mu - \alpha + \beta$ $E(Y_2) = \mu - \alpha + \beta$

  $E(Y_3) = \mu - \alpha + \beta$ $E(Y_4) = \mu - \alpha + \beta$

PROBLEM 7

We consider the random normal variable

$$Z = \begin{bmatrix} Y \\ X \end{bmatrix} = \begin{bmatrix} Y_1 \\ Y_2 \\ X_1 \\ X_2 \end{bmatrix}, \qquad E(Z) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \qquad D(Z) = \Sigma = \begin{bmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{bmatrix} = \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

Question 7.1

For $D(Z)$ to be regular, i.e. invertible, ρ must be in the range

Vælg en svarmulighed

- Don't know
- $-\frac{1}{2} < \rho < \frac{1}{2}$
- $-\frac{3}{4} < \rho < \frac{3}{4}$
- $-1 \leq \rho \leq 1$
- $-\frac{9}{10} < \rho < \frac{9}{10}$
- $-1 < \rho < 1$

Question 7.2

The first squared canonical correlation between Y and X is given by:

Vælg en svarmulighed

- $\frac{\rho^2}{\sqrt{3}}$
- 0
- Don't know
- $\sqrt{\frac{\rho^3}{3}}$
- ρ^2
- ρ

Question 7.3

The variable Y_1 explains the following fraction of the variance in Y_2

Vælg en svarmulighed

- ρ^2
- ρ
- Don't know
- 0
- $\sqrt{\frac{\rho^3}{3}}$
- $\frac{\rho^2}{\sqrt{3}}$

Question 7.4

The conditional mean $E\left(\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} | \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right)$ is:

Vælg en svarmulighed

- $\begin{bmatrix} -x_2\rho \\ -x_2\rho \end{bmatrix}$
- $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$
- $\begin{bmatrix} -x_1(\frac{\rho}{\rho^2-1} - \frac{\rho^3}{\rho^2-1}) \\ -x_1(\frac{\rho^2}{\rho^2-1} - \frac{\rho^4}{\rho^2-1}) \end{bmatrix}$
- $\begin{bmatrix} x_1\rho^2 \\ x_1\rho \end{bmatrix}$
- $\begin{bmatrix} -x_2(\frac{\rho}{\rho^2-1}) \\ -x_2(\frac{\rho^2}{\rho^2-1}) \end{bmatrix}$
- Don't know

Question 7.5

The conditional dispersion $D\left(\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \mid \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right)$ is:

Vælg en svarmulighed

- $\begin{bmatrix} 1-\rho^2 & \rho-\rho^3 \\ \rho-\rho^3 & 1-\rho \end{bmatrix}$
- $\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$
- $\begin{bmatrix} 1-\rho^4 & \frac{\rho-\rho^3}{\sqrt{\rho^2-1}} \\ \frac{\rho-\rho^3}{\sqrt{\rho^2-1}} & 1-\rho^2 \end{bmatrix}$
- $\begin{bmatrix} \frac{1-\rho^4}{\sqrt{\rho^2-1}} & \frac{\rho-\rho^3}{\sqrt{\rho^2-1}} \\ \frac{\rho-\rho^3}{\sqrt{\rho^2-1}} & 1 \end{bmatrix}$
- Don't know
- $\begin{bmatrix} 1-\rho^4 & \rho-\rho^3 \\ \rho-\rho^3 & 1-\rho^2 \end{bmatrix}$

Question 7.6

The squared partial correlation $\rho^2_{Y_1 Y_2 | X_1 X_2}$ is given by:

Vælg en svarmulighed

- $\frac{\rho^2}{\rho^2+1}$
- $\frac{\rho^2}{\sqrt{\rho^2+1}}$
- 0
- Don't know
- $\rho - \rho^3$
- $\frac{\rho-\rho^2}{(\rho^2-1)\,(\rho^4-1)}$

Question 7.7

The squared multiple correlation $\rho^2_{Y_1 | X_1 X_2}$ is

Vælg en svarmulighed

- ρ^2
- Don't know
- $1 - \rho^4$
- ρ^3
- ρ^4
- 0