STOCKHOLM UNIVERSITY
Department of Statistics
Mahmood Ul Hassan

EXAM IN MULTIVARIATE METHODS
March 22 2021

Time: 6 hours

The exam is for individual solving. It is an open-book exam, but you are not allowed to use the help of other students, friends, family, or similar. In case you need clarification, the teacher is available at Zoom. Time for Zoom meetings is 14:00-14:30 and 16:00:16:30.

Join Zoom Meeting from
https://stockholmuniversity.zoom.us/j/9255236581
Meeting ID: 925 523 6581

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

Question. 1 (6+2+3+2+3=16 Points)

For a data set with observations on two variables x_1 and x_2 the sample covariance matrix was found to be

$$S = \begin{bmatrix} 10 & 5 \\ 5 & 4 \end{bmatrix}$$

a) Using S, construct two principal components that are orthogonal to each other.
b) What proportion of variance is accounted by these principal components?
c) Using the constructed principal components in part (a) show that var(PC1)= λ_1 and var(PC2)= λ_2 , where λ_1 and λ_2 are eigenvalues.
d) Using the constructed principal components in part (a) compute the cov(PC1, PC2)
e) Compute the correlation of PC1 and PC2 with x_1 and x_2 .

Question. 2 (2+2+4+4+4=16 Points)

Given the following data

Observation	X_1	X_2	X_3
1	7	8	6
2	3	1	4
3	9	8	2
4	2	6	8

(a) Find the Euclidean distance between observation 2 and 4.
(b) Compute the sum of squared and cross products matrix for the variables X_1 and X_3 ($SSCP_{13}$).
(c) Find the var-cov matrix from the data.
(d) Compute the statistical distance between observation 1 and 4.
(e) Compute the Mahalanobis distance between observation 2 and 3.

Question. 3 (10+6=16 Points)

(a) For the following data

Observation	Y_1	Y_2	Gender
1	2	6	Male
2	2	6	Female
3	4	2	Male
4	4	3	Female
5	6	10	Male
6	8	5	Female
7	6	4	Female
8	5	6	Male
9	10	8	Male

a) Compute the $\mathbf{SSCP_b}$, $\mathbf{SSCP_w}$ and $\mathbf{SSCP_t}$ matrices.

b) Suppose $n_1=7$ and $n_2=8$ are observations in group-1 and group-2, respectively and

$$\text{Within-group covariance matrix for group-I} = S_1 = \begin{bmatrix} 20 & -5 \\ -5 & 8 \end{bmatrix}$$

Within-group SSCP matrix for group-II= $SSCP_2 = \begin{bmatrix} 100 & -5 \\ -5 & 10 \end{bmatrix}$

$\bar{X}_1 = \begin{bmatrix} 10 \\ 12 \end{bmatrix}$ and $\bar{X}_2 = \begin{bmatrix} 15 \\ 8 \end{bmatrix}$

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

Question. 4 (4+4+4+4=16 Points)

Observations on two variables were made for five subjects according to the following table.

Subject	Variable-1	Variable-2
1	5	4
2	6	8
3	5	3
4	2	6
5	4	10

a) Construct a similarity matrix containing squared Euclidean distances

b) Use the similarity matrix in part (a) and perform a cluster analysis with the following method

1) Single-linkage method.
2) Complete-linkage method.
3) Average linkage method.

Question. 5 (4+3+3+3+3=16 Points)

We have data sets with following variables:

Age: Age of the patient
Acid: Level of serum acid phosphate
X-ray: Result of x-ray examination (0=negative, 1=positive)
Size: Tumour size (0=small, 1=large)
Grade: Tumour grade (0=less serious, 1=more serious)
Involvement: Nodal involvement (0=no, 1=yes)

The data analytic task is to explore whether the independent variables can be used to predict the probability of nodal involvement in prostatic cancer. The following logistic models are fitted

Model-1

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.1994     0.7162  -1.675  0.09400 .
Size          1.7638     0.7483   2.357  0.01842 *
Xray          2.0550     0.7976   2.576  0.00998 **
log(Acid)     2.2922     1.1387   2.013  0.04412 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 70.252  on 52  degrees of freedom
Residual deviance: 48.986  on 49  degrees of freedom
```

Model-2

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.3308     0.5819  -0.569  0.56964
Xray          2.1020     0.7226   2.909  0.00363 **
log(Acid)     2.0363     1.1128   1.830  0.06728 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 70.252  on 52  degrees of freedom
Residual deviance: 55.272  on 50  degrees of freedom
```

Model-3

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.1701     0.3816  -3.066  0.00217 **
Xray          2.1817     0.6975   3.128  0.00176 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 70.252  on 52  degrees of freedom
```

```
Residual deviance: 59.001  on 51  degrees of freedom
```

a) Formulate the null and alternative hypothesis and perform testing of hypothesis using deviance statistic to compare the fitted model-1 with the model-2. Use $\alpha = 0.05$.
b) Formulate the null and alternative hypothesis and perform a test if model-3 is significantly better than the null model using deviance statistic. Use $\alpha = 0.10$.
c) What is the probability of nodal involvement in prostatic cancer for small tumour size with negative result of x-ray examination and 0.56 level of serum acid phosphate.
d) What is the odd of nodal involvement in prostatic cancer for large size of tumour with positive result of x-ray examination and 0.48 level of serum acid phosphate.
e) Classify the observations given in the table and compute the sensitivity and specificity.

Observation	Involvement	P-hat	Observation	Involvement	P-hat
1	0	0.05	16	0	0.26
2	0	0.07	17	0	0.26
3	0	0.44	18	0	0.18
4	1	0.07	19	0	0.31
5	0	0.12	20	0	0.34
6	0	0.10	21	1	0.72
7	1	0.48	22	1	0.26
8	0	0.16	23	0	0.83
9	0	0.22	24	0	0.40
10	0	0.06	25	1	0.90
11	0	0.13	26	1	0.41
12	0	0.56	27	1	0.41
13	1	0.83	28	1	0.87
14	1	0.16	29	1	0.91
15	0	0.18	30	1	0.96