STOCKHOLM UNIVERSITY
Department of Statistics
Ellinor Fackle-Fornius

EXAM IN MULTIVARIATE METHODS
November 1 2018

Time: 5 hours

Allowed aids: One A4 size sheet (with formulas, text etc.) brought by the student, pocket calculator, language dictionary.

The exam consists of five questions. To score maximum points on a question solutions need to be clear, detailed and well motivated.

Results will be announced no later than November 15.

Question 1. (16 points)

Define and describe the following tests:

a) χ^2 test for Confirmatory Factor Analysis

b) Multivariate Wilks' Λ for Discriminant Analysis

c) $-2LogL$ test for Logistic Regression

Question 2. (16 points)

We have data with scores on various Olympic decathlon events for 33 athletes. The events are: (1) 100m, (2) long jump, (3) shot putt, (4) high jump, (5) 400m, (6) 110m hurdles, (7) discus, (8) pole vault, (9) javelin, (10) 1500m. The throwing and jumping events are measured in metres and the running events in seconds. A Principal Components Analysis was performed after standardizing the data (that is, based on the correlation matrix) and the resulting outputs from PROC PRINCOMP are presented on page 2.

a) Draw a scree plot.

b) Explain how many principal components you would use to summarize the data.

c) Compute the loadings of the variables on the first PC. Interpret the first PC.

d) The analysis was performed on standardized data. What difference would you expect if the data had not been standardized? Which type of analysis would you prefer?

The PRINCOMP Procedure

Observations	33
Variables	10

Simple Statistics							
	_100m	longjump	shotputt	highjump	_400m	_110mhurdles	discus
Mean	11.19636364	7.133333333	13.97636364	1.982727273	49.27666667	15.04878788	42.35393939
StD	0.24332101	0.304340133	1.33199056	0.093983799	1.06966019	0.50676522	3.71913123

Simple Statistics			
	polevault	javelin	_1500m
Mean	4.739393939	59.43878788	276.0384848
StD	0.334420575	5.49599841	13.6570975

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	3.41823814	0.81184501	0.3418	0.3418
2	2.60639314	1.66309673	0.2606	0.6025
3	0.94329641	0.06527516	0.0943	0.6968
4	0.87802124	0.32139459	0.0878	0.7846
5	0.55662665	0.06539914	0.0557	0.8403
6	0.49122752	0.06063230	0.0491	0.8894
7	0.43059522	0.12379709	0.0431	0.9324
8	0.30679812	0.03984871	0.0307	0.9631
9	0.26694941	0.16509526	0.0267	0.9898
10	0.10185415		0.0102	1.0000

Eigenvectors										
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10
_100m	-.415882	0.148808	0.267472	0.088332	0.442314	-.030712	0.254398	-.663713	0.108395	0.109480
longjump	0.394051	-.152082	0.168949	0.244250	-.368914	0.093782	0.750534	-.141264	-.046139	0.055804
shotputt	0.269106	0.483537	-.098533	0.107763	0.009755	-.230021	-.110664	-.072506	-.422476	0.650737
highjump	0.212282	0.027898	0.854987	-.387944	0.001876	-.074544	-.135124	0.155436	0.102065	0.119412
_400m	-.355847	0.352160	0.189496	-.080575	-.146965	0.326929	0.141339	0.146839	-.650762	-.336814
_110mhurdles	-.433482	0.069568	0.126160	0.382290	0.088803	-.210491	0.272530	0.639004	0.207239	0.259718
discus	0.175792	0.503335	-.046100	-.025584	-.019359	-.614912	0.143973	-.009400	0.167241	-.534503
polevault	0.384082	0.149582	-.136872	-.143965	0.716743	0.347760	0.273266	0.276873	0.017664	-.065896
javelin	0.179944	0.371957	0.192328	0.600466	-.095582	0.437444	-.341910	-.058519	0.306196	-.130932
_1500m	-.170143	0.420965	-.222552	-.485642	-.339772	0.300324	0.186870	-.007310	0.456882	0.243118

Question 3. (16 points)

Consider the six-indicator two-factor model represented by the following equations:

$$A = 0.85F_1 + 0.12F_2 + U_A$$
$$B = 0.74F_1 + 0.07F_2 + U_B$$
$$C = 0.67F_1 + 0.18F_2 + U_C$$
$$D = 0.21F_1 + 0.93F_2 + U_D$$
$$E = 0.05F_1 + 0.77F_2 + U_E$$
$$F = 0.08F_1 + 0.62F_2 + U_F$$

The usual assumptions hold for the above model.

a) What are the pattern loadings of indicators A, C and E on the factors F_1 and F_2 when *i)* $Corr(F_1, F_2) = \phi_{12} = 0$ *ii)* $Corr(F_1, F_2) = \phi_{12} = 0.2$?

b) What are the structure loadings of indicators A, C and E on the factors F_1 and F_2 when *i)* $Corr(F_1, F_2) = \phi_{12} = 0$ *ii)* $Corr(F_1, F_2) = \phi_{12} = 0.2$?

c) Compute the correlations between indicators C and D when *i)* $Corr(F_1, F_2) = \phi_{12} = 0$ *ii)* $Corr(F_1, F_2) = \phi_{12} = 0.2$.

d) What percentage of the variance of indicator A is not accounted for by the common factors when *i)* $Corr(F_1, F_2) = \phi_{12} = 0$ *ii)* $Corr(F_1, F_2) = \phi_{12} = 0.2$?

Question 4. (16 points)

Consider the matrix of distances

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0 & & & \\ 13 & 0 & & \\ 52 & 65 & 0 & \\ 18 & 1 & 58 & 0 \end{pmatrix} \end{array}$$

Cluster the four items using each of the following procedures:

a) Single linkage

b) Complete linkage

c) Average linkage

Question 5. (16 points)

In order to identify the best sales prospects for a sales campaign, a riding-mower manufacturer is interested in classifying families as prospective riding-mower owners or nonowners on the basis of x_1 = income (in \$1000s) and x_2 = lot size (in 1000 ft^2). Data were collected and a logistic regression was fitted:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

	Estimate	SE	t-stat	p-value
(Intercept)	-29.2640	12.7889	-2.2882	0.0221
x1	0.1109	0.0543	2.0417	0.0412
x2	0.9638	0.4628	2.0825	0.0373

The following table displays observations on 12 riding-mower owners and 12 nonowners as well as the estimated probability to be an owner based on the logistic regression.

Riding-mower owners			Nonowners		
x_1	x_2	$\widehat{p}$	x_1	x_2	$\widehat{p}$
90.0	18.4	0.1746	105.0	19.6	0.7801
115.5	16.8	0.4333	82.8	20.8	0.4904
94.8	21.6	0.8873	94.8	17.2	0.1018
91.5	20.8	0.7163	73.2	20.4	0.1842
117.0	23.6	0.9984	114.0	17.6	0.5833
140.1	19.2	0.9916	79.2	17.6	0.0287
138.0	17.6	0.9524	89.4	16.0	0.0192
112.8	22.4	0.9921	96.0	18.4	0.2915
99.0	20.0	0.7284	77.4	16.4	0.0076
123.0	20.8	0.9881	63.0	18.8	0.0154
81.0	22.0	0.7148	81.0	14.0	0.0011
111.0	20.0	0.9103	93.0	14.8	0.0091

a) Interpret the parameter $\widehat{\beta}_1$.

b) What are the odds that a household with a \$60 000 income and a lot size of 20 000 ft^2 is an owner?

c) What is the classification of a household with a \$60 000 income and a lot size of 20 000 ft^2?

d) What is the minimum income that household with 16 000 ft^2 lot size should have before it is classified as an owner?

e) Classify the observations given in the table and compute the sensitivity and specificity of the classification.

Department of Statistics

Stockholms universitet

Correction sheet

Date: 1/11/2018

Room: Brunnsvikssalen

Exam: Multivariate Methods

Course: Multivariate Methods

Anonymous code: 0004-DWZ

☒ I authorise the anonymous posting of my exam, in whole or in part, on the department homepage as a sample student answer.

NOTE! ALSO WRITE ON THE BACK OF THE ANSWER SHEET

Mark answered questions

1	2	3	4	5	6	7	8	9	Total number of pages
X	X	X	X	X					9
Teacher's notes 4	12	12	16	15					

Points	Grade	Teacher's sign.
59	C	

+11 = 70

SU, DEPARTMENT OF STATISTICS

Room: Brunnsviksalen Anonymous code: 0004-DW2 Sheet number: 1

1 a) X^2-test for confirmatory factor analysis:

We do this test to evaluate the model fit.
To do this, we test if the difference between the Sample covariance matrix and the estimated covariance matrix equals zero.

2

b) –

–

c) -2 Log L test for Logistic regression:

this is a test for hypothesis $G=0$, $G\neq 0$ ✓

-2 log L gives us the deviance, but first we need the likelihood function:

$$L = \prod_{i=1}^{n} p(y_i) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i} = p^{\Sigma y_i}(1-p)^{n-\Sigma y_i}$$

$$l = \ln(L) = \Sigma y_i \ln(p) + (n-\Sigma y_i)\ln(1-p)$$

we then get the deviance:

$$-2\left[\Sigma y_i \ln(p) + [n-\Sigma y_i]\ln(1-p)\right]$$

then we calculate G, which is deviance (reduced model) - deviance (full model)

$$-2l(\text{reduced}) - (-2l(\text{full})) = 2l(\text{full}) - 2l\,\text{reduced}$$

G is then approx $X^2(n-q)$ ✓, and we use the X^2 distribution to find our limit.

G is used to compare 2 models while -2 LogL is used to assess overall fit

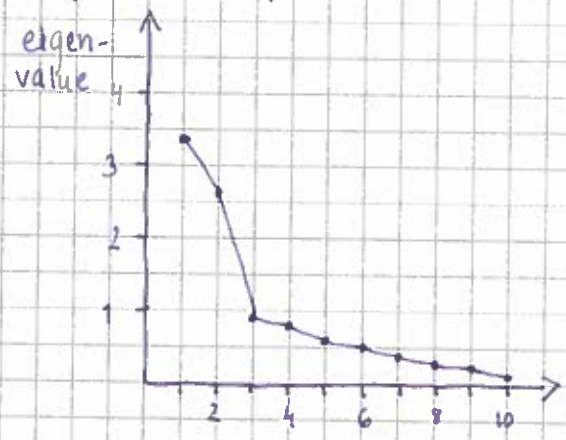
2

4

SU, DEPARTMENT OF STATISTICS

Room: Brunnsviksalen Anonymous code: 0004-DW2 Sheet number: 2

2 a) Scree plot



4

b) There is a few different ways to choose how many principal components to use to summarize the data.

Firstly, because we have standardized data, I could use the eigenvalue > 1 rule. In that case I would choose 2 principal components to use.

I could also look for an elbow in the scree plot, above we can see that this would give us 3 PC:s.

Lastly, we can decide upon a certain variance that we want to be explained. If we decide that we want 70% of the variance to be explained, we would end up with 4 PC:s.

4

c) We get the loadings of the first PC from the eigenvectors. ✓

Variable	Loading on first PC
100m	-0,415882
long jump	0,394051
Shot putt	0,269106
high jump	0,212282
400m	-0,355847
110m hurd.	-0,43848
discus	0,175792
polevault	0,384082
javelin	0,179944
1500 m	-0,170143

c) cont. Interpretation:

High loadings on a PC means that the variable is influential in forming the PC. Typically, we use values above 0.5 or below -0.5, though it's not really a rule.

In this case we do not have any values above 0.5, but if we look at the greatest numbers one can see that 100 m (-0,416), longjump (0,394), 110 m hurdles (-0,433) and polevault (0,384) is influential in forming PC 1.

Therefore, the interpretation of PC one is that it is a measure of the quality in jumping and lack of quality in running.

note that high value for running has a different meaning than for the other events

d)

If the data had not been standardized I would expect the eigenvectors to be more unequal. This is because when we standardize the data we give the variable equal weights, which is reflected in the eigenvalues.

What type of analysis you prefer depends on whether the variation of the variables reflects the importance of the different variables.

In this case, given that the events of decathlon is equally important, we should use standardized data.

1

3

12

3.

$A = 0{,}85 F_1 + 0{,}12 U_A$
$B = 0{,}74 F_1 + 0{,}07 U_B$
$C = 0{,}67 F_1 + 0{,}18 U_C$
$D = 0{,}21 F_1 + 0{,}93 U_D$
$E = 0{,}05 F_1 + 0{,}77 U_E$
$F = 0{,}08 F_1 + 0{,}62 U_F$

U_i ← Unique factor

a) i) if $\phi = 0$, then structure loading → pattern loading

Indicator	Pattern loading F_1	F_2
A	0,85	0
C	0,67	0
E	0,05	0

ii) if $\phi \neq 0$, (structure loading $= \lambda_{j1} + \lambda_{j2} \cdot \phi$, where $\lambda_{j1}, \lambda_{j2}$ is pattern loadings.)

Indicator	Pattern loading F_1	F_2	
A	0,85	0,17	$\lambda_{j2} = 0{,}85 \cdot 0{,}2 = 0{,}17$
C	0,67	0,134	$\lambda_{j2} = 0{,}67 \cdot 0{,}2 = 0{,}134$
E	0,05	0,01	$\lambda_{j2} = 0{,}05 \cdot 0{,}2 = 0{,}01$

pattern loading is the same as for i) 2

b) i)

Indicator	Structure loading F_1	F_2
A	0,85	0
C	0,67	0
E	0,05	0

ii)

Indicator	Structure loading F_1	F_2
A	0,884	0,34
C	0,6968	0,268
E	0,052	0,02

$\lambda_{A1} = 0{,}85 + 0{,}17 \cdot 0{,}2 = 0{,}884$
$\lambda_{A2} = 0{,}85 \cdot 0{,}2 + 0{,}17 =$

Följdfel

4

3 c) i) $\text{Corr}(C, D)$, $\text{Corr}(F_1, F_2) = 0$

$$\text{Corr}(X_i, X_j) = \lambda_{i1} \cdot \lambda_{j1} + \lambda_{i2} \cdot \lambda_{j2} + \underbrace{(\lambda_{i1} \cdot \lambda_{j2} + \lambda_{j1} \cdot \lambda_{i2})\,\phi}_{\text{disappears when } \phi = 0}$$

$$\text{Corr}(C, D) = \lambda_{c1} \cdot \lambda_{d1} + \lambda_{c2} \cdot \lambda_{d2} =$$

$$0{,}67 \cdot 0{,}21 + 0 \cdot 0 = 0{,}1344$$

ii) $\phi = 0{,}2$

$$\text{Corr}(C, D) = \lambda_{c1} \cdot \lambda_{D1} + \lambda_{c2} \cdot \lambda_{D2} + (\lambda_{c1} \cdot \lambda_{D2} + \lambda_{D1} \cdot \lambda_{c2})\,\phi =$$

$$= 0{,}67 \cdot 0{,}21 + 0 \cdot 0 + (0{,}67 \cdot 0 + 0{,}21 \cdot 0) \cdot 0{,}2 = 0{,}1344$$

4

d) $\text{Var}(A) = \lambda_A^2 + \text{Var}(U_A)$ för one-factor modul

i) $\text{Var}(A) = 0{,}85^2 + 0{,}12 = 0{,}8425$

$\frac{0{,}12}{0{,}8425} = 0{,}1424 \rightarrow 14{,}24\ \%$

Följdfel från uppg

ii) $0{,}85^2 + 0{,}17^2 + 0{,}12 = 0{,}8714$

$\frac{0{,}12}{0{,}8714} = 0{,}1377 \rightarrow 13{,}77\ \%$

2

(12)

SU, DEPARTMENT OF STATISTICS

Room: Brunnsviksalen Anonymous code: 0004-DW2 Sheet number: 6

4.

$$D = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & & & \\ 13 & 0 & & \\ 52 & 65 & 0 & \\ 18 & (1) & 58 & 0 \end{bmatrix} \end{array} \quad \text{Sym.}$$

a) Single-linkage - min-distance

As we can see in the matrix of distances, 1 is the shortest distance so we start by clustering 4 and 2 together.

then we calculate distances:

$d_{1(24)} = \min(d_{12}, d_{14}) = \min(13, 18) = 13$

$d_{3(24)} = \min(d_{23}, d_{34}) = \min(65, 58) = 58$

$$D = \begin{array}{c} \\ 1 \\ 3 \\ (24) \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 3 & (24) \end{array} \\ \begin{bmatrix} 0 & & \\ 52 & 0 & \\ (13) & 58 & 0 \end{bmatrix} \end{array} \quad \text{Sym.}$$

R

- Now the shortest distance is 13, we therefore form a new cluster consisting of 1, 2, 4 and calculate distances:

$d_{3(124)} = \min(d_{13}, d_{3(24)}) = \min(52, 58) = 52$

The solution is one cluster with 3, and one cluster with 1,2,4.

$$D = \begin{array}{c} \\ 3 \\ (124) \end{array} \begin{array}{c} \begin{array}{cc} 3 & (124) \end{array} \\ \begin{bmatrix} 0 & \\ 52 & 0 \end{bmatrix} \end{array} \quad \text{Sym}$$

R

5

b) Complete linkage - max-distance

In this case we will also start with the shortest distance 1, clustering 2 and 4 together.

calculating distances:

$d_{1(24)} = \max(d_{12}, d_{14}) = \max(13, 18) = 18$

$d_{3(24)} = \max(d_{23}, d_{34}) = \max(65, 58) = 65$

$$D = \begin{array}{c} \\ 1 \\ 3 \\ (24) \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 3 & (24) \end{array} \\ \begin{bmatrix} 0 & & \\ 52 & 0 & \\ (18) & 65 & 0 \end{bmatrix} \end{array}$$

R

→

4 b cont.) Now the shortest distance is 18 and we cluster 1,2,4 together and calculate distances:

$$d_{3(124)} = \max\left(d_{13}, d_{3(24)}\right) = \max(52, 65) = 65$$

$$D = \begin{array}{c} \\ 3 \\ (124) \end{array} \begin{array}{c} \begin{array}{cc} 3 & (124) \end{array} \\ \begin{bmatrix} 0 & \text{sym.} \\ 65 & 0 \end{bmatrix} \end{array}$$

R

Same clusering solution in b) as in a) but with different cluster distance

5

c) Average-linkage

Also now, we look for the shortest distance which is 1 and cluster 2 and 4 together.

Calculating distance:

$$d_{1(24)} = \frac{d_{12} + d_{14}}{2} = \frac{13 + 18}{2} = 15{,}5$$

$$d_{3(24)} = \frac{d_{23} + d_{34}}{2} = \frac{65 + 58}{2} = 61{,}5$$

$$D = \begin{array}{c} \\ 1 \\ 3 \\ (24) \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 3 & (24) \end{array} \\ \begin{bmatrix} 0 & & \text{sym} \\ 52 & 0 & \\ 15{,}5 & 61{,}5 & 0 \end{bmatrix} \end{array}$$

R

Now the shortest distance is 15,5, like in a) and b), we cluster 1,2,4 together and calculate the average distance:

$$d_{3(124)} = \frac{d_{13} + d_{23} + d_{34}}{3} = \frac{52 + 65 + 58}{3} = 58{,}3333$$

$$D = \begin{array}{c} \\ 3 \\ (124) \end{array} \begin{array}{c} \begin{array}{cc} 3 & (124) \end{array} \\ \begin{bmatrix} 0 & \text{sym.} \\ 58{,}33 & 0 \end{bmatrix} \end{array}$$

R

Also, Same solution as in a) and b) but with yet another cluster distance.

6

16

5. $\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

$\hat{\beta}_0 = -29,2640$

$\hat{\beta}_1 = 0,1109$

$\hat{\beta}_2 = 0,9638$

a) Interpretation of $\hat{\beta}_1$:

For every additional unit of X_1, the log(odds) of being a riding-mower owner increases with 0,1109, given that the rest remains unchanged.

3

b) $X_1 = \frac{60.000}{1000} = 60 \quad X_2 = \frac{20.000}{1000} = 20$

$\log(\text{odds}) = \log\left(\frac{p}{1-p}\right) = -29,264 + 0,1109 \cdot 60 + 0,9638 \cdot 20$

$\frac{p}{1-p} = e^{-29,264 + 0,1109 \cdot 60 + 0,9638 \cdot 20} = e^{-3,334}$

$\frac{p}{1-p} \approx 0,0357$, which is the odds

R 2

c) $p = 0,357\,(1-p) = 0,0357 - 0,0357p$

$1,0357p = 0,0357$

$\hat{p} = \frac{0,0357}{1,0357} \approx 0,0344 \rightarrow 3,44\,\%$ R

- the classification of a household with a $60.000 income and a lot size of 20.000 ft² is $\hat{p} = 0,0344$ ⇒ classification?

2

→

SU, DEPARTMENT OF STATISTICS

Room: Brunnsviksalen Anonymous code: 0004-DW2 Sheet number: 9

5. d) $X_1 = ?$ $X_2 = 16$

$\hat{p} > 0,5$ to be classified as an owner R

$$\log\left(\frac{0,5}{1-0,5}\right) = -29,264 + 0,1109 \cdot X_1 + 0,9638 \cdot 16$$

$$\frac{0,5}{0,5} = e^{-29,264 + 0,1109 \cdot X_1 + 0,9638 \cdot 16} = 1 \quad \vdots \quad e^0 = 1$$

which gives us: $-29,264 + 0,1109 \cdot X_1 + 0,9638 \cdot 16 = 0$

$$X_1 \cdot 0,1109 = 13,8432$$

$$X_1 = \frac{13,8432}{0,1109} \approx 124,82597$$ R

- The minimum income that a household with 16.000 ft^2 lot size should have to be classified as an owner is:

$ 124.826. R

4

e)

	Owner $\hat{p}$	class	Non-owner $\hat{p}$	class
1	0,1746	2	0,7801	1
2	0,4333	2	0,4904	2
3	0,8873	1	0,1018	2
4	0,7163	1	0,1842	2
5	0,9984	1	0,5833	1
6	0,9916	1	0,0287	2
7	0,9524	1	0,0192	2
8	0,9925	1	0,2915	2
9	0,7284	1	0,0076	2
10	0,9881	1	0,0154	2
11	0,7148	1	0,0011	2
12	0,9103	1	0,0091	2

R

TP = 10 TN = 10
FN = 2 FP = 2

Sensitivety:

$$\frac{TP}{TP+FN} = \frac{10}{10+2} = \frac{10}{12} = 0,8333$$ R

Specificity:

$$\frac{TN}{FP+TN} = \frac{10}{2+10} = \frac{10}{12} = 0,8333$$ R

4

15