STOCKHOLM UNIVERSITY
Department of Statistics
Ellinor Fackle-Fornius

EXAM IN MULTIVARIATE METHODS
March 27 2019

Time: 5 hours

Allowed aids: 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 April 10.

Question 1. (16 points)

Given the correlation matrix

$$\mathbf{R} = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}$$

a) Compute the eigenvalues of the correlation matrix. How are the eigenvalues related to principal components analysis?

b) What proportion of the variance is accounted for by the first principal component?

c) What is the difference between principal components analysis and exploratory factor analysis?

Question 2. (16 points)

The Motivational State Questionnaire (MSQ) was developed to study emotions in laboratory and field settings. Respondents were asked to indicate their current standing on scale for a number of variables/emotions. A confirmatory factor analysis with two factors was hypothesized based on a subset of these variables and the estimated pattern loadings are given in the following table. The correlation between the two factors was estimated to be 0.181.

	F_1 Energetic	F_2 Negative
Active	0.857	
Vigorous	0.776	
Wakeful	0.687	
Lively	0.890	
Jittery		0.538
Nervous		0.881
Scared		0.590

a) Draw a graph of the model.

b) What are the degrees of freedom for the model?

c) Fill out the missing numbers (marked by x) in the following estimated correlation matrix.

$$\hat{\mathbf{R}} = \begin{array}{l} \\ \text{Active} \\ \text{Vigorous} \\ \text{Wakeful} \\ \text{Lively} \\ \text{Jittery} \\ \text{Nervous} \\ \text{Scared} \end{array} \begin{array}{c} \begin{array}{ccccccc} \text{Active} & \text{Vigorous} & \text{Wakeful} & \text{Lively} & \text{Jittery} & \text{Nervous} & \text{Scared} \end{array} \\ \begin{pmatrix} x & & & & & & \\ 0.665 & x & & & & & \\ 0.589 & x & x & & & & \\ 0.763 & 0.691 & 0.612 & x & & & \\ 0.084 & 0.076 & 0.067 & 0.087 & x & & \\ 0.137 & 0.124 & 0.110 & x & 0.475 & x & \\ 0.092 & 0.083 & 0.074 & 0.095 & 0.318 & 0.520 & x \end{pmatrix} \end{array}$$

d) Describe one measure that can be used to evaluate the goodness of the fit of the model.

Question 3. (16 points)

Six retail chain stores were evaluated by a consumer panel on 10 service quality attributes. Based on their evaluations, the following similarity matrix containing squared Euclidean distances was computed.

Store #	1	2	3	4	5	6
1	0.00					
2	3.65	0.00				
3	46.81	13.29	0.00			
4	24.62	51.88	17.21	0.00		
5	39. 87	26.75	8.79	16.27	0.00	
6	82.48	40.05	18.91	6.20	65.22	0.00

a) Use the Single-Linkage method to perform a hierarchical clustering of the stores.

b) Draw the dendogram for the hierarchical clustering in a).

c) Describe the main steps of the k-means clustering algorithm.

Question 4. (16 points)

The salmon fishery is a valuable resource for both the United States and Canada. There are regulations such that Alaskan fishermen cannot catch too many Canadian salmon and vice versa. As an aid in regulating the catches samples of fish are taken during the harvest and the salmons are classified as coming from Alaskan or Canadian waters. This classification is based on the variables x_1=diameter of rings for the first-year freshwater growth (hundredths of an inch) and x_2=diameter of rings for the first-year marine growth (hundredths of an inch). Typically, the rings associated with freshwater/marine growth are smaller/larger for the Alaskan-born than for the Canadian-born salmon. For a data set with 50 Alaskan-born and 50 Canadian-born salmon we have the sample means

$$\bar{\mathbf{x}}_1 = \begin{pmatrix} 98.38 \\ 429.66 \end{pmatrix}, \quad \bar{\mathbf{x}}_2 = \begin{pmatrix} 137.46 \\ 366.62 \end{pmatrix}$$

and pooled sample covariance matrix

$$\mathbf{S}_{\text{pooled}} = \begin{pmatrix} 293.35 & -27.29 \\ -27.29 & 1146.20 \end{pmatrix}.$$

a) Calculate Fisher's linear discriminant function for this data set.

b) The linear discriminant function weights are derived by maximizing a certain function, which one?

c) Classification was performed with the derived linear discriminant function and resulted in the confusion matrix

		Predicted class	
		Alaskan	Canadian
True class	Alaskan	44	6
	Canadian	1	49

Calculate the false positive and the false negative rates of this classification, assuming "Alaskan" is the "positive" event.

Question 5. (16 points)

Consider again the salmon data described in Question 4. The probability that a salmon is Alaskan-born is now modeled via logistic regression with x_1=diameter of rings for the first-year freshwater growth (hundredths of an inch) and x_2=diameter of rings for the first-year marine growth (hundredths of an inch) as explanatory variables. The model fit is summarized in Table 1.

a) Why is logistic regression preferred over linear regression when the response variable is binary?

b) What is the estimated odds that a salmon with $x_1 = 118$ and $x_2 = 381$ is Alaskan-born?

c) What would be the classification of a salmon with $x_1 = 118$ and $x_2 = 381$?

d) Write the expression of the log-likelihood for the model.

(Intercept)	3.924
	(6.315)
x_1	0.126^{***}
	(0.036)
x_2	-0.048^{***}
	(0.015)
AIC	44.788
BIC	52.604
Log Likelihood	-19.394
Deviance	38.788
Num. obs.	100

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 1: Model fit summary

Formula Sheet for the Exam in Multivariate Methods

Vectors and matrices

- Length of a vector $\mathbf{a} = (a_1, a_2, \ldots, a_p)$

$$\|\mathbf{a}\| = \sqrt{a_1^2 + a_2^2 + \cdots + a_p^2}$$

- Determinant of a 2×2 matrix $\mathbf{A}$

$$\det(\mathbf{A}) = |\mathbf{A}| = a_{11}a_{22} - a_{12}a_{21}$$

- Inverse of a 2×2 matrix $\mathbf{A}$

$$\mathbf{A}^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}$$

- Eigenvalues are the roots of the characteristic equation

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \mathbf{0}$$

  For each eigenvalue the solution to

$$(\mathbf{A} - \lambda\mathbf{I})\,\mathbf{x} = \mathbf{0}$$

  gives the associated eigenvector $\mathbf{x}$

Distances

- Euclidean

$$D_{ik} = \sqrt{\sum_{j=1}^{p} (x_{ij} - x_{kj})^2}$$

- Statistical

$$SD_{ik} = \sqrt{\sum_{j=1}^{p} \left(\frac{x_{ij} - x_{kj}}{s_j}\right)^2}$$

- Mahalanobis

$$MD_{ik} = \sqrt{(\mathbf{x}_i - \mathbf{x}_k)^T \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_k)}$$

For $p = 2$

$$MD_{ik} = \sqrt{\frac{1}{1-r^2}\left[\frac{(x_{i1} - x_{k1})^2}{s_1^2} + \frac{(x_{i2} - x_{k2})^2}{s_2^2} - \frac{2r\,(x_{i1} - x_{k1})\,(x_{i2} - x_{k2})}{s_1 s_2}\right]}$$

Mean-correction and covariance

- Mean-corrected data

$$\underset{(n\times p)}{\mathbf{X}_m} = \{x_{ij}\} = \{X_{ij} - \bar{X}_j\}$$

- Covariance

$$\underset{(p\times p)}{\mathbf{S}} = \{s_{ij}\} = \left\{\frac{\sum_{i=1}^{n} x_{ij}x_{ik}}{n-1}\right\} = \frac{\mathbf{SSCP}}{df} = \frac{1}{n-1}\mathbf{X}_m^T\mathbf{X}_m$$

Group Analysis

- Total sum of squares and cross products

$$\mathbf{SSCP}_{\text{total}} = \mathbf{SSCP}_{\text{within}} + \mathbf{SSCP}_{\text{between}}$$

- Pooled within-group sum of squares and cross products

$$\mathbf{SSCP}_{\text{within}} = \sum_{\ell=1}^{g} \mathbf{SSCP}_{\ell}$$

- Pooled covariance matrix

$$\mathbf{S}_{\text{pooled}} = \frac{\mathbf{SSCP}_{\text{within}}}{n-g}$$

- Between-group sum of squares and cross products

$$\mathbf{SSCP}_{\text{between}} = \mathbf{SSCP}_{\text{total}} - \mathbf{SSCP}_{\text{within}}$$

For $g = 2$ groups

$$\mathbf{SSCP}_{\text{between}} = \frac{n_1 n_2}{n_1 + n_2} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T$$

Factor Analysis

- For the two-factor model

$$Var(x) = \lambda_1^2 + \lambda_2^2 + Var(\epsilon) + 2\lambda_1\lambda_2\phi$$

$$Cor(x, \xi_1) = \lambda_1 + \lambda_2\phi$$

$$Cor(x_j, x_k) = \lambda_{j1}\lambda_{k1} + \lambda_{j2}\lambda_{k2} + (\lambda_{j1}\lambda_{k2} + \lambda_{j2}\lambda_{k1})\,\phi$$

- RMSR for EFA

$$RMSR = \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=i+1}^{p} res_{ij}^2}{p(p-1)/2}}$$

- RMSR for CFA

$$RMSR = \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=i}^{p} (s_{ij} - \hat{\sigma}_{ij})^2}{p(p+1)/2}}$$

Two-Group Discriminant Analysis

- Maximize

$$\lambda = \frac{\gamma^T \mathbf{B} \gamma}{\gamma^T \mathbf{W} \gamma}$$

- Fisher's linear discriminant function

$$\gamma^T = (\mu_1 - \mu_2)^T \mathbf{\Sigma}^{-1}$$

- Wilks' Λ

$$\Lambda = \frac{|\mathbf{SSCP}_w|}{|\mathbf{SSCP}_t|}$$

$$F = \left(\frac{1-\Lambda}{\Lambda}\right)\left(\frac{n_1 + n_2 - p - 1}{p}\right) \sim F\left(p, n_1 + n_2 - p - 1\right)$$

- Classification based on decision theory: assign the observation to group 1 if

$$Z \geq \frac{\bar{Z}_1 + \bar{Z}_2}{2} + \ln\left[\frac{p_2 C(1|2)}{p_1 C(2|1)}\right]$$

Logistic regression

- Odds of the event $Y = 1$

$$odds = \frac{p}{1-p}$$

where

$$p = P(Y = 1)$$

- Probability of the event $Y = 1$ as a function of the explanatory variables

$$P(Y = 1|X_1, X_2, \ldots, X_k) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\cdots+\beta_k X_k)}}$$

Quadratic equation

- The roots of the quadratic equation $ax^2 + bx + c$

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Department of Statistics

Stockholms universitet

Correction sheet

Date: 27/03/2019

Room: Brunnsvikssalen

Exam: Multivariate Methods

Course: Multivariate Methods

Anonymous code: 0605 -HXF

☑ 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 (18)
Teacher's notes 15	13	14	15	8					

65 + 15

Points	Grade	Teacher's sign.
80	B	

SU, DEPARTMENT OF STATISTICS

Room: Brunnvik Anonymous code: 0005-HXF Sheet number: 1

27/03/19

Tentamen Multivariata Metoder

1.a) Eigenvalues are the roots of

$$\det(A-\lambda I)=0$$

Where A = Correlation matrix I = identity matrix

$$A-\lambda I = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix} - \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}$$

$$= \begin{pmatrix} 1-\lambda & 0.6 \\ 0.6 & 1-\lambda \end{pmatrix}$$

R

$$\det(A-\lambda I) = (1-\lambda)(1-\lambda) - 0.6\cdot 0.6$$
$$= (1-\lambda)^2 - 0.6^2$$
$$= (1-2\lambda+\lambda^2) - 0.36$$
$$= \lambda^2 - 2\lambda + 1 - 0.36$$
$$= \lambda^2 - 2\lambda + 0.64$$

R

$$\det(A-\lambda I)=0 \rightarrow \lambda^2 - 2\lambda + 0.64 = 0$$

$$\lambda = \frac{-(-2)}{2} \pm \sqrt{\frac{-2^2}{2} - 2\cdot 0.64}$$ ✓

$$\lambda = 1 \pm \sqrt{2 - 1.28}$$

$$\lambda = 1 \pm 0.8485$$

Svar

$$\lambda_1 = 1 + 0.8485 = 1.8485$$

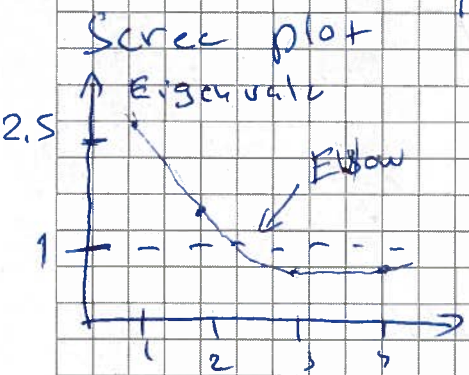
$$\lambda_2 = 1 - 0.8485 = 0.15145$$

1.A cont. →

1.a Cont)

$= Var(PC)$

Eigen values represent the proportion of variance that a principal component "captures". It is related to the total variance in the matrix. The ~~gott~~ goal of PCA is data - reduction. therefore, one can choose a (preferably low) number of principal componants to interpret the dataset. Usally, one plots the Eigenvalues and tries to find an "elbow" indicating decreasing importance of remaining Principal components. Otherwise, the "Eigenvalue over 1"-rule also works



→ Also, with the help of Eigen values, one can calculate the Eigen vector, which represent the weight of a certain variable in the new axis (Principal component analysis tries to find an axis with maximum variance). Once the new weights ant PC-scores are calculated, ~~int~~ interpretability is ensured by calculating loadings, which relate a variables "importance" in the Principal Component

$$l_{ij} = \frac{\sqrt{\lambda} \cdot w}{SE}$$

1 in the case of correlation matrix

7

SU, DEPARTMENT OF STATISTICS

Room: Brunnsvik Anonymous code: 0005-HXF Sheet number: 2

Q 1 b)

Because we are dealing with standardized data (correlation matrix) the variance of each variable is 1. The total variance in the data is the sum of its Eigenvalues, which therefore is $1 + 1 = 2$.

The proportion of variance is therefore

$$\frac{\text{Eigenvalue}}{\text{Total variance}} = \quad PC1 = \frac{1,8485}{2} \approx 0,92$$

$$PC2 = \frac{0,1515}{2} \approx 0,08$$

4

c) Principal components analysis is a data-reduction technique whilst EFA is a model.
PCA tries to reduce a number of variables to "it's essence" by calculating a new axis (and consequently new variable-scores) which maximizes variance. PCA can also be useful in calculating an index to capture an overall effect (like calculating inflation from the price-change of a basket of goods)

→

continued

c) EFA on the other hand tries to measure the importance of latent factors in a dataset and give a quantitative value to it's effect. For example deriving a students quantitative and literary ability from test scores in multiple exams.

PCA can be used to decide the number of latent factors (again with eigenvalues) but EFA tries then to model the loadings (pattern loadings) into a coherent structure.

4

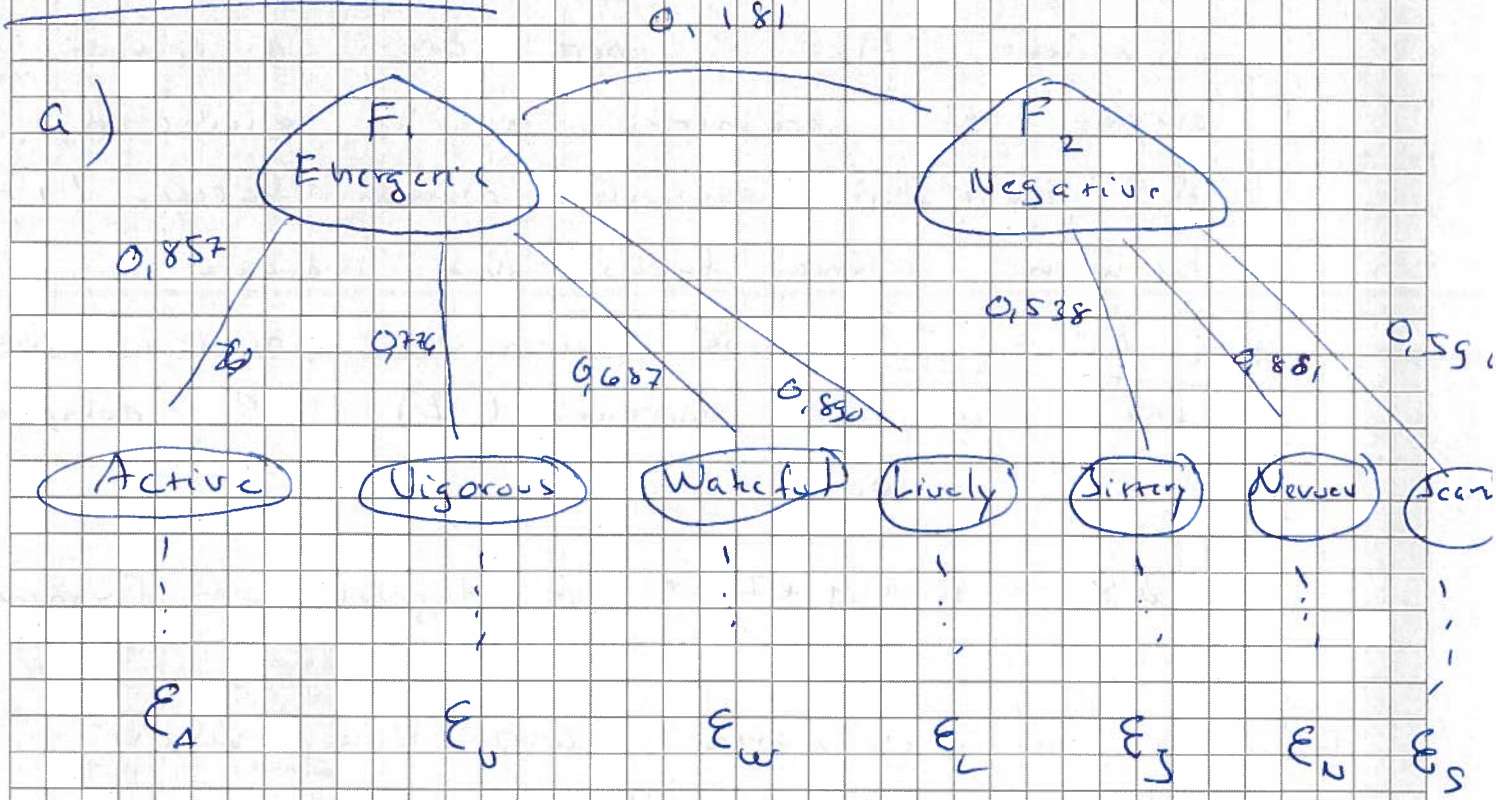
15

SU, DEPARTMENT OF STATISTICS

Room: Brunnsviks **Anonymous code:** 0005-HXF **Sheet number:** 3

Question 2

a)



Because of the structure where every indicator loads highly on one singular common factor, one can suspect a model with varimax rotation. The model yields

No rotations in CFA

$$\text{Active} = 0{,}857 F_1 + \varepsilon_A$$

$$\text{Vigorous} = 0{,}776 F_1 + \varepsilon_V$$

$$\text{Wakeful} = 0{,}687 F_1 + \varepsilon_W$$

$$\vdots$$

$$\text{Scared} = 0{,}590 F_2 + \varepsilon_S$$

2

b) In the depicted model, there are

$$\frac{p(p+1)}{2} = \frac{7 \cdot 8}{2} = 28$$ unique

elements. Also there are a certain amount of parameters to be estimated, the relanonship between common factors (1), between common fators and indicators ✓ (7 · 2 = 14) and between indicators and their unique variance (7). By comparing these we get

$$28 - (1 + 14 + 7) = 6$$ degrees of freedom ✓

2

c) In the estimated correlation matrix, the diagonal will show correlation between a variable and itself, therefore yielding 1. The remaining X's R

(Corr(Vigorous, Wakeful) and Corr(Nervous, lively) will need to take into account the oblique structure of the model, ~~using~~ that means considering the correlation between factors. Following formula will be used

$$\mathrm{Corr}(X_j, X_k) = \lambda_{j1}\lambda_{k1} + \lambda_{j2}\lambda_{k2} + (\lambda_{j1}\lambda_{k2} + \lambda_{j2}\lambda_{k1})\,\phi$$

SU, DEPARTMENT OF STATISTICS

Room: Brunnsviks Anonymous code: 0005-HXK Sheet number: 4

Question 2 c) continued

Corr (Vigorous, Wakeful) =

$= 0{,}776 \cdot 0{,}687 + 0 \cdot 0 + (0{,}776 \cdot 0 + 0{,}687 \cdot 0)\, 0{,}181$

$= 0{,}533$ R

Corr (Lively, Nervous) =

$= 0{,}890 \cdot 0 + 0 \cdot 0{,}881 + (0{,}890 \cdot 0{,}881 + 0 \cdot 0)\, 0{,}181$

$= 0{,}14192$ R

A proof of why $Corr(X_1, X_1) = 1$ is also shown

$$Corr(X_1, X_1) = \frac{Cov(X_1, X_1)}{\sigma_{X_1} \cdot \sigma_{X}} = \frac{Var(X_1)}{\sigma^2_{X_1}} = 1$$

6

d) One measure that can be used is Root mean square residual

$$RMSR = \sqrt{\frac{\sum\sum (s_{ij} - \hat{\sigma}_{ij})^2}{\frac{p(p+1)}{2}}}$$

which compares the standard deviances estimated by the model structure against the actual calculated variances / std in the data set.

→

~~Quest~~

One can also perform a χ^2-test against the model (with d.f as ~~pt~~ calculated before) to evaluate the fit

3

13

SU, DEPARTMENT OF STATISTICS

Room: Brunnuitng **Anonymous code:** 0005-HXF **Sheet number:** 5

Question 3

a) Single linkage means representing distance between clusters as the smallest distance between every member of the clusters

In the original matrix, (1 2) have the shortest distance and will become the first cluster

R

	(1 2)	3	4	5	6
(1 2)	0				
3	13,29	0			
4	24.62	17,21	0		
5	26,75	8,79	16,27	0	
6	40,05	18,91	6,2		0

5y4

R

Dist ((1,2) and 3) = min (d_{13}, d_{23}) = min (46,81, 13,29) = 13,29

Dist ((1,2) and 4) = min (d_{14}, d_{24}) = min (24.62, 51,88) = 24.62

Dist ((1,2) and 5) = min (d_{15}, d_{25}) = min (39.87, 26,75) = 26,75

Dist ((1,2) and 6) = min (d_{16}, d_{26}) = min (82,48, 40,05)
= 40,05

In the new similarity matrix, 4 and 6 yield the shortest squared euclidean distance and therefore become the new cluster

R

	(1 2)	3	5	(4,6)
(1 2)	0			
3	13,29	0		
5	26,75	8,79	0	
(4,6)	24,62	17,21	16,27	0

R

Dist ((1,2) and (4,6)) = min (d_{14} d_{16} d_{24} d_{26})
= min (24,62 , 82,48 51,88 40,07)
= 24,62

Dist (3 and (4,6)) = min (d_{34} d_{36}) = min (17,21 , 18,91)
= 17,21

Dist (5 and (4,6)) = min ($d_{~~35~~54}$ d_{56}) = min (16,27 , 63,22)
= 16,27

In the next step, (3 5) will become merger to a cluster as they have the lowest distance

R

	(1 2)	(3 5)	(4,6)
(1 2)	0		
(3 5)	8,79	0	
(4,6)	24,62	17,21	0

R

Dist ((1,2) and (35)) = min (d_{13} d_{35} d_{23} d_{25}) = min (46,81 , 8,79, 13,29, 24,75)
= 8,79

SU, DEPARTMENT OF STATISTICS

Room: Brunnsviks Anonymous code: 0005-HXF Sheet number: 6

Question 3 continued

Dist ((35) and (4,6)) = min (d_{34} d_{36} d_{54} d_{56})

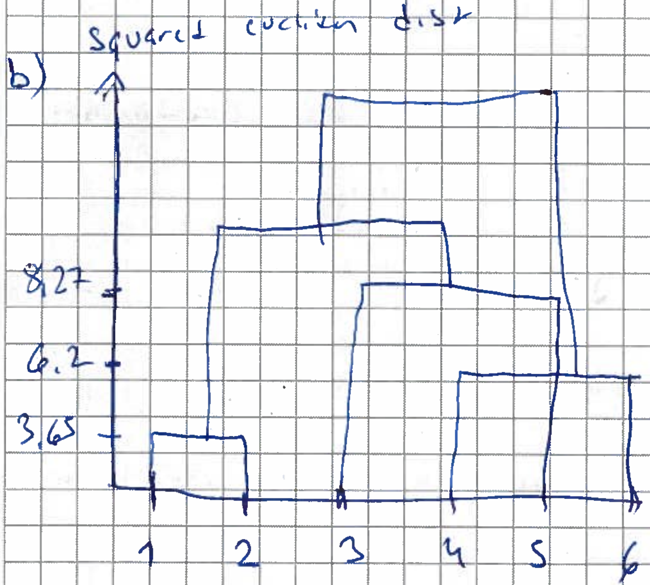
= min (17,21 18,91 16,27 63,12)

= 17,21 ✓ 16.27

Finally, we'd merge (12) and (35) together and then with (4,6) ending with one big cluster. But after step 3 all observations have their own cluster.

9

b)



} not so easy to read

2

c) ~~K-mean~~ K-means is a non-hierarchical clustering technique which assumes prior knowledge off the number of clusters. One starts off with ~~for~~ example 3 clusters and ~~dev~~ divides the data arbitrarily into 3 clusters. The centroid (mean) of every cluster is calculated as a next step. Then, every observation is compared to its clusters centroid and the 2 others. If it is closer (as euclidean distance) to another cluster, it changes cluster to the one with shortest distance. This procedure is then repeated (calculating centroid, ~~com~~ comparing distances, reclassifying) until no more reclassifications are made (which mean that each observation belongs to the cluster to which it has shortest euclidean distance).

1. Dividing observation in a ~~predifin~~ predefined number of clusters
2. Calculating centroids for every cluster
3. Calculating distance between observation and all centroids
4. Reclassify if distance proves to be shorter with another cluster until no more.

3

14

SU, DEPARTMENT OF STATISTICS

Room: Brunnsviken Anonymous code: 0005-HXR Sheet number: 7

Question 4

a) Fischer's linear function is calculated with

$$\gamma^T = (\mu_1 - \mu_2)^T \Sigma^{-1}$$

which in turn is estimated by

$$\hat{\gamma}^T = (\bar{x}_1 - \bar{x}_2)^T S^{-1}$$

R

S^{-1} is the inverse of the pooled sample covariance matrix

$$S^{-1} = \frac{1}{253{,}35 \cdot 1146{,}20 - (-27{,}29^2)} \begin{pmatrix} 1146{,}2 & -(-27{,}29) \\ -(-27{,}29) & 253{,}35 \end{pmatrix}$$

R

$$= 0{,}0000003 \cdot \begin{pmatrix} 1146{,}2 & 27{,}29 \\ 27{,}29 & 253{,}35 \end{pmatrix}$$

$$= \begin{pmatrix} 0{,}003416 & 0{,}000081 \\ 0{,}000081 & 0{,}000874 \end{pmatrix}$$

$$\gamma^T = (\bar{x}_1 - \bar{x}_2)^T \cdot S^{-1}$$

$$= \begin{pmatrix} 98{,}38 & -137{,}46 \\ 429{,}66 & -366{,}62 \end{pmatrix} \cdot S^{-1} = \begin{pmatrix} -39{,}08 \\ 63{,}04 \end{pmatrix}^T \cdot S^{-1}$$

$$= \underset{1 \times 2}{\begin{pmatrix} 63{,}04 & -39{,}08 \end{pmatrix}} \underset{2 \times 2}{\begin{pmatrix} 0{,}003416 & 0{,}000081 \\ 0{,}000081 & 0{,}000874 \end{pmatrix}}$$

R

= Kommer ge en 1x2 matris

a) $\begin{pmatrix} 63,04 \cdot 0,0034 + 0,000081 \cdot -39,08 \\ 63,04 \cdot 0,000081 + 0,000874 \cdot -39,08 \end{pmatrix}$

extra 0's

$= \begin{pmatrix} 0,2121 \\ -0,029 \end{pmatrix}$

$\gamma = 0,2121 \times X_1 + (-0,029) \cdot X_2$

7

b) It is derived by ~~maximi~~ maximizing

$$\lambda = \frac{\gamma^T B \gamma}{\gamma^T W \gamma}$$

Wher B represents the between group sum of squared (variance from the general mean of the data, a measure of difference) and W represents the within group sum of squared (variance from the groups internal mean, a measure of homogeneity). By ~~maximizing~~ maximizing the ratio, one tries to minimize to ~~the~~ variance in groups but maximize the difference between groups, therefor resulting in a ~~dist~~ function which discriminates well.

3

SU, DEPARTMENT OF STATISTICS

Room: Brunnsvikj Anonymous code: 0005-HXC Sheet number: 8

Question 4 c)

c) A confusion matrix shows events to non-events

Predicted

		Event	Non-Event
True	Event	True Positive	~~Error~~ False Negative
	Non-Event	False Positive	~~False~~ True Negative

if we assume "Alshan" to be the event, we calculate the FP and FN rate as

$$FP \text{ rate} = \frac{FP}{FP + TN} = \frac{1}{50} \quad R$$

$$FN \text{ rate} = \frac{FN}{TP + FN} = \frac{6}{50} \quad R$$

5

15

Question 5

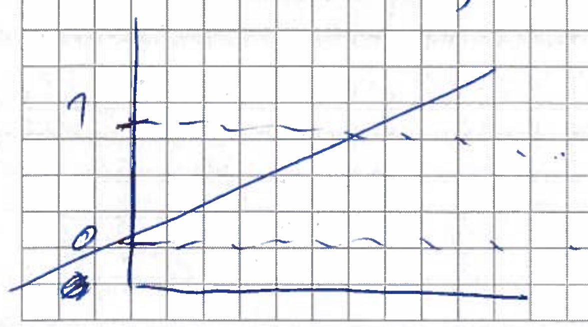
a) In a logistic regression, the variable of interest is probability of an event / classification / etc. Although linear regression has many nice features, creating a cap between [0, 1] for the dependent variable is not one of. Instead, we model the dependant variable as the log of the odds ~~odds~~, $\ln\left(\frac{p}{1-p}\right)$. By using using log odds, we ensure a value of [0, 1] and can easily "transform" the log odds into probability using

$$P(Y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\cdots\beta_i x_n)}}$$

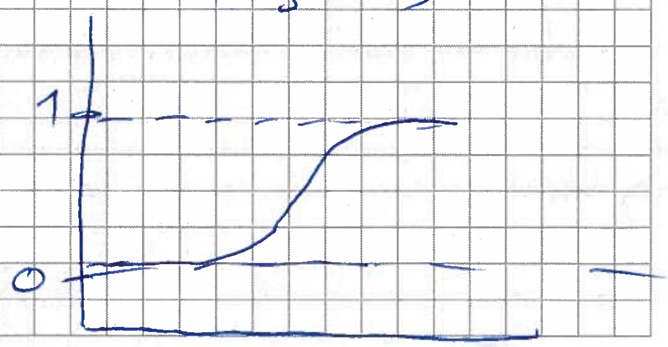
and subsequently still have the luxury of modelling the log odds as a linear function $\left(\ln\left(\frac{p}{1-p}\right) = \beta_0+\beta_1 x_1 \cdots \beta_n x_n\right)$

Graphically we can show this aswell.

Lin. Reg



Log. reg



3

SU, DEPARTMENT OF STATISTICS

Room: Brumullh Anonymous code: 0003-HXC Sheet number: 9

Question 5b)

our model is $\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x \dots$

$$\ln\left(\frac{p}{1-p}\right) = 3{,}924 + 0{,}126 x_1 - 0{,}048 x_2$$

for $x_1 = 118$ $x_2 = 381$ we get

$$\ln\left(\frac{p}{1-p}\right) = 3{,}924 + 0{,}126 \cdot 118 - 0{,}048 \cdot 381$$

$$= 0{,}504$$

$$\frac{p}{1-p} = e^{0{,}504} = 1{,}655$$

R

As Question 4 made clear the the event in this model is salmon being Alaskan (and non-event canadian) ~~that~~ it is 1,655 times more likely for a salmon with $x_1 = 118$ and $x_2 = 381$ to be Alaskan than canadian.

2

c) If we create a cut-off value at $p = 0{,}5$ for the classification, (meaning that to be classified as alaskan, the probability must be higher than 50%), a salmon with $x_1 = 118$ and $x_2 = 381$ would be classified a)

$$p(Y=1) = \frac{1}{1 + e^{-(3,324 + 118 \times 0,124 - 9048 \cdot 581)}}$$

$$= \frac{1}{1 + e^{-(0,504)}} = 0,6233$$ R

As the probability is higher than 0,5, it would be classified as R

Alaska

3

d) likelihood function is

$$L(B \mid X_1, X_2 \ldots X_n) = \prod \frac{1}{1 + e^{-(B_0 + B_1 X_1 + B_2 X_2)}}$$ ✓

$$L = \frac{1}{\prod (1 + e^{-(B_0 + B_1 X_1 + B_1 X_2)})}$$

$$\ln(L) = \ln(1) - \ln\left(\prod (1 + e^{-(B_0 + B_1 X_1 + B_1 X_2)})\right)$$

$$= \ln(1) - \Sigma \ln(1) + \Sigma \ln e^{-(B_0 + B_1 X_1 + B_2 X_2)}$$

$$= 0 - 0 + \Sigma -(B_0 + B_1 X_1 + B_2 X_2)$$

$$= -\Sigma B_0 + \Sigma B_1 X_1 - \Sigma B_2 X_2$$

$$= -\Sigma B_0 - B_1 \cdot \Sigma X_1 - B_2 \Sigma X_2$$

0

8