Written examination, date: 10. December 2018 *Page 1 of 25 pages* *Enclosure: 10 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	2	2	2	3
Question	1.1	1.2	1.3	1.4	1.5	1.6	2.1	2.2	2.3	3.1
Answer	5	4	1	2	2	1	2	3	5	3

Problem	3	3	3	3	3	4	4	4	4	4
Question	3.2	3.3	3.4	3.5	3.6	4.1	4.2	4.3	4.4	4.5
Answer	1	4	2	1	3	5	2	5	4	1

Problem	5	5	5	5	5	5	5	6	6	6
Question	5.1	5.2	5.3	5.4	5.5	5.6	5.7	6.1	6.2	6.3
Answer	2	2	4	1	2	4	3	3	1	2

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.

Remember to write your name, signature, and study number on the front page.

Please note, that there is one and only one correct answer to each question. Furthermore, some of the possible alternative answers may not make sense. When the text refers to SAS-output, the values may be rounded to fewer decimal places than in the output itself. The enclosures do not necessarily contain all the output generated by the given SAS programs. Please check that all pages of the exam paper and the enclosures are present.

Problem 1.

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

In the table below (source: http://www.statistikbanken.dk) we present some data related to hospital treatment for the five Danish regions that are responsible for healthcare. More specifically we give corresponding values of

1. No. of ambulant (outpatient) treatments (pr. 1000 capita)
2. No. of hospital admissions (pr. 1000 capita)
3. No. of bed days (pr. 1000 capita)
4. Fraction of population aged 65 or older
5. Sex

Region	Ambulant treatments (pr. 1000 capita)	Admissions (pr. 1000 capita)	Bed days (pr. 1000 capita)	Fraction of population aged 65 or older	Sex 1=male, 0 = female
RegionH	1077	216	667	0.15080411	1
RegionS	1142	282	790	0.205731552	1
RegionSyd	1447	190	615	0.192755409	1
RegionM	1072	192	559	0.173395829	1
RegionN	1069	177	628	0.195063255	1
RegionH	1490	248	703	0.18459817	0
RegionS	1395	294	770	0.23579325	0
RegionSyd	1859	204	604	0.222110388	0
RegionM	1459	208	549	0.198799601	0
RegionN	1449	194	629	0.224970663	0

We are interested in the differences between the regions. First we consider the model

$$[Ambulant \quad Admission \quad Bed\ days] = \mu + region_i + sex_j \ \ , i = 1 \dots 5,\ \ j = 1, 2$$

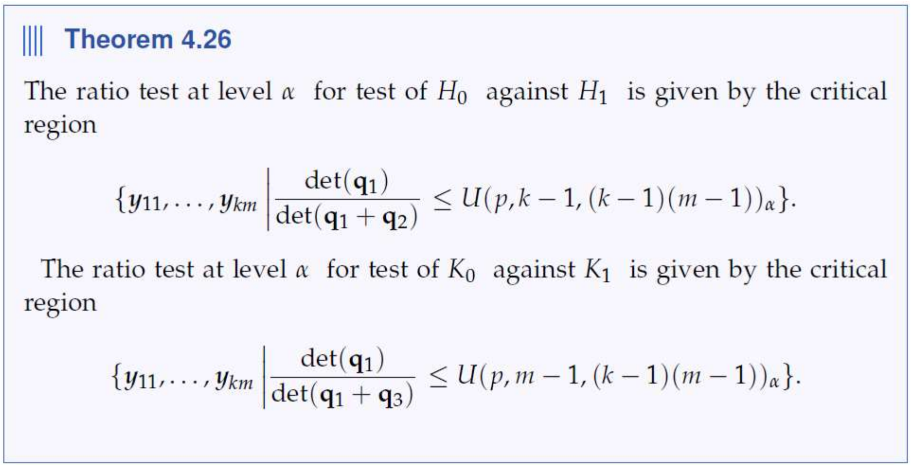
Question 1.1.

The usual test-statistic for no region effect has – under the null-hypothesis – the following distribution:

We identify the problem as a 2-side (2-way) MANOVA and use theorem 4.26.
We have p =3, k=5 and m=2

This yields:

$$U\big(p, k-1, (k-1)(m-1)\big) = U\big(3{,}5-1, (5-1)(2-1)\big) = U(3{,}4{,}4)$$



Theorem 4.26

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

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

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

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

Question 1.2.

The usual test-statistic for no sex effect is:

We use SAS and start by copying the data from the table:

```
data hospitaluse;
input region $ amb adm bed age65 sex; * 1 = men, 0 = females;
datalines;
RegionH 1077 216 667 0.15080411 1
RegionS 1142 282 790 0.205731552 1
RegionSyd 1447 190 615 0.192755409 1
RegionM 1072 192 559 0.173395829 1
RegionN 1069 177 628 0.195063255 1
RegionH 1490 248 703 0.18459817 0
RegionS 1395 294 770 0.23579325 0
RegionSyd 1859 204 604 0.222110388 0
RegionM 1459 208 549 0.198799601 0
RegionN 1449 194 629 0.224970663 0
;
```

We then run a MANOVA using PROC GLM

```
proc glm data = hospitaluse;
class region sex;
model amb adm bed = region sex;
manova h=_all_/printe printh;
run;
```

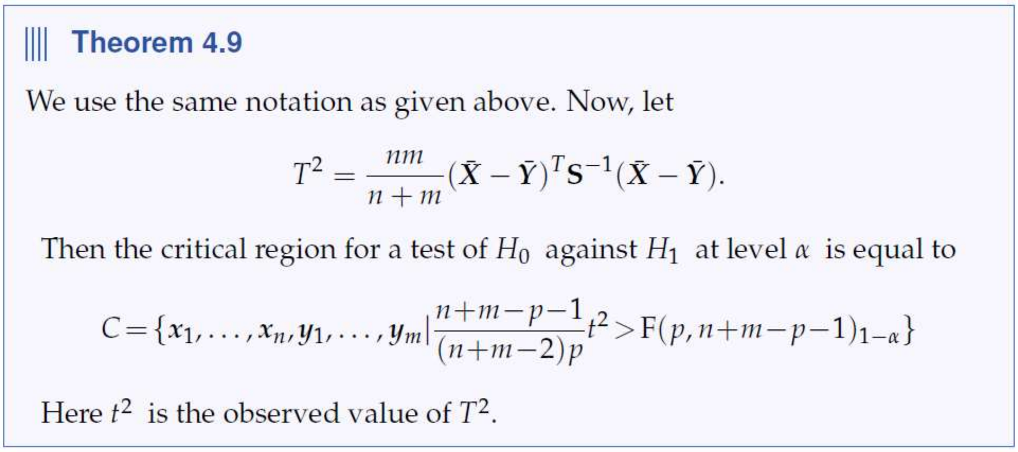
And find it in the output

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall sex Effect H = Type III SSCP Matrix for sex E = Error SSCP Matrix S=1 M=0.5 N=0					
Statistic	**Value**	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Wilks' Lambda	0.00210619	315.86	3	2	0.0032
Pillai's Trace	0.99789381	315.86	3	2	0.0032
Hotelling-Lawley Trace	473.79022218	315.86	3	2	0.0032
Roy's Greatest Root	473.79022218	315.86	3	2	0.0032

Question 1.3.

We now only consider RegionH and RegionS and the following variables $[Ambulant \quad Admission \quad Bed\ days]$. Note that Sex is now just considered as a replicate, i.e. we have two observations for each region. The usual test statistic for mean difference between RegionH and RegionS is:

We use



Theorem 4.9

We use the same notation as given above. Now, let

$$T^2 = \frac{nm}{n+m}(\bar{\mathbf{X}} - \bar{\mathbf{Y}})^T \mathbf{S}^{-1}(\bar{\mathbf{X}} - \bar{\mathbf{Y}}).$$

Then the critical region for a test of H_0 against H_1 at level α is equal to

$$C = \{x_1, \ldots, x_n, y_1, \ldots, y_m | \frac{n+m-p-1}{(n+m-2)p} t^2 > \mathrm{F}(p, n+m-p-1)_{1-\alpha}\}$$

Here t^2 is the observed value of T^2.

We already have the data. We now run the following script

```
proc discrim data=hospitaluse;
class region;
var amb adm bed;
run;
```

Generalized Squared Distance to region					
From region	**RegionH**	**RegionM**	**RegionN**	**RegionS**	**RegionSy**
RegionH	0	95.22985	221.89650	264.23327	274.47618

Generalized Squared Distance to region					
From region	RegionH	RegionM	RegionN	RegionS	RegionSy
RegionM	95.22985	0	307.67495	419.81110	338.89139
RegionN	221.89650	307.67495	0	964.28549	5.82609
RegionS	264.23327	419.81110	964.28549	0	1073
RegionSy	274.47618	338.89139	5.82609	1073	0

Inserting

$$T^2 = \frac{nm}{n+m} 264.23327 = \frac{2 \cdot 2}{2+2} 264.23327 = 264.23327$$

We now investigate how Age, ambulant treatments, and sex affect admissions and bed days with the following model

$$[Admission \quad Bed\ days] = [Age65 \quad Ambulant \quad Sex] \begin{bmatrix} \theta_{1,1} & \theta_{1,2} \\ \theta_{2,1} & \theta_{2,2} \\ \theta_{3,1} & \theta_{3,2} \end{bmatrix}$$

We test whether ambulant treatments and sex have any effect

$$\begin{bmatrix} \theta_{2,1} & \theta_{2,2} \\ \theta_{3,1} & \theta_{3,1} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

with the following model

$$\mathrm{H}_0\text{: } \boldsymbol{A} \begin{bmatrix} \theta_{1,1} & \theta_{1,2} \\ \theta_{2,1} & \theta_{2,2} \\ \theta_{3,1} & \theta_{3,2} \end{bmatrix} \boldsymbol{B}^T = \boldsymbol{C} \quad \text{vs.} \quad \mathrm{H}_1\text{: } \boldsymbol{A} \begin{bmatrix} \theta_{1,1} & \theta_{1,2} \\ \theta_{2,1} & \theta_{2,2} \\ \theta_{3,1} & \theta_{3,2} \end{bmatrix} \boldsymbol{B}^T \neq \boldsymbol{C}$$

Question 1.4.

In the above model A is equal to?

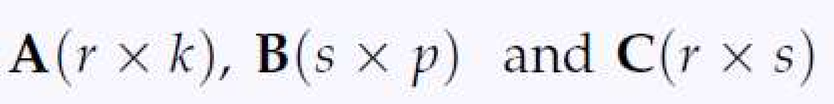
We need to select the two lower rows. Thus

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

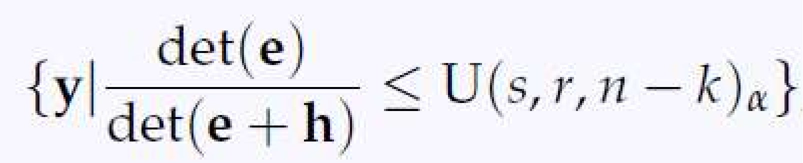
Question 1.5.

The usual test-statistic for the above model has – under the null-hypothesis – the following distribution:

We use 4.21



$$\mathbf{A}(r \times k),\ \mathbf{B}(s \times p)\ \text{ and } \mathbf{C}(r \times s)$$



$$\{\mathbf{y} | \frac{\det(\mathbf{e})}{\det(\mathbf{e}+\mathbf{h})} \leq U(s,r,n-k)_\alpha\}$$

A(r x k) = A(2 x 3)
C(r x s) = C(2 x 2)

Inserting

$$U(s,r,n-k) = U(2,2,10-3) = U(2,2,7)$$

Question 1.6.

The H matrix in the usual test statistic is? (hint: use the option '**INVERSE**' in the PROC GLM model statement to get $(X^TX)^{-1}$)

We run

```
proc glm data=hospitaluse;

model adm bed = age65 amb sex / noint solution INVERSE;
run;
```

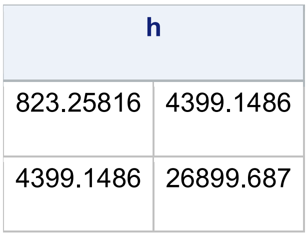
And find the inverse and parameters in the output. We then copy them into the following script

```
proc iml;
/*Reading the matrices corresponding to the estimation*/
invxx=
{149.48817279 -0.020476971 -3.656604767,
-0.020476971 2.8885874E-6 0.0004037433,
-3.656604767 0.0004037433 0.4022624992};
theta=
{1197.760990 3076.964542,
-0.018493 -0.007221,
13.028129 95.409017};

A={0 1 0,
0 0 1};
B={1 0,
0 1};
C={0 0,
0 0};
delta=A*theta*B`-C;
h=delta`*inv(A*invxx*A`)*delta;
print h;
```

```
run;
```

and get the result



h	
823.25816	4399.1486
4399.1486	26899.687

Problem 2.

We consider a three dimensional normally distributed random variable with mean

$$\mathrm{E}(\begin{bmatrix} X \\ Y \\ Z \end{bmatrix}) = \begin{bmatrix} \mu_x \\ \mu_y \\ \mu_z \end{bmatrix}$$

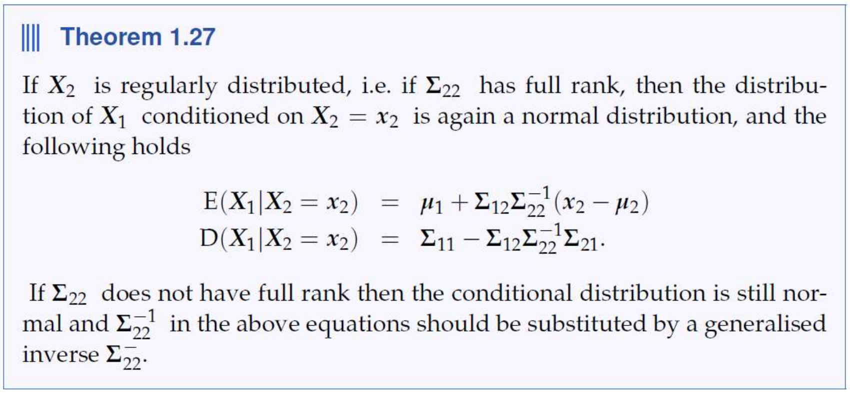
And dispersion

$$\mathrm{D}(\begin{bmatrix} X \\ Y \\ Z \end{bmatrix}) = \begin{bmatrix} 1 & \rho & \varphi \\ \rho & 1 & \rho \\ \varphi & \rho & 1 \end{bmatrix}$$

Question 2.1.

What is the expectation of Y given $\begin{bmatrix} X \\ Z \end{bmatrix} = \begin{bmatrix} x \\ z \end{bmatrix}$

We use



Theorem 1.27

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

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

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

We reorder the matrix

$$\mathrm{D}(\begin{bmatrix} Y \\ X \\ Z \end{bmatrix}) = \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \varphi \\ \rho & \varphi & 1 \end{bmatrix}$$

$$E(Y| \begin{bmatrix} X \\ Z \end{bmatrix} = \begin{bmatrix} x \\ z \end{bmatrix}) = \mu_y + [\rho \quad \rho] \begin{bmatrix} 1 & \varphi \\ \varphi & 1 \end{bmatrix}^{-1} \begin{bmatrix} x - \mu_x \\ z - \mu_z \end{bmatrix}$$

Question 2.2.

What is the dispersion of $\begin{bmatrix} X \\ Y \end{bmatrix}$ given Z=z

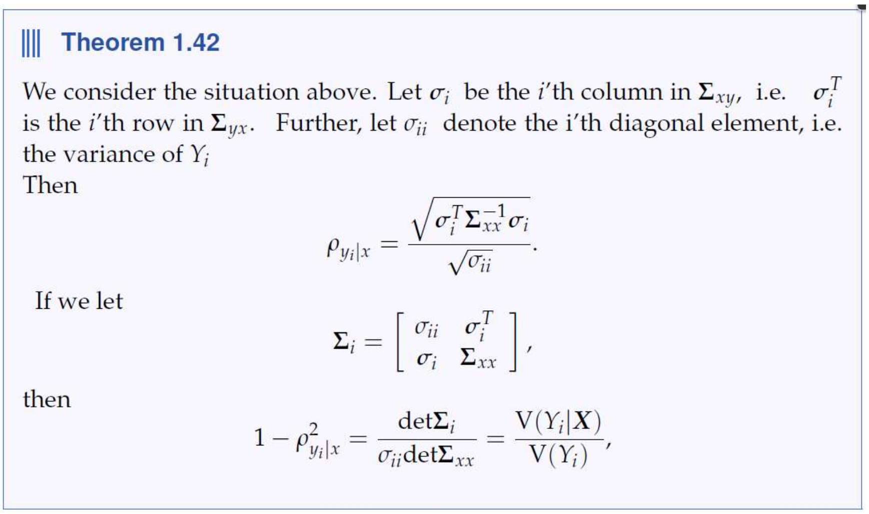
We again use Theorem 1.27

$$D\left(\begin{bmatrix} X \\ Z \end{bmatrix} \middle| Z = z\right) = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} - \begin{bmatrix} \varphi \\ \rho \end{bmatrix} 1^{-1} [\varphi \quad \rho] = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} - \begin{bmatrix} \varphi^2 & \rho\varphi \\ \rho\varphi & \rho^2 \end{bmatrix} = \begin{bmatrix} 1 - \varphi^2 & \rho - \rho\varphi \\ \rho - \rho\varphi & 1 - \rho^2 \end{bmatrix}$$

Question 2.3.

What is the squared maximum correlation of Y with a linear combination of X and Z

We use



||| **Theorem 1.42**

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

Then

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

If we let

$$\mathbf{\Sigma}_i = \begin{bmatrix} \sigma_{ii} & \sigma_i^T \\ \sigma_i & \mathbf{\Sigma}_{xx} \end{bmatrix},$$

then

$$1 - \rho_{y_i|x}^2 = \frac{\det \mathbf{\Sigma}_i}{\sigma_{ii} \det \mathbf{\Sigma}_{xx}} = \frac{\mathrm{V}(Y_i|\mathbf{X})}{\mathrm{V}(Y_i)},$$

We have $\Sigma_i = \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \varphi \\ \rho & \varphi & 1 \end{bmatrix}$ and $\Sigma_{xx} = \begin{bmatrix} 1 & \varphi \\ \varphi & 1 \end{bmatrix}$

$$\rho_{y|xz}^2 = 1 - \frac{\det \Sigma_i}{\sigma_{ii} \det \Sigma_{xx}} = 1 - \frac{2\varphi\rho^2 \; - \; 2\rho^2 \; - \; \varphi^2 \; + \; 1}{1 - \varphi^2}$$

Problem 3.

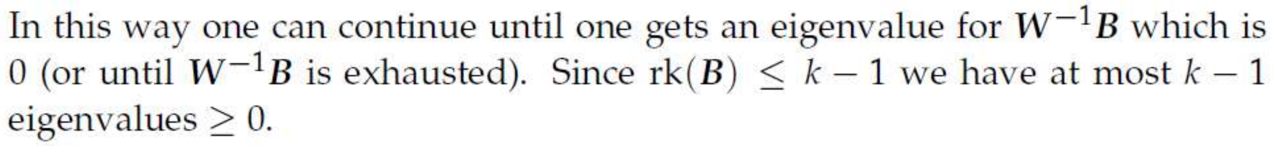
Enclosure A with SAS program and SAS output belongs to this problem. The data was compiled in order to investigate the use of different ingredients in three types of baking goods: cookies, pastries, and pizzas. For the three types, the use of 133 different ingredients was recorded. In total, 1931 recipes were investigated.

Background: The difference between cookies, pastries and pizza can lead to heated debates. Reddit user *u/everest4ever* compiled the following dataset by scraping http://www.foodnetwork.com/ to win an argument relating to a cookie competition in his office, where his cookies were beaten by the egg tarts of a colleague. His analysis showed that based on these data, egg tarts cannot be classified as cookies, and that the colleague should thus be disqualified. The reddit post can be found here:
https://www.reddit.com/r/dataisbeautiful/comments/7ke5a6/the_christmas_cookie_competition_at_my_office/

However, the background information is not relevant for the problem at hand.

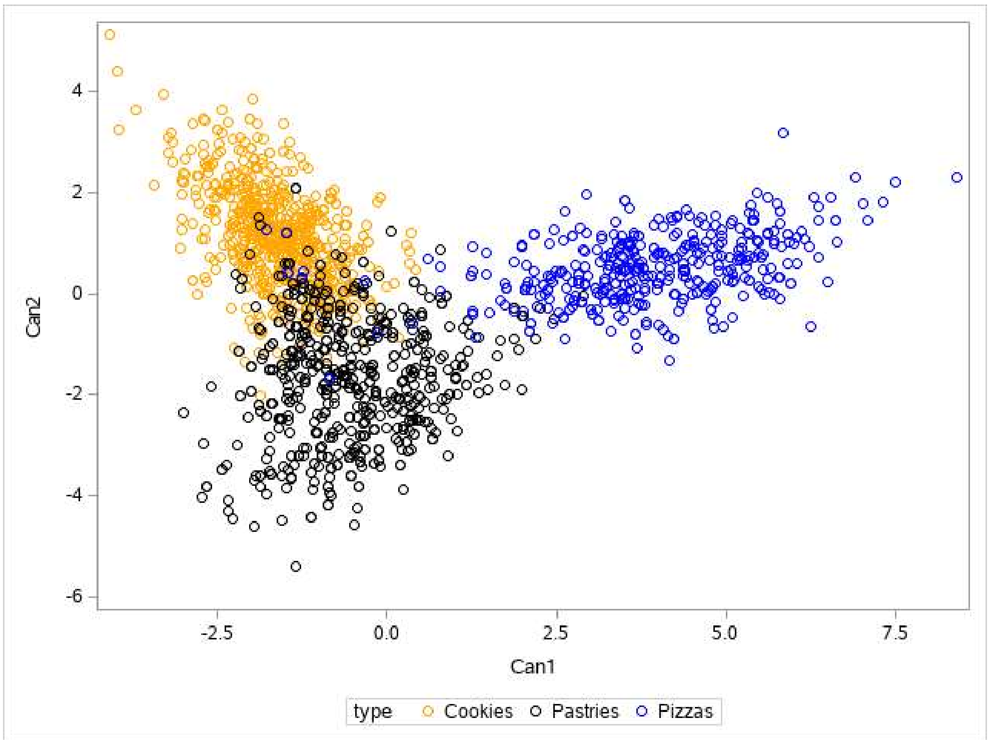
Question 3.1.

As we can only generate k-1 Canonical Discriminant Functions (CDF) the two plotted are all there is. See page 360.



In this way one can continue until one gets an eigenvalue for $\boldsymbol{W}^{-1}\boldsymbol{B}$ which is 0 (or until $\boldsymbol{W}^{-1}\boldsymbol{B}$ is exhausted). Since $\text{rk}(\boldsymbol{B}) \leq k-1$ we have at most $k-1$ eigenvalues ≥ 0.

Further, using all CDF is equivalent to LDA. I.e. if the groups are separated based on CDF, they will also be by LDA as they are equivalent in this case. Based on the plot of the first 2 canonical discriminant functions we can conclude



The data has discriminative value, but a linear method will not give a perfect separation

Question 3.2.

The first canonical function clearly has the most information, while the second has very little discriminative power with regards to cookies and pizza. We find the 4 numerically largest variables from CAN1

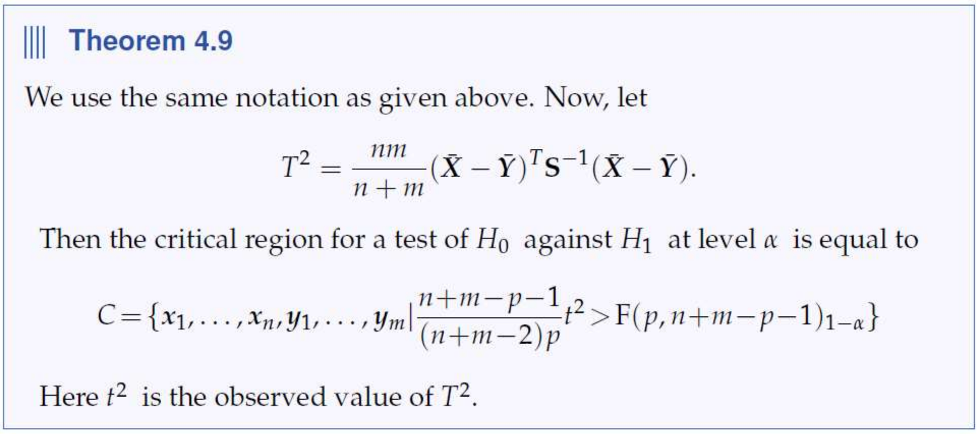
Variable	Can1
sugar	-0.39475
leaves	-0.2709
cookies	-0.23706
spinach	-0.22616
eggs	-0.21292
butter	-0.21007
ketchup	-0.17146
cookiedough	-0.15533
ginger	-0.14748
coconut	-0.14254
....	....
sausage	0.126159
mayonnaise	0.131727
tomatoes	0.131727
salt	0.132274
basil	0.172087
garlic	0.18226
oil	0.224803
cheese	0.284828
dough	0.356623
yeast	0.486825

I.e. Yeast, sugar, dough, and cheese

Question 3.3.

First we test whether there is a difference in mean value given these three variables. The usual test-statistic for this has – under the null-hypothesis – the following distribution

We use theorem



Theorem 4.9

We use the same notation as given above. Now, let

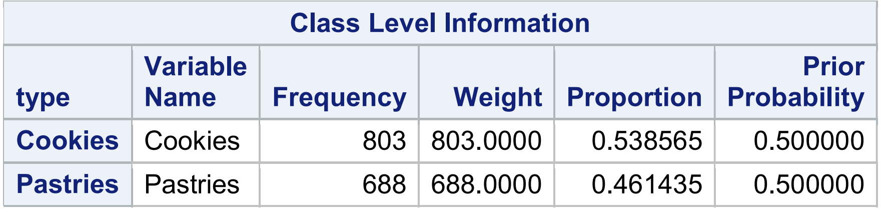
$$T^2 = \frac{nm}{n+m}(\bar{\mathbf{X}} - \bar{\mathbf{Y}})^T \mathbf{S}^{-1}(\bar{\mathbf{X}} - \bar{\mathbf{Y}}).$$

Then the critical region for a test of H_0 against H_1 at level α is equal to

$$C = \{x_1, \ldots, x_n, y_1, \ldots, y_m | \frac{n+m-p-1}{(n+m-2)p} t^2 > \mathrm{F}(p, n+m-p-1)_{1-\alpha}\}$$

Here t^2 is the observed value of T^2.

We find in the output



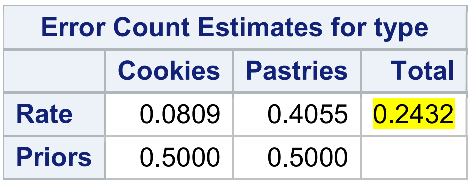
Class Level Information					
type	Variable Name	Frequency	Weight	Proportion	Prior Probability
Cookies	Cookies	803	803.0000	0.538565	0.500000
Pastries	Pastries	688	688.0000	0.461435	0.500000

Thus p=3, n=803, m=688. We insert

$$F(p, n+m-p-1) = F(3,803+688-3-1) = F(3,1487)$$

Question 3.4.

What is the misclassification rate.

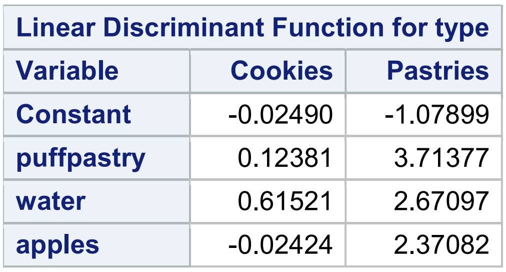


Error Count Estimates for type			
	Cookies	Pastries	Total
Rate	0.0809	0.4055	0.2432
Priors	0.5000	0.5000	

Question 3.5.

A new recipe needs to be classified. In the ingredient list we find $[puffpastry\ water\ apples] = [1\ 0\ 0]$, i.e. puffpastry but not apples or water.

We find in the output



Linear Discriminant Function for type		
Variable	Cookies	Pastries
Constant	-0.02490	-1.07899
puffpastry	0.12381	3.71377
water	0.61521	2.67097
apples	-0.02424	2.37082

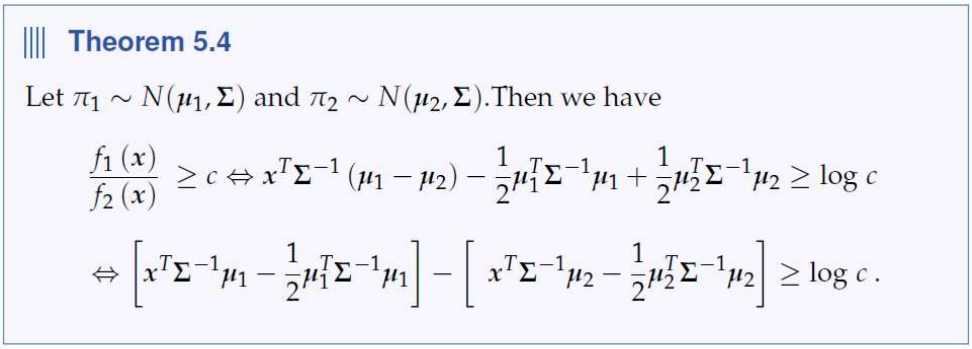
For cookies we have $S_{cookies}: -0.0249 + 0.12381 = 0.0989$

For pastries we have $S_{pastries}: -1.07899 + 3.71377 = 2.6348$

Question 3.6.

We consider the Linear Discriminant Function for classifying between cookies and pastries using a subset of the variables: puffpastry water apples. We classify as cookies if the function is positive. Using equal priors but a loss of ten for classifying pastry wrong we get:

We use

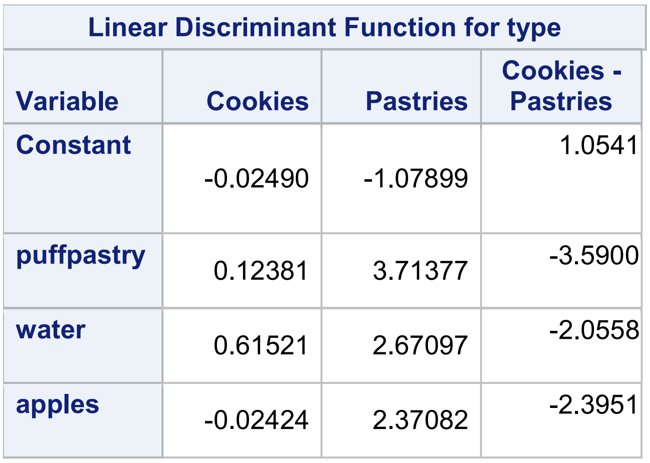


Theorem 5.4

Let $\pi_1 \sim N(\mu_1, \Sigma)$ and $\pi_2 \sim N(\mu_2, \Sigma)$.Then we have

$$\frac{f_1(x)}{f_2(x)} \geq c \Leftrightarrow x^T\Sigma^{-1}(\mu_1-\mu_2) - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 \geq \log c$$

$$\Leftrightarrow \left[x^T\Sigma^{-1}\mu_1 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1\right] - \left[x^T\Sigma^{-1}\mu_2 - \frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2\right] \geq \log c\,.$$

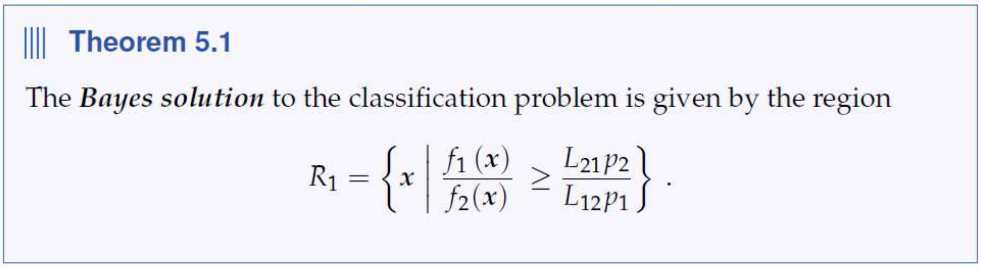


Linear Discriminant Function for type

Variable	Cookies	Pastries	Cookies - Pastries
Constant	-0.02490	-1.07899	1.0541
puffpastry	0.12381	3.71377	-3.5900
water	0.61521	2.67097	-2.0558
apples	-0.02424	2.37082	-2.3951

$$[puffpastry\ water\ apples]\begin{bmatrix}-3.5900\\-2.0558\\-2.3951\end{bmatrix} + 1.0541 > \log c$$

We then need to adjust the right hand side



Theorem 5.1

The *Bayes solution* to the classification problem is given by the region

$$R_1 = \left\{x \,\middle|\, \frac{f_1(x)}{f_2(x)} \geq \frac{L_{21}p_2}{L_{12}p_1}\right\}\,.$$

$$[puffpastry\ water\ apples]\begin{bmatrix}-3.5900\\-2.0558\\-2.3951\end{bmatrix} + 1.0541 > \log\frac{10}{1}$$

$$[puffpastry\ water\ apples]\begin{bmatrix}-3.5900\\-2.0558\\-2.3951\end{bmatrix}+1.0541>2.3026$$

$$[puffpastry\ water\ apples]\begin{bmatrix}-3.5900\\-2.0558\\-2.3951\end{bmatrix}-1.2485>0$$

Problem 4.

Enclosure B with SAS program and SAS output belongs to this problem. We consider data giving the rates (pr. 1000 capita) of different types of crimes and the prevalence of different types of unemployment benefits and social security (pr. 1000 capita) for the 98 municipalities (kommuner) in Denmark (Source http://www.statistikbanken.dk)

We consider the following variables for crime

SAS-name	Meaning
C2	Sexual crimes
C19	Murder
C21	Simple violence
C22	Serious violence
C23	Especially serious violence
C30	Threats
C53	Robbery
C55	Vandalism
C64	Sale of narcotics
C74	Weapon possession

And for social security and benefits

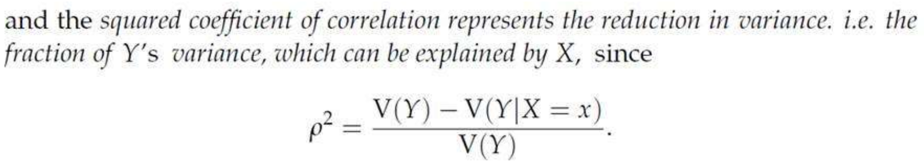
SAS-name	Meaning
S1	Educational benefits (SU)
S3	Unemployment benefit
S4	Social security
S19	Flexjob (state supported jobs)
S33	Integration benefits
S36	Sickness benefit

We shall now investigate the relations between the crime rates and the social benefits by means of a Canonical Correlation Analysis.

Question 4.1.

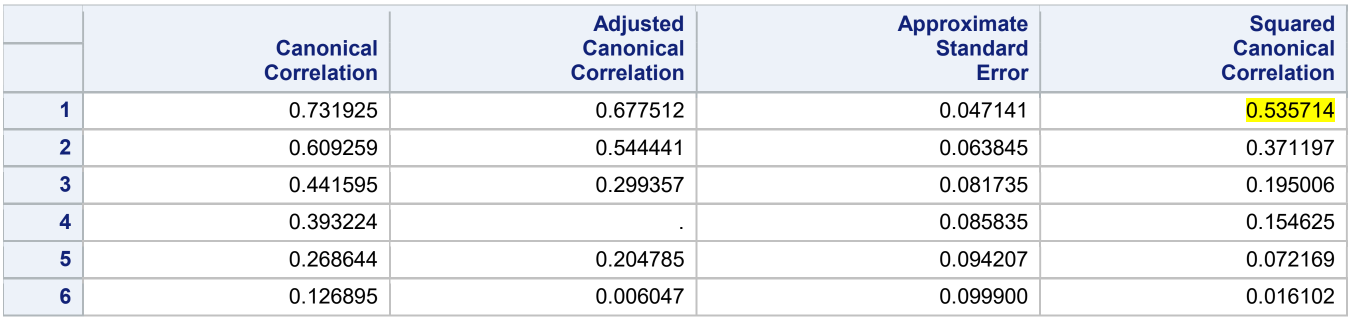
The first canonical correlation describes what fraction of the variation between V1 and W1

The squared correlation is the degree of variance explained, see 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

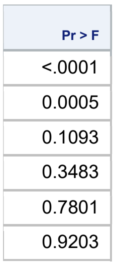
$$\rho^2 = \frac{\mathrm{V}(Y) - \mathrm{V}(Y|\mathrm{X} = x)}{\mathrm{V}(Y)}.$$



	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation
1	0.731925	0.677512	0.047141	0.535714
2	0.609259	0.544441	0.063845	0.371197
3	0.441595	0.299357	0.081735	0.195006
4	0.393224	.	0.085835	0.154625
5	0.268644	0.204785	0.094207	0.072169
6	0.126895	0.006047	0.099900	0.016102

Question 4.2.

How many canonical correlations are significant at the 5% level?



Pr > F
<.0001
0.0005
0.1093
0.3483
0.7801
0.9203

2 canonical correlations

Question 4.3.

The third canonical variate V3 can be interpreted as

A contrast between Threats and Weapon Possesion against simple violence. Can be seen from the standardized coefficients, as well as the correlations

Question 4.4.

From the relation between V1 and W1, we see a clear link between robberies and the number of people on educational benefits. One may speculate on whether students that have run out of money, may be tempted to commit a robbery, or whether an underlying socioeconomic factor is the reason for this. We investigate this further. What is the correlation between C53 (robberies) and S1 (educational benefits) when we condition on S3 (Unemployment benefit)?

We use the formula from page 34

$$\rho_{y_1y_2|x} = \frac{\rho_{y_1y_2} - \rho_{y_1x}\rho_{y_2x}}{\sqrt{(1-\rho_{y_1x}^2)(1-\rho_{y_2x}^2)}}.$$

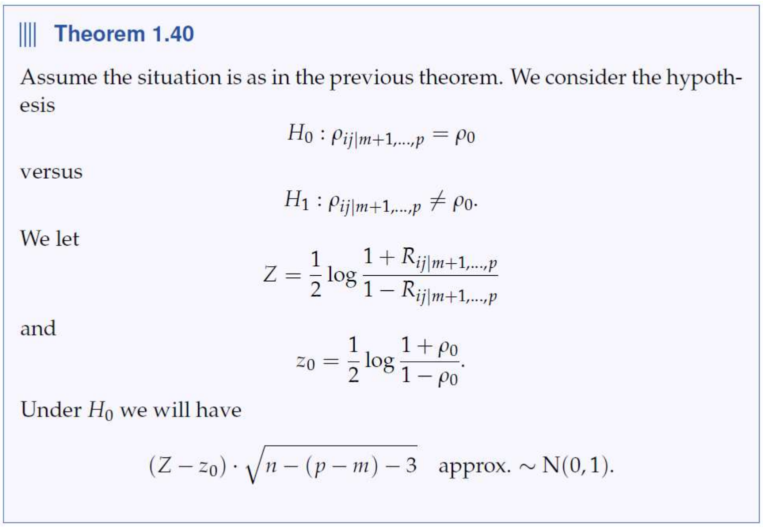
We insert from the correlation matrix

$$\rho_{C53,S1|S3} = \frac{\rho_{C53,S1} - \rho_{C53,S3}\rho_{S1,S3}}{\sqrt{(1-\rho_{C53,S3}^2)(1-\rho_{S1,S3}^2)}} = \frac{0.60196 - 0.52351 \cdot 0.59661}{\sqrt{(1-0.52351^2)(1-0.59661^2)}} = 0.4236$$

Question 4.5.

The 95% confidence interval for the correlation between C53 (robberies) and S1 (educational benefits) 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

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

Also shown in example 1.41.
We have n=98, we insert

$$P\left(-1.96 < (Z-z)\sqrt{98-0-3} < 1.96\right) \approx 95\%$$
$$P(-1.96 - 9.7468Z < -9.7468z < 1.96 - 9.7468Z) \approx 95\%$$
$$P(\text{Z} - 0.2011 < z < Z + 0.2011) \approx 95\%$$

We find

$$Z = \frac{1}{2}\log\frac{1+0.60196}{1-0.60196} = 0.6962$$

And we get the z-limits

$$[0.4951, 0.8973\ ]$$

We then need to transform it

$$\left[\frac{e^{2\cdot0.4951}-1}{e^{2\cdot0.4951}+1}, \frac{e^{2\cdot0.8973}-1}{e^{2\cdot0.8973}+1}\right] = [0.4583, 0.7150]$$

Problem 5.

Enclosure C with SAS program and SAS output belongs to this problem. We consider the relation between overall satisfaction with life, and the satisfaction with personal economy, family life, social relations and work, in 7 different income groups. The data is based on a questionnaire, where 0 means ‘not satisfied at all’, and 10 means ‘perfectly satisfied’. Data is from http://www.statistikbanken.dk

The variables are

SAS-name	Meaning
life	Overall satisfaction with life
econ	Satisfaction with economical situation
familiy	Satisfaction with family life
social	Satisfaction with social relations
work	Satisfaction with work

The observation are from the following income groups

Observation	Income group (DKK)
1	0-99.999
2	100.000-199.999
3	200.000-299.999
4	300.000-399.999
5	400.000-499.999
6	500.000-599.999
7	600.000 +

We consider two models: M1 and M2.

M1 with all variables

$$life = \mu + \beta_1 \cdot econ + \beta_2 \cdot family + \beta_3 \cdot social + \beta_4 \cdot work + \epsilon$$

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

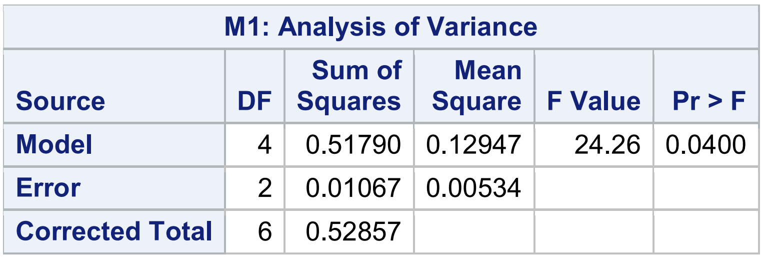
M2 is a model where we have performed stepwise model selection.

Question 5.1.

What is the usual test statistic for M1 vs M2

We find the ANOVA tables in the output

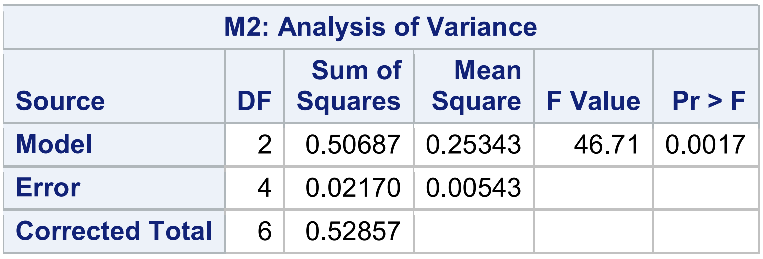
M1:



M1: Analysis of Variance

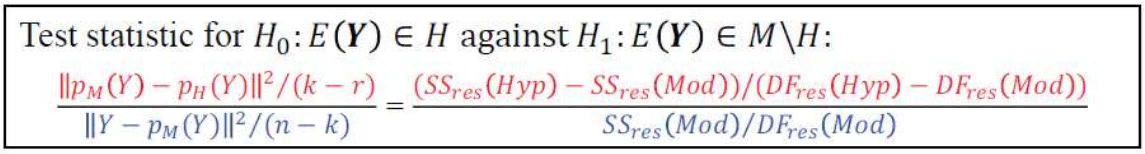
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	0.51790	0.12947	24.26	0.0400
Error	2	0.01067	0.00534		
Corrected Total	6	0.52857			

M2:



M2: Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	0.50687	0.25343	46.71	0.0017
Error	4	0.02170	0.00543		
Corrected Total	6	0.52857			

We have



Test statistic for $H_0: E(\boldsymbol{Y}) \in H$ against $H_1: E(\boldsymbol{Y}) \in M \backslash H$:

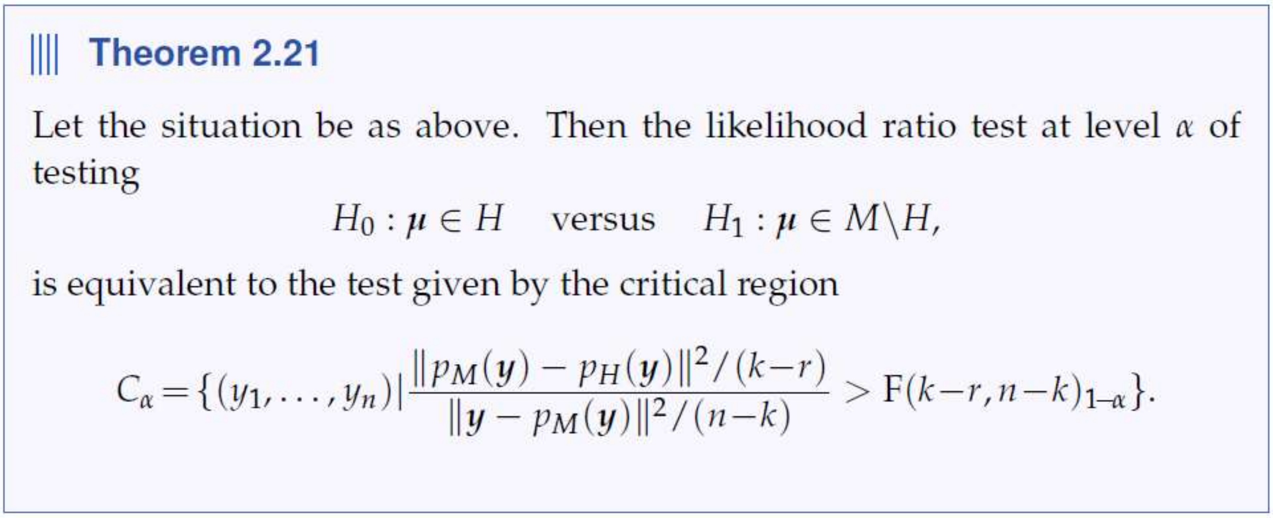
$$\frac{\|p_M(Y) - p_H(Y)\|^2/(k-r)}{\|Y - p_M(Y)\|^2/(n-k)} = \frac{(SS_{res}(Hyp) - SS_{res}(Mod))/(DF_{res}(Hyp) - DF_{res}(Mod))}{SS_{res}(Mod)/DF_{res}(Mod)}$$

$$F = \frac{(0.02170 - 0.01067)/(4-2)}{0.01067/2} = 1.0337$$

Question 5.2.

The usual test-statistic for M1 vs M2 has – under the null-hypothesis – the following distribution

We use



Theorem 2.21

Let the situation be as above. Then the likelihood ratio test at level α of testing

$$H_0 : \boldsymbol{\mu} \in H \quad \text{versus} \quad H_1 : \boldsymbol{\mu} \in M \backslash H,$$

is equivalent to the test given by the critical region

$$C_\alpha = \{(y_1, \ldots, y_n) | \frac{\|p_M(\boldsymbol{y}) - p_H(\boldsymbol{y})\|^2/(k-r)}{\|\boldsymbol{y} - p_M(\boldsymbol{y})\|^2/(n-k)} > \mathrm{F}(k-r, n-k)_{1-\alpha}\}.$$

And readily gets the answer F(2,2)

Question 5.3.

What is the reduction in the fraction of variance described when moving from model M1 to M2

We find R^2 for the two models M1: 0.9798, M2: 0.9589

Answer = 0.9798 – 0.9589 = 0.0209

We now only consider model M2

Question 5.4.

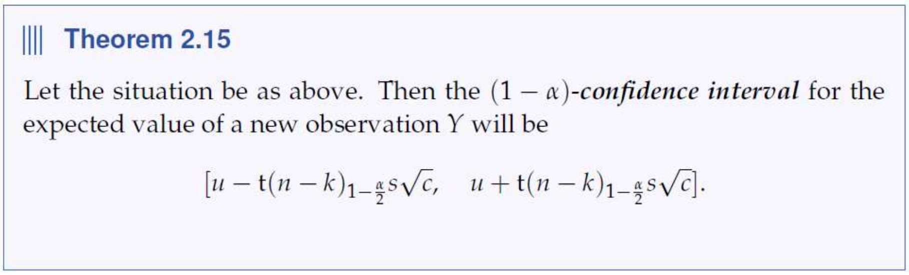
What is the leverage of observation 1

Found in output 0.5288

Question 5.5.

What is the 95% confidence interval for observation 7

We 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

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

We insert

$$\left[8.1408 - t(7-3)_{0.975}\sqrt{0.0054 \cdot 0.8248}\,, \qquad 8.1408 + t(7-3)_{0.975}\sqrt{0.0054 \cdot 0.8248}\right]$$

$$\left[8.1408 - 2.776\sqrt{0.0054 \cdot 0.8248}\,, \qquad 8.1408 + 2.776\sqrt{0.0054 \cdot 0.8248}\right]$$

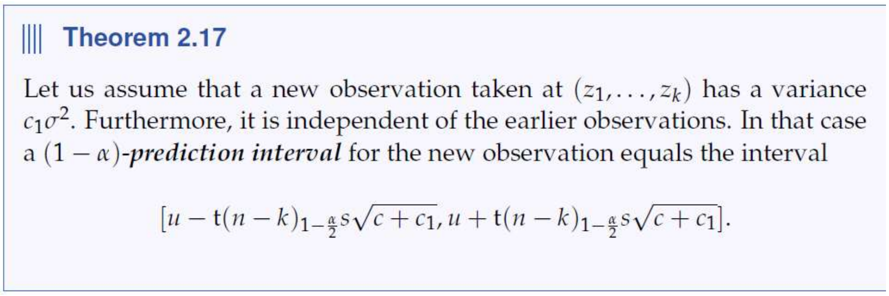
$$[\,7.9555, \qquad 8.3261]$$

$$[\,7.96, \qquad 8.33]$$

Question 5.6.

What is the 95% prediction interval for observation 7

We use

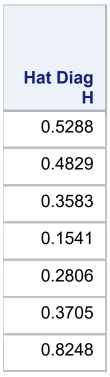


Theorem 2.17

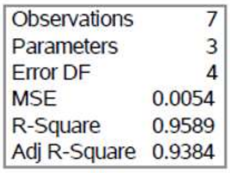
Let us assume that a new observation taken at $(z_1, \ldots, z_k)$ has a variance $c_1\sigma^2$. Furthermore, it is independent of the earlier observations. In that case a $(1-\alpha)$-***prediction interval*** for the new observation equals the interval

$$[u - \mathrm{t}(n-k)_{1-\frac{\alpha}{2}} s\sqrt{c+c_1}, u + \mathrm{t}(n-k)_{1-\frac{\alpha}{2}} s\sqrt{c+c_1}].$$

We have from the output



Hat Diag H
0.5288
0.4829
0.3583
0.1541
0.2806
0.3705
0.8248



Observations	7
Parameters	3
Error DF	4
MSE	0.0054
R-Square	0.9589
Adj R-Square	0.9384

$$\left[8.1408 - t(7-3)_{0.975}\sqrt{0.0054 \cdot 0.8248 + 0.0054 \cdot 1}\,, \qquad 8.1408 + t(7-3)_{0.975}\sqrt{0.0054 \cdot 0.8248 + 0.0054 \cdot 1}\right]$$

$$\left[8.1408 - 2.776\sqrt{0.0054 \cdot 0.8248 + 0.0054 \cdot 1}\,, \qquad 8.1408 + 2.776\sqrt{0.0054 \cdot 0.8248 + 0.0054 \cdot 1}\right]$$

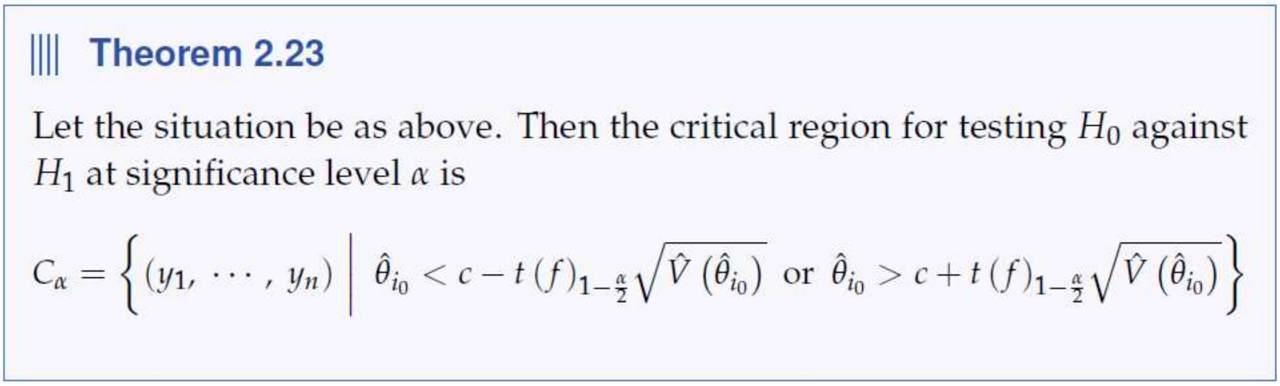
$$[\,7.87, \qquad 8.42]$$

Question 5.7.

What is the 95% confidence interval for the econ parameter

We use from page 108

$$\hat{V}\left(\hat{\theta}_{i_0}\right) = \hat{\sigma}^2 \left\{ \left(x^T \Sigma^{-1} x \right)^{-1} \right\}_{ii}$$



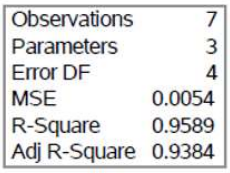
Theorem 2.23

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

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

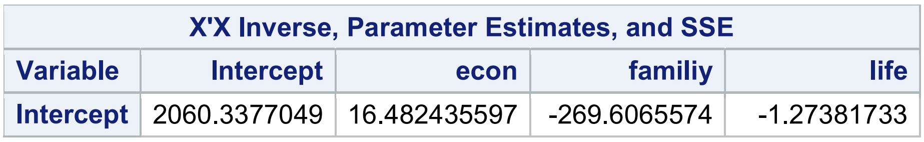
Where $f = n - \operatorname{rk}(x)$

We have from the output



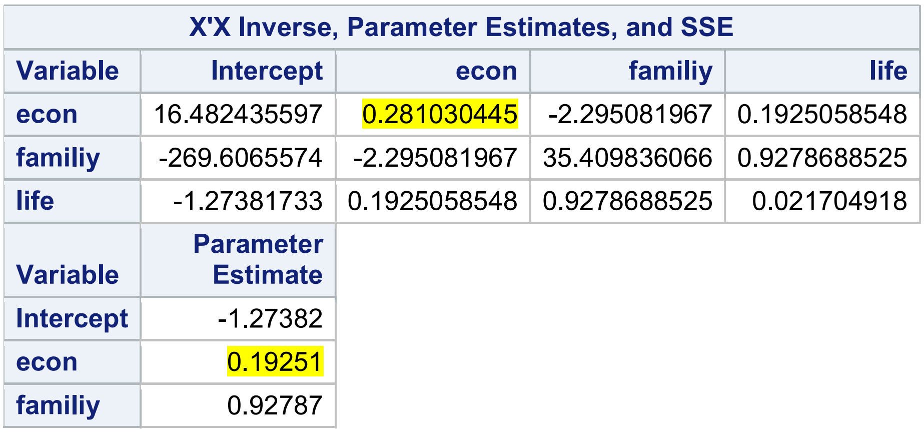
Observations	7
Parameters	3
Error DF	4
MSE	0.0054
R-Square	0.9589
Adj R-Square	0.9384

And



X'X Inverse, Parameter Estimates, and SSE

Variable	Intercept	econ	familiy	life
Intercept	2060.3377049	16.482435597	-269.6065574	-1.27381733



X'X Inverse, Parameter Estimates, and SSE

Variable	Intercept	econ	familiy	life
econ	16.482435597	0.281030445	-2.295081967	0.1925058548
familiy	-269.6065574	-2.295081967	35.409836066	0.9278688525
life	-1.27381733	0.1925058548	0.9278688525	0.021704918

Variable	Parameter Estimate
Intercept	-1.27382
econ	0.19251
familiy	0.92787

We start by finding

$$V(\theta_i) = 0.0054 \cdot 0.281030445 = 0.0015$$

We then have

$$\left[0.19251 - t(7-3)_{0.975}\sqrt{0.0015}\,, \qquad 0.19251 + t(7-3)_{0.975}\sqrt{0.0015}\right]$$

$$\left[0.19251 - 2.776\sqrt{0.0015}\,, \qquad 0.19251 + 2.776\sqrt{0.0015}\right]$$

$$[0.0844, \qquad 0.3007]$$

$$[0.084, \qquad 0.301]$$

Problem 6.

Enclosure D with SAS program and SAS output belongs to this problem. We again consider the data from problem 4, but now only the crime variables.

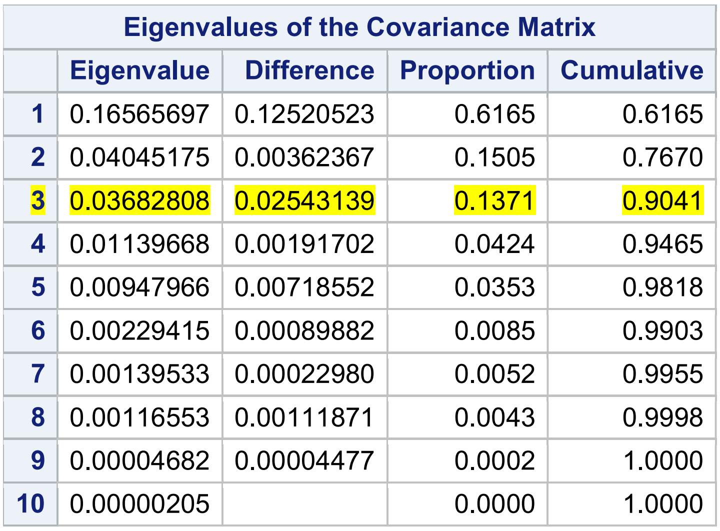
We consider the following variables for crime

SAS-name	Meaning
C2	Sexual crimes
C19	Murder
C21	Simple violence
C22	Serious violence
C23	Especially serious violence
C30	Threats
C53	Robbery
C55	Vandalism
C64	Sale of narcotics
C74	Weapon possession

We seek to investigate the underlying patterns in crime by running a principal component analysis on all crime variables.

Question 6.1.

How many factors do we need to account for 90 % of the variance
We find in the output



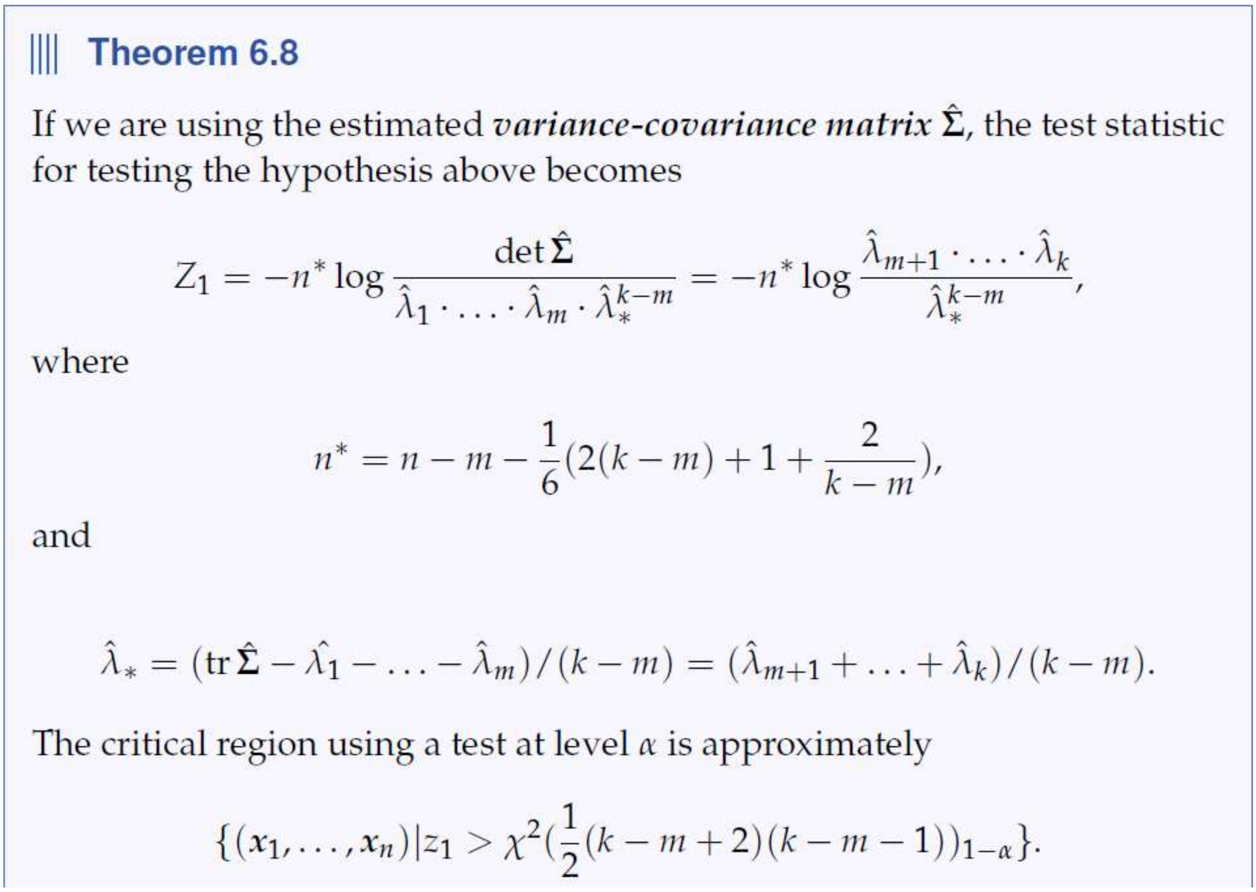
Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	0.16565697	0.12520523	0.6165	0.6165
2	0.04045175	0.00362367	0.1505	0.7670
3	0.03682808	0.02543139	0.1371	0.9041
4	0.01139668	0.00191702	0.0424	0.9465
5	0.00947966	0.00718552	0.0353	0.9818
6	0.00229415	0.00089882	0.0085	0.9903
7	0.00139533	0.00022980	0.0052	0.9955
8	0.00116553	0.00111871	0.0043	0.9998
9	0.00004682	0.00004477	0.0002	1.0000
10	0.00000205		0.0000	1.0000

Question 6.2.

The usual test statistic for the last 4 eigenvalues being equal is

We use



Theorem 6.8

If we are using the estimated ***variance-covariance matrix*** $\hat{\boldsymbol{\Sigma}}$, the test statistic for testing the hypothesis above becomes

$$Z_1 = -n^* \log \frac{\det \hat{\boldsymbol{\Sigma}}}{\hat{\lambda}_1 \cdot \ldots \cdot \hat{\lambda}_m \cdot \hat{\lambda}_*^{k-m}} = -n^* \log \frac{\hat{\lambda}_{m+1} \cdot \ldots \cdot \hat{\lambda}_k}{\hat{\lambda}_*^{k-m}},$$

where

$$n^* = n - m - \frac{1}{6}(2(k-m) + 1 + \frac{2}{k-m}),$$

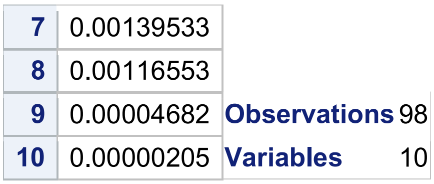
and

$$\hat{\lambda}_* = (\operatorname{tr} \hat{\boldsymbol{\Sigma}} - \hat{\lambda}_1 - \ldots - \hat{\lambda}_m)/(k-m) = (\hat{\lambda}_{m+1} + \ldots + \hat{\lambda}_k)/(k-m).$$

The critical region using a test at level α is approximately

$$\{(x_1, \ldots, x_n) | z_1 > \chi^2(\frac{1}{2}(k-m+2)(k-m-1))_{1-\alpha}\}.$$

We find the eigenvalues and number of observations



7	0.00139533		
8	0.00116553		
9	0.00004682	**Observations**	98
10	0.00000205	**Variables**	10

$$n^* = 98 - 6 - \frac{1}{6}\left(2(10-6) + 1 + \frac{2}{10-6}\right) = 90.4167$$

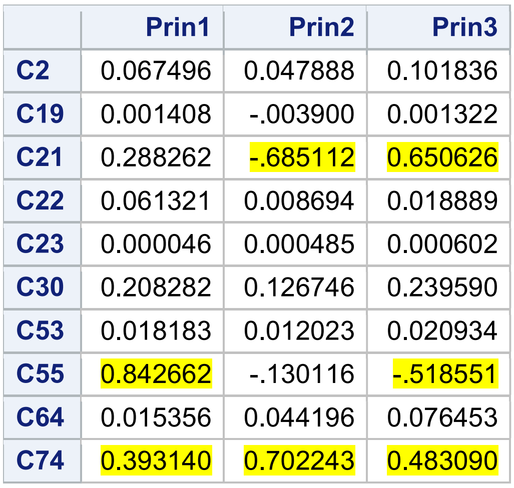
$$Z_2 = -90.4167 \log \frac{0.00139533 \cdot 0.00116553 \cdot 0.00004682 \cdot 0.00000205}{\left[\frac{(0.00139533 + 0.00116553 + 0.00004682 + 0.00000205)}{4}\right]^4}$$

$$= -90.4167 \log \frac{1.5609 \cdot 10^{-16}}{(6.5243 \cdot 10^{-4})^4} = -90.4167 \log(8.6148 \cdot 10^{-4}) = 638.0580$$

Question 6.3.

From the score plots we see at least four clear outliers: observation 36 (Guldborgsund), 39 (Lolland), 59 (Fanø), and 75 (Samsø). Out of these, find the two where problems with vandalism (C55) are *lowest*.

We look at the principal components

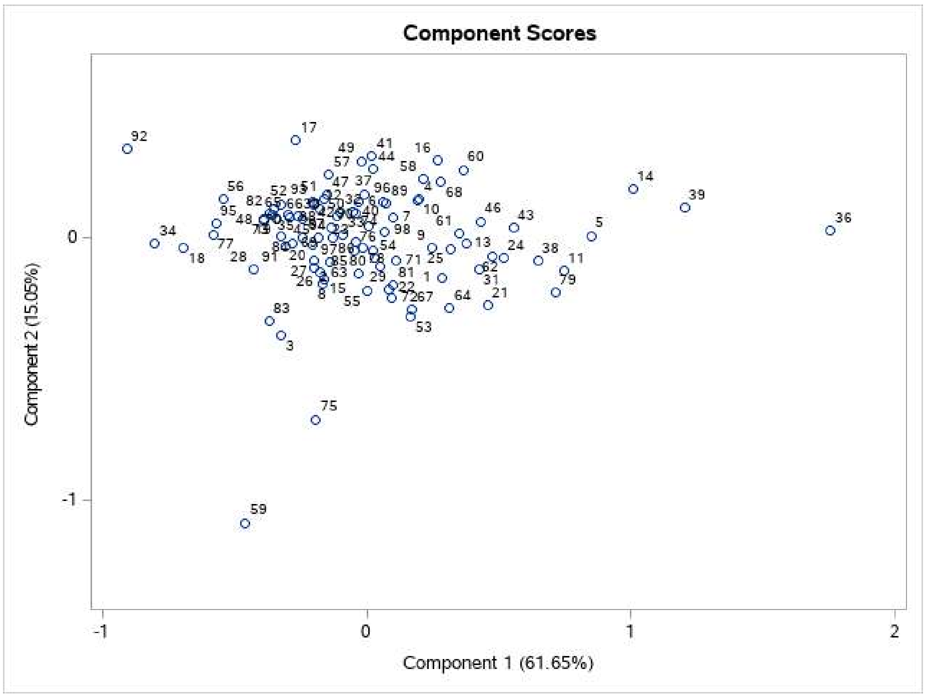


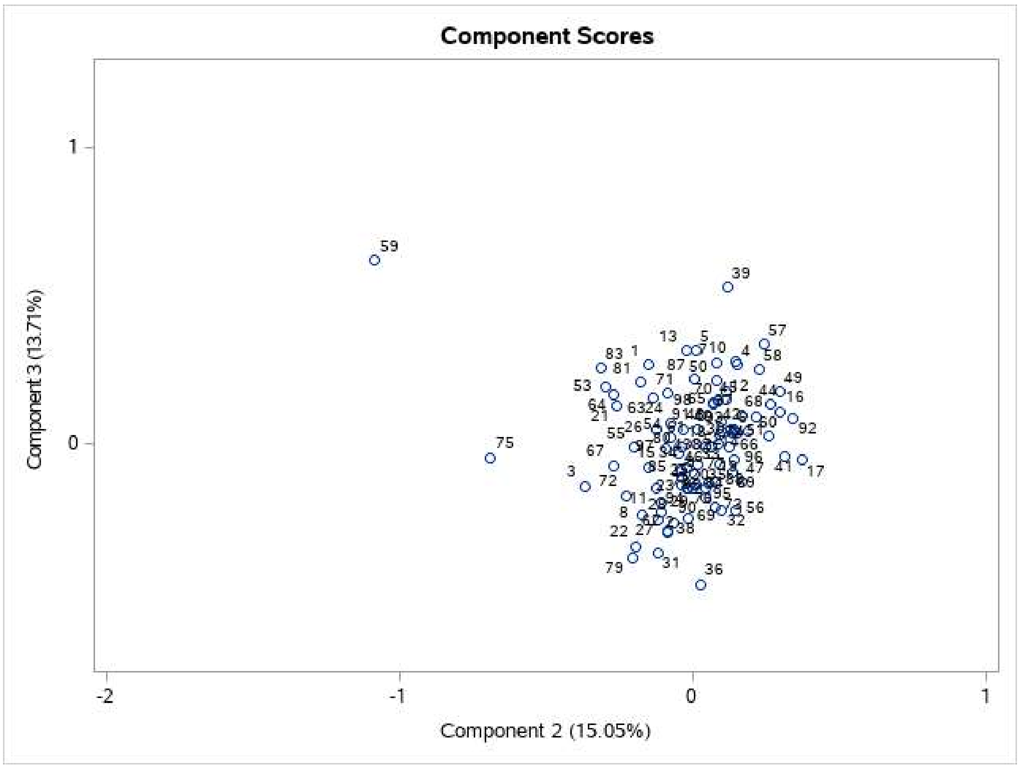
	Prin1	Prin2	Prin3
C2	0.067496	0.047888	0.101836
C19	0.001408	-.003900	0.001322
C21	0.288262	-.685112	0.650626
C22	0.061321	0.008694	0.018889
C23	0.000046	0.000485	0.000602
C30	0.208282	0.126746	0.239590
C53	0.018183	0.012023	0.020934
C55	0.842662	-.130116	-.518551
C64	0.015356	0.044196	0.076453
C74	0.393140	0.702243	0.483090

So for the problem to be *lowest*, we need a low score on component 1.
We read from the score plots

Observation	Component 1	Component 2	Component 3
Guldborgsund – 36	1.8	0	-0.5
Lolland – 39	1.2	0.1	0.5
Fanø – 59	-0.5	-1.1	0.6
Samsø - 75	-0.1	-0.6	0

This seems to indicate that Fanø and Samsø has the lowest problem with vandalism





LAST PAGE:
END OF THE EXAM SET