Written examination, date: 8th of December 2020 *Page 1 of 27 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

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 SAS

```
data Q1_1;
```

```
cdf1 = 2*cdf('t',-3.8424,355);
run;
```

0.0001538526

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
|

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

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

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

We find in the output

Eigenvalues of the Correlation Matrix: Total = 10 Average = 1				
	Eigenvalue	Difference	Proportion	Cumulative
1	2.26647957	0.96098520	0.2266	0.2266
2	1.30549437	0.15283051	0.1305	0.3572
3	1.15266386	0.11150352	0.1153	0.4725
4	1.04116034	0.06092892	0.1041	0.5766
5	0.98023142	0.07271249	0.0980	0.6746
6	0.90751893	0.19532513	0.0908	0.7654
7	0.71219379	0.00671489	0.0712	0.8366
8	0.70547891	0.07507545	0.0705	0.9071
9	0.63040345	0.33202808	0.0630	0.9702
10	0.29837537		0.0298	1.0000

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

and find

Variance Explained by Each Factor		
Factor1	Factor2	Factor3
2.2664796	1.3054944	1.1526639

Giving 0.22664796

ANSWER 5

Question 1.6.

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

We have page 405

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

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

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

The answer is thus $0.67677^2=0.4580$

ANSWER 5

Question 1.7.

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

We find on page 404

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

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 the output

Final Communality Estimates: Total = 4.724638									
age	traveltime	studytime	famrel	freetime	goout	Dalc	Walc	health	absences
0.63684447	0.12128610	0.29212136	0.63242873	0.52865610	0.47140938	0.64611442	0.73696820	0.21659766	0.44221138

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:

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

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.

Question 2.1.

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

We start by reading the data into SAS and running the regression:

```
data examdata;
input age traveltime goout health absences G1 G3;
datalines;
18 2 4 3 6 5 6
17 1 3 3 4 5 6
15 1 2 3 10 7 10
15 1 2 5 2 15 15
16 1 2 5 4 6 10
16 1 2 5 10 15 15
16 1 4 3 0 12 11
17 2 4 1 6 6 6
15 1 2 1 0 16 19
15 1 1 5 0 14 15
15 1 3 2 0 10 9
15 3 2 4 4 10 12
15 1 3 5 2 14 14
15 2 3 3 2 10 11
15 1 2 3 0 14 16
16 1 4 2 4 14 14
16 1 3 2 6 13 14
16 3 2 4 4 8 10
17 1 5 5 16 6 5
16 1 3 5 4 8 10
run;


proc reg data=examdata plots=all;
model G3 = age traveltime goout health absences G1 / cli clb clm
influence;
run;
```

We find in the output:

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	6.90255	7.51908	0.92	0.3753	-9.34144	23.14654
age	1	0.01645	0.46579	0.04	0.9724	-0.98983	1.02272
traveltime	1	-0.10128	0.41806	-0.24	0.8123	-1.00445	0.80188
goout	1	-1.23724	0.36628	-3.38	0.0049	-2.02853	-0.44595
health	1	-0.23517	0.20714	-1.14	0.2768	-0.68267	0.21234
absences	1	0.02455	0.07920	0.31	0.7615	-0.14656	0.19565
G1	1	0.82181	0.09529	8.62	<.0001	0.61594	1.02767

Age is the first variable eliminated. This is also confirmed if we ran:

```
proc reg data=examdata;
model G3 = age traveltime goout health absences G1 / selection=backward;
run;
```

ANSWER 1

Question 2.2.

The observation with the highest leverage is:

We find in the SAS-output

Output Statistics				
Residual	RStudent	Hat Diag H	Cov Ratio	DFFITS
0.4021	0.4673	0.4440	2.7759	0.4176
-0.8709	-0.9570	0.3426	1.5918	-0.6908
0.1338	0.1895	0.6309	4.6462	0.2477
-0.7739	-0.7598	0.1980	1.5715	-0.3776
1.5568	1.8724	0.3411	0.4419	1.3471
-0.9867	-1.3248	0.5309	1.4360	-1.4094
-0.2717	-0.2794	0.2983	2.3850	-0.1822
-0.8736	-0.9442	0.3217	1.5632	-0.6502
1.5127	1.8660	0.3746	0.4706	1.4441
-1.1402	-1.2468	0.3032	1.0713	-0.8225
-2.0840	-2.7119	0.2975	0.0878	-1.7646
0.2535	0.3006	0.4716	3.1447	0.2840
0.2852	0.2912	0.2874	2.3394	0.1849
0.2033	0.2028	0.2561	2.2984	0.1190
0.6267	0.5886	0.1396	1.6677	0.2371
0.7514	0.7727	0.2679	1.7027	0.4674
0.2868	0.2797	0.2191	2.1428	0.1481
-0.1194	-0.1340	0.4138	2.9562	-0.1126
-0.0425	-0.0625	0.6592	5.1275	-0.0870
1.1505	1.1676	0.2025	1.0339	0.5883

Observation 19

ANSWER 5

Question 2.3.

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

Can easily be found by changing the alphs value in the SAS procedure

```
proc reg data=examdata alpha=0.01;
model G3 = age traveltime goout health absences G1 / cli clb clm
influence;
run;
```

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	99% CL Mean		99% CL Predict	
1	6	5.5979	0.7454	3.3526	7.8432	1.5486	9.6472
2	6	6.8709	0.6547	4.8986	8.8432	2.9664	10.7754
3	10	9.8662	0.8886	7.1896	12.5427	5.5627	14.1696

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:

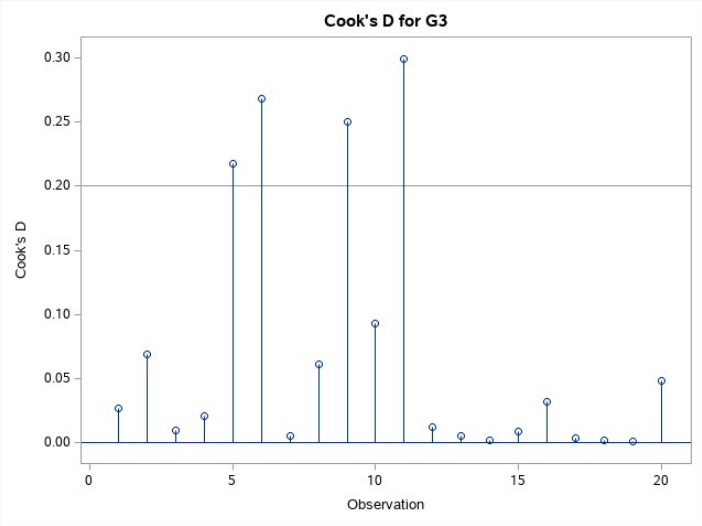
We find DFFITS in the output

Output Statistics

RStudent	Hat Diag H	Cov Ratio	DFFITS
0.4673	0.4440	2.7759	0.4176
-0.9570	0.3426	1.5918	-0.6908
0.1895	0.6309	4.6462	0.2477
-0.7598	0.1980	1.5715	-0.3776
1.8724	0.3411	0.4419	1.3471
-1.3248	0.5309	1.4360	-1.4094
-0.2794	0.2983	2.3850	-0.1822
-0.9442	0.3217	1.5632	-0.6502
1.8660	0.3746	0.4706	1.4441
-1.2468	0.3032	1.0713	-0.8225
-2.7119	0.2975	0.0878	-1.7646
0.3006	0.4716	3.1447	0.2840
0.2912	0.2874	2.3394	0.1849
0.2028	0.2561	2.2984	0.1190
0.5886	0.1396	1.6677	0.2371
0.7727	0.2679	1.7027	0.4674
0.2797	0.2191	2.1428	0.1481
-0.1340	0.4138	2.9562	-0.1126
-0.0625	0.6592	5.1275	-0.0870
1.1676	0.2025	1.0339	0.5883

Observation 11

Cooks D yields the same result



ANSWER 4

Question 2.5.

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

We run an RSQUARE selection.

```
proc reg data=examdata;
model G3 = age traveltime goout health absences G1 / selection=rsquare;
```

```
run;
```

2	0.2430	health absences
2	0.0684	traveltime health
3	0.9413	goout health G1
3	0.9363	traveltime goout G1
3	0.9362	goout absences G1
3	0.9362	age goout G1

Alternatively, it can be solved by brute force comparison of the five options.

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:

We find in the output

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	16.59606478	2.39490942	6.93	<.0001
age	-0.19583833	0.13543729	-1.45	0.1491
traveltime	-0.31984571	0.24404167	-1.31	0.1909
studytime	0.31949646	0.20742868	1.54	0.1244
famrel	0.10395988	0.19169663	0.54	0.5880
freetime	0.09191867	0.17645260	0.52	0.6028
goout	-0.39698744	0.17621951	-2.25	0.0249
Dalc	-0.02372773	0.23876157	-0.10	0.9209
Walc	-0.14039605	0.18389425	-0.76	0.4457
health	-0.19434803	0.11960391	-1.62	0.1051
absences	-0.06844792	0.02078401	-3.29	0.0011

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

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

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

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

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.

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

This gives us

Model H_i	Dim(H_i)	$SS_{res}(H_i)$	$DF_{res}(H_i)$	Variation $SS_{res}(H_i) - SS_{res}(H_{i-1})$	DF	Test statistic	p-value
H_1	10	3309.097611	346				
H_2	7	3476.833242	349	167.7356	3	$\frac{167.7356\ /3}{3309.097611\ /346}$	6.6358e-04
H_3	5	3501.835318	351	25.0021	2	$\frac{25.0021\ /2}{3476.833242\ /349}$	0.0032
H_4	1	3649.480274	355	147.6450	4	$\frac{147.6450\ /4}{3501.835318\ /351}$	0.0058
H_5	1	3709.04762	356	59.5673	1	$\frac{59.5673\ /1}{3649.480274\ /355}$	0.0166

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:

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

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

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

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

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

We need **q1** and **q3**, we find in the output

q1:

E = Error SSCP Matrix					
	ProteinB	TCAIndexM	TCAIndexB	TCAM	TCAB
ProteinB	550.75940273	1935.9518735	-243.0044024	310.13759519	392.63201427
TCAIndexM	1935.9518735	11609.48673	9408.2290888	1933.3853057	1994.4878555
TCAIndexB	-243.0044024	9408.2290888	58232.80745	1741.7388603	2690.9404775
TCAM	310.13759519	1933.3853057	1741.7388603	341.46534636	332.75476611
TCAB	392.63201427	1994.4878555	2690.9404775	332.75476611	444.32920438

q3:

H = Type III SSCP Matrix for Treatment					
	ProteinB	TCAIndexM	TCAIndexB	TCAM	TCAB
ProteinB	1.0829176098	19.325203335	106.61214681	5.6577643303	6.2339760868
TCAIndexM	19.325203335	344.86786487	1902.5467833	100.96561835	111.24840373
TCAIndexB	106.61214681	1902.5467833	10495.85836	557.00119376	613.72866021
TCAM	5.6577643303	100.96561835	557.00119376	29.559309892	32.569760819
TCAB	6.2339760868	111.24840373	613.72866021	32.569760819	35.886809391

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

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:

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

q1:

E = Error SSCP Matrix					
	ProteinB	TCAIndexM	TCAIndexB	TCAM	TCAB
ProteinB	550.75940273	1935.9518735	-243.0044024	310.13759519	392.63201427
TCAIndexM	1935.9518735	11609.48673	9408.2290888	1933.3853057	1994.4878555
TCAIndexB	-243.0044024	9408.2290888	58232.80745	1741.7388603	2690.9404775
TCAM	310.13759519	1933.3853057	1741.7388603	341.46534636	332.75476611
TCAB	392.63201427	1994.4878555	2690.9404775	332.75476611	444.32920438

q3:

H = Type III SSCP Matrix for Country					
	ProteinB	TCAIndexM	TCAIndexB	TCAM	TCAB
ProteinB	22.348871964	1.5570432036	49.763973738	8.8822477378	19.120285507
TCAIndexM	1.5570432036	0.1084790115	3.4670500246	0.6188251244	1.3321079761
TCAIndexB	49.763973738	3.4670500246	110.80886257	19.777997917	42.574917757
TCAM	8.8822477378	0.6188251244	19.777997917	3.530125592	7.5990910398
TCAB	19.120285507	1.3321079761	42.574917757	7.5990910398	16.358110533

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

We find the confusion matrix in the output:

The DISCRIM Procedure
Classification Summary for Calibration Data: HOME.HERRING
Resubstitution Summary using Linear Discriminant Function

Number of Observations and Percent Classified into Country

From Country	1	2	3	Total
1	29 44.62	16 24.62	20 30.77	65 100.00
2	8 13.79	50 86.21	0 0.00	58 100.00
3	7 7.45	5 5.32	82 87.23	94 100.00
Total	44 20.28	71 32.72	102 47.00	217 100.00
Priors	0.33333	0.33333	0.33333	

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.

We find in the output for the full model

Generalized Squared Distance to Country			
From Country	1	2	3
1	0	2.13508	2.63739
2	2.13508	0	6.96931
3	2.63739	6.96931	0

And for the reduced model

Generalized Squared Distance to Country			
From Country	1	2	3
1	0	0.70913	1.51850
2	0.70913	0	2.46381
3	1.51850	2.46381	0

We have n1=65 and n2 = 58 FROM

Class Level Information					
Country	Variable Name	Frequency	Weight	Proportion	Prior Probability
1	_1	65	65.0000	0.299539	0.333333
2	_2	58	58.0000	0.267281	0.333333
3	_3	94	94.0000	0.433180	0.333333

The full model has p=7 variables and the reduced has q=5

We insert

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

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 π_i	Classified as not π_i	Sum
Nature π_i	$tp_i = N_{ii}$ = # true positive	$fn_i = N_{i\cdot} - N_{ii}$ = # false negative	$P_i = N_{i\cdot}$ = # from π_i
Nature not π_i	$fp_i = N_{\cdot i} - N_{ii}$ = # false positive	$tn_i = N_{\cdot\cdot} - N_{i\cdot} - N_{\cdot i} + N_{ii}$ = # true negative	$NN_i = N_{\cdot\cdot} - N_{i\cdot}$ = # in all classes but π_i
Sum	$CP_i = N_{\cdot i}$ = # clas. as π_i	$CN_i = N_{\cdot\cdot} - N_{\cdot i}$ = # clas. as not π_i	$TN = N_{\cdot\cdot}$ = 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	$\frac{1}{k}\sum_{i=1}^{k}\frac{tp_i+tn_i}{TN} = \frac{2}{kN_{\cdot\cdot}}\sum_{i=1}^{k}N_{ii} + \frac{(k-2)}{k} = 1 - \frac{2}{k}\frac{1}{N_{\cdot\cdot}}\sum_{i\neq j}N_{ij}$
Average class error rate, misclassification rate	$\frac{1}{k}\sum_{i=1}^{k}\frac{fp_i+fn_i}{TN} = \frac{2}{k} - \frac{2}{k}\frac{1}{N_{\cdot\cdot}}\sum_{i=1}^{k}N_{ii} = \frac{2}{k}\frac{1}{N_{\cdot\cdot}}\sum_{i\neq j}N_{ij}$
Average class precision	$\frac{1}{k}\sum_{i=1}^{k}\frac{tp_i}{tp_i+fp_i} = \frac{1}{k}\sum_{i=1}^{k}\frac{N_{ii}}{N_{\cdot i}}$
Average class sensitivity	$\frac{1}{k}\sum_{i=1}^{k}\frac{tp_i}{tp_i+fn_i} = \frac{1}{k}\sum_{i=1}^{k}\frac{N_{ii}}{N_{i\cdot}}$
Average class specificity	$\frac{1}{k}\sum_{i=1}^{k}\frac{tn_i}{fp_i+tn_i} = \frac{1}{k}\sum_{i=1}^{k}\frac{N_{\cdot\cdot}-N_{i\cdot}-(N_{\cdot i}-N_{ii})}{N_{\cdot\cdot}-N_{i\cdot}}$

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

We find in the output

Number of Observations and Percent Classified into Country

From Country	1	2	3	Total
1	32 49.23	16 24.62	17 26.15	65 100.00
2	7 12.07	46 79.31	5 8.62	58 100.00
3	19 20.21	9 9.57	66 70.21	94 100.00
Total	58 26.73	71 32.72	88 40.55	217 100.00
Priors	0.33333	0.33333	0.33333	

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

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

Country3: sensitivity = $\frac{66}{66+19+9} = 0.7021$

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 find in the output

Generalized Squared Distance to Country			
From Country	1	2	3
1	0	0.70913	1.51850
2	0.70913	0	2.46381
3	1.51850	2.46381	0

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