Written examination, date: 8th of December 2020 *Page 1 of 18 pages* *Enclosure: 12 pages*
Course name: Multivariate Statistics
Course number: 02409
Aids allowed: All
Exam duration: 4 hours
Weighting: The questions are given equal weight

There is a total of 30 questions for the 5 problems.

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										

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										

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										

The possible answers for each question are numbered from 1 to 6.

The answers must be uploaded. See “AnswerSheetExam02409.txt”
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.

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.

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 patterns 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

1 ☐ $]0.5, 1]$

2 ☐ $]0.1, 0.5]$

3 ☐ $]0.01, 0.1]$

4 ☐ $]0.001, 0.01]$

5 ☐ $]0.0001, 0.001]$

6 ☐ Don't know.

Question 1.2.

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

1 ☐ 0.64492

2 ☐ -0.0542

3 ☐ 0.3973

4 ☐ -0.1289

5 ☐ -0.3098

6 ☐ Don't know.

Question 1.3.

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

1 ☐ $[-0.2906, -0.0168]$

2 ☐ $[-0.5454, 0.2404]$

3 ☐ $[-0.2117, -0.0933]$

4 ☐ $[-0.2827, -0.0168]$

5 ☐ $[-0.3035, -0.0015]$

6 ☐ Don't know.

Question 1.4.

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

1 ☐ 0.4725

2 ☐ 1/3

3 ☐ 0.1153

4 ☐ $\frac{1.152}{2.266+1.305}$

5 ☐ 1.152

6 ☐ Don't know.

Question 1.5.

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

1 ☐ $\frac{2.266}{2.266+1.305+1.153}$

2 ☐ 0.961

3 ☐ 0.28267+0.26172+0.40541+0.08102+0.39461+0.61816+0.79279+0.84050+0.15239+0.22610

4 ☐ 0.7555

5 ☐ 0.2266

6 ☐ Don't know.

Question 1.6.

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

1 ☐ 0.52865610

2 ☐ 0.67677

3 ☐ 0.8691

4 ☐ $\frac{0.67677}{1.3977638}$

5 ☐ 0.4580

6 ☐ Don't know.

Question 1.7.

Using the rotated factor model the uniqueness of *freetime* is:

1 ☐ 0.52865610

2 ☐ 0.4713439

3 ☐ 0.67677

4 ☐ $0.19640^2 + 0.67677^2 + (-0.17908)^2$

5 ☐ $\frac{0.52865610}{2.0513441+1.3977638+1.2755300}$

6 ☐ Don't know.

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 and age
D: Mainly an average of age and absences
E: Mainly an average of 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 factor model (UF) to the rotated (RF) we get the following interpretations of the three factors:

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

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

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

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

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

6 ☐ Don’t know.

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. It is not important to understand the meaning of the variables in detail.

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:

1 ☐ age

2 ☐ traveltime

3 ☐ goout

4 ☐ health

5 ☐ absences

6 ☐ Don't know.

Question 2.2.

The observation with the highest leverage is:

1 ☐ 2

2 ☐ 3

3 ☐ 9

4 ☐ 11

5 ☐ 19

6 ☐ Don't know.

Question 2.3.

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

1 ☐ [7.95; 11.79]

2 ☐ [6.78; 12.95]

3 ☐ [5.56; 14.17]

4 ☐ [7.19; 12.54]

5 ☐ 9.8662 ± 0.8886

6 ☐ Don't know.

Question 2.4.

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

1 ☐ 2

2 ☐ 3

3 ☐ 9

4 ☐ 11

5 ☐ 19

6 ☐ Don't know.

Question 2.5.

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

1 ☐ age goout G1

2 ☐ traveltime health absences

3 ☐ age health G1

4 ☐ traveltime health G1

5 ☐ goout health G1

6 ☐ Don't know.

Problem 3.

Enclosure B with SAS program and SAS output belongs to this problem.

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:

1 ☐ 0.0208

2 ☐ 0.0011

3 ☐ <.0001

4 ☐ -3.29

5 ☐ 0.1078

6 ☐ Don’t know.

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 null-hypothesis - the following distribution

1 ☐ t(11)

2 ☐ t(346)

3 ☐ F(11)

4 ☐ F(10)

5 ☐ t(10)

6 ☐ Don't know.

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

1 ☐ M1

2 ☐ M2

3 ☐ M3

4 ☐ M4

5 ☐ M5

6 ☐ Don't know.

Problem 4.

Enclosure C with SAS program and SAS output belongs to this problem.

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 'julefrokost'. 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. In total we have 308 observations, but only the 217 complete observations are used in this problem.

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:

1 ☐ 0.765609

2 ☐ 1935.9518735

3 ☐ 0.9540

4 ☐ 0.70283311

5 ☐ 17.76

6 ☐ Don't know.

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:

1 ☐ ProteinB

2 ☐ TCAIndexM

3 ☐ TCAIndexB

4 ☐ TCAM

5 ☐ TCAB

6 ☐ Don’t know.

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

1 ☐ 8

2 ☐ 16

3 ☐ 44

4 ☐ 56

5 ☐ 102

6 ☐ Don’t know.

Question 4.4.

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

1 ☐ $\frac{308+217-7-}{7-5} \cdot \frac{2.13508-0.70913}{(308+217)(308+217-2)/(308\cdot 217)+0.70913}$

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

3 ☐ $\frac{308+217-7-}{7-5} \cdot \frac{6.96931-2.63739}{(308+217)(308+217-2)/(308\cdot 217)+2.63739}$

4 ☐ $\frac{65+58-7-1}{7(65+58\quad)} \cdot \frac{65\cdot 58}{65+58} 2.13508$

5 ☐ $\frac{65+58-7-1}{7-3} \cdot \frac{2.13508-0.70913}{(65+58)(65+58-2)/(65\cdot 58)+0.70913}$

6 ☐ Don’t know.

Question 4.5.

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

1 ☐ $[0.6666, 0.6666, 0.6666]$

2 ☐ $[0.4923, 0.7931, 0.7021]$

3 ☐ $[0.3333, 0.3333, 0.3333]$

4 ☐ $[0.5077, 0.2069, 0.2979]$

5 ☐ $[0.2995, 0.2673, 0.4332]$

6 ☐ Don’t know.

Question 4.6.

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

1 ☐ $\frac{65+58-7-1}{7-3} \cdot \frac{2.13508-0.70913}{(65+58)(65+58\quad)/(65\cdot 58)+0.70913}$

2 ☐ $\frac{65+94-5-1}{5(65+94-2)} \cdot \frac{65\cdot 94}{65+94} 1.51850$

3 ☐ $\frac{65+58-7-1}{7(65+58\quad)} \cdot \frac{65\cdot 58}{65+58} 0.70913$

4 ☐ $\frac{65\cdot 58}{65+5} \cdot 0.70913$

5 ☐ $\frac{65+58-5-1}{5(65+58\quad)} \cdot \frac{65\cdot 58}{65+58} 0.70913$

6 ☐ Don’t know.

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?

1 ☐ $|\rho| < \frac{1}{2}$

2 ☐ $-\frac{1}{4} < \rho < \frac{1}{4}$

3 ☐ $\rho > \frac{1}{4}$

4 ☐ $-\sqrt{2} < \rho < \sqrt{2}$

5 ☐ $\rho < \frac{1}{2}$

6 ☐ Don't know

Question 5.2.

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

1 ☐ $2 - 2\rho$

2 ☐ $2 - \rho$

3 ☐ 2

4 ☐ $2 + \rho$

5 ☐ $2 + 2\rho$

6 ☐ Don't know

Question 5.3.

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

1 ☐ -2ρ

2 ☐ $-\rho$

3 ☐ 0

4 ☐ ρ

5 ☐ 2ρ

6 ☐ Don't know

Question 5.4.

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

1 ☐ -2ρ

2 ☐ $-\rho$

3 ☐ 0

4 ☐ ρ

5 ☐ 2ρ

6 ☐ Don't know

Question 5.5.

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

1 ☐ $\rho(x_1 - x_{2)}$

2 ☐ $-2\rho x_1$

3 ☐ 0

4 ☐ ρx_1

5 ☐ $\rho(x_1 + x_{2)}$

6 ☐ Don't know

Question 5.6.

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

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

2 ☐ $\rho x_1 - \rho x_2$

3 ☐ $2\rho x_1 + 2\rho x_2$

4 ☐ x_2

5 ☐ $-\frac{\rho^2}{1-\rho^2}x_1$

6 ☐ Don't know

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

1 ☐ $\frac{\rho^2}{1-\rho^2}$

2 ☐ $\frac{\rho^2}{(1-\rho)^2}$

3 ☐ $\frac{1-2\rho^2}{1-\rho^2}$

4 ☐ 0

5 ☐ $2\rho^2$

6 ☐ Don't know

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

1 ☐ $\frac{\rho^2}{2(1-\rho)^2}$

2 ☐ $\frac{\rho^2}{1-\rho^2}$

3 ☐ $\frac{(1-2\rho^2)^2}{1-4\rho^2}$

4 ☐ $\frac{4\rho^2}{4(1-\rho)^2}$

5 ☐ $\frac{\rho^2}{(1-\rho)^2}$

6 ☐ Don't know

LAST PAGE:
END OF THE EXAM SET

SAS-PROGRAM

```
proc corr data=studentFA noprob;
var age traveltime studytime famrel freetime goout dalc walc health
absences;
run;

proc factor data=studentFA nfactors=3 rotate=varimax plots=all;
var age traveltime studytime famrel freetime goout Dalc Walc health
absences;
run;
```

Some SAS-outputs have been omitted or truncated

The CORR Procedure

10 Variables:	age traveltime studytime famrel freetime goout Dalc Walc health absences

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
age	357	16.65546	1.26826	5946	15.00000	22.00000
traveltime	357	1.43137	0.68607	511.00000	1.00000	4.00000
studytime	357	2.04202	0.83190	729.00000	1.00000	4.00000
famrel	357	3.95518	0.88572	1412	1.00000	5.00000
freetime	357	3.24650	1.01160	1159	1.00000	5.00000
goout	357	3.09804	1.09078	1106	1.00000	5.00000
Dalc	357	1.49580	0.91989	534.00000	1.00000	5.00000
Walc	357	2.33053	1.29497	832.00000	1.00000	5.00000
health	357	3.54902	1.40264	1267	1.00000	5.00000
absences	357	6.31653	8.18762	2255	0	75.00000

Pearson Correlation Coefficients, N = 357

	age	traveltime	studytime	famrel	freetime	goout	Dalc	Walc	health	absences
age	1.00000	0.10672	0.00045	0.06623	0.00289	0.12804	0.14202	0.12084	-0.04969	0.21558
traveltime	0.10672	1.00000	-0.09583	-0.02357	-0.00794	0.03717	0.15421	0.13942	0.00132	0.00463
studytime	0.00045	-0.09583	1.00000	0.05212	-0.15253	-0.04789	-0.19982	-0.24760	-0.07279	-0.07454
famrel	0.06623	-0.02357	0.05212	1.00000	0.13463	0.03073	-0.07953	-0.12664	0.10804	-0.05808
freetime	0.00289	-0.00794	-0.15253	0.13463	1.00000	0.28352	0.20940	0.13276	0.08648	-0.07049
goout	0.12804	0.03717	-0.04789	0.03073	0.28352	1.00000	0.28176	0.44432	-0.00958	0.05659
Dalc	0.14202	0.15421	-0.19982	-0.07953	0.20940	0.28176	1.00000	0.64492	0.08888	0.10479
Walc	0.12084	0.13942	-0.24760	-0.12664	0.13276	0.44432	0.64492	1.00000	0.11168	0.12310
health	-0.04969	0.00132	-0.07279	0.10804	0.08648	-0.00958	0.08888	0.11168	1.00000	-0.02912
absences	0.21558	0.00463	-0.07454	-0.05808	-0.07049	0.05659	0.10479	0.12310	-0.02912	1.00000

The FACTOR Procedure

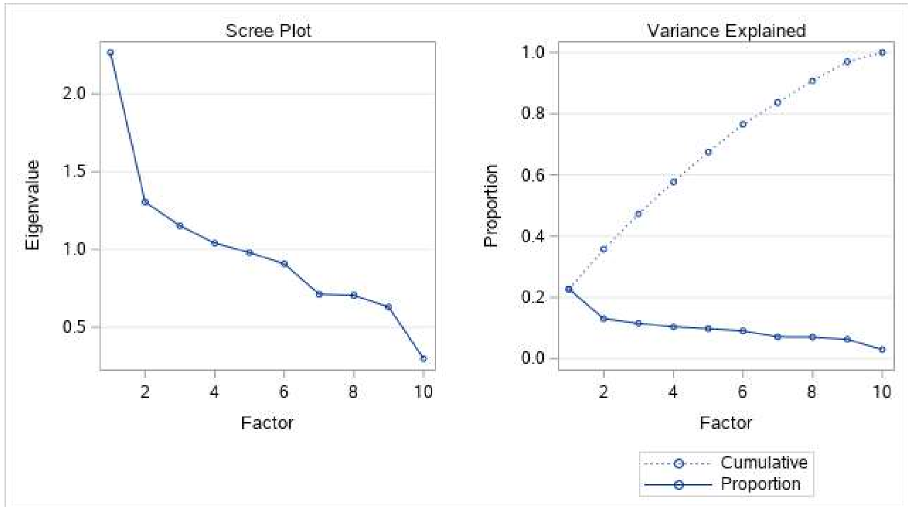
Input Data Type	Raw Data
Number of Records Read	357
Number of Records Used	357
N for Significance Tests	357

The FACTOR Procedure
Initial Factor Method: Principal Components

Prior Communality Estimates: ONE

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

3 factors will be retained by the NFACTOR criterion.



The FACTOR Procedure

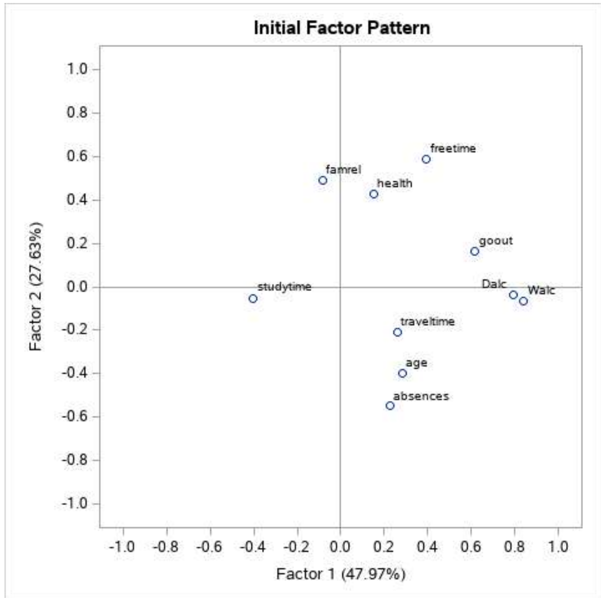
Initial Factor Method: Principal Components

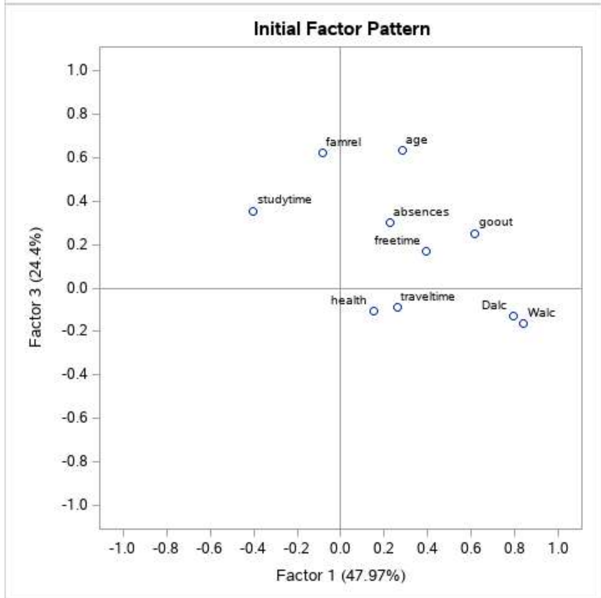
Factor Pattern			
	Factor1	**Factor2**	**Factor3**
age	0.28267	-0.39690	0.63199
traveltime	0.26172	-0.21130	-0.09024
studytime	-0.40541	-0.05649	0.35295
famrel	-0.08102	0.48920	0.62173
freetime	0.39461	0.58670	0.16946
goout	0.61816	0.16092	0.25177
Dalc	0.79279	-0.03856	-0.12691
Walc	0.84050	-0.06796	-0.16095
health	0.15239	0.42747	-0.10315
absences	0.22610	-0.54942	0.29871

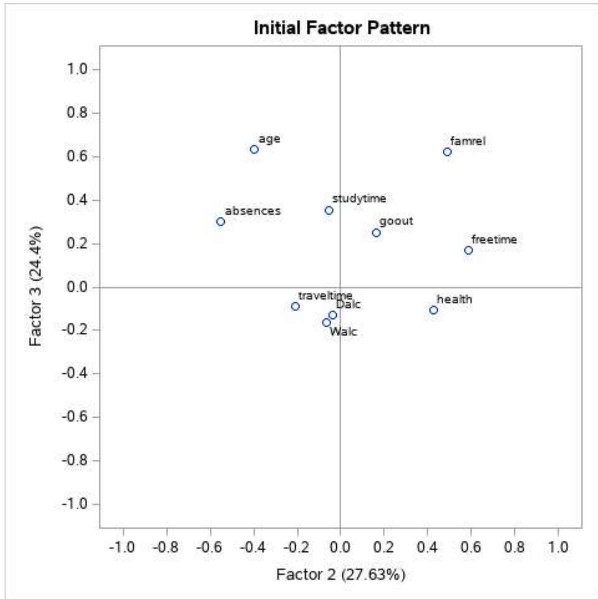
Variance Explained by Each Factor		
Factor1	**Factor2**	**Factor3**
2.2664796	1.3054944	1.1526639

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

The FACTOR Procedure
Initial Factor Method: Principal Components







The FACTOR Procedure
Rotation Method: Varimax

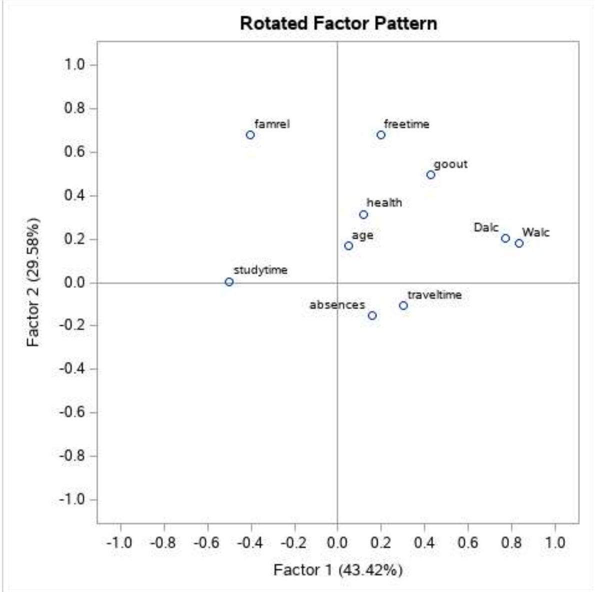
Orthogonal Transformation Matrix			
	1	2	3
1	0.89658	0.38213	0.22388
2	-0.14768	0.73453	-0.66232
3	-0.41753	0.56076	0.71499

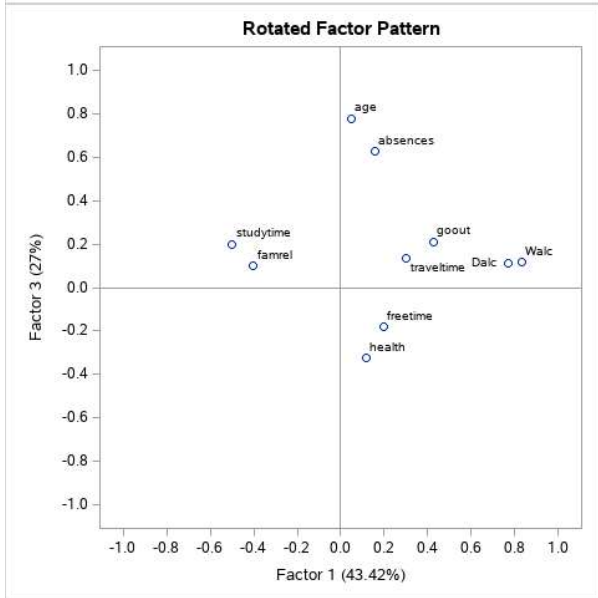
Rotated Factor Pattern			
	Factor1	Factor2	Factor3
age	0.04818	0.17088	0.77803
traveltime	0.30353	-0.10580	0.13401
studytime	-0.50251	0.00151	0.19901
famrel	-0.40448	0.67701	0.10239
freetime	0.19640	0.67677	-0.17908
goout	0.42535	0.49560	0.21182
Dalc	0.76949	0.20346	0.11229
Walc	0.83082	0.18101	0.11810
health	0.11657	0.31438	-0.32276
absences	0.15913	-0.14966	0.62809

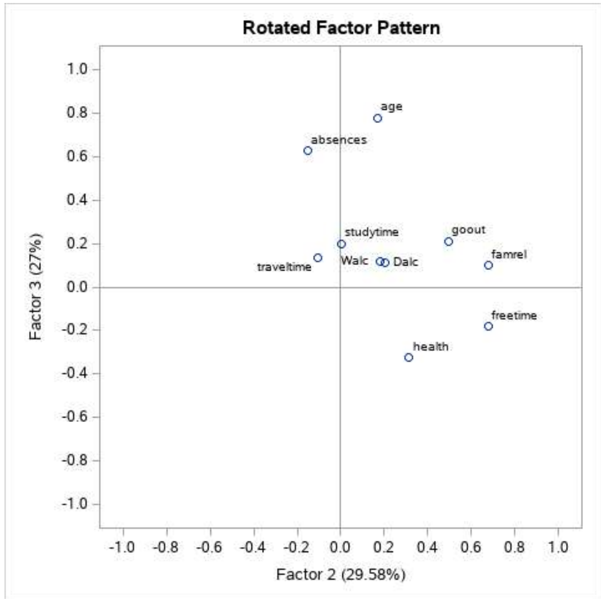
Variance Explained by Each Factor		
Factor1	Factor2	Factor3
2.0513441	1.3977638	1.2755300

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

The FACTOR Procedure
Rotation Method: Varimax






Rotated Factor Pattern
age
absences
studytime
goout
famrel
traveltime
Walc
Dalc
freetime
health
Factor 3 (27%)
Factor 2 (29.58%)
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0

SAS-PROGRAM

```
title 'MODEL 1';
proc glm data=studentFA;
model g3 = age traveltime studytime famrel freetime goout dalc walc
health absences;
run;

title 'MODEL 2';
proc glm data=studentFA;
model g3 = traveltime studytime famrel freetime goout dalc walc;
run;

title 'MODEL 3';
proc glm data=studentFA;
model g3 = studytime famrel goout dalc walc;
run;

title 'MODEL 4';
proc glm data=studentFA;
model g3 = studytime;
run;

title 'MODEL 5';
proc glm data=studentFA;
model g3 = ;
run;
```

Some SAS-outputs have been omitted or truncated

MODEL 1

The GLM Procedure

Number of Observations Read	357
Number of Observations Used	357

The GLM Procedure

Dependent Variable: G3

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	399.950009	39.995001	4.18	<.0001
Error	346	3309.097611	9.563866		
Corrected Total	356	3709.047619			

R-Square	Coeff Var	Root MSE	G3 Mean
0.107831	26.83618	3.092550	11.52381

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

MODEL 2

The GLM Procedure

Number of Observations Read	357
Number of Observations Used	357

The GLM Procedure

Dependent Variable: G3

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	232.214378	33.173483	3.33	0.0019
Error	349	3476.833242	9.962273		
Corrected Total	356	3709.047619			

R-Square	Coeff Var	Root MSE	G3 Mean
0.062608	27.38944	3.156307	11.52381

Source	DF	Type I SS	Mean Square	F Value	Pr > F
traveltime	1	36.9308058	36.9308058	3.71	0.0550
studytime	1	51.3892487	51.3892487	5.16	0.0237
famrel	1	3.2294859	3.2294859	0.32	0.5695
freetime	1	0.2773349	0.2773349	0.03	0.8676
goout	1	112.7131838	112.7131838	11.31	0.0009
Dalc	1	16.0314376	16.0314376	1.61	0.2054
Walc	1	11.6428807	11.6428807	1.17	0.2804

MODEL 3

The GLM Procedure

Number of Observations Read	357
Number of Observations Used	357

The GLM Procedure

Dependent Variable: G3

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	207.212301	41.442460	4.15	0.0011
Error	351	3501.835318	9.976739		
Corrected Total	356	3709.047619			

R-Square	Coeff Var	Root MSE	G3 Mean
0.055867	27.40932	3.158598	11.52381

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	12.28659530	1.03473479	11.87	<.0001
studytime	0.35528304	0.20856081	1.70	0.0894
famrel	0.08413867	0.19148081	0.44	0.6606
goout	-0.37150477	0.17256117	-2.15	0.0320
Dalc	-0.09069022	0.23847608	-0.38	0.7040
Walc	-0.22933510	0.18535636	-1.24	0.2168

MODEL 4

The GLM Procedure

Number of Observations Read	357
Number of Observations Used	357

The GLM Procedure

Dependent Variable: G3

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	59.567345	59.567345	5.79	0.0166
Error	355	3649.480274	10.280226		
Corrected Total	356	3709.047619			

R-Square	Coeff Var	Root MSE	G3 Mean
0.016060	27.82308	3.206279	11.52381

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	10.51972622	0.45032212	23.36	<.0001
studytime	0.49171158	0.20427144	2.41	0.0166

MODEL 5

The GLM Procedure

Number of Observations Read	357
Number of Observations Used	357

The GLM Procedure

Dependent Variable: G3

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	47408.95238	47408.95238	4550.38	<.0001
Error	356	3709.04762	10.41867		
Uncorrected Total	357	51118.00000			

R-Square	Coeff Var	Root MSE	G3 Mean
0.000000	28.00981	3.227797	11.52381

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	11.52380952	0.17083313	67.46	<.0001

SAS-PROGRAM

```
proc glm data=home.herring;
model ProteinB TCAIndexM TCAIndexB TCAM TCAB = country treatment;
manova h=_all_/printe printh;
run;



title 'Full model';
proc discrim data=home.herring pool=yes;
class country;
var Water ashm ProteinB TCAIndexM TCAIndexB TCAM  TCAB;

run;
title 'Reduced model';
proc discrim data=home.herring pool=yes;
class country;
var ProteinB TCAIndexM TCAIndexB TCAM  TCAB;
run;
```

Some SAS-outputs have been omitted or truncated

The GLM Procedure

Number of Observations Read	308
Number of Observations Used	217

The GLM Procedure
Multivariate Analysis of Variance

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

Partial Correlation Coefficients from the Error SSCP Matrix / Prob > |r|

DF = 214	ProteinB	TCAIndexM	TCAIndexB	TCAM	TCAB
ProteinB	1.000000	0.765609 <.0001	-0.042909 0.5315	0.715155 <.0001	0.793693 <.0001
TCAIndexM	0.765609 <.0001	1.000000	0.361841 <.0001	0.971043 <.0001	0.878158 <.0001
TCAIndexB	-0.042909 0.5315	0.361841 <.0001	1.000000	0.390594 <.0001	0.529015 <.0001
TCAM	0.715155 <.0001	0.971043 <.0001	0.390594 <.0001	1.000000	0.854277 <.0001
TCAB	0.793693 <.0001	0.878158 <.0001	0.529015 <.0001	0.854277 <.0001	1.000000

The GLM Procedure
Multivariate Analysis of Variance

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

Characteristic Roots and Vectors of: E Inverse * H, where H = Type III SSCP Matrix for Country E = Error SSCP Matrix						
		Characteristic Vector V'EV=1				
Characteristic Root	Percent	ProteinB	TCAIndexM	TCAIndexB	TCAM	TCAB
0.38994346	100.00	0.02210266	-0.04061828	-0.00064423	0.16785187	0.05557009
0.00000000	0.00	-0.14751606	0.00282201	-0.00987792	-0.04876092	0.22055605
0.00000000	0.00	0.02490980	0.01413748	-0.00063096	-0.06161947	0.00000000
0.00000000	0.00	-0.03491092	-0.00335833	0.00085636	0.08363119	0.00000000
0.00000000	0.00	0.03260163	0.00908450	0.00485126	-0.11080220	0.00000000

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall Country Effect H = Type III SSCP Matrix for Country E = Error SSCP Matrix S=1 M=1.5 N=104					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.71945373	16.38	5	210	<.0001
Pillai's Trace	0.28054627	16.38	5	210	<.0001
Hotelling-Lawley Trace	0.38994346	16.38	5	210	<.0001
Roy's Greatest Root	0.38994346	16.38	5	210	<.0001

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

Characteristic Roots and Vectors of: E Inverse * H, where H = Type III SSCP Matrix for Treatment E = Error SSCP Matrix						
		Characteristic Vector V'EV=1				
Characteristic Root	Percent	ProteinB	TCAIndexM	TCAIndexB	TCAM	TCAB
0.42281288	100.00	0.00494347	-0.03159396	0.00183861	0.17100089	0.01898749
0.00000000	0.00	-0.13854175	-0.00456579	-0.00992409	-0.02062099	0.22665498
0.00000000	0.00	0.04157964	0.00213342	-0.00018250	-0.01180662	0.00000000
0.00000000	0.00	-0.05857993	0.02630193	-0.00094123	-0.06089089	0.00000000
0.00000000	0.00	0.02780152	0.01536603	0.00445726	-0.14179744	0.00000000

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall Treatment Effect H = Type III SSCP Matrix for Treatment E = Error SSCP Matrix S=1 M=1.5 N=104					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.70283311	17.76	5	210	<.0001
Pillai's Trace	0.29716689	17.76	5	210	<.0001
Hotelling-Lawley Trace	0.42281288	17.76	5	210	<.0001
Roy's Greatest Root	0.42281288	17.76	5	210	<.0001

Full model

The DISCRIM Procedure

Total Sample Size	217	DF Total	216
Variables	7	DF Within Classes	214
Classes	3	DF Between Classes	2

Number of Observations Read	308
Number of Observations Used	217

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

Pooled Covariance Matrix Information	
Covariance Matrix Rank	Natural Log of the Determinant of the Covariance Matrix
7	6.91182

The DISCRIM Procedure

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

Linear Discriminant Function for Country			
Variable	1	2	3
Constant	-214.06754	-199.48854	-231.01164
Water	3.67582	3.34010	3.94909
AshM	9.84759	10.21063	10.25664
ProteinB	21.67761	20.88335	21.94040
TCAIndexM	-5.38804	-4.96943	-6.44518
TCAIndexB	1.61922	1.61593	1.57136
TCAM	44.14465	40.87644	49.30540
TCAB	-40.31472	-38.83249	-39.47448

Full model

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	

Error Count Estimates for Country				
	1	2	3	Total
Rate	0.5538	0.1379	0.1277	0.2731
Priors	0.3333	0.3333	0.3333	

Reduced model

The DISCRIM Procedure

Total Sample Size	217	DF Total	216
Variables	5	DF Within Classes	214
Classes	3	DF Between Classes	2

Number of Observations Read	308
Number of Observations Used	217

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

Pooled Covariance Matrix Information	
Covariance Matrix Rank	Natural Log of the Determinant of the Covariance Matrix
5	4.48347

The DISCRIM Procedure

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

Linear Discriminant Function for Country			
Variable	1	2	3
Constant	-77.32998	-74.11137	-76.68533
ProteinB	26.82560	26.12167	27.32768
TCAIndexM	0.73678	0.89514	0.05843
TCAIndexB	1.93025	1.90983	1.90263
TCAM	0.21604	-1.64104	2.77655
TCAB	-37.45547	-36.11821	-36.43240

Reduced model

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	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	

Error Count Estimates for Country				
	1	2	3	Total
Rate	0.5077	0.2069	0.2979	0.3375
Priors	0.3333	0.3333	0.3333	