University of Liege

Faculty of Medicine

# Multivariate statistical analysis

Adelin Albert
Professor emeritus

English edition 2018

# Preface

Multivariate statistics covers all methods that deal with the analysis of observations made on several variables simultaneously. Therefore it refers intensively to vectors and more generally to matrix calculus. It is obvious that most problems and applications encountered in practice involve at least two variables and are therefore multidimensional. The most renowned example in the literature is the Fisher iris dataset which consists of the measurements of four variables, respectively sepal length, sepal width, petal length and petal width, made on 50 iris setosa, 50 iris versicolor and 50 iris virginica flowers. Amazingly, this dataset which is reproduced in Annex can be used to illustrate most of the multivariate statistical methods described in this book.

Multivariate statistics requires at least some basic knowledge of univariate statistics. This can be found in any elementary or advanced textbook. Multivariate statistical methods were mainly developed during the $20^{\text{th}}$ century but their intensive use has only been made possible by the advent of computers. Actually, as the number of variables to be studied gets larger, statistical calculations become longer, more complicated and eventually intractable manually. Today computers can solve complex multivariate statistical problems in a few seconds. Moreover, the advent of user-friendly statistical packages such as SAS, SPSS, Stata and R, has made multivariate methods available to a large research community of users. Nonetheless, despite the fact that statistical packages have become cheaper and easier to use, their appropriate utilization in practice requires to have some knowledge about what these multivariate methods can do and what they cannot do. This was the primary objective when preparing these lecture notes for students and researchers.

This book is structured in 8 chapters. Chapter 1 is a rapid overview of the basic notions of matrix calculus needed to present, characterize and solve multivariate statistical problems. In Chapter 2, we briefly introduce the concepts of population, sample, variable, type of variable (quantitative, qualitative and binary), and data. Then we define the $n \times p$ data matrix

which results from the observation on $n$ subjects or objects (rows) of $p$ variables (columns). The representation of multivariate observations by different techniques will also be addressed. In Chapter 3 we will extend the classical concepts of mean and variance to the multivariate setting. In particular, we introduce the notion of variance-covariance matrix and of correlation matrix. The Mahalanobis distance between two points (observations) of the $p$-dimensional space, which generalizes the classical Euclidean distance by accounting for the associations between variables, will also be defined.

Chapter 4 is concerned with the first important multivariate statistical method, namely principal component analysis (PCA). Indeed, this method can provide a two-dimensional graphical representation of data points of the multidimensional space with a minimal loss of information. In other words, it gives us a photography of the data matrix. In this chapter we also briefly mention the more recent biplot method which provides in a similar way a graphical representation of the variables studied.

Chapter 5 looks at the relation and correlation between a "dependent" quantitative variable and a set of so-called "independent" or "explanatory" variables. This is the method of multiple regression (MR) and correlation. In many statistics books, this method is seen as a univariate method in the sense that the explanatory variables are considered as factors of an experiment which are fixed by the investigator (factorial design) and only the dependent variable is observed. We decided to consider it as a multivariate statistical method because it requires matrix calculus but also because it becomes a multivariate problem when all the variables, dependent and explanatory, are observed simultaneously, as it is often the case.

In Chapter 6 we refer to one of the most utilized methods of multivariate statistics. Logistic regression (LR) studies the association between a binary dependent variable (occurrence or non-occurrence of an event) and a vector of variables. The notion of the odds ratio (OR) will be introduced. We shall also give a short description of ordinal logistic regression (OLR) which extends logistic regression to the case where the dependent variable is no longer binary but ordinal (ordered categories).

Chapter 7 is dedicated to survival analysis. Basically, the methods described here concern the study of the association between a lifetime or time-to-event variable and a set of covariates. This requires the definition of the survival curve and its Kaplan-Meier estimation, of the hazard function and of the hazard ratio (HR) The chapter will also focus on the most celebrated Cox "proportional hazards (PH)" regression model. This method which has a wide range of applications is one of the most cited in the literature.

Finally, Chapter 8 handles one of the oldest multivariate statistical problem, namely discriminant analysis (DA). We describe how two or several

populations can be discriminated (separated) based on a vector of variables. This can be done by the Fisher linear discriminant function or its multiple group extension. We also show how "canonical discriminant analysis" allows to display the populations on a 2-dimensional plane, as in principal component analysis. Interestingly, discriminant analysis can also be seen as the method of classifying a multivariate observation (subject or object) of unknown origin among two or several populations with a minimal risk of error. This approach will be briefly described.

The Annexes reproduce three datasets which will be used to illustrate the multivariate statistical methods described in the text : (1) Fisher iris dataset, (2) Severe head injury patients dataset, and (3) Rectal cancer patients dataset. They also contain four of the most frequently used statistical tables, namely the Gaussian (Normal), Chi-squared, Student $t$ and Snedecor $F$ distributions .

The author expresses his deepest gratitude to his former teaching assistants, Anne-Françoise Donneau, Laurence Seidel and Sophie Vanbelle. He thanks Danièle Bartholoméus and Anna Marchetta for their expert secretarial assistance.

Liège, French version, March 2006
English version, October 2018
Adelin Albert

# Chapter 1

# Basics of matrix calculus

## 1.1 Matrices

### 1.1.1 Definition

A *matrix* $\underset{\sim}{A}$ of dimension $r \times c$ is a rectangular table with $r$ rows and $c$ columns consisting of numbers. Let $a_{ij}$ denote the matrix element $\underset{\sim}{A}$ at the intersection of the i$^{\text{th}}$ row and the j$^{\text{th}}$ column. Then the full matrix writes

$$\underset{\sim}{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1c} \\ a_{21} & a_{22} & \dots & a_{2c} \\ \dots & \dots & \dots & \dots \\ a_{r1} & a_{r2} & \dots & a_{rc} \end{pmatrix}.$$

When $r = c$, the matrix is *square*. Moreover if $a_{ij} = a_{ji}$ for all $i \neq j$, the matrix is said to be *symmetric*. The trace of a square matrix is the sum of its diagonal elements, $tr\underset{\sim}{A} = a_{11} + \ldots + a_{rr}$.

### 1.1.2 Vectors and scalars

- When $c = 1$ (single column), the matrix reduces to a *column vector* and the second subscript is no longer needed. The vector merely writes,

$$\underset{\sim}{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_r \end{pmatrix}$$

  and has dimension $r$.

- When $r = 1$ (one single line), the matrix reduces to a *row vector* of dimension $c$ denoted

$$\underset{\sim}{a}^T = (a_1, a_2, \ldots, a_c).$$

Since by definition, a vector is always a column vector, the particular notation of the vector above indicates that the vector is *transposed*.

Thus, an $r \times c$ dimensional matrix $\underset{\sim}{A}$ is composed of $c$ column vectors of dimension $r$ or of $r$ row vectors of dimension $c$

$$\underset{\sim}{A}_{r\times c} = (\underset{\sim}{a}_1, \ldots, \underset{\sim}{a}_c) = \begin{pmatrix} \underset{\sim}{a}_1^T \\ \vdots \\ \underset{\sim}{a}_r^T \end{pmatrix}.$$

- When $r = c = 1$, the matrix reduces to a *scalar* $a_{11}$. Thus, a scalar is $1 \times 1$ dimensional matrix.

### 1.1.3 Examples

The $8 \times 3$ matrix below displays the observations made on 8 subjects of 3 variables, height (cm), weight (kg) and arm circumference (cm)

$$\begin{pmatrix} 167.9 & 71.8 & 30.0 \\ 183.8 & 75.1 & 29.4 \\ 172.9 & 58.0 & 26.0 \\ 175.5 & 58.4 & 25.7 \\ 176.4 & 67.7 & 27.9 \\ 168.5 & 75.2 & 31.7 \\ 178.0 & 67.3 & 27.4 \\ 178.0 & 71.3 & 29.0 \end{pmatrix}.$$

The $3 \times 3$ square symmetric matrix represents the table of correlations between the 3 variables

$$\begin{pmatrix} 1.0 & 0.049 & -0.30 \\ 0.049 & 1.0 & 0.93 \\ -0.30 & 0.93 & 1.0 \end{pmatrix}.$$

The vector (167.9, 71.8, 30.0) gives the observations of the three variables for subject #1. It is a $1 \times 3$ matrix.

Lastly, the following $8 \times 1$ matrix

$$\begin{pmatrix} 71.8 \\ 75.1 \\ 58.0 \\ 58.4 \\ 67.7 \\ 75.2 \\ 67.3 \\ 71.3 \end{pmatrix}$$

represents the vector of observations of variable #2 (weight) for the 8 subjects.

### 1.1.4 Special matrices

- The *transpose* of matrix $A$ of dimension $r \times c$ is obtained by transposing the rows and columns of $\underset{\sim}{A}$. Specifically, it is the $c \times r$ dimensional matrix denoted $\underset{\sim}{A}^T$. Thus, the transpose of the $2 \times 3$ matrix

$$\underset{\sim}{A} = \begin{pmatrix} 167.9 & 71.8 & 30.0 \\ 183.8 & 75.1 & 29.4 \end{pmatrix}$$

is the $3 \times 2$ matrix

$$\underset{\sim}{A}^T = \begin{pmatrix} 167.9 & 183.8 \\ 71.8 & 75.1 \\ 30.0 & 29.4 \end{pmatrix}.$$

It is easily seen that transposing a symmetric matrix leaves the matrix unchanged. In other terms,

$$\underset{\sim}{A}^T = \underset{\sim}{A}$$

- A *diagonal matrix* is a symmetric matrix whose elements are all zeroes except those on the diagonal,

$$\underset{\sim}{D} = \begin{pmatrix} d_{11} & 0 & \dots & 0 \\ 0 & d_{22} & \dots & 0 \\ 0 & 0 & \dots & d_{rr} \end{pmatrix}$$

or simply

$$\text{diag}(d_1, d_2, \dots, d_r).$$

- The *identity* matrix of dimension $r \times r$ is the diagonal matrix whose diagonal elements are all equal to 1. It is denoted by $\underset{\sim}{I}_r$ or $\underset{\sim}{I}$. This matrix generalizes the unit number "1".

- The *null* matrix is the matrix whose elements are all zeroes, whether rectangular or square. It writes $\underset{\sim}{0}$.

## 1.2 Operations on matrices

### 1.2.1 Addition

The sum of two matrices $\underset{\sim}{A}$ and $\underset{\sim}{B}$ of equal dimension $r \times c$ is a matrix $\underset{\sim}{C}$ of dimension $r \times c$

$$\underset{\sim}{A} + \underset{\sim}{B} = \underset{\sim}{C}$$

where $c_{ij} = a_{ij} + b_{ij}$.

### 1.2.2 Subtraction

The subtraction of two matrices $\underset{\sim}{A}$ and $\underset{\sim}{B}$ of equal dimension $r \times c$ is a matrix $\underset{\sim}{C}$ of dimension $r \times c$

$$\underset{\sim}{A} - \underset{\sim}{B} = \underset{\sim}{C}$$

where $c_{ij} = a_{ij} - b_{ij}$.

### 1.2.3 Multiplication by a scalar

The product of an $r \times c$ matrix $\underset{\sim}{A}$ by a scalar $k$ is the matrix

$$\underset{\sim}{C} = k\underset{\sim}{A}$$

where $c_{ij} = ka_{ij}$. Thus all elements of the matrix $\underset{\sim}{A}$ are multiplied by the scalar.

### 1.2.4 Product of two matrices

The product of $r \times c$ matrix $\underset{\sim}{A}$ by $r' \times c'$ matrix $\underset{\sim}{B}$ is only possible if the number of columns of the first matrix is equal to the number of rows of the

second ($c = r'$). This yields a matrix $\underset{\sim}{C}$ whose number of rows is that of matrix $\underset{\sim}{A}$ and number of columns that of matrix $\underset{\sim}{B}$. Specifically,

$$\underset{\sim}{A}_{r\times c} \cdot \underset{\sim}{B}_{c\times c'} = \underset{\sim}{C}_{r\times c'}$$

where $c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \ldots + a_{ic}b_{cj} = \sum_{k=1}^{c} a_{ik} \cdot b_{kj}$.

Thus the product of two matrices $\underset{\sim}{A}$ and $\underset{\sim}{B}$, denoted by $A.B, A \times B$ or $AB$, is done "rows of $\underset{\sim}{A}$ by columns of $\underset{\sim}{B}$"; specifically $c_{ij} = \underset{\sim}{a}_i^T \cdot \underset{\sim}{b}_j$.

As an illustration, consider the two matrices

$$\underset{\sim}{A}_{2\times 3} = \begin{pmatrix} 1 & 0 & 2 \\ 3 & 5 & 1 \end{pmatrix} \quad \text{and} \quad \underset{\sim}{B}_{3\times 2} = \begin{pmatrix} 1 & 1 \\ 0 & 0 \\ 2 & 1 \end{pmatrix}.$$

The product $\underset{\sim}{A} \times \underset{\sim}{B}$ is possible and yields a $2 \times 2$ dimensional matrix $\underset{\sim}{C}$. As an example, element $c_{11}$ is obtained by multiplying the first row of $\underset{\sim}{A}$ by the first column of $\underset{\sim}{B}$, i.e. $c_{11} = 1 \times 1 + 0 \times 0 + 2 \times 2 = 5$. Similarly, $c_{12} = 1 \times 1 + 0 \times 0 + 2 \times 1 = 3$, $c_{21} = 3 \times 1 + 5 \times 0 + 1 \times 2 = 5$ and $c_{22} = 3 \times 1 + 5 \times 0 + 1 \times 1 = 4$.

Thus,

$$\underset{\sim}{C} = \underset{\sim}{A} \times \underset{\sim}{B} = \begin{pmatrix} 5 & 3 \\ 5 & 4 \end{pmatrix}.$$

Remarks

- The product of two square matrices $\underset{\sim}{A}$ and $\underset{\sim}{B}$ of equal dimensions is always possible. In general, however, this product is not commutative: $\underset{\sim}{A} \cdot \underset{\sim}{B} \neq \underset{\sim}{B} \cdot \underset{\sim}{A}$.

- Let $\underset{\sim}{a}$ a vector of length $r$. Then, the product

$$\underset{\sim}{a}^T \cdot \underset{\sim}{a} = a_1^2 + a_2^2 + \ldots + a_r^2$$

is called the *scalar product* because it yields a scalar (also termed the *norm* of $\underset{\sim}{a}$ or $\|\underset{\sim}{a}\|$), whereas the *matrix* product

$$\underset{\sim}{a} \cdot \underset{\sim}{a}^T = \begin{pmatrix} a_1^2 & a_1a_2 & \ldots & a_1a_r \\ a_2a_1 & a_2^2 & \ldots & a_2a_r \\ \ldots & \ldots & \ldots & \ldots \\ a_ra_1 & a_ra_2 & \ldots & a_r^2 \end{pmatrix}$$

leads to a symmetric matrix of dimension $r \times r$.

- When the scalar product of two vectors $\underset{\sim}{a}$ and $\underset{\sim}{b}$ is equal to zeo, $\underset{\sim}{a}^T \underset{\sim}{b} = 0$, the vectors are *orthogonal*.

- For square matrices $\underset{\sim}{A}$, it is easily seen that

$$\underset{\sim}{A} \cdot \underset{\sim}{I} = \underset{\sim}{I} \cdot \underset{\sim}{A} = \underset{\sim}{A}$$

- Finally note that for any $r \times c$ dimensional matrix $\underset{\sim}{A}$, the product $\underset{\sim}{A} \cdot \underset{\sim}{A}^T$ is a symmetric matrix of dimension $r \times r$ and the product $\underset{\sim}{A}^T \cdot \underset{\sim}{A}$ a symmetric matrix of dimension $c \times c$.

### 1.2.5 Generalization

Matrices of same dimensions $r \times c$ can be added and subtracted at will or linear combinations of such matrices can be computed such as

$$k_1 \underset{\sim}{A}_1 + k_2 \underset{\sim}{A}_2 + \ldots + k_n \underset{\sim}{A}_n$$

where $k_i$ are scalars.

The product of several matrices is feasible as long as adjacent matrices in the product fullfil the condition that the number of columns of each equals the number of rows of the next one. Thus, the triple product

$$\underset{\sim}{A}_{5\times 3} \cdot \underset{\sim}{B}_{3\times 8} \cdot \underset{\sim}{C}_{8\times 7}$$

is possible and gives a $5 \times 7$. matrix.

## 1.3 Matrix inversion

### 1.3.1 Definition

To some extent, matrix inversion allows the "division" of matrices. Remember that the inverse of a real number $k \neq 0$ writes $\frac{1}{k}$ or $k^{-1}$ so that $k \cdot k^{-1} = 1$. Thus, dividing two numbers, $k_1$ and $k_2$, actually corresponds to multiplying one by the inverse of the other, i.e. $k_1/k_2 = k_1 \times k_2^{-1}$. The division is a special case of multiplication.

Matrix inversion typically applies to sqaure matrices. Thus, the inverse of the $r \times r$ square matrix $\underset{\sim}{A}$ is the matrix of same dimension denoted by $\underset{\sim}{A}^{-1}$, such that

$$\underset{\sim}{A} \cdot \underset{\sim}{A}^{-1} = \underset{\sim}{A}^{-1} \cdot \underset{\sim}{A} = \underset{\sim}{I}$$

Inverting a matrix is by no means an easy task, in particular for high dimensional matrices. Matrix inversion requires computing its determinant.

### 1.3.2 Determinant

The *determinant* of the square matrix $\underset{\sim}{A}$ of dimension $r \times r$ is the scalar noted $dtm\underset{\sim}{A}$ or $|\underset{\sim}{A}|$ and, for whichever $1 \leq i \leq r$,, given by the expression

$$|\underset{\sim}{A}| = a_{i1}A_{i1} + a_{i2}A_{i2} + \ldots + a_{ir}A_{ir}$$

where $A_{ij}$ is the *cofactor* associated with element $a_{ij}$, specifically the quantity

$$A_{ij} = (-1)^{i+j}\mathcal{M}_{ij}$$

where $\mathcal{M}_{ij}$ is the *minor* associated with $a_{ij}$, that is the determinant of the matrix obtained by deleting in matrix $\underset{\sim}{A}$ the i$^{\text{th}}$ row and the j$^{\text{th}}$ column.

By convention, the determinant of a scalar $a$ is the scalar itself: $|a| = a$.

As an example, the determinant of matrix $\underset{\sim}{A}$ of dimension $2 \times 2$ is (taking $i = 1$)

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}$$

Likewise, it is easily seen that (taking $i = 1$)

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12}\begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13}\begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}$$

$$= a_{11}a_{22}a_{33} + a_{21}a_{32}a_{13} + a_{31}a_{12}a_{23} - a_{31}a_{22}a_{13} - a_{33}a_{21}a_{12} - a_{32}a_{23}a_{11}.$$

### 1.3.3 Inverse calculation

The *inverse* of matrix $\underset{\sim}{A}$ of dimension $r \times r$ is given by the expression

$$\underset{\sim}{A}^{-1} = \frac{1}{|\underset{\sim}{A}|}\begin{pmatrix} A_{11} & A_{12} & \ldots & A_{1r} \\ A_{21} & A_{22} & \ldots & A_{2r} \\ \ldots & \ldots & \ldots & \ldots \\ A_{r1} & A_{r2} & \ldots & A_{rr} \end{pmatrix}^T$$

where the matrix on the right-hand side is the matrix of cofactors which has been transposed. The element $(i, j)$ of matrix $\underset{\sim}{A}^{-1}$ is sometimes written $a^{ij} = A_{ji}/|\underset{\sim}{A}|$.

It is immediately seen that if $|\underset{\sim}{A}| = 0$, the inverse $\underset{\sim}{A}^{-1}$ does not exist and the matrix $\underset{\sim}{A}$ is *singular* (or *not invertible*). Further, when $\underset{\sim}{A}$ is symmetric, $\underset{\sim}{A}^{-1}$ is also symmetric.

### 1.3.4 Examples

The determinants and inverses of the *matrices* below are easily verified:

$$\underset{\sim}{A} = \begin{pmatrix} 2 & 3 \\ 2 & 5 \end{pmatrix} \quad |\underset{\sim}{A}| = 4 \quad \underset{\sim}{A}^{-1} = \begin{pmatrix} 1.25 & -0.75 \\ -0.5 & 0.5 \end{pmatrix}$$

$$\underset{\sim}{A} = \begin{pmatrix} 1.0 & 0.24 & 0.58 \\ 0.24 & 1.0 & 0.73 \\ 0.58 & 0.73 & 1.0 \end{pmatrix} \quad |\underset{\sim}{A}| = 0.2763 \quad \underset{\sim}{A}^{-1} = \begin{pmatrix} 1.690 & 0.664 & -1.465 \\ 0.664 & 2.401 & 2.138 \\ -1.465 & 2.138 & 3.410 \end{pmatrix}$$

$$\underset{\sim}{D} = \mathrm{diag}(4, 5, -2) \quad |\underset{\sim}{D}| = -40 \quad \underset{\sim}{D}^{-1} = \mathrm{diag}\left(\frac{1}{4}, \frac{1}{5}, -\frac{1}{2}\right).$$

## 1.4 Rank of a matrix

The *rank* of a matrix $\underset{\sim}{A}$ of dimension $r \times c$ corresponds to the dimension of the largest square sub-matrix whose determinant is not equal to 0 that can be found in it.

$$0 \leq \text{ Rank } (\underset{\sim}{A}) \leq \min(r, c)$$

Thus, the rank of the $3 \times 3$ matrix

$$\underset{\sim}{A} = \begin{pmatrix} 14 & 3 & 8 \\ 2 & 0 & 2 \\ 2 & 3 & -4 \end{pmatrix}$$

is less than or equal to 3. However, it cannot be equal to 3 since $|\underset{\sim}{A}| = 0$. It is in fact equal to 2; indeed, for instance, the determinant of the $2 \times 2$ square sub-matrix $\begin{pmatrix} 14 & 3 \\ 2 & 0 \end{pmatrix}$ differs from 0.

The rank of a $r \times c$ dimensional matrix $\underset{\sim}{A}$ can also be defined as the maximum number of columns or rows of the matrix *linearly independent*. The $c$ vectors $\underset{\sim}{\ell}_1, \underset{\sim}{\ell}_2, \ldots, \underset{\sim}{\ell}_c$ are linearly independent if it is not possible to find coefficients $k_1, \ldots, k_c$ not all equal to 0 such that

$$k_1 \underset{\sim}{\ell}_1 + k_2 \underset{\sim}{\ell}_2 + \ldots + k_c \underset{\sim}{\ell}_c = \underset{\sim}{0}.$$

In the example above, the rank of $\underset{\sim}{A}$ cannot be 3 since the 3 columns of the matrix are not linearly independent. Indeed, by choosing $k_1 = 1, k_2 = -2$ and $k_3 = -1$, we have

$$1 \times \begin{pmatrix} 14 \\ 2 \\ 2 \end{pmatrix} - 2 \times \begin{pmatrix} 3 \\ 0 \\ 3 \end{pmatrix} - 1 \times \begin{pmatrix} 8 \\ 2 \\ -4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

By contrast, the matrix is of rank 2 since for example columns 1 and 2 of the matrix are linearly independent. Indeed, it is impossible to find $k_1$ and $k_2$ not both equal to 0 such that

$$k_1 \times \begin{pmatrix} 14 \\ 2 \\ 2 \end{pmatrix} + k_2 \times \begin{pmatrix} 3 \\ 0 \\ 3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Thus, the rank of $\underset{\sim}{A}$ equals 2.

Remarks

- When vectors are not linearly independent, they are said to be *linearly dependent*. There is a linear relationship between them and anyone can be expressed as a linear function of the others.
- The square matrix $\underset{\sim}{A}$ of dimension $r \times r$ is singular if and only if it is not of full rank. Therefore, when rank $\underset{\sim}{A} < r$, then $|\underset{\sim}{A}| = 0$ and conversely.
- The rank of a square matrix evidences the true dimension of the matrix. If the matrix is not of full rank, it entails basically redundant information.

## 1.5 Quadratic forms

Let $\underset{\sim}{A}$ a symmetric matrix of dimension $r \times r$ and $\underset{\sim}{b}$ a vector of dimension $r$, then the triple product

$$\underset{\sim}{b}^T \cdot \underset{\sim}{A} \cdot \underset{\sim}{b}$$

yields a scalar called a *quadratic form*.

The matrix $\underset{\sim}{A}$ or quadratic form $\underset{\sim}{b}^T \cdot \underset{\sim}{A} \cdot \underset{\sim}{b}$ is *positive definite(pd)* if and only if

$$\underset{\sim}{b}^T \cdot \underset{\sim}{A} \cdot \underset{\sim}{b} > 0 \quad \forall \underset{\sim}{b} \neq \underset{\sim}{0}$$

By contrast, if $\underset{\sim}{b}^T \cdot \underset{\sim}{A} \cdot \underset{\sim}{b} \geq 0 \quad \forall \underset{\sim}{b} \neq 0$, the quadratic form is said to be *positive semidefinite (psd).*

## 1.6 Eigen values and vectors

### 1.6.1 Eigen values

Let $\underset{\sim}{A}$ be a square matrix of dimension $r \times r$ and $\lambda$ a scalar. The *eigen values* also called *latent roots* of matrix $\underset{\sim}{A}$ are the solutions of the caracteristic equation

$$|\underset{\sim}{A} - \lambda \underset{\sim}{I}| = 0$$

i.e. of the polynomial of order $r$

$$|\underset{\sim}{A} - \lambda \underset{\sim}{I}| = b_0 + b_1\lambda + b_2\lambda^2 + \ldots + b_r\lambda^r$$

In the equation above, it can be shown that $b_0 = |\underset{\sim}{A}|$, $b_{r-1} = (-1)^{r-1} tr\underset{\sim}{A}$ and $b_r = (-1)^r$.

Denote by $\lambda_1, \lambda_2, \ldots, \lambda_r$, the $r$ eigen values of matrix $\underset{\sim}{A}$, solutions of the polynomial equation, which can be real or complex numbers. When matrix $\underset{\sim}{A}$ is symmetric, as often the case in statistics, all eigen values are real numbers.

It is easily seen that

$$tr\underset{\sim}{A} = \lambda_1 + \lambda_2 + \ldots + \lambda_r = \sum_{i=1}^{r} \lambda_i$$

$$|\underset{\sim}{A}| = \lambda_1 \times \lambda_2 \times \ldots \times \lambda_r = \prod_{i=1}^{r} \lambda_i$$

If matrix $\underset{\sim}{A}$ is *positive definite*, $\lambda_i > 0\ \forall i$. If $\underset{\sim}{A}$ is *positive semi-definite*, then at least one eigen value $\lambda_i$ is equal to 0 and the matrix cannot be inverted because $|\underset{\sim}{A}| = 0$. It is also seen that the rank of matrix $\underset{\sim}{A}$ is equal to the number of its non null eigen values.

The eigen values are in some sense the building blocks of matrix $\underset{\sim}{A}$.

### 1.6.2 Eigen vectors

For each eigen value $\lambda_i\,(i = 1, \ldots, r)$, there is a corresponding *eigen vector* $\underset{\sim}{x_i}$ obtained by solving the equations

$$\underset{\sim}{A}\underset{\sim}{x_i} = \lambda_i \underset{\sim}{x_i}$$

or equivalently

$$(\underset{\sim}{A} - \lambda_i \underset{\sim}{I})\underset{\sim}{x_i} = \underset{\sim}{0}$$

The solution $\underset{\sim}{x_i} \neq 0$ exists because the matrix $(\underset{\sim}{A} - \lambda_i \underset{\sim}{I})$ cannot be inverted because its determinant is null. Otherwise, one would get the trivial solution $\underset{\sim}{x_i} = \underset{\sim}{0}$.

However the solution $\underset{\sim}{x_i} \neq \underset{\sim}{0}$ is not unique and it is common to normalize the vector $\underset{\sim}{x_i}$ to unity, specifically $\|\underset{\sim}{x_i}\| = \underset{\sim}{x_i}^T \cdot \underset{\sim}{x_i} = 1$.

The eigen vectors $\underset{\sim}{x_i}$ and $\underset{\sim}{x_j}$ associated to different eigen values $\lambda_i \neq \lambda_j$ are orthogonal, namely

$$\underset{\sim}{x_i}^T \cdot \underset{\sim}{x_j} = 0 \quad \forall i \neq j$$

The eigen vectors define a cartesian space. It is easily seen that $\underset{\sim}{A} = \lambda_1 \underset{\sim}{x_1} \cdot \underset{\sim}{x_1}^T + \ldots + \lambda_r \underset{\sim}{x_r} \cdot \underset{\sim}{x_r}^T$.

Consider for instance the symmetric matrix of dimension $2 \times 2$ :

$$\underset{\sim}{A} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}.$$

The eigen values are solutions of the second order equation

$$\begin{vmatrix} 1-\lambda & 1 \\ 1 & 2-\lambda \end{vmatrix} = (1-\lambda)(2-\lambda) - 1$$
$$= 1 - 3\lambda + \lambda^2 = 0$$

and write $\lambda_1 = (3+\sqrt{5})/2 \simeq 2.618$ and $\lambda_2 = (3-\sqrt{5})/2 \simeq 0.382$. Observe that $\lambda_1 + \lambda_2 = tr\underset{\sim}{A} = 3$ and that $\lambda_1 \times \lambda_2 = |\underset{\sim}{A}| = 1$.

The eigen vector associated with the eigen value $\lambda_1 = 2.618$ can be derived from the equation

$$\begin{pmatrix} -1.618 & 1 \\ 1 & -0.618 \end{pmatrix} \quad \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

whose solution is $x_1 = k$ and $x_2 = 1.618k\,(k \neq 0)$. To obtain a unique solution, the vector is normed to unity by taking $k = 0.5257$. Hence $\underset{\sim}{x}_1^T = (0.5257, 0.8507)$.

Likewise, the eigen vector associated with eigen value $\lambda_2 = 0.382$ is obtained from equation

$$\begin{pmatrix} 0.618 & 1 \\ 1 & 1.618 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

and writes $x_1 = k$ and $x_2 = -0.618k\,(k \neq 0)$. To obtain a unique solution, the vector is normed to unity by taking $k = 0.8507$. Thus, $\underset{\sim}{x}_2^T = (0.8507, -0.5257)$.

It is easy to check that eigen vectors $\underset{\sim}{x}_1$ and $\underset{\sim}{x}_2$ are orthogonal and that (up to rounding errors)

$$\underset{\sim}{A} = 2.618 \times \begin{pmatrix} 0.5257 \\ 0.8507 \end{pmatrix} (0.5257,\ 0.8507) + 0.382 \times \begin{pmatrix} 0.8507 \\ -0.5257 \end{pmatrix} (0.8507,\ -0.5257)$$

## 1.7 Matrix equation

Matrix calculus can also be helpful in writing and solving systems of equations. Thus, if $\underset{\sim}{A}$ is a square non singular matrix of dimension $r \times r$, $\underset{\sim}{x}$ and $\underset{\sim}{b}$ vectors of length $r$, one can easily solve the system of equations :

$$\underset{\sim}{A} \cdot \underset{\sim}{x} = \underset{\sim}{b}$$

by multiplying both sides of the equality by the inverse of matrix $\underset{\sim}{A}$ so that

$$\underset{\sim}{x} = \underset{\sim}{A}^{-1} \cdot \underset{\sim}{b}.$$

As an illustration, the set of 3 equations with 3 unknowns

$$\begin{aligned} 8x_1 + 7x_2 - 3x_3 &= 14 \\ 7x_1 - 2x_2 + 4x_3 &= 8 \\ x_1 + 2x_2 - 5x_3 &= -1 \end{aligned}$$

can be written in matrix form

$$\begin{pmatrix} 8 & 7 & -3 \\ 7 & -2 & 4 \\ 1 & 2 & -5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 14 \\ 8 \\ -1 \end{pmatrix}.$$

The inverse of matrix $\underset{\sim}{A}$ can be determined as described previously. The determinant of $\underset{\sim}{A}$ is equal to $|\underset{\sim}{A}| = 241$ and the matrix of cofactors writes

$$\begin{pmatrix} 2 & 39 & 16 \\ 29 & -37 & -9 \\ 22 & -53 & -65 \end{pmatrix}.$$

Thus, the inverse matrix writes (after transposing the matrix of cofactors)

$$\underset{\sim}{A}^{-1} = \frac{1}{241} \begin{pmatrix} 2 & 29 & 22 \\ 39 & -37 & -53 \\ 16 & -9 & -65 \end{pmatrix} = \begin{pmatrix} 0.00830 & 0.120 & 0.0913 \\ 0.162 & -0.154 & -0.220 \\ 0.0664 & -0.0373 & -0.270 \end{pmatrix}.$$

The vector $\underset{\sim}{x}$, solution of the original equations, is given by the systmem of equations

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0.00830 & 0.120 & 0.0913 \\ 0.162 & -0.154 & -0.220 \\ 0.0664 & -0.0373 & -0.270 \end{pmatrix} \begin{pmatrix} 14 \\ 8 \\ -1 \end{pmatrix}.$$

Thus, $x_1 = 0.00830 \times 14 + 0.120 \times 8 + 0.0913 \times (-1) = 0.985$. Similarly, $x_2 = 1.256$ and $x_3 = 0.901$.

Finally, $\underset{\sim}{x}^T = (0.985, 1.256, 0.901)$.

# Chapter 2

# Matrix of observations

## 2.1 Introduction

In multivariate statistics, the starting point is generally a matrix of observations where the rows corresponds to subjects and the columns to variables. Before defining formally this matrix, there is a need to recall some basic notions in statistics, such as population, sample, variables and data. We shall present the various types of variables et the transformations they may undergo before constructing the matrix of observations. The chapter ends with some graphical representations of multivariate data.

## 2.2 Population and sample

In any statistical problem, there is a need to clearly define the population under study from which a sample will be drawn.

### 2.2.1 Population

A population is a collection (set) of individuals (or objects) who have a least one characteristic in common. Le $N$ denote the size of the population. This number is often assumed to be infinite ($N \approx \infty$), so that it does not appear in the mathematical formulas. Classical examples are the student population of a university, the population of cities in a country, or the population of hospital patients.

It is of importance to clearly define the population under study before starting any data analysis. Note that the basic element or item (subject or object) of the population on which observations or measurements will be

made is called the *statistical unit* In the preceding examples, the statistical unit will be the student, the city and the patient, respectively.

### 2.2.2 Sample

In practice as the population is infinite (or at least extremely large) it cannot be studied completely. Thus, we have to draw a sample, that is a subset, from the population. Let $n$ be the sample size which by definition is finite. Sampling from a population is by no way an easy task. Rigorous sampling methods are required, otherwise the samples will not be representative of the population and biased conclusions may be drawn.

### 2.2.3 Sampling

Sampling is a random mechanism (process) to draw samples from a population. There are many sampling schemes but the most popular is "simple random sampling" which fulfills the following criteria :

1. the sample size $n$ is fixed beforehand;
2. successive draws are made at random (by chance) and independently of each other (no links between draws);
3. successive draws are made from an unchanged population (draws cannot modify the population).

As populations are generally infinite, criterion (3) is always satisfied. Otherwise the element drawn has to be put back (replaced) in the population; this implies that the population remains invariant but also that the same item may be drawn several times. Other sampling methods comprise, e.g. stratified sampling, systematic sampling or cluster sampling. Today, random sampling is done by computer.

## 2.3 Variables

### 2.3.1 Definition

In statistics, variables are theoretical concepts. They correspond to the characteristics (factors) which will be observed or measured on the sample elements. As for the population under study, they have to be clearly defined beforehand. Referring to the population of university students, we may be

interested in the following variables: age, gender, race, faculty, graduation score. Considering cities, variables may include number of inhabitants, socieconomic level, air pollution concentration, prevalence of respiratory affections. For hospital patients, one may be interested in age, gender, kind of disease, length and costs of hospital stay, outcome at discharge.

Variables can be of different types but basically they are either quantitative or qualitative (categorical.

### 2.3.2 Quantitative variable

A quantitative variable expresses a quantity. Quantitative variables are generally measured by means of an equipement or assessed by counting. In both cases however values are numerical and can be immediately reported in the observation matrix.

When variables are measured, we say they are "quantitative continuous" as they can take in theory any possible value in a given interval, thus an infinity of values. Measures are generally associated with units whereas counts are just numbers. Examples are height (cm) of a student, ozone concentration ($mg/mm^3$) in a city, cholesterol level (g/l) of a patient.

By contrast, count variables are "quantitative discrete" as they can only take a finite number of distinct values; number of exams passed successfully by students, number of hospitals in cities, number of doctor visits for patients.

### 2.3.3 Binary variable

A binary variable has only two outcomes, e.g. absent or present, success or failure, man or woman, healthy or diseased. Classically, the outcomes are coded 0 and 1. Thus, 0 = absence and 1 = presence, 0 = success and 1 = failure, 0 = man and 1 = woman, 0 = healthy and 1 = diseased. The solution of a statistical problem does not depend on how the two outcomes are coded, but just remember what was 0 and what was 1. Interestingly, binary variables may considered as quantitative discrete variables and statistical calculations can be performed on their values 0 and 1.

### 2.3.4 Qualitative variable

Qualitative variables express a quality (feature). In general they are not measured but observed. Values taken by a qualitative variable are called categories, classes or modalities, they are textual not numeric ! Therefore they are also called "nominal" but also "categorical". Examples are the race of a student (caucasian, african, asian), the size of a city (small, medium,

large), the status of a patient (single, married, divorced, widow). When the modalities of a qualitative variable are "ordered", the variable is said "qualitative ordinal". This was the case for the size of the city (small, medium, large) or the stage of a patient's cancer (I to IV). Modalities of a qualitative variable can be numbered but in no way should calculations be made on these values as the numbering is just arbitrary (e.g. 1= caucasian, 2=african, 3=asian). This is less true for ordinal qualitative variables where calculations can be done, although with caution (e.g. 1= small, 2 = medium, 3 = large). In general, the categories (modalities) of a qualitative variable are mutually exclusive in the sense that a population element cannot belong to two categories; thus the size of a city cannot be small and large at the same time. There are situations where the modalities of a qualitative variable are not disjoint (mutually exclusive). This often occurs in multiple choice questionnaire where for a single question several answers may be possible. As an example, consider the qualitative variable "organ affected"; for a given patient, several choices are possible among the following ones: heart, lung, kidney, brain, stomach, bowel, liver, pancreas, prostate/ovaries. To analyze data of this type, it is advised to define as many binary variables as there are modalities of the variable.

Finaly, note that a binary variable is also a qualitative variable with two modalities. Since two modalities are always ordered, a binary variable is also ordinal.

### 2.3.5 Categorized variable

It happens that a continuous variable can not be measured as such or will not be studied as such. This applies for example when considering confidential information such as monthly salary (EUR) of employees, size of practice (number of patients seen per week) for a doctor. In such situations it is common to divide the range of values of the quantitative variable into categories, so that the quantitative variable is "categorized" and is being treated as a categorical, usually ordinal, qualitative variable. The variable "salary" may be categorized as follows: less than 500 EUR/month, 500 to 1000 EUR/month, 1000 to 2000 EUR/month, 2000 to 3000 EUR/months, more than 3000 EUR/month. The physcian's size of practice could be splitted as follows: $< 30$ patients/week, 30-60 patients/week, 60-90 patients/week, and $> 90$ patients/week. It is worth mentioning that categorizing a quantitative variable necessarily entails a loss of information.

## 2.4 Data

Data should not be confused with variables. While variables are theoretical concepts, data are real observations. Data result from the measurement or counting of a quantitative variable or assessment of a binary or qualitative variable. Thus, considering the variable weight (kg), a datum corresponds to the measure of weight in a particular individual (e.g. 73 kg). Likewise, if the size of a given city is large, then "large" is a datum not a variable.

In multivariate statistics, all data should be numerical, i.e. numbers on which calculations (addition, subtraction, multiplication, division) will be performed. In this respect, quantitative variables should not be a problem, as they always yield numbers. This also applies to binary variables when a code 0 or 1 has been given to the two outcomes.

When it comes to qualitative variables, the problem is less straightforward as data are generally names or text. Numbers can be given to each modality category) of the variable but since this is done in an arbitrary way no calculations can be made on these numbers, except as already mentioned when the variable is ordinal (ordered categories).

To handle qualitative variables numerically, they should be replaced by a set of binary variables. We already mentioned that if the modalities are not mutually exclusive (multiple choice questions), then a binary variable is associated with each modality and the binary variable takes the value 1 if the modality is present and 0 if it is not. Thus there are as many binary variables as there are modalities of the qualitative variable. By contrast, if modalities are mutually exclusive, then one category is selected as the "reference category", and a binary variable is associated with each of the other categories. Thus, there are as many binary variables as there are modalities of the qualitative variable minus one. As an illustration, consider the variable race (caucasian, african, asian) and choose the category "caucasian" as the reference one, then define the binary variable "african" being equal to 1 if the subject is african and 0 otherwise, and the binary variable "asian" being equal to 1 if the subject is asian and 0 otherwise. In this context, the variable race has been expanded into two binary variables "african" and "asian". For an asian subject, the data will be african=0 and asian=1. By contrast, a caucasian subject will be numerically coded african=0 and asian=0. Thus it suffices to define two binary variables to perfectly make numerical the variable race. It should be noted that the choice of the reference category in ways impacts the ultimate solution of the problem.

The process of replacing a qualitative variable by binary variables is called a Boolean process because binary variables are also called *Boolean variables* (Boole algebra $0-1$). By this process, all data become numerical whether

from quantitative or qualitative variables.

## 2.5 Matrix of observations

Hereafter, $\underset{\sim}{X}^T = (X_1, \ldots, X_p)$ denotes the vector of the $p$ "numerical" variables of the multivariate problem. This presumes that qualitative variables have been appropriately replaced by the corresponding binary variables.

### 2.5.1 Definition

The *observation matrix* (or data matrix) is the matrix obtained by the observation of the $p-$variate vector $\underset{\sim}{X}^T = (X_1, \ldots, X_p)$ in a sample of $n$ subjects or objects drawn randomly from the population under study.

The observation matrix has therefore $n$ rows (subjects or objects) and $p$ columns (the variables). It is an $n \times p$ dimensional matrix written

$$\underset{\sim}{X}_{n\times p} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}. \tag{2.1}$$

In this matrix, the element $x_{ij}$ represents the observation made on subject (or object) $i$ for variable $X_j (i = 1, \ldots, n; j = 1, \ldots, p)$. In general the matrix is rectangular since there are more subjects than variables ($n > p$). All elements of matrix (2.1) are real numbers.

The dimensions of the matrix, respectively $n$ and $p$, are fundamental characteristics of the multivariate problem.

Matrix $\underset{\sim}{X}_{n\times p}$ is complete when none of the $x_{ij}$ observations is missing; otherwise it is incomplete and this raised difficulties for subsequent calculations.

### 2.5.2 Examples

Consider a random sample of 10 students drawn from the pouplation of all students of the university and suppose we are interested in the following variables: age, gender, race and final grade (< 5, 5-7, 7-9 and 9-10). The variable race was replaced by two binary variables "african" and "asian". The categories of the final grade variable were numbered sequentially 1 to 4 as they are ordered. Variables age (years) and gender (0=male, 1=female) are fine. The matrix of observations writes

$$\underset{\sim}{X}_{10\times 5} = \begin{pmatrix} 18 & 0 & 1 & 0 & 1 \\ 19 & 1 & 1 & 0 & 1 \\ 24 & 1 & 1 & 0 & 4 \\ 21 & 0 & 0 & 0 & 3 \\ 22 & 0 & 0 & 1 & 2 \\ 26 & 0 & 1 & 0 & 3 \\ 19 & 1 & 1 & 0 & 2 \\ 17 & 1 & 1 & 0 & 1 \\ 20 & 1 & 0 & 0 & 2 \\ 20 & 1 & 1 & 0 & 4 \end{pmatrix}$$

where students correspond to rows 1 to 10 and variables to 5 columns: specifically $X_1$ = age (years), $X_2$ = gender (0=male, 1=female), $X_3$ = race african (0=no, 1=yes), $X_4$ = race asian (0=no, 1=yes), $X_5$ = final grade (1 = less than 5, 2 = 5-7; 3 = 7-9, 4 = 9-10). Thus, subject #2 is 19 years old, female, african and has a final score <5. Alternatively, subject #4, aged 21 years, male, caucasian, has a final score between 7 and 9.

Several data matrices are given in the Annex. In particular, the well-known Fisher iris dataset. This dataset consists of three data matrices, one for each species of iris flowers (setosa, versicolor and virginica). R.A. Fisher took a random sample of 50 flowers from each species and studied four variables: sepal length ($X_1$), sepal width ($X_2$), petal length ($X_3$) and petal width ($X_4$). For each species, we have an observation matrix of dimension $50 \times 4$ since $n = 50$ and $p = 4$. By aligning the three matrices under each other, we can create a single observation matrix but this requires defining a 3-categorical qualitative variable to distinguish the three iris species. Taking the "setosa" category as the reference class, we then create two binary variables as follows: versicolor (0=no, 1=yes) and virginica (0=no, 1=yes). The final matrix would then have 150 rows and 6 columns.

### 2.5.3 Cluster of points

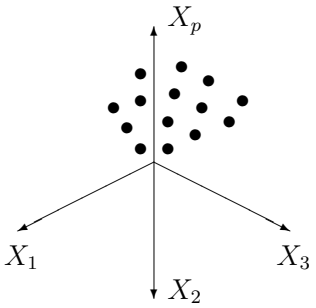
By associating each variable $X_1, \ldots, X_p$ with an oriented axis, we define a referential of othogonal axes in the $p$-dimensional Euclidean space $\mathbb{R}^p$, also called observation space. Then the observation of vector $\underset{\sim}{X}$ for subject $i$

$$\underset{\sim}{x}_i^T = \begin{pmatrix} x_{i1}, & \ldots, & x_{ip} \end{pmatrix} \tag{2.2}$$

namely the $i$-th row of observation matrix $\underset{\sim}{X}_{n\times p}$, can be considered as a point of $\mathbb{R}^p$. In other terms, to each observation matrix $\underset{\sim}{X}_{n\times p}$ there corresponds a

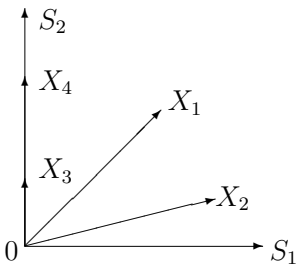
cluster (cloud) of $n$ points in the $p$-dimensional space $\mathbb{R}^p$. There are as many points in the cloud as there are rows in the data matrix.

$$\underset{\sim}{X}_{n\times p} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots \\ x_{i1} & x_{i2} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \quad \Leftrightarrow$$




This is just an extension of plotting the observations of two variables on a 2-dimensional graph (scatter plot) and similarly for three variables in the 3-dimensional space. Beyond 3 variables, a classical representation is no longer possible but it can be figured out. Conversely, we can associate an axis to each subject (row) ($i = 1, \dots, n$) and consider the $\mathbb{R}^n$ Euclidean space. The observation of variable $X_j$ for each subject, namely the j-th column vector $(x_{1j}, x_{2j}, \dots, x_{nj})^T$, would correspond to a point in the space $\mathbb{R}^n$. Thus, there will be $p$ points in space $\mathbb{R}^n$, one for each variable of the multivariate problem. This representation is somewhat more difficult to grasp in practice. As an example, the observation of four variables on two subjects ($S_1$ and $S_2$)gives the figure below :

$$\underset{\sim}{X}_{2\times 4} = \begin{pmatrix} 3 & 5 & 0 & 0 \\ 4 & 1 & 2 & 6 \end{pmatrix} \quad \Leftrightarrow$$




In conclusion, a matrix of observations $\underset{\sim}{X}_{n\times p}$ corresponds to a cluster (cloud) of $n$ points in $\mathbb{R}^p$ or to $p$ points in space $\mathbb{R}^n$. In space $\mathbb{R}^p$, points close to each other have a similar profile, whereas in $\mathbb{R}^n$ variables pointing in the same (or opposite) direction are positively (negatively) correlated. In Biplot graphical representations, when variables are orthogonal they are usually viewed as non-correlated.

## 2.6 Graphical display of multivariate data

Any multivariate statistical analysis should start by analyzing each variable separately first. However, statisticians have always strived to get a graphical display of multivariate data.

As an illustration of such methods, we shall consider the scores (range: 0-20) obtained by 6 students in the following courses : $X_1$=Chemistry (Ch), $X_2$=Physics (Ph), $X_3$=Mathematics (Ma), $X_4$=Psychology (Ps) and $X_5$=Botany (Bo). Data are given in the table below:

| Student | Ch | Ph | Ma | Ps | Bo |
|---|---|---|---|---|---|
| 1 | 19 | 19 | 15 | 14 | 16 |
| 2 | 15 | 13 | 12 | 14 | 12 |
| 3 | 12 | 10 | 11 | 15 | 11 |
| 4 | 5 | 5 | 12 | 14 | 14 |
| 5 | 4 | 3 | 8 | 11 | 9 |
| 6 | 1 | 0 | 2 | 9 | 10 |

There are various ways to represent the pofile of each student. It is convenient to transform all observations so they range in the interval [0-1] by using the formula

$$Y = \frac{X - min}{max - min}$$

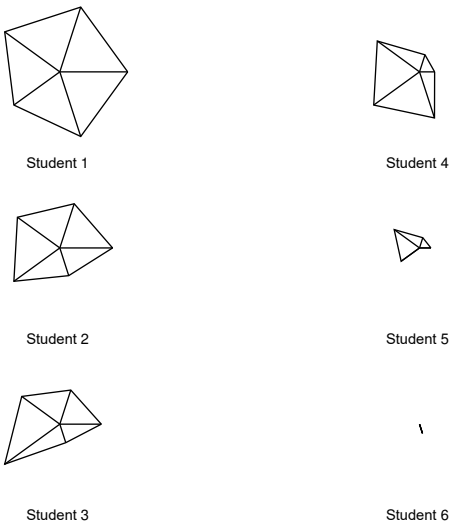
where $min$ et $max$ represent respectively the lowest and highest score in each course.

### 2.6.1 Glyphs

A *glyph* is a fixed circle with $p$ arrows (rays) pointing outside the circle, the length of these arrows corresponding to the value of each variable. There will be a glyph for each of the $n$ subjects and these can be compared visually.

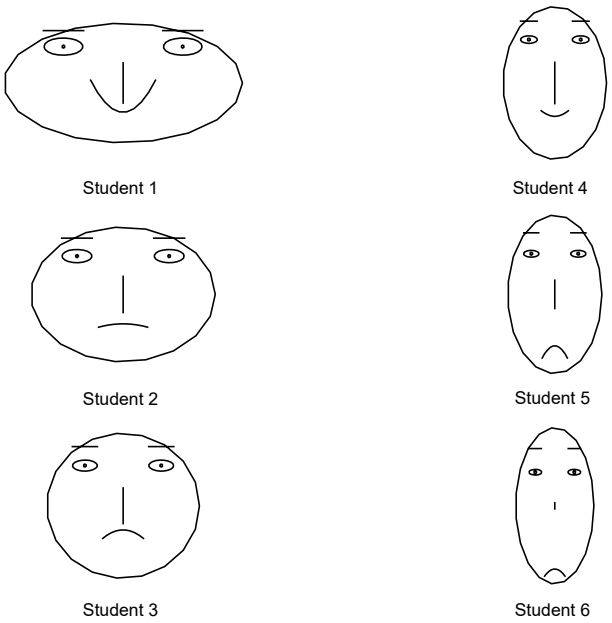
### 2.6.2 Stars

A *star* is like a glyph representation except that rays of equal length are displayed within the circle. The values of the variables determine on each ray a point and these points are joined to form a polygon looking like a star.



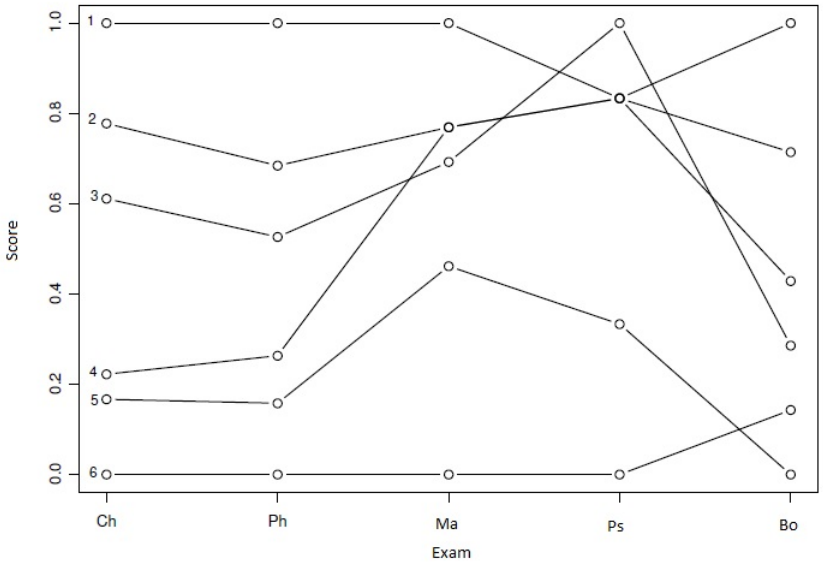

### 2.6.3 Chernoff faces

The method of *Chernoff faces* consists in associating with each variable of the multivariate problem one characteristic of a human face such as length of nose, shape of face, size of eyes, or of ears and so on. Unfortunately by interchanging variables with characteristics, completely different faces are obtained for the same subject; therefore it is wise to try a few times to obtain the best representation.



### 2.6.4 Profiles

This method is quite similar to the star representation except that variables are represented next to each other on the X-axis. Values of the variables are joined by linear segments, hence defining the *profile* of each student. Clearly if the order of the variables displayed changes, the profile will also be completely different.

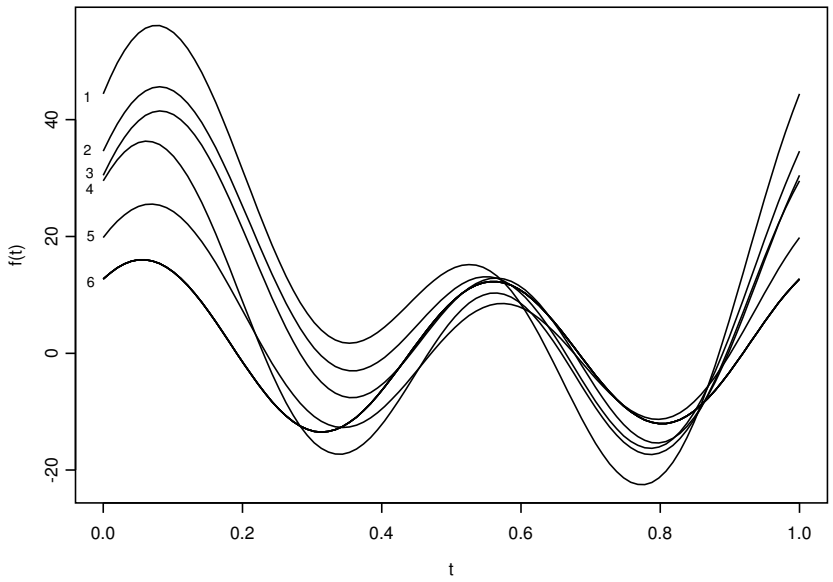



### 2.6.5 Fourier graphs

This approach due to Andrews consists in representing the multivariate observation $\underset{\sim}{x}_i^T = (x_{i1}, \ldots, x_{ip})$ by the finite Fourier series

$$f_i(t) = x_{i1}/\sqrt{2} + x_{i2} \sin t + x_{i3} \cos t + x_{i4} \sin 2t + x_{i5} \cos 2t + \ldots$$

and to plot the curve $f_i(t)$ in the interval $-\pi < t < \pi$ (or equivalently by replacing $t$ by $2\pi t$ in the interval $0 < t < 1$).




## 2.7 Outlying multivariate data

As in univariate statistics, an observation can be outlying from a multivariate standpoint. An outlier is an observation that does not really come from the same population as the other observations. It departs markedly from the other ones. It evidences an atypical profile.

Such outlying observations can seriously affect data analysis. It is therefore important to highlight such outliers, to correct them if possible or to discard them from the sample. Graphical representations described in section 2.6. can sometimes help identifying outliers and extreme multivariate observations. Other approaches will also be presented later to cope with this problem.

# Chapter 3

# Mean and dispersion

## 3.1 Introduction

In this chapter, the matrix of observations will be summarized by its mean vector and dispersion matrix. This is the first step in any multivariate statistical analysis. According to R.A. Fisher's definition, statistics is at first the discipline that concerns data reduction methods. We introduce the concept of the mean vector which extends the classical concept of (arithmetic) mean value to the multidimensional setting. Then we define the dispersion matrix which measures multivariate variability in a given sample, expanding therewith the notion of variance. Since, in practice, the dispersion matrix is not easy to interpret, we introduce the correlation matrix. The chapter ends with the defintion of the Mahalanobis distance which allows to calculate the distance between any multivariate observation and the mean point by taking into account the associations between the different variables. This distance generalizes the univariate standard deviate distance.

## 3.2 Mean vector

Consider the observation matrix $\underset{\sim n\times p}{X}$ defined in Chapter 2. The rows of this matrix correponds to the $p$-variate observations on $n$ subjects. The matrix can also be written

$$\underset{\sim}{X}_{n\times p} = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ \dots & \dots & \dots \\ x_{i1} & \dots & x_{ip} \\ \dots & \dots & \dots \\ x_{n1} & \dots & x_{np} \end{pmatrix} = \begin{pmatrix} \underset{\sim}{x}_1^T \\ \dots \\ \underset{\sim}{x}_i^T \\ \dots \\ \underset{\sim}{x}_n^T \end{pmatrix} \tag{3.1}$$

where $\underset{\sim}{x}_i^T = (x_{i1}, \dots, x_{ip})$ is the observation of vector $\underset{\sim}{X}^T = (X_1, \dots, X_p)$ for subject $i$.

Considering only observations of variable $X_j$, i.e. the column $j$ of matrix (3.1), then the mean of these observations is

$$\bar{x}_j = \frac{1}{n}\sum_{k=1}^{n} x_{kj} \quad (j = 1, \dots, p) \tag{3.2}$$

which consists of summing the observations of column $j$ and dividing by the sample size $n$.

By definition, the mean vector of the multivariate sample is the vector of the univariate mean values of the $p$ variables. Thus,

$$\underset{\sim}{\bar{x}}^T = (\bar{x}_1, \dots, \bar{x}_p) \tag{3.3}$$

In matrix notation, the expression (3.3) becomes

$$\underset{\sim}{\bar{x}} = \frac{1}{n}\sum_{k=1}^{n} \underset{\sim}{x}_k \tag{3.4}$$

by analogy with the univariate formula (3.2). The $n$ vectors of observations are added and the sum is divided by $n$.

In the $p$ dimensional space $\mathbb{R}^p$, the mean point is (physically) the gravity center of the cloud of points, since $\sum_k (\underset{\sim}{x}_k - \underset{\sim}{\bar{x}}) = \underset{\sim}{0}$. It is a location parameter as it determines the position of the cluster of points in $\mathbb{R}^p$.

## 3.3 Dispersion matrix

The mean vector provides no information about the dispersion of points in the $p$-dimensional space. In this context, we introduce the dispersion matrix also called the variance-covariance matrix.

Considering two variables $X_i$ and $X_j$, we define the covariance which measures the association between these two variables. The covariance $cov(X_i, X_j)$ is given by

$$s_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j) \tag{3.5}$$

or

$$s_{ij} = \frac{1}{n-1}\left\{\sum_{k=1}^{n} x_{ki}x_{kj} - \frac{\left(\sum_{k=1}^{n} x_{ki}\right)\left(\sum_{k=1}^{n} x_{kj}\right)}{n}\right\}. \tag{3.6}$$

A positive (negative) covariance indicates an increasing (decreasing) relation bteween the two variables, whereas a covariance close to 0 corresponds to the absence of relationship.

When computing the covariance of a variable $X_i$ with itself, i.e. $cov(X_i, X_i)$, the formulas (3.5) and (3.6) simplify and become respectively :

$$s_{ii} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)^2 \tag{3.7}$$

and

$$s_{ii} = \frac{1}{n-1}\left\{\sum_{k=1}^{n} x_{ki}^2 - \frac{\left(\sum_{k=1}^{n} x_{ki}\right)^2}{n}\right\}. \tag{3.8}$$

It turns out to be the variance $s_i^2 = var(X_i)$ of variable $X_i$. Thus, the variance is a special case of the covariance. Of note, a variance is never negative.

By definition, the dispersion matrix $\underset{\sim}{S}$ is the square matrix of dimension $p \times p$ whose diagonal elements are the variances of the $p$ variables and the off-diagonal elements the covariances. It is therefore called the *variance-covariance matrix* or simply the *covariance matrix*. We have

$$\underset{\sim}{S} = \begin{pmatrix} s_{11} & s_{12} & \dots & s_{1p} \\ s_{21} & s_{22} & \dots & s_{2p} \\ \dots & \dots & \dots & \dots \\ s_{p1} & s_{p2} & \dots & s_{pp} \end{pmatrix}. \tag{3.9}$$

Since $s_{ij} = s_{ji}$, the matrix $\underset{\sim}{S}$ is symmetric.

In matrix notation, the expression (3.9) also writes

$$\underset{\sim}{S} = \frac{1}{n-1}\sum_{k=1}^{n}(\underset{\sim}{x}_k - \underset{\sim}{\bar{x}})(\underset{\sim}{x}_k - \underset{\sim}{\bar{x}})^T \tag{3.10}$$

or

$$\underset{\sim}{S} = \frac{1}{n-1}\left\{\sum_{k=1}^{n} \underset{\sim}{x}_k \underset{\sim}{x}_k^T - \frac{\left(\sum_{k=1}^{n} \underset{\sim}{x}_k\right)\left(\sum_{k=1}^{n} \underset{\sim}{x}_k\right)^T}{n}\right\} \tag{3.11}$$

by analogy with the univariate formulas (3.5) and (3.6).

The dispersion matrix charaterizes the shape of the observed cloud around the mean point $\underset{\sim}{\bar{x}}$ in $\mathbb{R}^p$.

To calculate the dispersion matrix, we need the sum of squares of elements in each column of the observation matrix and the sum of cross-products of column elements taken 2 by 2. This leads to the matrix $\underset{\sim}{A}$, called the matrix of *sums of squares and cross-products*, whose generic element writes

$$a_{ij} = \sum_{k=1}^{n} x_{ki}x_{kj} - \frac{\left(\sum_{k=1}^{n} x_{ki}\right)\left(\sum_{k=1}^{n} x_{kj}\right)}{n} \qquad (i,j = 1,\ldots,p) \tag{3.12}$$

Using relation (3.6), we obtain

$$s_{ij} = \frac{1}{n-1}\, a_{ij}$$

and hence

$$\underset{\sim}{S} = \frac{1}{n-1}\, \underset{\sim}{A} \tag{3.13}$$

To simplify the notation, let by convention

$$\sum x_j = \sum_{k=1}^{n} x_{kj} \qquad (j = 1,\ldots,p)$$

$$\sum x_j^2 = \sum_{k=1}^{n} x_{kj}^2 \qquad (j = 1,\ldots,p)$$

$$\sum x_i x_j = \sum_{k=1}^{n} x_{ki}x_{kj} \qquad (i,j = 1,\ldots,p; i < j)$$

denote respectively the sum of the elements of column $j$, the sum of squares of elements of column $j$ and the sum of products of elements of columns $i$ and $j$.

Then,

$$\begin{aligned} \bar{x}_i &= \sum x_i/n \\ a_{ij} &= \sum x_i x_j - \frac{\left(\sum x_i\right)\left(\sum x_j\right)}{n} \quad (i,j = 1,\ldots,p) \\ s_{ij} &= a_{ij}/(n-1). \end{aligned} \tag{3.14}$$

## 3.4 Correlation matrix

The variance-covariance matrix $\underset{\sim}{S}$ is important in theory but its interpretation in practical applications is difficult. Indeed, covariances can be negative or positive, and their units are the product of the units of the two corresponding variables, and likewise for variances. This is the reason to introduce the notion of correlation between the two variables $X_i$ and $X_j$, say $\text{corr}(X_i, X_j)$, defined as follows :

$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii} \cdot s_{jj}}} \qquad (i, j = 1, \ldots, p) \tag{3.15}$$

or equivalently, using the same notation as in (3.14)

$$r_{ij} = \frac{\sum x_i x_j - \frac{\left(\sum x_i\right)\left(\sum x_j\right)}{n}}{\sqrt{\left[\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right]\left[\sum x_j^2 - \frac{\left(\sum x_j\right)^2}{n}\right]}} \tag{3.16}$$

Thus, to calculate the correlation between the two variables, five quantities are needed, respectively $\sum x_i, \sum x_j, \sum x_i^2, \sum x_j^2$ and $\sum x_i x_j$. Note that when $i = j, r_{ii} = 1$ $(i = 1, \ldots, p)$.

By definition, the *correlation matrix* $\underset{\sim}{R}$ is the square matrix of dimension $p \times p$ containing all correlations between variables taken 2 by 2. We have

$$\underset{\sim}{R} = \begin{pmatrix} 1 & r_{12} & \ldots & r_{1p} \\ r_{21} & 1 & \ldots & r_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ r_{p1} & r_{p2} & \ldots & 1 \end{pmatrix} \tag{3.17}$$

On the diagonal all elements are equal to 1 and off-diagonal there are the correlations. Since $r_{ij} = r_{ji}$, the matrix is symmetric.

In matrix notation, it can be shown that

$$\underset{\sim}{R} = \underset{\sim}{D}^{-1} \underset{\sim}{S} \underset{\sim}{D}^{-1} \tag{3.18}$$

where $\underset{\sim}{D} = \text{diag}(\sqrt{s_{11}}, \ldots, \sqrt{s_{pp}})$ is the diagonal matrix of standard deviations of the variables since $\sqrt{s_{ii}} = \sqrt{s_i^2} = s_i$ is the standard deviation of variable $X_i$. In conclusion, knowing $\underset{\sim}{S}$, we can calculate $\underset{\sim}{R}$ but the inverse is only possible if we know matrix $\underset{\sim}{D}$.

Remember that a correlation is a dimensionless number always comprised between –1 and +1. A positive (negative) correlation indicates an increasing (decreasing) linear relation between two variables, whereas a null correlation indicates the absence of any linear relationship.

## 3.5 Examples

Consider the observations matrix of dimension $8 \times 3$ given in section 1.1.3. It displays the observation on 8 subjects of 3 variables, $X_1$ = height (cm), $X_2$ = weight (kg) and $X_3$ = arm circumference (cm).

$$\underset{\sim}{X}_{8\times 3} = \begin{pmatrix} 167.9 & 71.8 & 30.0 \\ 183.8 & 75.1 & 29.4 \\ 172.9 & 58.0 & 26.0 \\ 175.5 & 58.4 & 25.7 \\ 176.4 & 67.7 & 27.9 \\ 168.5 & 75.2 & 31.7 \\ 178.0 & 67.3 & 27.4 \\ 178.0 & 71.3 & 29.0 \end{pmatrix}.$$

In the 3-dimensional space, this matrix represents a cluster (cloud) of 8 points. We shall calculate the mean vector $\underset{\sim}{\bar{x}}$, the dispersion matrix $\underset{\sim}{S}$ and the correlation matrix $\underset{\sim}{R}$.

Here $n = 8$ and $p = 3$.

Let us first calculate the sums, sums of squares and sums of cross-products of the observations, in total $p + p + p(p-1)/2 = 9$ quantities.

$$\begin{aligned} \textstyle\sum x_1 &= 167.9 + 183.8 + \ldots + 178.0 = 1401.0 \\ \textstyle\sum x_2 &= 71.8 + 75.1 + \ldots + 71.3 = 544.8 \\ \textstyle\sum x_3 &= 30.0 + 29.4 + \ldots + 29.0 = 227.1 \end{aligned}$$

$$\begin{aligned} \textstyle\sum x_1^2 &= (167.9)^2 + (183.8)^2 + \ldots + (178.0)^2 = 245544.7 \\ \textstyle\sum x_2^2 &= (71.8)^2 + (75.1)^2 + \ldots + (71.3)^2 = 37421.12 \\ \textstyle\sum x_3^2 &= (30.0)^2 + (29.4)^2 + \ldots + (29.0)^2 = 6475.91 \end{aligned}$$

$$\begin{aligned} \textstyle\sum x_1 x_2 &= (167.9 \times 71.8) + \ldots + (178.0 \times 71.3) = 95420.28 \\ \textstyle\sum x_1 x_3 &= (167.9 \times 30.0) + \ldots + (178.0 \times 29.0) = 39748.68 \\ \textstyle\sum x_2 x_3 &= (71.8 \times 30.0) + \ldots + (71.3 \times 29.0) = 15555.21 \end{aligned}$$

The mean vector writes

$$\begin{aligned} \underset{\sim}{\bar{x}}^T &= \left(\tfrac{1401}{8}, \tfrac{544.8}{8}, \tfrac{227.1}{8}\right) \\ &= (175.1, 68.1, 28.4) \end{aligned}$$

To obtain the dispersion matrix, we have to determine first matrix $\underset{\sim}{A}$ of dimension $3 \times 3$ by using expressions (3.12)

$$a_{11} = \sum x_1^2 - \tfrac{(\sum x_1)^2}{n} = 245544.7 - \frac{(1401)^2}{8} = 194.60$$

$$a_{12} = \sum x_1 x_2 - \tfrac{(\sum x_1)(\sum x_2)}{n} = 95420.28 - \frac{1401 \times 544.8}{8} = 12.18$$

$$a_{13} = \sum x_1 x_3 - \tfrac{(\sum x_1)(\sum x_3)}{n} = 39748.68 - \frac{1401 \times 227.1}{8} = -22.208$$

$$a_{22} = \sum x_2^2 - \tfrac{(\sum x_2)^2}{n} = 37421.12 - \frac{(544.8)^2}{8} = 320.24$$

$$a_{23} = \sum x_2 x_3 - \tfrac{(\sum x_2)(\sum x_3)}{n} = 15555.21 - \frac{544.8 \times 227.1}{8} = 89.70$$

$$a_{33} = \sum x_3^2 - \tfrac{(\sum x_3)^2}{n} = 6475.91 - \frac{(227.1)^2}{8} = 29.109$$

Thus, matrix $\underset{\sim}{A}$ writes

$$\underset{\sim}{A} = \begin{pmatrix} 194.6 & 12.18 & -22.208 \\ 12.18 & 320.24 & 89.7 \\ -22.208 & 89.7 & 29.109 \end{pmatrix}.$$

Finally, matrix $\underset{\sim}{S}$ is obtained by dividing matrix $\underset{\sim}{A}$ by $n - 1 = 7$ and we have

$$\underset{\sim}{S} = \begin{pmatrix} 27.799 & 1.74 & -3.1725 \\ 1.74 & 45.749 & 12.814 \\ -3,1725 & 12.814 & 4.1584 \end{pmatrix}.$$

To obtain matrix $\underset{\sim}{R}$, we need to calculate the correlations between all

pairs of variables by using formula (3.15):

$$r_{12} = \frac{1.74}{\sqrt{27.799 \times 45.749}} = 0.0488$$

$$r_{13} = \frac{-3.1725}{\sqrt{27.799 \times 4.1584}} = -0.2951$$

$$r_{23} = \frac{12.814}{\sqrt{45.749 \times 4.1584}} = 0.9291$$

Thus, the correlation matrix writes

$$\underset{\sim}{R} = \begin{pmatrix} 1.0 & 0.0488 & -0.2951 \\ 0.0488 & 1.0 & 0.9291 \\ -0.2951 & 0.9291 & 1.0 \end{pmatrix}$$

## Fisher iris dataset

The calculation of the matrix of sums of squares and products, the dispersion matrix and the correlation matrix applied to Fisher iris setosa observation matrix ($n = 50, p = 4$) yields the following results :

$$\underset{\sim}{\bar{x}}^T = (50.1, 34.3, 14.6, 2.46)$$

$$\underset{\sim}{A} = \begin{pmatrix} 608.825 & & & \\ 486.178 & 704.081 & & \\ 80.164 & 57.33 & 147.784 & \\ 50.617 & 45.57 & 29.743 & 54.439 \end{pmatrix}$$

$$\underset{\sim}{S} = \begin{pmatrix} 12.425 & & & \\ 9.922 & 14.369 & & \\ 1.636 & 1.170 & 3.016 & \\ 1.033 & 0.930 & 0.607 & 1.111 \end{pmatrix}$$

$$\underset{\sim}{R} = \begin{pmatrix} 1.00 & & & \\ 0.743 & 1.00 & & \\ 0.267 & 0.177 & 1.00 & \\ 0.278 & 0.233 & 0.332 & 1.00 \end{pmatrix}$$

Observe that the standard deviations of the 4 variables are obtained by taking the square root of the diagonal elements of $\underset{\sim}{S}$. Therefore, $s_1 =$

$3.52, s_2 = 3.79, s_3 = 1.74$ and $s_4 = 1.05$. Finally, the inverse matrix of $\underset{\sim}{S}$ is

$$\underset{\sim}{S}^{-1} = \begin{pmatrix} 0.189 & & & \\ -0.124 & 0.156 & & \\ -0.045 & 0.011 & 0.388 & \\ -0.048 & -0.021 & -0.179 & 1.060 \end{pmatrix}$$

## 3.6 Mahalanobis distance

Consider a multivariate observation $\underset{\sim}{x}^T = (x_1, \ldots, x_p)$. Remember that the (squared) Euclidean distance between $\underset{\sim}{x}$ and the mean vector $\underset{\sim}{\bar{x}}$ of the cluster of points is given by

$$\begin{aligned} d^2(\underset{\sim}{x}) &= (\underset{\sim}{x} - \underset{\sim}{\bar{x}})^T(\underset{\sim}{x} - \underset{\sim}{\bar{x}}) \\ &= \sum_{j=1}^{p}(x_j - \bar{x}_j)^2 \end{aligned} \tag{3.19}$$

namely the sum of the squared differences beteween the components of the two vectors. Unfortunately, this distance does not account for the associations between the different variables.

By definition, the *Mahalanobis distance* between $\underset{\sim}{x}$ and $\underset{\sim}{\bar{x}}$ is the quadratic form

$$D^2(\underset{\sim}{x}) = (\underset{\sim}{x} - \underset{\sim}{\bar{x}})^T \underset{\sim}{S}^{-1}(\underset{\sim}{x} - \underset{\sim}{\bar{x}}) \tag{3.20}$$

where $\underset{\sim}{S}^{-1}$ is the inverse of the dispersion matrix. In the univariate case ($p = 1$), the expressions (3.19) and (3.20) become respectively

$$\begin{aligned} d^2(x) &= (x - \bar{x})^2 \\ D^2(x) &= \left(\frac{x - \bar{x}}{s}\right)^2 \end{aligned}$$

It is seen that Mahalanobis distance accounts for the dispersion of sample values, $s$ being the standard deviation, whereas the Euclidean distance does not.

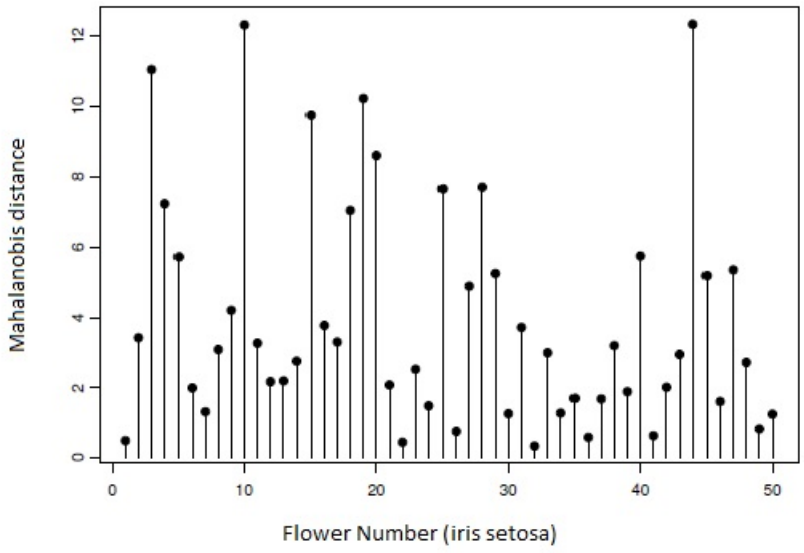
In the bivariate case ($p = 2$), points located at the same Euclidean distance from the mean ($d^2 = c$) lie on a circle, while those located at the same Mahalanobis distance ($D^2 = c$) fall on an ellipse. Likewise, for 3 dimensions ($p = 3$), instead of a sphere it is an ellipsoid.

When computing the Mahalanobis distance for all sample observations $\underset{\sim}{x}_i$ $(i = 1, \ldots, n)$, i.e.

$$D^2\left(\underset{\sim}{x}_i\right) = \left(\underset{\sim}{x}_i - \underset{\sim}{\bar{x}}\right)^T \underset{\sim}{S}^{-1} \left(\underset{\sim}{x}_i - \underset{\sim}{\bar{x}}\right),$$

distances can be ranked in increasing order of magnitude, from smallest to largest. This shows which observations are close to the mean and those departing substantially from it.

The figure below plots the Mahalanobis distance for each flower of the iris setosa dataset. Some flowers are markedly distant from the sample mean value. Flower 44 is the farthest ($D^2 = 12.33$) and flower 32 the closest ($D^2 = 0.34$).




For large $n$, Mahalanobis distance follows a chi-squared distribution on $p$ degrees of freedom. Thus, a multivariate observation may be considered as *extreme* or *outlying* when

$$D^2\left(\underset{\sim}{x}\right) \geq Q_{\mathcal{X}^2}(0.99; p) \tag{3.21}$$

with $Q_{\mathcal{X}^2}(0.99; p)$ the 99% quantile (percentile) of the chi-squared on $p$ degrees of freedom. This threshold is only exceeded by 1% of observations. See the chi-square distribution table in the appendix.

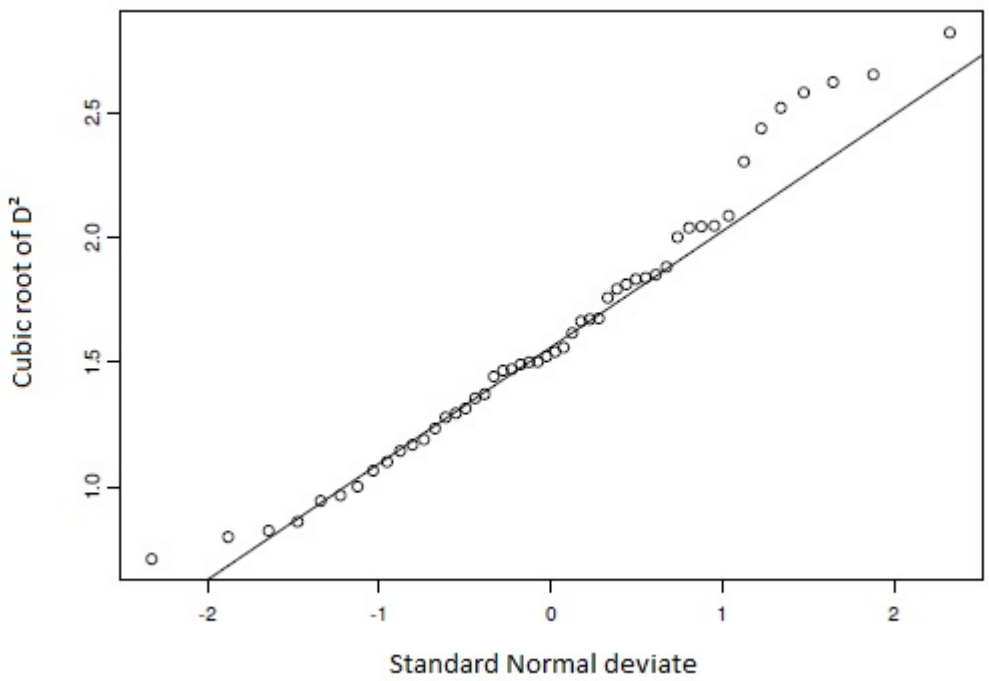
For small $n$, the following critical threshold should be used

$$D^2\left(\underset{\sim}{x}\right) \geq \frac{p(n^2-1)\ Q_F\ (0.99; p, n-p)}{n(n-p)} \tag{3.22}$$

with $Q_F(0.99; p, n-p)$ the 99% quantile (percentile) of the Snedecor $F$ distribution on $p$ and $n-p$ degrees of freedom. For example, with $n = 50$ and $p = 4$, the critical level is equal to 13.28 using the first formula and to $(4 \times 2499 \times 3.76)/(50 \times 46) = 16.33$ using the second formula.

The considerations above should be taken with some caution because if the multivariate sample contains extreme or outlying observations, they will not only affect the mean vector $\bar{x}$ and the dispersion matrix $\underset{\sim}{S}$ but also the Mahalanobis distance.

Some authors have suggested to use the transform $\sqrt[3]{D^2(x)}$ to normalize the distribution of the Mahalanobis distances. Thus, plotting these values on a Gaussian probability scale, they should distribute along a straight line. Isolated points or points departing from the straight line at the upper end will generally indicate the presence of outliers or the non-normality of the distribution. This is illustrated in the figure below for the Fisher iris setosa data.



## 3.7 Rao paradox

For a Normally distributed variable, the 95% reference interval which contains 95% of the values is determined by the boundaries $\bar{x} \pm 1.96s$. Thus, an observation $x$ is said to be *abnormal* if it falls outside the interval, i.e. if $|x - \bar{x}|/s \geq 1.96$.

In the multivariate situation, when the joint distribution of the $p$ variables $X_1, \ldots, X_p$ is Normal, we can determine the reference interval for each variable separately, as just described, i.e.

$$\bar{x}_i \pm 1.96 s_i \qquad (i = 1, \ldots, p)$$

with $\bar{x}_i$ and $s_i$ being the mean and standard deviation of variable $X_i$. These $p$ reference intervals form a "rectangular box" in the $p$-dimensional space $\mathbb{R}^p$. An observation $\underset{\sim}{x}$ may be considered *normal* if $|x_i - \bar{x}_i|/s_i \leq 1.96$ $(i = 1, \ldots, p)$, that is if it falls within the box.

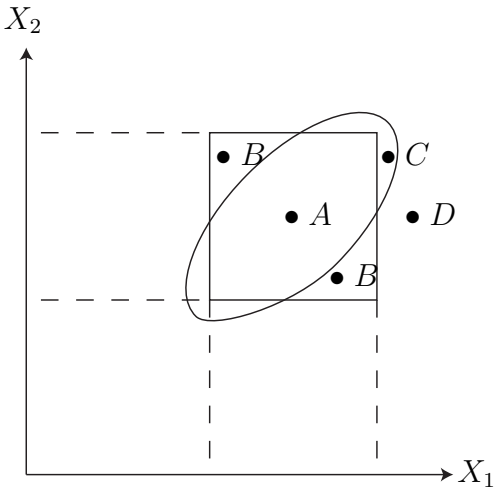
This approach however does not account for the relations between the variables studied. Therefore it is preferable to define an ellipsoidal reference region as mentioned previously. Then, an observation will considered *normal* if it falls within the ellipsoid containing 95% of the population, i.e. if

$$\left(\underset{\sim}{x} - \underset{\sim}{\bar{x}}\right)^T \underset{\sim}{S}^{-1} \left(\underset{\sim}{x} - \underset{\sim}{\bar{x}}\right) \leq Q_{\mathcal{X}^2}(0.95; p)$$

and abnormal if it falls outside the ellisoid (or ellipse in the case of two variables).

When comparing the rectangle and the ellipse, we are faced with the paradox of Rao in the sense that an observation can fall outside the ellise (it is "abnormal" from a multivariate standpoint) but within the rectangular box (it is therefore "normal" for each variable separately). This paradoxical situation can be explained by the associations between variables. Thus, considering weight and height, an individual may have a normal height and weight but when combining the two variables the individual may be considered as "abnormal" (e.g. obesity, thinness). By contrast, a subject may exhibit excessive height and weight, but when combined the subject looks in perfect harmony.

As seen in the figure below, several profiles are possible: normal from both univariate and multivariate perspective (case A), normal at the univariate level but abnormal at the multivariate level (case B), abnormal at the univariate level but normal at the multivariate level (case C), abnormal at both univariate and multivariate levels (case D).




## 3.8 Final remark

The mean vector $\bar{x}$ and the variance-covariance matrix $\underset{\sim}{S}$, as well as the correlation matrix $\underset{\sim}{R}$, can be calculated regardless of the dimension of the observation matrix $\underset{\sim}{X}_{n \times p}$. However, when $n \leq p$, that is when there are as many or fewer observations (rows) than variables (columns), the matrices $\underset{\sim}{S}$ and $\underset{\sim}{R}$ are singular. Thus, their determinant is equal to zero and matrices cannot be inverted. In consequence, the Mahalanobis distance cannot be computed either because $\underset{\sim}{S}^{-1}$ does not exist.

In general, it is highly recommended to have more observations than variables. Some have suggested to have at least 10 times more subjects than variables ($n/p \geq 10$) but this is just a heuristic rule.

Finally, note that from a population standpoint, the mean of vector $\underset{\sim}{X}$ writes $\underset{\sim}{\mu}$ and the variance-covariance matrix $\underset{\sim}{\sum}$. As in univariate statistics, theoretical (population) parameters are designated by Greek letters.

# Chapter 4

# Principal component analysis

## 4.1 Introduction

This chapter is dedicated to one of the most important multivariate statistical method, called *principal component analysis (PCA)*. It is basically a descriptive method designed to explore the structure of multvariate data. We first introduce PCA based on the variance-covariance matrix and thereafter PCA based on the correlation matrix. We end the chapter by briefly mentioning the *biplot* method.

## 4.2 Objectives

Statisticians have always strived to get a 2-dimensional representation of a cluster of points located in the $p$-dimensional space. While the human brain can easily grasp data distributed in the 3-dimensional space or possibly in the 4-dimensional space (like in relativity), it can hardly do so for higher spaces.

Principal component analysis of a multivariate dataset pursue several objectives.

1. Obtain a graphical representation of a multivariate sample, i.e. of the corresponding cloud of points in the $p$-dimensional space, with a minimal loss of information.

2. Highlight in the $p$-dimensional space some potential outlying (extreme or atypical) observations or clearly identified subsets of observations, known as "clusters".

3. Determine the actual dimension of the multivariate problem at hand, because in general variables are correlated and may contain redundant information.

4. Reduce the number of variables studied and replace them with a more limited number of factors which are weighted sums of the original variables. These factors can then be used in subsequent analyses.

In addition, the biplot method provides a graphical representation of the variables, thus enhancing exploratory multivariate data analysis, in particular by disentangling the relationships between variables.

## 4.3 Problem definition

Consider a vector $\underset{\sim}{X}^T = (X_1, \ldots, X_p)$ of $p$ variables and a random sample of size $n$ representing a matrix of observations $\underset{\sim}{X}_{n \times p}$. No restrictions are made for $n$ and $p$. Thus we may have $n < p$. Let $\underset{\sim}{\bar{x}}$ and $\underset{\sim}{S}$, respectively denote the mean vector and variance-covariance matrix of the sample, as described in Chapter 3. In the $p$-dimensional space $\mathbb{R}^p$, the sample can be seen as a set of points whose gravity center is $\underset{\sim}{\bar{x}}$ and shape given by $\underset{\sim}{S}$.

The statistical problem consists in finding a new vector of $p$ components, say $\underset{\sim}{Y}^T = (Y_1, \ldots, Y_p)$, with the following properties :

(i) variables $Y_i$, called *principal components* (PC), are weighted sums (i.e. linear combinations) of the original variables

$$Y_i = \underset{\sim}{a_i}^T \underset{\sim}{X} = a_{1i}X_1 + \ldots + a_{pi}X_p \tag{4.1}$$

where the weights (coefficients) are normed to unity, that is $a_{1i}^2 + \ldots + a_{pi}^2 = 1$, or equivalently, $||\underset{\sim}{a_i}|| = 1$ $(i = 1, \ldots, p)$

(ii) variables $Y_i$ have decreasing variability (variances),

$$\text{var } (Y_1) \geq \text{ var } (Y_2) \geq \ldots \geq \text{ var } (Y_p) \geq 0 \tag{4.2}$$

(iii) variables $Y_i$ are uncorrelated,

$$\text{corr } (Y_i, Y_j) = 0 \quad \forall i \neq j \tag{4.3}$$

In general we will only use the first two principal components, $Y_1$ et $Y_2$, because quite often the other components only add little complementary information. Thus, we reduce the problem from a $p$-dimensional to a 2-dimensional space, which was one of the major objectives of PCA.

Remember that, given a vector $\underset{\sim}{X}^T = (X_1, \ldots, X_p)$ of $p$ variables and a vector of numbers $\underset{\sim}{a}^T = (a_1, \ldots, a_p)$, the product $\underset{\sim}{a}^T\underset{\sim}{X} = a_1X_1 + \ldots + a_pX_p$ is a scalar, that is a univariate variable. Thus the scalar product (which is a weighted sum) allows a point $\underset{\sim}{x}^T = (x_1, \ldots, x_p)$ of the $p$-dimensional space to be visualized as a point $y = \underset{\sim}{a}^T\underset{\sim}{x}$ on the straight line, that is a number.

## 4.4 Principle of the method

### 4.4.1 First principal component

To derive $Y_1$, the first principal component, we need to find a vector of weights $\underset{\sim}{a}_1^T = (a_{11}, \ldots, a_{p1})$, normed to unity $||\underset{\sim}{a}_1|| = 1$, such that variable $Y_1 = \underset{\sim}{a}_1^T\underset{\sim}{X}$ has maximum variance. In other terms, if we calculate $Y_1$ for all $n$ multivariate observations $\underset{\sim}{x}_i$ $(i = 1, \ldots, n)$, we get $n$ values, say

$$\left\{ y_{11} = \underset{\sim}{a}_1^T\underset{\sim}{x}_1, \ldots, y_{n1} = \underset{\sim}{a}_1^T\underset{\sim}{x}_n \right\} \tag{4.4}$$

whose variance

$$V_1^2 = \sum_{k=1}^{n} (y_{k1} - \bar{y}_1)^2 / (n-1), \tag{4.5}$$

with $\bar{y}_1 = \sum_{k=1}^{n} y_{k1}/n$ the mean of the $n$ values, has to maximized.

To put it in another way, we need to look at the data cloud and find out in which direction the variability is the largest. Imagine a straight line crossing the data cloud on which we project all data points. Clearly points on the straight line will be more or less dispersed. Now if we rotate the line in all directions until the dispersion of projected points is maximal, then the corresponding direction gives in fact the first principal component $Y_1$. This direction also minimizes the sum of squared distances between observed and projected points (least square principle introduced by K. Pearson).

To find vector $\underset{\sim}{a}_1$, we need to determine the largest eigen value (latent root) of dispersion matrix $\underset{\sim}{S}$; in other terms, we need to solve the polynomial equation

$$|\underset{\sim}{S} - \lambda \underset{\sim}{I}| = 0 \tag{4.6}$$

Denoting $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$, the $p$ solutions of the polynomial of degree $p$ listed in decreasing order of magnitude, then $\underset{\sim}{a}_1$ is solution of the equation

$$\underset{\sim}{S}\underset{\sim}{a}_1 = \lambda_1 \underset{\sim}{a}_1 \tag{4.7}$$

so that $\underset{\sim}{a}_1$ is the eigen vector associated with the largest eigen value $\lambda_1$ of $\underset{\sim}{S}$ and normalized to unity.

At this point, we established the first principal component $Y_1 = \underset{\sim}{a}_1^T \underset{\sim}{X}$ and its variance is $V_1^2 = \lambda_1$.

### 4.4.2 Second principal component

To derive the second principal component $Y_2$, we need to find the vector of weights $\underset{\sim}{a}_2^T = (a_{12}, \ldots, a_{p2})$, normalized to unity $\|\underset{\sim}{a}_2\| = 1$, such that the variance of $Y_2 = \underset{\sim}{a}_2^T \underset{\sim}{X}$ is maximal after that of $Y_1$ and such that the correlation between $Y_1$ and $Y_2$ is equal to zero.

If we compute $Y_2$ for all $n$ multivariate observations $\underset{\sim}{x}_i$ $(i = 1, \ldots, n)$, we obtain $n$ values

$$\left\{ y_{12} = \underset{\sim}{a}_2^T \underset{\sim}{x}_1, \ldots, y_{n2} = \underset{\sim}{a}_2^T \underset{\sim}{x}_n \right\} \tag{4.8}$$

The variance of these values, say

$$V_2^2 = \sum_{k=1}^{n} (y_{k2} - \bar{y}_2)^2/(n-1) \tag{4.9}$$

with $\bar{y}_2 = \sum_{k=1}^{n} y_{k2}/n$ the mean of the $n$ values, should be maximized but less or equal to $V_1^2$.

Besides, the correlation between $Y_1$ and $Y_2$, say

$$r_{12} = \frac{\sum_{k=1}^{n} (y_{k1} - \bar{y}_1)(y_{k2} - \bar{y}_2)}{V_1 \cdot V_2}, \tag{4.10}$$

should be equal to 0 (knowing that $\underset{\sim}{a}_1$ is already known !).

From a geometrical standpoint, we look for the second most dispersed direction but orthogonal (perpendicular) to the first one.

To find $\underset{\sim}{a}_2$, we need to solve the equation

$$\underset{\sim}{\underset{\sim}{S}}\underset{\sim}{a}_2 = \lambda_2 \underset{\sim}{a}_2 \tag{4.11}$$

where $\lambda_2$ is the second largest eigen value of $\underset{\sim}{S}$. Therefore, $\underset{\sim}{a}_2$ is the eigen vector associated with $\lambda_2$ normalized to unity and with variance $V_2^2 = \lambda_2$. Moreover, $r_{12} = 0$.

### 4.4.3 Graphical display

The first two principal components, $Y_1$ et $Y_2$, define a referential of perpendicular axes and therefore a "principal" plane on which the $n$ multivariate observations $\underset{\sim}{x}_i$ $(i = 1, \ldots, n)$ can be represented. It suffices to plot on this plane the $n$ points

$$\begin{array}{c} (y_{11}, y_{12}) \\ (y_{21}, y_{22}) \\ \ldots \ \ldots \\ (y_{n1}, y_{n2}) \end{array} \tag{4.12}$$

defined in (4.4) and (4.8).

We obtain this way a 2-dimensional representation of the cloud of points located in $\mathbb{R}^P$. This was one goal of PCA.

### 4.4.4 Quality of the representation

How reliable is the 2-dimensional representation of the multivariate cluster? There should be indeed some loss of information in the reduction process (just like for a photography of a 3-dimensional object). So what is the "quality" of the picture ?

The process of constructing the principal components can proceed beyond the first two elements. Thus, to find the third principal component, $Y_3 = \underset{\sim}{a}_3^T \underset{\sim}{x}$, we need to solve the equation $\underset{\sim}{\underset{\sim}{S}}\underset{\sim}{a}_3 = \lambda_3 \underset{\sim}{a}_3$, where $\lambda_3$ is the third largest eigen value of $\underset{\sim}{S}$. This procedure can be pursued up to the $p$-th and last principal component $Y_p = \underset{\sim}{a}_p^T \underset{\sim}{X}$, where $\underset{\sim}{a}_p$ is the solution of equation $\underset{\sim}{\underset{\sim}{S}}\underset{\sim}{a}_p = \lambda_p \underset{\sim}{a}_p$ and $\lambda_p$ the smallest eigen value of $\underset{\sim}{S}$. If $\lambda_p = 0$, then the principal component $Y_p$ is a constant and has no statistical interest anymore since its variance $V_p^2 = 0$.

To assess the quality of the first two principal components, $Y_1$ et $Y_2$, we proceed as follows.

The total variability of the $p$ principal components is equal to $\lambda_1 + \lambda_2 + \ldots + \lambda_p$. But it is known that the sum of the eigen values of a matrix is equal to the trace of the matrix, that is the sum of its diagonal elements. Hence,

$$\lambda_1 + \ldots + \lambda_p = tr \underset{\sim}{S} = s_{11} + s_{22} + \ldots + s_{pp} \tag{4.13}$$

The quantity $s_{11} = s_1^2$ which is the variance of $X_1$ should not be confused with $V_1^2 = \lambda_1$ which is the variance of $Y_1$; and likewise for the other elements. The sums are the same but not the individual terms !

The contribution of the first principal component to the reduction of the multivariate problem can be evaluated by the ratio

$$\frac{\lambda_1}{\lambda_1 + \ldots + \lambda_p} = \frac{\lambda_1}{\mathrm{tr}\underset{\sim}{S}} \tag{4.14}$$

Expression (4.14) assesses the part (proportion) of variability explained by $Y_1$ with respect to the total variability. The closer the ratio to 1, the more important principal component $Y_1$ and the less the other ones. The same applies to each principal component $Y_i$ by computing the ratio $\lambda_i/\mathrm{tr}\ \underset{\sim}{S}$.

The quality of the first two principal components combined is given by the ratio

$$Q = \frac{\lambda_1 + \lambda_2}{\mathrm{tr}\ \underset{\sim}{S}} \tag{4.15}$$

and conversely the loss of information writes

$$\mathcal{P} = 1 - Q = \frac{\lambda_3 + \ldots + \lambda_p}{\mathrm{tr}\ \underset{\sim}{S}} \tag{4.16}$$

Of note, in practice we know the sum of the eigen values before computing them because it is equal to $\mathrm{tr}\underset{\sim}{S}$. Today statistical packages allow you to calculate all principal components or preferably to fix the number of principal components you want to use. If, at some point in the calculation process, one eigen value is equal to 0, then all the following eigen values will also be equal to 0. This indicates that there are too many variables in the multivariate problem and that some variables are linear combinations of the others. For example, if $X_1$ is the length of a rectangle, $X_2$ its width and $X_3$ its perimeter, one eigen value will be null since $X_3 = 2(X_1 + X_2)$.

## 4.5 Interpretation of principal components

Principal components $Y_1$ et $Y_2$ can also be written

$$\begin{aligned} Y_1 &= \underset{\sim}{a}_1^T(\underset{\sim}{X} - \underset{\sim}{\bar{x}}) \\ Y_2 &= \underset{\sim}{a}_2^T(\underset{\sim}{X} - \underset{\sim}{\bar{x}}) \end{aligned} \tag{4.17}$$

by subtracting the mean to obtain a graphical representation of the $n$ points on the plane $(Y_1, Y_2)$ centered on the origin.

There is a need to given sense to principal components, in other words to interpret them. This is a difficult exercise, because it can be very subjective and prone to fallacy. Prudency should always prevail. When variables $X_1, \ldots, X_p$ are biometric characteristics (for example, subject or object dimensions), then in general $Y_1$ represents the size of the objects, whereas $Y_2$ features their shape.

To facilitate the interpretation of principal components, it is advised to calculate the correlations between the principal components and the original variables. It should be remembered that principal components are not correlated to each other. It is easily shown that the correlation between principal component $Y_j$ and variable $X_i$ is given by the expression

$$\text{corr}(X_i, Y_j) = \frac{a_{ij}\sqrt{\lambda_j}}{s_i}(i, j = 1 \ldots p) \tag{4.18}$$

with $\lambda_j$ the j-th eigen value of $\underset{\sim}{S}$, $a_{ij}$ the coefficient (weight) of variable $X_i$ in the principal component $Y_j$ and $s_i$ the standard deviation of variable $X_i$.

Thus, knowing the correlations between the first principal component and each of the variables $X_1, \ldots, X_p$, we may admit that the principal component essentially represents those variables which are highly correlated with it (whether positively or negatively). By contrast, it does not express those variables weakly correlated with it.

Principal component analysis based on the dispersion matrix $\underset{\sim}{S}$ becomes problematic when the variances on the diagonal of the matrix are markedly different (factor 1 to 100). Indeed, the variables with a large variance prevail (in the total sum) and PCA overestimates the weights of these variables, leaving the other ones almost as unimportant. In such situations, it is recommended to perform PCA on the matrix of correlations $\underset{\sim}{R}$.

## 4.6 PCA on the correlation matrix

### 4.6.1 Standard deviate vector

Principal component analysis on the correlation matrix $\underset{\sim}{R}$ is based on the fact that matrix $\underset{\sim}{R}$ is the variance-covariance matrix of the vector of standard deviates $\underset{\sim}{Z}^T = (Z_1, \ldots, Z_p)$, where

$$Z_j = \frac{X_j - \bar{x}_j}{s_j} \quad (j = 1, \ldots, p.) \tag{4.19}$$

In matrix notation, we have

$$\underset{\sim}{Z} = \underset{\sim}{D}^{-1}(\underset{\sim}{X} - \underset{\sim}{\bar{x}}) \tag{4.20}$$

with $\underset{\sim}{D} = \text{diag}\ (s_1, \ldots, s_p)$ and $\underset{\sim}{\bar{x}}$ the mean vector.

Thus, if we calculate the dispersion matrix of the standard deviates of the observations, namely

$$\underset{\sim_{n\times p}}{Z} = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ z_{n1} & z_{n2} & \cdots & z_{np} \end{bmatrix} \tag{4.21}$$

with $z_{ij} = (x_{ij} - \bar{x}_j)/s_j$, it is seen that $\underset{\sim}{S}_z = \underset{\sim}{R}$ and that $\underset{\sim}{\bar{z}} = \underset{\sim}{0}$.

### 4.6.2 Derivation of principal components

To derive the principal components,

$$Y_j = \underset{\sim}{a}_j^T \underset{\sim}{Z} \quad (j = 1, \ldots, p) \tag{4.22}$$

we need to find the eigen values (in decreasing order of magnitude) of matrix $\underset{\sim}{R}$ and the corresponding eigen vectors.

Denote again by $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ the $p$ eigen values of matrix $\underset{\sim}{R}$, i.e. the solutions of the polynomial equation

$$|\underset{\sim}{R} - \lambda \underset{\sim}{I}| = 0 \tag{4.23}$$

The first principal component $Y_1$ writes

$$Y_1 = \underset{\sim}{a}_1^T \underset{\sim}{Z}$$

where $\underset{\sim}{a}_1$ is the solution of the matrix equation $\underset{\sim}{\underset{\sim}{R}}\underset{\sim}{a}_1 = \lambda_1 \underset{\sim}{a}_1$.

The second principal component is given by $Y_2 = \underset{\sim}{a}_2^T \underset{\sim}{Z}$, where $\underset{\sim}{a}_2$ is the solution of the matrix equation $\underset{\sim}{\underset{\sim}{R}}\underset{\sim}{a}_2 = \lambda_2 \underset{\sim}{a}_2$. This procedure goes on until all solutions have been obtained.

### 4.6.3 Quality

The quality of the first two principal components is given by the expression

$$Q = \frac{\lambda_1 + \lambda_2}{\mathrm{tr}\underset{\sim}{R}} \tag{4.24}$$

with $\mathrm{tr}\underset{\sim}{R} = \lambda_1 + \ldots + \lambda_p$. Besides, since all diagonal elements of the correlation matrix $\underset{\sim}{R}$ are equal to 1, we have $\mathrm{tr}\underset{\sim}{R} = p$, the number of variables of the multivariate problem. It follows that expression (4.24) becomes

$$Q = \frac{\lambda_1 + \lambda_2}{p} \tag{4.25}$$

and the loss of information writes

$$\mathcal{P} = 1 - Q = \frac{\lambda_3 + \ldots + \lambda_p}{p} \tag{4.26}$$

### 4.6.4 Interpretation

To interpret the principal components $Y_1, \ldots, Y_p$, we calculate as before the correlations between each of them with the original variables $X_1, \ldots, X_p$ by observing that $\mathrm{corr}(Y_j, X_i) = \mathrm{corr}(Y_j, Z_i)$ since $Z_i$ is a linear transform of $X_i$. Consequently,

$$\mathrm{corr}(Y_j, X_i) = a_{ij}\sqrt{\lambda_j} \quad (i, j = 1, \ldots, p) \tag{4.27}$$

When performing PCA on correlation matrix $\underset{\sim}{R}$, all variables are equally important since their variances are equal to 1 through the standard deviate transformation $\underset{\sim}{Z}$.

It should be remarked that there is no relationship between principal components derived from the dispersion matrix $\underset{\sim}{S}$ and those from the correlation matrix $\underset{\sim}{R}$. This may seem somewhat surprising as there is a relationship between the two matrices.

## 4.7 Example : Fisher iris dataset

We shall illustrate the use of PCA on the Fisher iris setosa dataset (see Annex I). The observation matrix consists of $n = 50$ iris setosa flowers on which $p = 4$ variables were measured : $X_1$ = sepal length, $X_2$ = sepal width, $X_3$ = petal length and $X_4$ = petal width. Thus we have an observation matrix of dimension $n \times p = 50 \times 4$.

### 4.7.1 PCA on the variance-covariance matrix

Consider first PCA on the variance-covariance matrix $\underset{\sim}{S}$. As described in section 3.5.

$$\underset{\sim}{S} = \begin{pmatrix} 12.425 & & & \\ 9.922 & 14.369 & & \\ 1.636 & 1.170 & 3.016 & \\ 1.033 & 0.930 & 0.607 & 1.111 \end{pmatrix}.$$

The trace (sum of diagonal elements, i.e. sum of the variances) of matrix $\underset{\sim}{S}$ is equal to $tr\underset{\sim}{S} = 30.92$. The eigen values and eigen vectors of $\underset{\sim}{S}$ are given below. It is seen that the first eigen value $\lambda_1 = 23.65$ already accounts for more than 75% of the total variance. Further, based on the first two eigen values, we obtain a 2-dimensional representation of the 4-dimensional observation sample with a quality of 88.4%.

| Eigen value | Quality (%) | Cumulative quality |
|---|---|---|
| $\lambda_1 = 23.65$ | 76.47 | 76.47 |
| $\lambda_2 = 3.69$ | 11.94 | 88.41 |
| $\lambda_3 = 2.68$ | 8.67 | 97.08 |
| $\lambda_4 = 0.90$ | 2.92 | 100.0 |

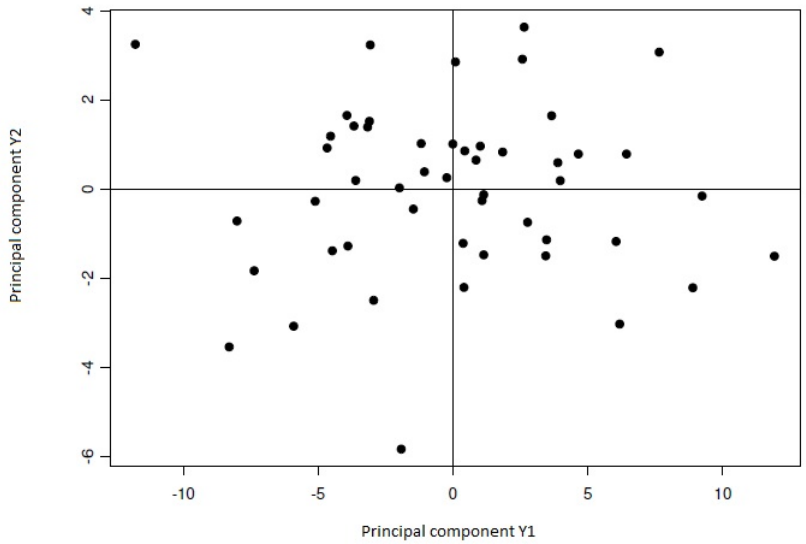
| | Eigen vector | | | |
|---|---|---|---|---|
| Variable | $\underset{\sim}{a_1}$ | $\underset{\sim}{a_2}$ | $\underset{\sim}{a_3}$ | $\underset{\sim}{a_4}$ |
| $X_1$ | 0.669 | -0.598 | 0.440 | 0.036 |
| $X_2$ | 0.734 | 0.621 | -0.275 | 0.020 |
| $X_3$ | 0.097 | -0.490 | -0.832 | 0.234 |
| $X_4$ | 0.0636 | -0.131 | -0.195 | -0.970 |

As an illustration, the first principal component writes $Y_1 = 0.669(X_1 - 50.1) + 0.734(X_2 - 34.3) + 0.097(X_3 - 14.6) + 0.0636(X_4 - 2.46)$ and the second is given by $Y_2 = -0.598(X_1 - 50.1) + 0.621(X_2 - 34.3) - 0.490(X_3 - 14.6) - 0.131(X_4 - 2.46)$.

The correlation matrix between variables and principal components is given below. Observe that the first principal component is positively correlated with all 4 variables but expresses basically only the sepals (which have large variances). By contrast, the second principal component opposes petal width against the other variables.

| Correlations between variables and principal components | | | | |
|---|---|---|---|---|
| | $Y_1$ | $Y_2$ | $Y_3$ | $Y_4$ |
| $X_1$ | 0.923 | 0.326 | -0.204 | -0.010 |
| $X_2$ | 0.942 | -0.315 | 0.119 | -0.005 |
| $X_3$ | 0.270 | 0.542 | 0.785 | -0.131 |
| $X_4$ | 0.293 | 0.239 | 0.303 | 0.875 |

The figure below displays the 50 iris setosa flowers in the 2-dimensional space defined by the first two principal components $Y_1$ et $Y_2$. As a reminder, the mean has been subtracted for each variable so that the points are centered around the origin.



### 4.7.2 PCA on the correlation matrix

Results of PCA applied to the correlation matrix $\underset{\sim}{R}$ are reported below. The correlation matrix is given by (see section 3.5)

$$\underset{\sim}{R} = \begin{pmatrix} 1.00 & & & \\ 0.743 & 1.00 & & \\ 0.267 & 0.177 & 1.00 & \\ 0.278 & 0.233 & 0.332 & 1.00 \end{pmatrix}$$

The first principal component represents 51% of the total variability since $tr\underset{\sim}{R} = 4$, and when adding the second principal component we reach a quality of over 75%. Thus, we have quite a satisfactory representation of the iris data since the loss of information is less than 25%.

| Eigen value | Quality (%) | Cumulative quality |
|---|---|---|
| $\lambda_1 = 2.06$ | 51.46 | 51.46 |
| $\lambda_2 = 1.02$ | 25.55 | 77.02 |
| $\lambda_3 = 0.67$ | 16.70 | 93.71 |
| $\lambda_4 = 0.25$ | 6.29 | 100.00 |

| | Eigen vector | | | |
|---|---|---|---|---|
| Variable | $\underset{\sim}{a}_1$ | $\underset{\sim}{a}_2$ | $\underset{\sim}{a}_3$ | $\underset{\sim}{a}_4$ |
| $X_1$ | 0.604 | -0.335 | 0.067 | 0.720 |
| $X_2$ | 0.576 | -0.441 | 0.001 | -0.689 |
| $X_3$ | 0.375 | 0.627 | 0.677 | -0.087 |
| $X_4$ | 0.403 | 0.548 | -0.733 | -0.015 |

The matrix of correlations between variables and principal components is given below:

| Correlations between variables and principal components | | | | |
|---|---|---|---|---|
| | $Y_1$ | $Y_2$ | $Y_3$ | $Y_4$ |
| $X_1$ | 0.867 | -0.339 | 0.055 | 0.361 |
| $X_2$ | 0.826 | -0.446 | 0.001 | -0.345 |
| $X_3$ | 0.539 | 0.634 | 0.553 | -0.044 |
| $X_4$ | 0.578 | 0.554 | -0.599 | -0.007 |

The first principal component writes

$$Y_1 = 0.604\ Z_1 + 0.576\ Z_2 + 0.375\ Z_3 + 0.403\ Z_4,$$

where $Z_1 = (X_1 - 50.1)/3.52, Z_2 = (X_2 - 34.3)/3.79, Z_3 = (X_3 - 14.6)/1.74$ and $Z_4 = (X_4 - 2.46)/1.05$.

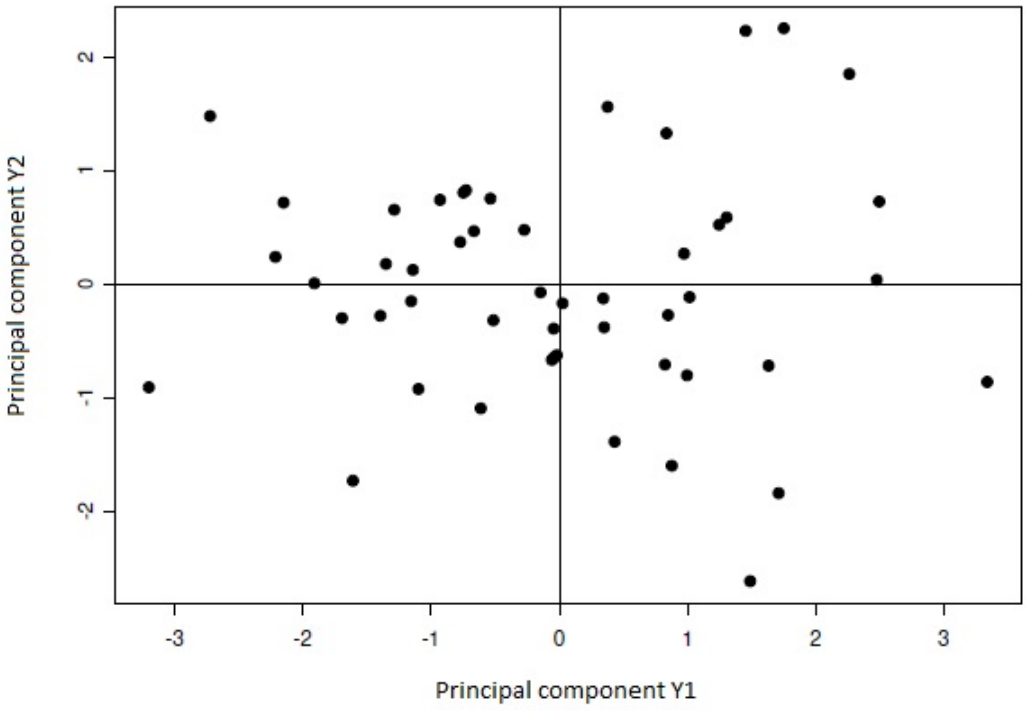
It is a weighted sum of the four variables and all coefficients are positive; the larger each sepal and petal measures, the greater the principal component. The latter is therefore a "size" parameter.

The second principal component

$$Y_2 = -0.335\ Z_1 - 0.441\ Z_2 + 0.627\ Z_3 + 0.548\ Z_4$$

is a "contrast" between petals and sepals. It therefore opposes sepals and petals and can therefore be interpreted as a "shape" parameter.

The figure below displays the 50 iris setosa flowers in the plane defined by the first two principal components (size and shape). The X-axis represents the size of the flowers, so that when moving from left to right, the flowers become increasingly bigger. By contrast on the Y-scale when moving top down, sepals become smaller and smaller but petal larger and larger. Thus their shape is changing.



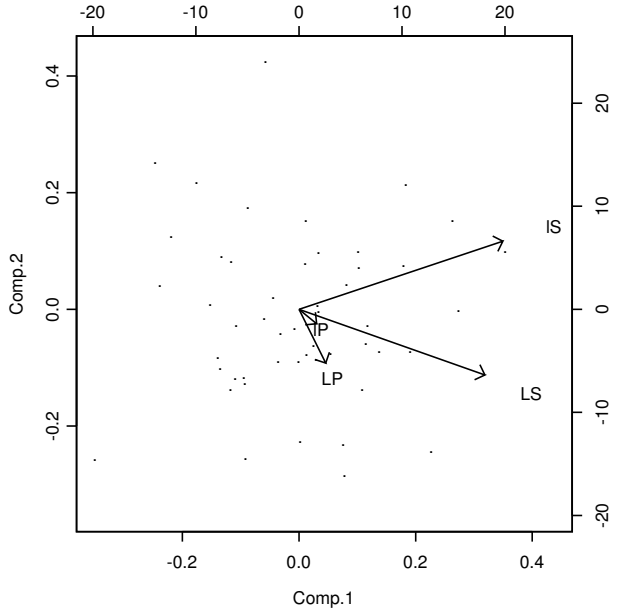
## 4.8 Biplot

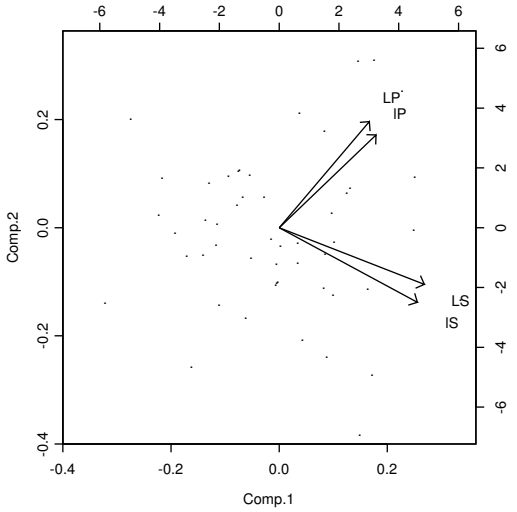
The biplot method allows representing on the same plane both the $n$ observations and the $p$ variables. There is actually a double system of (X,Y) axes, hence the name of *biplot*. Focussing on variables, each variable is represented by an oriented vector of a given length, starting at the origin (0,0) and pointing in a given direction. More generally, biplots may be interpreted as follows :

- variables pointing in the same direction are positively correlated;
- variables pointing in opposite directions are negatively correlated;
- variables pointing in perpendicular (orthogonal) directions are not correlated.

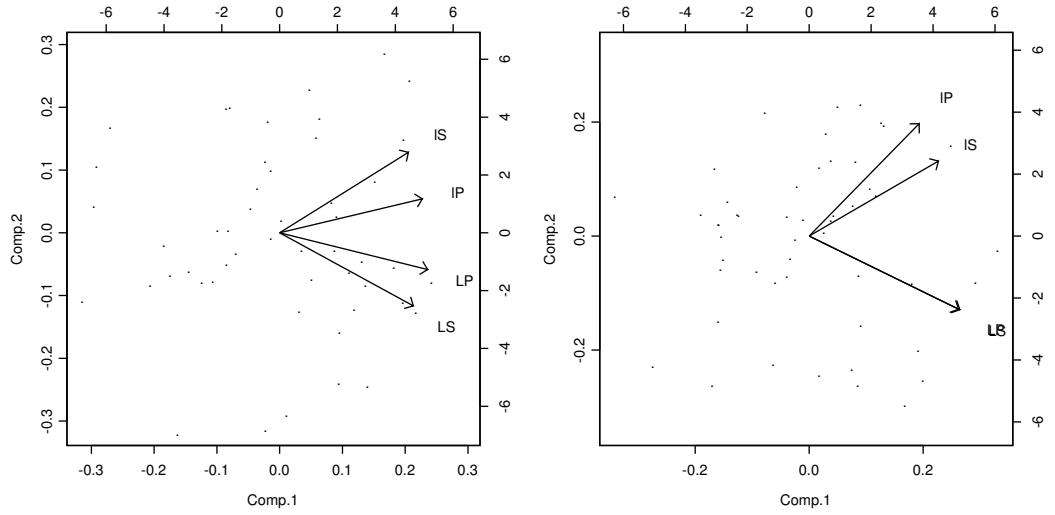
In addition, when moving in the direction of the vector, one goes from low to high values of the variable. As for the length of the vector, it represents the importance of the variable. As for PCA, the biplot metod applies to both the dispersion matix $\underset{\sim}{S}$ and the correlation matrix $\underset{\sim}{R}$. The latter however is usually preferred for the same reasons as PCA.

The figures below illustrate the biplot method applied to Fisher iris setosa dataset based on matrices $\underset{\sim}{S}$ and $\underset{\sim}{R}$, respectively.






It is seen that variables related to sepals (width and length) are highly correlated with each other as they point in the same direction. The same applies for length and width of petals. However, sepal and petal chracateristics are perpendicular and therefore unrelated to each other. Amazingly, biplots based on matrix $R$ for iris versicolor and iris virginica flowers (see figures below) demonstrate that sepal length is closely correlated with petal length and likewise for sepal and petal width. By contrast, lengths and widths are not related to each other. This highlights marked differences in the shape of iris setosa on the one hand, iris versicolor and virginica on the other hand.



# Chapter 5

# Multiple regression and correlation

## 5.1 Introduction

Multiple regression (MR) is one of the most used method in statistics. It is not immediately seen as being part of multivariate statistics because it studies the relationship (association) between the mean of a (dependent) random variable and a set of (independent) experimental factors (covariates) whose values have been fixed by the investigator. For instance, in a laboratory experiment, the length of a chemical reaction may be measured for given values of the pH, the temperature and the concentration of a solution. The experimental conditions are fixed (specified) but the duration of the reaction is observed as its value is not known in advance. In this chapter, we consider multiple regression from a multivariate statistics perspective as it makes extensive use of matrix calculus. We shall study its properties and look at associated hypothesis testing. When all variables (dependent and independent) are observed simultaneously, we shall introduce the concept of multiple correlation. We close the chapter with a few words on variable selection methods in multiple regression.

## 5.2 Problem setting

Consider the vector of $p+1$ variables $(Y, \underset{\sim}{X}^T) = (Y, X_1, \ldots, X_p)$ where $Y$ is the "dependent" variable and $\underset{\sim}{X}$ the vector of "independent" variables. This wording is quite unfortunate as the variables $X_1, \ldots, X_p$ are generally not independent of each other, to the contrary they are most often correlated.

The adjective "explanatory" would be a better term for variables of vector $\underset{\sim}{X}$. They are also often called factors, covariates or criteria !

Here we assume that variable $Y$ is quantitative (continuous) and follows a Normal distribution. There are no restrictions about the number and the nature (type) of explanatory variables $X_1, \ldots, X_p$.

There is a need to distinguish between two situations:

1. Values of vector $\underset{\sim}{X}$ are fixed and variable $Y$ is observed. This is typically a case of *multiple regression*.

2. Variable $Y$ and vector $\underset{\sim}{X}$ are observed together. This is typically a case for *multiple correlation* but also multiple regression.

In both situations, data are presented as an $n \times (p+1)$-dimensional matrix

$$\underset{\sim}{X}_{n\times(p+1)} = \begin{bmatrix} y_1 & x_{11} & \ldots & x_{1p} \\ y_2 & x_{21} & \ldots & x_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ y_n & x_{n1} & \ldots & x_{np} \end{bmatrix} \tag{5.1}$$

## 5.3 Multiple regression

### 5.3.1 Model definition

Let us first consider multiple regression (MR). Basically we look for a relation between the dependent variable $Y$ and the independent variables $X_1, \ldots, X_p$. In statistics, this usually means the assumption of a linear model, specifically

$$E(Y|\underset{\sim}{X} = \underset{\sim}{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

or

$$E(Y|\underset{\sim}{x}) = \beta_0 + \underset{\sim}{\beta}^T \underset{\sim}{x} \tag{5.2}$$

where $E(Y|\underset{\sim}{X} = \underset{\sim}{x})$ is the mean of $Y$ for any fixed value of the vector $\underset{\sim}{X} = \underset{\sim}{x}$, $\underset{\sim}{\beta}^T = (\beta_1, \ldots, \beta_p)$ is the vector of regression coefficients and $\beta_0$ the constant term or "intercept". Note that the linear relation concerns the mean of $Y$ as it would be impossible for all values of $Y$for a given $\underset{\sim}{X}$ to be on the hyperplane.

Indeed equation (5.2) defines a hyperplane in the $p+1$-dimensional space $\mathbb{R}^{p+1}$. When $p = 1$, this is a straight line $E(Y|x) = \beta_0 + \beta_1 x$ and when $p = 2$ this is the equation of a plane $E(Y|\underset{\sim}{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

Equation (5.2) defines the "Multiple regression of $Y$ on $\underset{\sim}{X}$".

When the regression coefficients $\beta_0, \beta_1, \ldots, \beta_p$ are known, then, for any given $\underset{\sim}{X} = \underset{\sim}{x}$, we can calculate the mean of $Y$ and hence get a prediction of the value $Y$ for that $\underset{\sim}{x}$.

In terms of individual values of $Y$, equation (5.2) can be written

$$Y|\underset{\sim}{x} = \beta_0 + \underset{\sim}{\beta}^T \underset{\sim}{x} + \epsilon \tag{5.3}$$

where $\epsilon$ is a "random" error term with mean 0 and variance equal to $\sigma^2$ for all $\underset{\sim}{x}$. The term "$\beta_0 + \underset{\sim}{\beta}^T \underset{\sim}{x}$" is the fixed (deterministic) component of the model and $\epsilon$ the random (unexplained) component.

Referring to observation matrix (5.1), equation (5.3) can also be written :

$$\begin{aligned} y_i &= \beta_0 + \underset{\sim}{\beta}^T \underset{\sim}{x}_i + \epsilon_i \qquad (i = 1, \ldots, n) \\ &= \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i \end{aligned} \tag{5.4}$$

where $\underset{\sim}{\epsilon}^T = (\epsilon_1, \ldots, \epsilon_n)$ is the vector of random errors, also called residuals, namely the differences between the observed values $y_i$ and the model $\beta_0 + \underset{\sim}{\beta}^T \underset{\sim}{x}_i$ since

$$\epsilon_i = y_i - (\beta_0 + \underset{\sim}{\beta}^T \underset{\sim}{x}_i) \qquad (i = 1, \ldots, n) \tag{5.5}$$

The linear model (5.2) comprises $p + 1$ unknowns which are the parameters of the model, respectively $\beta_0, \beta_1, \ldots, \beta_p$. These parameters need to be estimated from the data matrix (5.1).

### 5.3.2 Parameter estimation

To estimate the parameters of the model (5.2), we use the *least square method*. Specifically, we search for the values of $\beta_0$ and $\underset{\sim}{\beta}$ which minimize the sum of squares of the residuals, i.e. $\underset{\sim}{\epsilon}^T \underset{\sim}{\epsilon} = \epsilon_1^2 + \ldots + \epsilon_n^2$.

Given equation (5.5), we have to find, over the space of parameter values $\beta_0, \underset{\sim}{\beta}$, the minimum of the function

$$\sum_{i=1}^{n} \left[ y_i - (\beta_0 + \underset{\sim}{\beta}^T \underset{\sim}{x}_i) \right]^2 \tag{5.6}$$

This is done by taking the derivatives of the function with respect to all parameters and make them equal to 0. It should be observed that by

derivating the function by $\beta_0$, we get

$$\sum_{i=1}^{n}\left[y_i - (\beta_0 + \underset{\sim}{\beta}^T \underset{\sim}{x}_i)\right] = 0$$

or

$$\sum_{i=1}^{n} y_i - n\beta_0 - \underset{\sim}{\beta}^T \sum_{i=1}^{n} \underset{\sim}{x}_i = 0$$

Dividing each term by $n$, it is seen that the solution $\beta_0$ has to satisfy the equation

$$\beta_0 = \bar{y} - \underset{\sim}{\beta}^T \underset{\sim}{\bar{x}} \tag{5.7}$$

By replacing $\beta_0$ by its value (5.7) in expression (5.6), the function writes

$$\sum_{i=1}^{n}\left[(y_i - \bar{y}) - \underset{\sim}{\beta}^T(\underset{\sim}{x}_i - \underset{\sim}{\bar{x}})\right]^2 \tag{5.8}$$

The derivative of the function (5.8) by vector $\underset{\sim}{\beta}$ gives the matrix equation

$$\underset{\sim}{A} \cdot \underset{\sim}{\beta} = \underset{\sim}{c} \tag{5.9}$$

where $\underset{\sim}{A}$ is the matrix of sums of squares and cross-products of vector $\underset{\sim}{X}$ (see equation (3.12)) and $\underset{\sim}{c}$ the vector of sums of cross-products of $Y$ and each variable of vector $\underset{\sim}{X}$. In other terms, $\underset{\sim}{A} = (n-1)\,\underset{\sim}{S}$, where $\underset{\sim}{S}$ is the dispersion matrix of vector $\underset{\sim}{X}$ and $\underset{\sim}{c}^T = (c_1, \ldots, c_p)$ where $c_j = \sum_{i=1}^{n}(y_i - \bar{y})(x_{ij} - \bar{x}_j)$.

In practice, we need to calculate $\sum x_i, \sum x_i^2, \sum x_i x_j$ as we did for matrix $\underset{\sim}{S}$ but also $\sum y, \sum y^2$ et $\sum y x_i$. Then, the elements of matrix $\underset{\sim}{A}$ and of vector $\underset{\sim}{c}$ are respectively equal to

$$a_{ij} = \sum x_i x_j - \frac{(\sum x_i)\,(\sum x_j)}{n}$$

and (5.10)

$$c_j = \sum y x_j - \frac{(\sum y)\,(\sum x_j)}{n}$$

$(i, j = 1, \ldots, p)$.

Finally, the solution $\hat{\underset{\sim}{\beta}}$ also denoted $b$ of equation (5.9) can be obtained by inverting matrix $\underset{\sim}{A}$ and calculating the expression

$$\underset{\sim}{b} = \underset{\sim}{A}^{-1}\underset{\sim}{c} \tag{5.11}$$

Using relation (5.7), the intercept is given by

$$b_0 = \bar{y} - \underset{\sim}{b}^T \underset{\sim}{\bar{x}} \tag{5.12}$$

Thus, the estimated multiple regression equation writes

$$\hat{Y} = b_0 + \underset{\sim}{b}^T \underset{\sim}{x} \tag{5.13}$$

### 5.3.3 Analysis of variance

In multiple regression, the total variability of the observations of variable $Y$ can be decomposed in two parts : one related to the model (hyperplane) and one related to the dispersion of data points around the hyperplane (residuals).

The total sum of squares can be written, using (5.13) and letting $\hat{Y}_i = b_0 + \underset{\sim}{b}^T \underset{\sim_i}{x}$ ,

$$\begin{aligned} \sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(y_i - \hat{Y}_i + \hat{Y}_i - \bar{y})^2 \\ &= \sum_{i=1}^{n}\left(y_i - \hat{Y}_i\right)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{y})^2 \end{aligned} \tag{5.14}$$

because the double product vanishes by summation.

The term on the left-hand side is the *total sum of squares, SST*

$$\text{SST} = \sum y^2 - \frac{\left(\sum y\right)^2}{n} \tag{5.15}$$

The second term on the right-hand side is the *regression sum of squares, SSR*

$$\text{SSR} = \underset{\sim}{b}^T \underset{\sim}{c} = b_1 c_1 + \ldots + b_p c_p \tag{5.16}$$

Lastly the first term of the rhs is the *residual sum of squares* or error sum of squares noted SSE. Thus,

$$\text{SSE} = \text{SST} - \text{SSR} \tag{5.17}$$

The degrees of freedom associated with each sum of squares are respectively equal to $n-1$ for SST, $p$ for SSR and $n-p-1$ for SSE. By deviding each sum of squares by its degrees of freedom, we derive the mean squares. In particular,

$$\text{MSR} = \text{SSR/p} \tag{5.18}$$

is the *mean square due to the regression*, that is the amount of total variability explained by the model (5.2), and

$$s^2 = \text{SSE}/(n-p-1) \tag{5.19}$$

is the *residual variability* not explained by the model. It corresponds to the estimate of $\sigma^2$ defined in the linear model and which measures the variability of observations $y_1, \ldots, y_n$ around the hyperplane. If all observations were located on the hyperplane, then $s^2$ would be equal to 0.

If $\underset{\sim}{\beta} = \underset{\sim}{0}$, that is if the mean of $Y$ does not depend on vector $\underset{\sim}{X}$, the ratio

$$F = \frac{\text{MSR}}{s^2} \tag{5.20}$$

is distributed as a Snedecor $F$ test on $p$ and $n-p-1$ degrees of freedom.

It follows that to test the null hypothesis $H_0 : \underset{\sim}{\beta} = \underset{\sim}{0}$ against the alternative hypothesis $H_1 : \underset{\sim}{\beta} \neq \underset{\sim}{0}$, we will reject $H_0$ at the $\alpha$ significance level if

$$F \geq Q_F(1-\alpha; p, n-p-1) \tag{5.21}$$

with $Q_F(1-\alpha; p, n-p-1)$ the $(1-\alpha)$-quantile of the Snedecor $F$ distribution on $p$ and $n-p-1$ degrees of freedom.

The analysis of variance (ANOVA) table is displayed as follows:

| Variability | Sum of squares | Degres of freedom | Mean square | $F$ |
|---|---|---|---|---|
| Regression | SSR | $p$ | MSR | $F = \frac{\text{MSR}}{s^2}$ |
| Residual | SSE | $n-p-1$ | $s^2$ | |
| Total | SST | $n-1$ | | |

In conclusion, the analysis of variance allows us to see whether the multiple regression makes sense, that is whether $\underset{\sim}{\beta}$ is not equal to 0. Remark that $\underset{\sim}{\beta} \neq \underset{\sim}{0}$ does not imply that $\beta_i \neq 0$ pour $\forall i$.

### 5.3.4 Utility of explanatory variables

If the null hypothesis $H_0 : \underset{\sim}{\beta} = \underset{\sim}{0}$ is rejected, we may assess the utility of each of the $p$ independent (explanatory) variables of the model.

To test the hypothesis $H_0 : \beta_i = 0$ versus $H_1 : \beta_i \neq 0$ $(i = 1, \ldots, p)$, it suffices to calculate the criterion

$$t = \frac{b_i}{\text{SE}(b_i)} \tag{5.22}$$

distributed under $H_0$ as a Student $t$ on $n-p-1$ degrees of freedom. In this expression, the standard error of $b_i$ is given by

$$\mathrm{SE}(b_i) = s\sqrt{a^{ii}} \tag{5.23}$$

where $s = \sqrt{s^2}$ and $a^{ii} = (\underset{\sim}{A}^{-1})_{ii}$, is the diagonal element of the inverse matrix used in (5.11). Indeed, it is easy to show that var $\underset{\sim}{b} = s^2\underset{\sim}{A}^{-1}$.

We reject $H_0 : \beta_i = 0$ if

$$|t| \geq Q_t(1 - \frac{\alpha}{2}; n-p-1) \tag{5.24}$$

with $Q_t(1-\frac{\alpha}{2}; n-p-1)$ the $(1-\frac{\alpha}{2})$-quantile of the Student $t$ distribution on $n-p-1$ degrees of freedom; otherwise we don't reject $H_0$.

When the null hypothesis is not rejected, we can assume $\beta_i = 0$, so that variable $X_i$ is no longer needed in the model. The latter can therefore be simplified.

### 5.3.5 Quality of the multiple regression

The quality of the multiple regression model (5.2) can be assessed by the criterion (see ANOVA table)

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} \tag{5.25}$$

called the *multiple coefficient of determination*. We have $0 \leq R^2 \leq 1$. and the closer $R^2$ to 1, the better the quality of the regression model since SSR is close to SST and SSE $\approx 0$.

$R^2$ represents the proportion (percentage) of the variability of the observations of $Y$ explained by the explanatory variables $X_1, \ldots, X_p$.

### 5.3.6 Prediction precision

Once the parameters of the model have been estimated, we can predict $Y$ for a given value $\underset{\sim}{X} = \underset{\sim}{x}$ by using equation (5.13), namely $\hat{Y} = b_0 + \underset{\sim}{b}^T\underset{\sim}{x} = \bar{y} + \underset{\sim}{b}^T(\underset{\sim}{x} - \underset{\sim}{\bar{x}})$. The statistical precision of this estimation, i.e. its standard error SE $(\hat{Y})$, is given by the square root of expression

$$\text{var } \hat{Y} = s^2\left\{\frac{1}{n} + (\underset{\sim}{x} - \underset{\sim}{\bar{x}})^T\underset{\sim}{A}^{-1}(\underset{\sim}{x} - \underset{\sim}{\bar{x}})\right\} \tag{5.26}$$

Then, the 95% confidence interval for $E(Y|\underset{\sim}{x})$ writes

$$\hat{Y} \pm 1.96 \text{ SE}(\hat{Y}) \tag{5.27}$$

Its is easily seen that the confidence interval is minimum when $\underset{\sim}{X} = \underset{\sim}{\bar{x}}$ as the second term of (5.26) vanishes and SE $(\hat{Y}) = s/\sqrt{n}$.

## 5.4 Multiple correlation coefficient

### 5.4.1 Definition

Suppose that $Y$ and $\underset{\sim}{X}$ are observed together (covariates are no longer fixed), indicating that variable $Y$ and vector $\underset{\sim}{X}$ are random. Then, all variables are on the same level and none is really priviledged.

Nonetheless, we can still calculate as before the regression of $Y$ on $\underset{\sim}{X}$, specifically

$$E(Y|\underset{\sim}{x}) = \beta_0 + \underset{\sim}{\beta}^T \underset{\sim}{x}.$$

This expression defines a new variable, denoted by $\hat{Y} = E(Y|\underset{\sim}{x})$, which is the linear prediction of $Y$ from $\underset{\sim}{X} = \underset{\sim}{x}$.

The classical correlation coefficient between variable $Y$ and variable $\hat{Y}$ (itself derived from vector $\underset{\sim}{X}$) is called the *multiple correlation coefficient* between $Y$ and $\underset{\sim}{X}$.

By definition, we have

$$R = \text{corr } (Y, \hat{Y}) \tag{5.28}$$

In practical applications, we estimate the multiple regression of $Y$ on $\underset{\sim}{X}$ from the observation matrix (5.1), that is $\hat{Y} = b_0 + \underset{\sim}{b}^T \underset{\sim}{x}$, and we calculate $\hat{Y}$ for each observation. Thus, we can define the matrix of observed and predicted values of $Y$:

$$\begin{pmatrix} y_1 & \hat{Y}_1 \\ y_2 & \hat{Y}_2 \\ \dots & \dots \\ y_n & \hat{Y}_n \end{pmatrix} \tag{5.29}$$

The correlation between the two columns (observed vs. predicted) writes, since $\bar{\hat{Y}} = \bar{y}$,

$$\hat{R} = \text{corr}\,(y, \hat{Y}) = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{Y}_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \cdot \sum_{i=1}^{n}(\hat{Y}_i - \bar{y})^2}} \tag{5.30}$$

It can be shown that the numerator of (5.30) and the second term under the square root at the denominator are both equal to $\underset{\sim}{b}^T \underset{\sim}{c}$. Consequently,

$$\hat{R} = \frac{\underset{\sim}{b}^T \underset{\sim}{c}}{\sqrt{\sum(y - \bar{y})^2 \cdot \underset{\sim}{b}^T \underset{\sim}{c}}} = \sqrt{\frac{\underset{\sim}{b}^T \underset{\sim}{c}}{\sum(y - \bar{y})^2}}$$

Knowing that $\text{SST} = \sum(y - \bar{y})^2$ and $\text{SSR} = \underset{\sim}{b}^T \underset{\sim}{c}$, we finally have

$$\hat{R} = \sqrt{\frac{\text{SCR}}{\text{SCT}}} \tag{5.31}$$

By definition, $0 \leq \hat{R} \leq 1$, and the closer the multiple correlation coefficient to 1, the better the correlation between $Y$ and $\underset{\sim}{X}$.

### 5.4.2 Hypthosis testing

We may ask ourselves whether the multiple correlation coefficient between $Y$ and $\underset{\sim}{X}$ is null, in which case there is no linear relationship between $Y$ and $\underset{\sim}{X}$.

To test the null hypothesis $H_0 : R = 0$ versus $H_1 : R \neq 0$, we have to assume that the joint distribution of $(Y, \underset{\sim}{X})$ is multinormal, and therefore that the components of vector $\underset{\sim}{X}$ are quantitative and normal.

Then, under $H_0$, the criterion

$$F = \frac{n - p - 1}{p} \frac{\hat{R}^2}{1 - \hat{R}^2} \tag{5.32}$$

follows a Snedecor $F$ distribution on $p$ and $n - p - 1$ degrees of freedom. It is immediately seen that this is the same $F$ test used in the analysis of variance above. Thus it suffices to replace in equation (5.32) $\hat{R}$ by its expression (5.31).

We reject $H_0$ if $F \geq Q_F(1 - \alpha; p, n - p - 1)$, otherwise we do not reject $H_0$.

### 5.4.3 Remark

Consider a vector $\underset{\sim}{X}^T = (X_1, \ldots, X_p)$ observed on a sample of subjects (objects) of size $n$. If we isolate variable $X_1$ and consider it as the dependent variable $Y$, we can compute the multiple regression of $X_1$ on the $(p-1)$-dimensional vector $(X_2, \ldots, X_p)$ and the corresponding multiple correlation coefficient $\hat{R}_1$ as described previously. This operation can be repeated for each variable of vector $\underset{\sim}{X}$. In total, we can calculate $p$ multiple regressions and $p$ multiple correlation coefficients $\hat{R}_1, \hat{R}_2, \ldots, \hat{R}_p$. It should be stressed that there is no real relationship between them. Finally, remark that the multiple correlation coefficient $\hat{R}$ should not be confused with the estimated correlation matrix $\underset{\sim}{R}$ of vector $X_1, \ldots, X_p$.

## 5.5 Variable selection methods

In multiple regression as in other multivariate statistical techniques, one strives to keep only the relevant variables and to eliminate those bringing little or only redundant information. This requires variable selection methods. Several approaches are possible.

The first approach consists in calculating the multiple regression of $Y$ on $\underset{\sim}{X}$, i.e. $\hat{Y} = b_0 + b_1 x_1 + \ldots + b_p x_p$, and to test the statistical significance of each regression coefficient as described in section 5.3.4. Non significant variables (coefficients) will be eliminated and the final multiple regression will be recalculated on the remaining variables.

At the other extreme, we could perform multiple regression on all possible subsets of variables of vector $\underset{\sim}{X}$ and keep the best one. This approach is feasible as long as the number of variables $p$ is reasonable but becomes intractable for large values, as the number of possibilities is close to $2^p$. Thus, for $p = 30$, there are a billion possible distinct subsets.

### 5.5.1 Forward variable selection

The forward variable selection technique proceeds in several steps. At the first step, all simple linear regression of $Y$ on each of the variables $X_1, \ldots, X_p$. are calculated and we select the best one, namely the one yielding the highest coefficient of determination $\hat{R}^2$. Denote by $X_{(1)}$ this variable.

Next, all multiple regressions of $Y$ on $X_{(1)}$ and each of the $p-1$ remaining variables are calculated. We select (retain) the variable yielding the greatest $\hat{R}^2$. Let $X_{(2)}$ be this variable.

Then, compute multiple regressions of $Y$ on all triplets $(X_{(1)}, X_{(2)}, X_i)$ where $X_i$ is any of the $p-2$ remaining variables and retain the one yielding the highest multiple coefficient of determination. The same procedure is repeated for quadruplets, quintuplets, etc.

The selection process stops at step $k$ when none of the $p-k$ remaining variables significantly improves $\hat{R}^2$ of the $k$ selected variables $X_{(1)}, \ldots, X_{(k)}$. The stopping criterion used to test the lack of significant improvement is

$$F = \frac{\mathrm{SCR}_{(k+1)} - \mathrm{SCR}_{(k)}}{s^2_{(k+1)}}$$

distributed as a Snedecor $F$ test on 1 and $n-k-2$ degrees of freedom. The process stops as soon as $F < Q_F(1-\alpha; 1, n-k-2)$ for each of the $p-k$ variables added.

### 5.5.2 Backward variable selection

Backward variable selection proceeds the other way round. To start, we calculate the (full) multiple regression of $Y$ on all variables $X_1, \ldots, X_p$. and the corresponding $\hat{R}^2$. Then, at the first step of the selection procedure, each of the $p$ variables is alternatively discarded from the analysis, hence yieding $p$ multiple regressions based on $p-1$ variables and the $p$ corresponding multiple coefficients of determination $\hat{R}^2$. The variable that will be eliminated is the one with the greatest $\hat{R}^2$, that is the variable leading to the minimal loss of information. Let $X_{(1)}$ denote this variable. At the second step, we calculate all multiple regressions of $Y$ by eliminating in addition to $X_{(1)}$ alternatively each of the $p-1$ remaining variables and computing the corresponding $\hat{R}^2$. We eliminate variable denoted $X_{(2)}$ yiedling the greatest $\hat{R}^2$. The selection process pursues the same way until the step where the elimination of any of the remaining variables leads to a statistically significant deterioration of $\hat{R}^2$.

Specifically, the selection process stops at step $k$ when any of the $p-k$ remaining variables implies a significant drop of $\hat{R}^2$ when discarded from the multiple regression. This is based on the criterion

$$F = \frac{\mathrm{SCR}_{(k)} - \mathrm{SCR}_{(k-1)}}{s^2_{(k)}}$$

distributed as a Snedecor $F$ on 1 and $n-k-1$ degrees of freedom. The selection process stops as soon as $F \geq Q_F(1-\alpha; 1, n-k-1)$ for any of the $k$ variables still in the multiple regression.

### 5.5.3 "Stepwise" variable selection

The "stepwise" variable selection method combines forward and backward procedures. The method proceeds forward but, at each selection step, before proceeding any further, it checks whether among the variables already selected any variable should be discarded (backward selection).

## 5.6 Example

### 5.6.1 Data

To illustrate multiple regression, consider the data of cholesterol (mg/dl), weight (kg) and systolic blood pressure (SBP, mmHg) in 11 male subjects aged 14 to 24 years.

| Subject | Cholesterol (mg/dl) | Weight (kg) | SBP (mmHg) |
|---|---|---|---|
| 1 | 162.2 | 51.0 | 108 |
| 2 | 158.0 | 52.9 | 111 |
| 3 | 157.0 | 56.0 | 115 |
| 4 | 155.0 | 56.5 | 116 |
| 5 | 156.0 | 58.0 | 117 |
| 6 | 154.1 | 60.1 | 120 |
| 7 | 169.1 | 58.0 | 124 |
| 8 | 181.0 | 61.0 | 127 |
| 9 | 174.9 | 59.4 | 122 |
| 10 | 180.2 | 56.1 | 121 |
| 11 | 174.0 | 61.1 | 125 |

### 5.6.2 Multiple regression

Let us first calculate the multiple regression of cholesterol ($Y$) on weight ($X_1$) and SBP ($X_2$).
The sums of observations and sums of squares and cross-products of observations are :

$$\begin{array}{lll} \sum y = 1821.5 & \sum x_1 = 630.1 & \sum x_2 = 1306 \\ \sum y^2 = 302723.51 & \sum yx_1 = 104467.79 & \sum yx_2 = 216682 \\ & \sum x_1^2 = 36197.45 & \sum x_1x_2 = 74983.3 \\ & & \sum x_2^2 = 155410 \end{array}$$

Then $\bar{y} = 165.59$
$\bar{x_1} = 57.282$
$\bar{x_2} = 118.73$

The elements of matrix $\underset{\sim}{A}_{2\times 2}$ are given by

$$a_{11} = \sum x_1^2 - \frac{(\sum x_1)^2}{n} = 36197.45 - \frac{(630.1)^2}{11} = 104.18$$

$$a_{12} = \sum x_1 x_2 - \frac{(\sum x_1)(\sum x_2)}{n} = 74983.3 - \frac{630.1 \times 1306}{11} = 173.25$$

$$a_{22} = \sum x_2^2 - \frac{(\sum x_2)^2}{n} = 155410 - \frac{(1306)^2}{11} = 352.18$$

Thus, matrix $\underset{\sim}{A}$ writes

$$\underset{\sim}{A} = \begin{pmatrix} 104.18 & 173.25 \\ 173.25 & 352.18 \end{pmatrix}$$

and the elements of vector $\underset{\sim}{c}$ are respectively

$$c_1 = \sum y x_1 - \frac{(\sum y)(\sum x_1)}{n} = 104467.79 - \frac{1821.5 \times 630.1}{11} = 128.96$$

$$c_2 = \sum y x_2 - \frac{(\sum y)(\sum x_2)}{n} = 216682 - \frac{1821.5 \times 1306}{11} = 420.27$$

so that

$$\underset{\sim}{c} = \begin{pmatrix} 128.96 \\ 420.27 \end{pmatrix}.$$

Then we need to solve the matrix equation (5.11)

$$\begin{aligned} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} &= \begin{pmatrix} 104.18 & 173.25 \\ 173.25 & 325.18 \end{pmatrix}^{-1} \begin{pmatrix} 128.96 \\ 420.27 \end{pmatrix} \\ &= \begin{pmatrix} 0.0528 & -0.0260 \\ -0.0260 & 0.0156 \end{pmatrix} \begin{pmatrix} 128.96 \\ 420.27 \end{pmatrix} \end{aligned}$$

so that $b_1 = -4.1039$ et $b_2 = 3.2121$.

The intercept is given by

$$\begin{aligned} b_0 &= \bar{y} - \underset{\sim}{b}^T \underset{\sim}{\bar{x}} \\ &= 19.30 \end{aligned}$$

Thus the multiple regression equation writes

$$\hat{Y} = 19.30 - 4.10\ x_1 + 3.21\ x_2$$

or more explicitely

$$\text{Cholesterol} = 19.30 - 4.10\times \text{ weight} + 3.21\times \text{ SBP}.$$

### 5.6.3 Analysis of variance

$$SCT = \sum y^2 - \frac{(\sum y)^2}{n} = 302723.51 - \frac{(1821.5)^2}{11} = 1099.67$$

$$SCR = \underset{\sim}{b}^T \underset{\sim}{c} = -4.1039 \times 128.96 + 3.2121 \times 420.27 = 820.71$$

$$SCE = SCT - SCR = 1099.67 - 820.71 = 278.96$$

The ANOVA table is given below

| Variability | Sum of squares | Degrees of freedom | Mean square | F |
|---|---|---|---|---|
| Regression | 820.71 | 2 | 410.36 | 11.77 |
| Residual | 278.96 | 8 | $s^2$=34.866 | |
| Total | 1099.67 | 10 | | |

The Snedecor $F$ test on 2 and 8 degrees of freedom is equal to 11.77. Since the 95% percentile $Q_F(0.95; 2, 8) = 4.46$ and since $F = 11.77 > 4.46$, the null hypothesis $H_0 : \underset{\sim}{\beta} = \underset{\sim}{0}$ is rejected and it can be concluded that the multiple regression of cholesterol on weight and SBP makes sense.

### 5.6.4 Usefulness of explanatory variables

By using equation (5.23) and noting that $s = \sqrt{34.866} = 5.9047$, the standard errors of the regression coefficients are respectively equal to

$$\begin{aligned} SE(b_1) &= 5.9047 \times \sqrt{0.0528} = 1.3568 \\ SE(b_2) &= 5.9047 \times \sqrt{0.0156} = 0.7375 \end{aligned}$$

By comparing the absolute values of the Student $t$ test to the two-sided 5% critical level, namely $Q_t(0.975; 8) = 2.31$, we conclude that both regression coefficients are statistically significant, indicating that both variables are useful in the regression predicting cholesterol. Indeed, the Student t-tests write

$$\text{Weight:} \qquad t = \frac{b_1}{SE(b_1)} = \frac{-4.1039}{1.3568} = -3.02 \ (p = 0.016)$$
$$\text{SBP:} \qquad t = \frac{b_2}{SE(b_2)} = \frac{3.2121}{0.7375} = 4.36 \ (p = 0.0024)$$

both greater than 2.31 in aboslute value.

### 5.6.5 Quality of the regression

Using equation (5.25), we have

$$R^2 = \frac{SCR}{SCT} = \frac{820.71}{1099.67} = 0.746$$

Thus, 74.6% of the variability of cholesterol values are explained by weight and SBP. Globally, 25.4% remain unexplained.

### 5.6.6 Prediction and confidence interval

Consider a subject with a body weight of 55 kg and an SBP of 120 mmHg. The predicted value of cholesterol is equal to

$$\begin{aligned} \text{Cholesterol} &= 19.30 - 4.10 \times 55 + 3.21 \times 120 \\ &= 179.0 \text{ mg/dl} \end{aligned}$$

The sampling variability of this prediction is given by equation (5.26),

$$\text{Var } \hat{Y} = 34.866 \left\{ \frac{1}{11} + h^2 \right\}$$

where

$$\begin{aligned} h^2 &= (55 - 57.28, 120 - 118.73) \begin{pmatrix} 0.0528 & -0.0260 \\ -0.0260 & 0.0156 \end{pmatrix} \begin{pmatrix} 55 - 57.28 \\ 120 - 118.75 \end{pmatrix} \\ &= (-2.28, 1.27) \begin{pmatrix} 0.0528 & -0.0260 \\ -0.0260 & 0.0156 \end{pmatrix} \begin{pmatrix} -2.28 \\ 1.27 \end{pmatrix} \\ &= 0.4497 \end{aligned} \tag{5.33}$$

It follows that Var $\hat{Y} = 18.849$ and SE($\hat{Y}$)=4.3415. We can conclude that for subject with weight of 55 kg and SBP of 120 mmHg, the true (unknown) mean cholesterol level (mg/dl) is comprised with 95% confidence in the interval $\hat{Y} \pm 1.96 SE(\hat{Y}) = 179.03 \pm 1.96 \times 4.3415 = [170.5 - 187.5]$.

### 5.6.7 Multiple correlation

The multiple correlation of cholesterol on weight and SBP is

$$\hat{R}_1 = \sqrt{0.746} = 0.86$$

Likewise we can calculate the multiple coefficient of correlation of weight on cholesterol and SBP ($\hat{R}_2 = \sqrt{0.92} = 0.96$) and that of SBP on cholesterol and weight ($\hat{R}_3 = \sqrt{0.95} = 0.97$) since all 3 variables were observed together and none of them was really favored.

# Chapter 6

# Logistic regression

## 6.1 Introduction

Logistic regression (LR) belongs to regression methods on several variables. It applies when the "dependent" variable is binary (dichotomous) and no longer continuous. This is a frequent situation in medical applications, as for example predicting the patient's outcome (death or survival, remission or non-remission) from a set of demographic, clinical or biochemical covariates (the "independent variables"). The goal is to predict outcome before it actually occurs. Can we predict the success or failure of a surgical operation based on the patient clinical status? Can we assess the chances of success of a health prevention program based on charateristics of the pouplation in which it will be conducted? We shall introduce the logistic regression model and describe the maximum likelihood (ML) estimation method to estimate the model parameters.The variable selection techniques will also be briefly outlined as we did for multiple regression. We close the chapter by presenting ordinal logistic regression (OLR) which generalizes the LR method. Examples will illustrate the methodology.

## 6.2 Model definition

Let $Y$ denote a "dependent" binary variable(0 or 1) and $\underset{\sim}{X}^T = (X_1, \ldots, X_p)$ a vector of "independent" variables also called "covariates". No restriction is made whatsoever regarding the number and type of variables $X_1, \ldots, X_p$. The problem consists in studying the relationship between the mean of $Y$ and the vector $\underset{\sim}{X}$ as in multiple regression.

As a reminder, the mean of a binary variable $E(Y)$ is a proportion,

$$E(Y) = \pi \tag{6.1}$$

which is also the probability that $Y$ is equal to 1, $P[Y = 1]$. Likewise, for any value $\underset{\sim}{X} = \underset{\sim}{x}$ , we have

$$\begin{aligned} E(Y|\underset{\sim}{x}) &= P\left[Y = 1|\underset{\sim}{X} = \underset{\sim}{x}\right] \\ &= \pi(\underset{\sim}{x}) \end{aligned} \tag{6.2}$$

Moreover, since a proportion is always comprised between 0 and 1,

$$0 \leq \pi(\underset{\sim}{x}) \leq 1 \quad \forall \underset{\sim}{x} \tag{6.3}$$

In trying to establish a relation between the mean of $Y$ and vector $\underset{\sim}{X}$, clearly the approach used in multiple regression (see section (5.2)), namely the model $E(Y|\underset{\sim}{x}) = \pi(\underset{\sim}{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$, is not applicable because the left-hand side is a proportion $\pi(\underset{\sim}{x})$ confined in the interval [0, 1] as seen in (6.3), whereas the right-hand side $\beta_0 + \underset{\sim}{\beta}^T \underset{\sim}{x}$ varies basically between $-\infty$ and $+\infty$. We have to find another approach. The idea is to transform the term on the left-hand side $\pi(\underset{\sim}{x})$ to free it from the interval [0, 1] and expand it in the interval $]-\infty, +\infty[$. This transformation defines the logistic model.

Let us apply to $\pi(\underset{\sim}{x})$ the *logit* transform

$$\text{logit } \pi(\underset{\sim}{x}) = \log \frac{\pi(\underset{\sim}{x})}{1 - \pi(\underset{\sim}{x})} \tag{6.4}$$

with "log" denoting the Nepearian logarithm often abbreviated "ln". Then, the *logistic regression* model writes without any ambiguity

$$\log \frac{\pi(\underset{\sim}{x})}{1 - \pi(\underset{\sim}{x})} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \tag{6.5}$$

Note that $1 - \pi(\underset{\sim}{x}) = P\left[Y = 0|\underset{\sim}{x}\right]$.

By taking the exponential of the two members in (6.5) and solving the equation with respect to $\pi(\underset{\sim}{x})$, the logistic model can also be written

$$\pi(\underset{\sim}{x}) = \frac{e^{\beta_0 + \underset{\sim}{\beta}^T \underset{\sim}{x}}}{1 + e^{\beta_0 + \underset{\sim}{\beta}^T \underset{\sim}{x}}} \tag{6.6}$$

where $\beta_0 + \underset{\sim}{\beta}^T \underset{\sim}{x} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$.

The name "logistic" in LR is because the function $f(t) = e^t/(1+e^t), t \in \mathbb{R}$, is called the logistic function in mathematics. In equation (6.6), terms on both sides of the equality sign lie in the interval [0, 1]. The model (6.6), linking the mean of $Y$ to vector $\underset{\sim}{X}$, is no longer linear but a logistic function of a linear combination (weighted sum) of the variables $X_1, \ldots, X_p$. It defines the "logistic regression of $Y$ on $\underset{\sim}{X}$".

Of note, the logistic regression model can also be written

$$\pi(\underset{\sim}{x}) = \frac{1}{1 + e^{-(\beta_0 + \underset{\sim}{\beta}^T \underset{\sim}{x})}} \tag{6.7}$$

It is seen that if $\beta_0 + \underset{\sim}{\beta}^T \underset{\sim}{x} = 0, \pi(\underset{\sim}{x}) = P[Y = 1|\underset{\sim}{x}] = 0.5$.

## 6.3 Maximum likelihood estimation

Consider the observation matrix

$$\underset{\sim}{X}_{n\times(p+1)} = \begin{pmatrix} y_1 & x_{11} & \ldots & x_{1p} \\ y_2 & y_{21} & \ldots & x_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ y_n & x_{n1} & \ldots & x_{np} \end{pmatrix} \tag{6.8}$$

where the observations $y_i = 0$ or 1 for $\forall i$.

As in multiple regression we shall discern two situations, depending whether variables $X_1, \ldots, X_p$ have been fixed by the investigator and variable $Y$ observed, or whether variable $Y$ and vector $\underset{\sim}{X}$ have been observed jointly (together). Consider the latter situation.

To estimate the unknown parameters (coefficients) $\beta_0, \beta_1, \ldots, \beta_p$, the least square method is not applicable because variable $Y$ is no longer Normal but binary. The alternative approach is to use the maximum likelihood (ML) estimation method.

By definition, the *likelihood* of observation $\left(y_i, \underset{\sim}{x}_i^T\right)$, specifically row $i$ of matrix (6.8), writes successively

$$\begin{aligned} L(y_i, \underset{\sim}{x}_i^T) &= L(y_i|\underset{\sim}{x}_i) \cdot L(\underset{\sim}{x}_i) \\ &= \left[\frac{e^{\beta_0 + \underset{\sim}{\beta}^T \underset{\sim}{x}_i}}{1 + e^{\beta_0 + \underset{\sim}{\beta}^T \underset{\sim}{x}_i}}\right]^{y_i} \left[\frac{1}{1 + e^{\beta_0 + \underset{\sim}{\beta}^T \underset{\sim}{x}_i}}\right]^{1-y_i} \cdot L(\underset{\sim}{x}_i) \end{aligned} \tag{6.9}$$

where $L(\underset{\sim}{x}_i)$ is the marginal likelihood of vector $\underset{\sim}{x}_i$ which does not depend on the unknown parameters $\beta_0$ and $\underset{\sim}{\beta}$.

The likelihood of the entire sample (6.8) is the product of the likelihood of each observation $\left(y_i, \underset{\sim}{x}_i^T\right)$ so that

$$\begin{aligned} L\left(\beta_0, \underset{\sim}{\beta}^T\right) &= \prod_{i=1}^{n} L(y_i, \underset{\sim}{x}_i) \\ &= \prod_{i=1}^{n} L(y_i|\underset{\sim}{x}_i) \cdot L(\underset{\sim}{x}_i) \end{aligned} \tag{6.10}$$

By taking the logarithm of both members of equation (6.10) and accounting for expression (6.9), we get the logarithm of the likelihood, namely

$$\begin{aligned} & l\left(\beta_0, \underset{\sim}{\beta}^T\right) \\ &= \log L\left(\beta_0, \underset{\sim}{\beta}^T\right) \\ &= \sum_{i=1}^{n}\left[y_i \log \frac{e^{\beta_0+\underset{\sim}{\beta}^T\underset{\sim}{x}_i}}{1+e^{\beta_0+\underset{\sim}{\beta}^T\underset{\sim}{x}_i}} + (1-y_i)\log\frac{1}{1+e^{\beta_0+\underset{\sim}{\beta}^T\underset{\sim}{x}_i}}\right] + \sum_{i=1}^{n}\log L(\underset{\sim}{x}_i) \end{aligned} \tag{6.11}$$

For a given data matrix (6.8), the function $l\left(\beta_0, \underset{\sim}{\beta}^T\right)$ does only depend on the unknown parameters. We can even forget about the second term of (6.11) as it does not depend on $\underset{\sim}{\beta}$ so that

$$l\left(\beta_0, \underset{\sim}{\beta}^T\right) = \sum_{i=1}^{n}\left[y_i \log \frac{e^{\beta_0+\underset{\sim}{\beta}^T\underset{\sim}{x}_i}}{1+e^{\beta_0+\underset{\sim}{\beta}^T\underset{\sim}{x}_i}} + (1-y_i)\log\frac{1}{1+e^{\beta_0+\underset{\sim}{\beta}^T\underset{\sim}{x}_i}}\right] + C \tag{6.12}$$

The ML estimates $\hat{\beta}_0$ and $\underset{\sim}{\hat{\beta}}$ (or $b_0, \underset{\sim}{b}$) are those parameter values which maximize function (6.12). Thus, we have to maximize a function of $p+1$ unknown parameters. This cannot be solved "analytically" (i.e. by a simple formula as in multiple regression) but by a numerical approach, for example using the Newton-Raphson iterative procedure. This is only feasible with a computer.

The computer supplies the ML estimates $b_0, b_1, \ldots, b_p$ and their standard errors $SE(b_0)$, $SE(b_1), \ldots, SE(b_p)$. More generally, the computer program

also gives an estimation of the (asymptotic) dispersion matrix of the estimators $b_0$ et $\underset{\sim}{b}$. Then, the prediction equation writes

$$\hat{\pi}(\underset{\sim}{x}) = \hat{P}[Y = 1|\underset{\sim}{x}] = \frac{e^{b_0+b_1x_1+\ldots+b_px_p}}{1 + e^{b_0+b_1x_1+\ldots+b_px_p}} \tag{6.13}$$

and can be used in practice.

Note : Suppose for a moment that the multivariate sample (6.8) has been obtained conditionnally on $\underset{\sim}{X} = \underset{\sim}{x}$ because covariates were fixed by the investigator (first sampling situation). Then, it turns out that we get exactly the same likelihood function as in (6.12) and the same estimates as if variable $Y$ and vector $\underset{\sim}{X}$ had been observed together (second sampling situation).

## 6.4 Tests on the model

### 6.4.1 Global approach

As in multiple regression, the question arises as to the global utility of vector $\underset{\sim}{X}$ by testing the null hypothesis

$$H_0 : \underset{\sim}{\beta} = \underset{\sim}{0} \text{ versus } H_1 : \underset{\sim}{\beta} \neq \underset{\sim}{0} \tag{6.14}$$

This can be done by computing the likelihood ratio test

$$LR = 2(\hat{l}_p - \hat{l}_0), \tag{6.15}$$

where $\hat{l}_p$ is the maximum likelihood based on the entire set of variables and $\hat{l}_0$ that based only on the intercept $\beta_0$, distributed as a chi-square test on $p$ degrees of freedom under the null hypothesis. If $LR \geq Q_{\chi^2}(1-\alpha; p)$, $H_0$ is rejected and the logistic model makes sense. Otherwise, we can assume that $\underset{\sim}{\beta} = \underset{\sim}{0}$ and there is no association between $Y$ and $\underset{\sim}{X}$.

### 6.4.2 Utility of covariates

When vector $\underset{\sim}{\beta}$ is not equal to 0, we like to verify which of the covariates $X_1, \ldots, X_p$ are really useful in the model by testing the hypothesis $H_0 : \beta_i = 0$ versus $H_1 : \beta_i \neq 0$ for each single variable $(i = 1, \ldots, p)$.

For this purpose we calculate the criterion

$$Z_i = \frac{b_i}{\text{SE}(b_i)} \tag{6.16}$$

or its square $Z_i^2$, distributed as a chi-square test on 1 degree of freedom. Thus, the hypothesis $\beta_i = 0$ is rejected if

$$Z_i^2 \geq Q_{\chi^2}(1-\alpha;1) \tag{6.17}$$

As in general $\alpha = 0.05$, the rejection level is equal to $Q_{\chi^2}(0.95;1) = 3.84$.

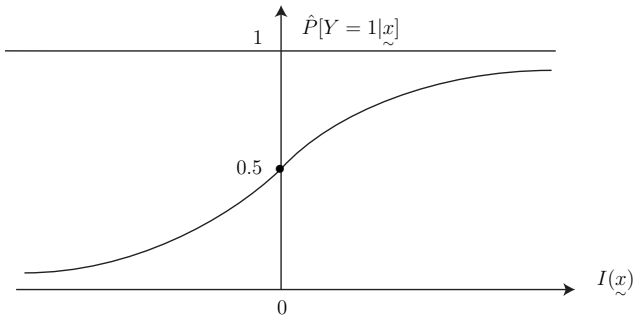
### 6.4.3 Quality of the logistic model

To assess the quality of the logistic regression model, we consider all pairs of observations $(y_i,\ y_j)$, with $y_i = 0$ and $y_j = 1$. If we denote by $n_0$ the number of observations with $y_i = 0$ and by $n_1$ the number of observations with $y_j = 1$, it turns out there are in total $N = n_0 \times n_1$ distinct pairs.

Next we compute for each pair $(y_i,\ y_j)$, the probabilities $\hat{\pi}(\underset{\sim}{x}_i)$ and $\hat{\pi}(\underset{\sim}{x}_j)$ by using Eq.(6.13). Since $y_i < y_j$, we would expect $\hat{\pi}(\underset{\sim}{x}_i) < \hat{\pi}(\underset{\sim}{x}_j)$. If so, we say there is *concordance*, otherwise there is *discordance*. When $\hat{\pi}(\underset{\sim}{x}_i) = \hat{\pi}(\underset{\sim}{x}_j)$, pairs are said to be *tied*. Let $N_c$ be the number of "concordant" pairs. Then , the quality of the logistic regression model can be measured by the percentage of concordance, say $(N_c/N) \times 100\%$. The percentages of discordance and ties are also determined.

### 6.4.4 Prediction

Equation (6.13) allows us to predict for any new multivariate observation $\underset{\sim}{x}$ the probability that $Y = 1$, specifically the expected proportion of subjects (objects) with value $Y = 1$ among all subjects (objects) with the same characteristics $\underset{\sim}{x}$. Interestingly, this prediction is based on a scalar $I(\underset{\sim}{x}) = b_0 + \underset{\sim}{b}^T \underset{\sim}{x}$, which is a weighted sum of the components of vector $\underset{\sim}{X}$. By plotting on a graph $\hat{P}[Y = 1|\underset{\sim}{x}]$ against the values of $I(\underset{\sim}{x})$ we obtain a logistic curve.




In sum, starting from multivariate observation $\underset{\sim}{X} = \underset{\sim}{x}$, we calculate $I(\underset{\sim}{x})$ and by using equation (6.13) also written

$$\hat{P}\left[Y = 1|\underset{\sim}{x}\right] = \frac{e^{I(\underset{\sim}{x})}}{1 + e^{I(\underset{\sim}{x})}} \tag{6.18}$$

we get an estimate of the probability of the event $Y = 1$. Observe that $I(\underset{\sim}{x})$ is often viewed as a *risk index*.

As in multiple regression, we can determine a 95% confidence interval for $\hat{P}[Y = 1|\underset{\sim}{x}]$. To do this, we first calcualte a confidence interval for $I(\underset{\sim}{x})$, whose boundaries are $I_1(\underset{\sim}{x})$ and $I_2(\underset{\sim}{x})$. These can be obtained by using the estimated asymptotic variance-covariance matrix of estimates $b_0$ and $\underset{\sim}{b}$, notée $\underset{\sim}{\hat{V}}$ derived by the computer program. Then, using equation (6.6) and substituting on the right-hand side $I_1(\underset{\sim}{x})$ and then $I_2(\underset{\sim}{x})$ we finally obtain the confidence interval for the probability.

## 6.5 Variable selection procedures

The variable selection problem also concerns logistic regression as we would like to keep only the variables that are useful. Several procedures exist. One would be to keep only those variables for which $\beta_i \neq 0$ (see section 6.4.2.). Another one would be to consider all possible subsets of variables and perform logistic regression on each of them. This is almost impossible given the high number of such possible subsets. Besides, computation time would increase exponentially. Therefore it is advised to use as in multiple regression forward (ascending), backward (descending) or stepwise (combining forward and backward) variable selection procedures.

### 6.5.1 Ascending selection (forward)

At first, we perform logistic regression of $Y$ on each variable $X_i$ separately and choose the best one, specifically that with greatest maximized likelihood (6.12). Let $X_{(1)}$ denote the corresponding variable and $\hat{l}_1 = \hat{l}(b_0, b_1)$ its maximum likelihood.

Then we perform all logistic regressions of $Y$ on two variables, by including $X_{(1)}$ and each of the remaining variables one by one. We will choose the variable $X_{(2)}$ which together with $X_{(1)}$ yields the greatest maximized likelihood, say $\hat{l}_2 = \hat{l}(b_0, b_1, b_2)$. Thereafter, we calculate all logistic regressions of $Y$ on all triplets $(X_{(1)}, X_{(2)}, X_i)$ where $X_i$ is successively one of the $p-2$ remaining variables. The procedure continues with quadruplets, quintruplets, sextuplets, etc.

The process stops at step $k$ when none of the $p-k$ remaining variables significantly improves the maximum likelihood of the $k$ variables selected. The criterion used for this writes

$$\chi^2_{(1)} = 2\left(\hat{l}_{k+1} - \hat{l}_k\right) \tag{6.19}$$

distributed as a chi-square test on 1 degree of freedom. Hence, the procedure stops at step $k$ if $\chi^2_{(1)} < 3.84$ for each of the $p-k$ remaining (non selected) variables.

### 6.5.2 Descending selection (backward)

Proceed as in multiple regression by starting with the full regression model of $p$ variables. Calculate the maximum likelihood $\hat{l}_p$. Next, discard among variables $X_1, \ldots, X_p$ that for which the maximum likelihood, $\hat{l}_{p-1}$, decreases the least when compared to $\hat{l}_p$. Let $X_{(1)}$ denote this variable. Then proceed the same way for the $p-1$ remaining variables and search for the one which yields the lowest decrease of the maximum likelihood $\hat{l}_{p-2}$ when discarded. Let $X_{(2)}$ be this variable. One goes on until the process stops.

The process stops at step $k$ when any of the $p-k$ remaining variables significantly deteriorates the maximum likelihood $\hat{l}_{p-k}$ when discarded from the regression. In other terms, the process stops at step $k$ if

$$\chi^2_{(1)} = 2\left(\hat{l}_k - \hat{l}_{k-1}\right) \tag{6.20}$$

is statistically significant ($\chi^2_{(1)} \geq 3.84$) each time a variable is discarded among the $p-k$ still retained.

### 6.5.3 Stepwise selection

The "stepwise" variable selection procedure combines ascending and descending principles. At each step, we look for the variable which improves most the maximum likelihood function but before moving to the next step it checks whether among the already selected variables any should be discarded.

## 6.6 Odds ratio

In epidemiology, the *odds ratio* (OR) is used to measure the association between a risk factor (variable $X$) and the presence/absence of a disease (binary variable $Y$). The odds ratio related to variable $X_i (i = 1, \ldots, p)$ and adjusted for the other variables is given by the expression

$$OR_i = e^{\beta_i} \tag{6.21}$$

Obviously if $\beta_i = 0$, there is no association between the risk factor $X_i$ and the disease ($OR_i = 1$). By contrast, if $\beta_i > 0$ or $\beta_i < 0$, there is a positive ($OR_i > 1$) or negative ($OR_i < 1$) association between them.

To estimate $OR_i$ we apply logistic regression and we get

$$\widehat{OR_i} = e^{b_i} \tag{6.22}$$

This estimation is interpreted in exactly the same way as the theoretical value $OR_i$.

Since $SE(b_i)$ is known, we can determine a 95% confidence interval for $OR_i$ by calculating the quantities $b_{i1} = b_i - 1.96 \times SE(b_i)$ and $b_{i2} = b_i + 1.96 \times SE(b_i)$ and also $\widehat{OR}_{i1} = e^{b_{i1}}$ and $\widehat{OR}_{i2} = e^{b_{i2}}$. The 95% confidence interval (IC 95%) then writes

$$\widehat{OR}_{i1} \leq OR_i \leq \widehat{OR}_{i2} \quad (i = 1, \ldots, p) \tag{6.23}$$

If this interval contains the value of 1, there is no significant association between the risk factor $X_i$ and the disease $Y$. Otherwise, the association is statistically significant. Thus, in practice it suffices to verify whether the confidence interval contains the value 1 or not .

## 6.7 Example

Data listed in Annex II correspond to 60 severe brain injury (trauma) patients. For each patient, the following variables were recorded: 6-month

outcome ($Y$) coded in three categories (1 = low or moderate disability, 2 = severe disability or persistent vegetative state, 3 = death), age (years) at time of trauma, and cerebro-spinal fluid (CSF) CK-BB isoenzyme (UI/l) assayed 24h after trauma. To illustrate the use of logistic regression, categories 1 and 2 of the outcome were combiend to create a binary outcome variable $Y = 0$ (survival) and $Y = 1$ (death). There were 27 subjects (45%) alive and 33 subjects (55%) deceased.

LR applied to the data yielded the following results:

| Parameter | ML estimation | Standard error | $Z^2$ | p-value |
|---|---|---|---|---|
| Intercept | $b_0 = -2.419$ | 0.7467 | 10.5 | 0.0012 |
| Age | $b_1 = 0.0502$ | 0.0228 | 4.85 | 0.028 |
| CK-BB | $b_2 = 0.00534$ | 0.00181 | 8.72 | 0.0032 |

Thus the risk index is given by the expression

$$I(\underset{\sim}{x}) = -2.42 + 0.0502 \times \text{age} + 0.00534 \times \text{CK-BB}$$

Both covariates, age and CK-BB, were significant predictors of the patient's outcome; the older the patient and the higher his/her CSF CK-BB level, the higher the risk of death. The likelihood ratio test showed that vector $\underset{\sim}{\beta}$ was not equal to 0. Indeed, $LR = 23.33$ on 2 degrees of freedom ($p < 0.0001$).

To assess the quality of the model, observe that there is a total of $27 \times 33 = 891$ pairs of observations of the two groups. Of these, 736 were "concordant" (82.6%), 152 "discordant" (17.1%) and 3 ties (0.3%). The model can therefore be considered as satisfactory.

Now consider a 45 years old patient with a CSF CK-BB level of 300 UI/l. The patient's risk score amounts

$$\begin{aligned} I(\underset{\sim}{x}) &= -2.42 + 0.0502 \times 45 + 0.00534 \times 300 \\ &= 1.44 \end{aligned}$$

and the corresponding probability of death is

$$P[Y = 1|\underset{\sim}{x}] = \frac{e^{1.44}}{1 + e^{1.44}} = 0.81$$

Thus, this patient has 81% chances of not surviving and 19% chances of surviving his/her brain injury.

Note that the odd ratios related to age and CK-BB are respectively equal to 1.05 and 1.01. The corresponding 95% confidence intervals are $1.0055 < OR_1 < 1.0995$ and $1.0018 < OR_2 < 1.0089$. Both do not cover the value $OR = 1$, hence confirming the significant association of the two variables with outcome.

## 6.8 Ordinal logistic regression

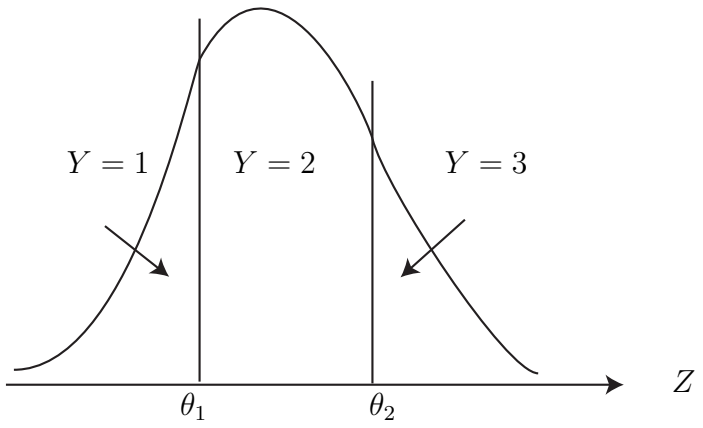
### 6.8.1 Problem definition

In frequent settings, the dependent variable $Y$ is qualitative ordinal meaning that its "categories" (modalities) are ordered $m_1 < m_2 < \ldots < m_k$ (e.g. cancer staging). We mentioned that in such a situation the categories may be numbered from 1 to $k$. Underneath variable $Y$, there is so to say a hidden continuous variable $Z$ and specific "threshold levels" $\theta_1, \ldots, \theta_{k-1}$ which define variable $Y$. Thus,

$$Y = i \text{ if } \theta_{i-1} < Z \leq \theta_i \quad (i = 1, \ldots, k) \tag{6.24}$$

By definition, $\theta_0 = -\infty$ and $\theta_k = +\infty$. This is exactly what happens when a patient's diseased condition is assessed as minor, moderate, or severe. Considering the 3 ordered categories $m_1 < m_2 < m_3$, we can write

$$\begin{array}{lll} Y = 1 & \text{if} & Z \leq \theta_1 \\ Y = 2 & \text{if} & \theta_1 < Z \leq \theta_2 \\ Y = 3 & \text{if} & \theta_2 < Z \end{array} \tag{6.25}$$




In practice, $Z$ is unknown and can only be observed through the categories by means of variable $Y$. Basically, this principle also applies to the logistic

model in the sense that there are only $k = 2$ categories and hence only one threshold (cut-off) $\theta$. Then, $Y = 0$ if $Z \leq \theta$ and $Y = 1$ if $Z > \theta$.

### 6.8.2 The ordinal logistic model

Suppose that we observe the $p + 1$ dimensional vector $(Y, \underset{\sim}{X})$ where $Y$ is an ordinal variable with $k$ modalities. We can calculate the probability of occurrence of each modality $P[Y = i] = \pi_i$ $(i = 1, \ldots, k)$. Obviously, $\pi_1 + \pi_2 + \ldots + \pi_k = 1$. Likewise, we can work conditionnally on $\underset{\sim}{X} = \underset{\sim}{x}$ and write

$$\begin{aligned} &P\left[Y = i|\underset{\sim}{x}\right] = \pi_i(\underset{\sim}{x}) \\ &\pi_1(\underset{\sim}{x}) + \pi_2(\underset{\sim}{x}) + \ldots + \pi_k(\underset{\sim}{x}) = 1 \ \forall \underset{\sim}{x} \end{aligned} \tag{6.26}$$

The categories of $Y$ are defined by the thresholds $\theta_1, \ldots, \theta_{k-1}$ on the scale of the "latent" variable $Z$.

Then the ordinal logistic regression (OLR) model assumes that ($k = 3$ for simplicity)

$$\begin{aligned} P\left[Y = 1|\underset{\sim}{x}\right] &= \frac{e^{\theta_1 - \underset{\sim}{\beta}^T \underset{\sim}{x}}}{1 + e^{\theta_1 - \underset{\sim}{\beta}^T \underset{\sim}{x}}} \\ P\left[Y = 2|\underset{\sim}{x}\right] &= \frac{e^{\theta_2 - \underset{\sim}{\beta}^T \underset{\sim}{x}}}{1 + e^{\theta_2 - \underset{\sim}{\beta}^T \underset{\sim}{x}}} - \frac{e^{\theta_1 - \underset{\sim}{\beta}^T \underset{\sim}{x}}}{1 + e^{\theta_1 - \underset{\sim}{\beta}^T \underset{\sim}{x}}} \\ P\left[Y = 3|\underset{\sim}{x}\right] &= 1 - \frac{e^{\theta_2 - \underset{\sim}{\beta}^T \underset{\sim}{x}}}{1 + e^{\theta_2 - \underset{\sim}{\beta}^T \underset{\sim}{x}}} \end{aligned} \tag{6.27}$$

and more generally

$$P\left[Y = i|\underset{\sim}{x}\right] = \frac{e^{\theta_i - \underset{\sim}{\beta}^T \underset{\sim}{x}}}{1 + e^{\theta_i - \underset{\sim}{\beta}^T \underset{\sim}{x}}} - \frac{e^{\theta_{i-1} - \underset{\sim}{\beta}^T \underset{\sim}{x}}}{1 + e^{\theta_{i-1} - \underset{\sim}{\beta}^T \underset{\sim}{x}}} \qquad (i = 1, \ldots, k) \tag{6.28}$$

Note that in (6.27) and (6.28) the sum of the probabilities is equal to 1. The logistic function $f(t) = e^t/(1 + e^t)$ can also be clearly discerned in each case. Finally remember that $\underset{\sim}{\beta}^T \underset{\sim}{x} = \beta_1 x_1 + \ldots + \beta_p x_p$.

The parameters of the model are the $k-1$ thresholds $\theta_1, \ldots, \theta_{k-1}$ and the regression coefficients of the variables $\beta_1, \ldots, \beta_p$. There is no intercept $\beta_0$, the latter being replaced by the cut-off values. In total, $p + k - 1$ unknown parameters need to be estimated. Observe that, when $k = 2$ (logistic model), we retrieve the $p + 1$ parameters $\theta_1 = \beta_0$ and $\underset{\sim}{\beta}$.

### 6.8.3 Maximum likelihood estimation

The observation matrix has the same form as in (6.3), but the values of $y_i$ are equal to $1, 2, \ldots, k$.

As in logistic regression, we have to maximize the likelihood of the sample or its logarithm

$$l = \log L = \sum_{i=1}^{n} \log P\left[Y_i = y_i | \underset{\sim}{x}_i\right] + C \tag{6.29}$$

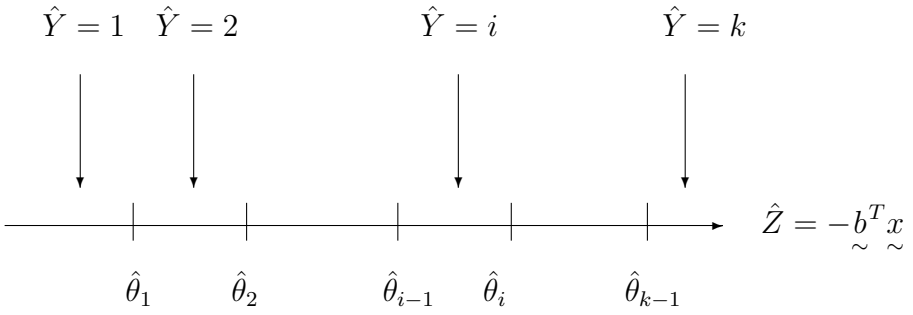
where the probabilities have to be replaced by their expressions (6.28). For given $y_i$ and $\underset{\sim}{x}_i$, the likelihood does only depend on the parameters $\underset{\sim}{\theta}$ and $\underset{\sim}{\beta}$, say $l(\theta_1, \ldots, \theta_{k-1}, \beta_1, \ldots, \beta_p)$. By using a computer program, we search for values $\hat{\theta}_1, \ldots, \hat{\theta}_{k-1}, \beta_1, \ldots, b_p$ which maximize (6.29) and denote by $\hat{l}(\underset{\sim}{\hat{\theta}}, \underset{\sim}{b})$ the maximum. The program also gives the standard errors $\mathrm{SE}(\hat{\theta}_i)$ and $\mathrm{SE}(b_i)$ of the estimates $\hat{\theta}_i$ and $b_i$. Then, equations (6.28) become

$$\hat{P}\left[Y = i | \underset{\sim}{x}\right] = \frac{e^{\hat{\theta}_i - \underset{\sim}{b}^T \underset{\sim}{x}}}{1 + e^{\hat{\theta}_i - \underset{\sim}{b}^T \underset{\sim}{x}}} - \frac{e^{\hat{\theta}_{i-1} - \underset{\sim}{b}^T \underset{\sim}{x}}}{1 + e^{\hat{\theta}_{i-1} - \underset{\sim}{b}^T \underset{\sim}{x}}} \qquad (i = 1, \ldots, k) \tag{6.30}$$

### 6.8.4 Prediction

Consider a new subject with known vector $\underset{\sim}{x}$ but unknown value of $Y$ and let us calculate the probabilities of each category of $Y$ for this subject by using equations (6.30). Then we shall assign the subject to the category with highest probability.

In fact, the univariate index $I(\underset{\sim}{x}) = -\underset{\sim}{b}^T \underset{\sim}{x}$ (note that all signs of the $b_i$ have been changed !) may be interpreted as a predictor of the latent (underlying) unobservable variable $Z$, i.e. a sort of risk indicator.



It suffices to look where the value $\hat{Z} = -\underset{\sim}{b}^T \underset{\sim}{x}$ is located with respect to the cut-offs $\hat{\theta}_1, \ldots, \hat{\theta}_{k-1}$ and hence to determine the category $Y$.

### 6.8.5 Other remarks

In ordinal logistic regression, as in LR, the quality of the regression can be appraised by computing the percentage of concordance for all pairs $(y_i,\ y_j)$ with $y_i < y_j$.

Denoting by $n_1,\ n_2, ...,\ n_k$ the observed number of subject in each modality of $Y$, there are $N = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} n_i n_j$ pairs. A pair $(y_i,\ y_j)$ will be concordant if $\hat{\pi}_i(\underset{\sim}{x_i}) < \hat{\pi}_j(\underset{\sim}{x_j})$, tied if equal and discordant otherwise. The probabilities $\hat{\pi}_i(\underset{\sim}{x_i})$ and $\hat{\pi}_j(\underset{\sim}{x_j})$ are derived from equation (6.13).

The variable selection techniques described previously can also be applied here, whether ascending, descending or stepwise.

Finally it is possible to test the equality of two adjacent thresholds, e.g. $H_0 : \theta_i = \theta_{i+1}$. If the null hypothesis is not rejected, then the two thresholds as well as the two categories can be merged, thus simplifying the logistic model (6.30). The two merged categories are said to be indistinguishable.

### 6.8.6 Example

We may apply the ordinal logistic regression model to the brain injury patient data given in Annex II. We consider a 3-category ordinal variable $Y$, denoted "Favorable evolution" for category $Y = 1$, "Unfavorable evolution" for category $Y = 2$, and "Death" category $Y = 3$. There are respectively 19, 8 and 33 subjects in the three categories defined.

The model entails 4 parameters, namely two thresholds $\theta_1$ and $\theta_2$ and the regression coefficients $\beta_1$ and $\beta_2$ for the two variables age and CK-BB.

Applying OLR to the data gives the following results:

| Parameter | ML estimation | Standard error | $Z^2$ | p-value |
|---|---|---|---|---|
| Threshold 1 | $\hat{\theta_1} = 1.778$ | 0.6709 | 7.02 | 0.0081 |
| Threshold 2 | $\hat{\theta_2} = 2.629$ | 0.725 | 13.1 | 0.0003 |
| Age | $b_1 = -0.0498$ | 0.0215 | 5.37 | 0.0204 |
| CK-BB | $b_2 = -0.00601$ | 0.00188 | 10.2 | 0.0014 |

The risk index (latent variable Z) writes

$$\begin{aligned} \hat{Z} &= -(-0.0498 \times \text{age} - 0.00601 \times \text{CK-BB}) \\ &= 0.0498 \times \text{age} + 0.00601 \times \text{CK-BB} \end{aligned}$$

which can be compared to the two threshold levels 1.78 and 2.63.

Again, note that both variables age and CK-BB are significant predictors of outcome $Y$. The likelihood ratio test which compares the full model with the two covariates to the restricted model including only the two thresholds is equal to $LR = 27.46$ on 2 degrees of freedom ($p < 0.0001$).

The quality of the model is given by the percentage of concordance obtained from the $N = 19 \times 8 + 19 \times 33 + 8 \times 33 = 1043$ possible pairs from any of the three groups. We found 846 concordant pairs (81.1%), 194 discordant pairs (18.6%) and 3 cases of tied pairs (0.3%).

Now consider anew the example of the 45 years old trauma patient with a CSF CK-BB level of 300 UI/l. The patient's associated risk index is equal to

$$\hat{Z} = 0.0498 \times 45 + 0.00601 \times 300 = 4.04$$

This value is beyond the second threshold since $4.04 > 2.63$. Thus the patient falls in the "Death" category. More precisely the probabilities associated with each category given by equations (6.27) are respectively equal to

$$\begin{aligned}
\hat{\pi}_1(\underset{\sim}{x}) &= \frac{e^{1.78-4.04}}{1+e^{1.78-4.04}} = \frac{e^{-2.26}}{1+e^{-2.26}} = 0.09 \\
\hat{\pi}_2(\underset{\sim}{x}) &= \frac{e^{2.63-4.04}}{1+e^{2.63-4.04}} - \hat{\pi}_1(x) \\
&= \frac{e^{-1.41}}{1+e^{-1.41}} - 0.09 = 0.20 - 0.09 = 0.11 \\
\hat{\pi}_3(\underset{\sim}{x}) &= 1 - \hat{\pi}_1(\underset{\sim}{x}) - \hat{\pi}_2(\underset{\sim}{x}) \\
&= 1 - 0.09 - 0.11 = 0.80
\end{aligned}$$

Compared to the probabilities obtained by logistic regression, ORL probabilities better differentiate categories death and unfavorable evolution.

# Chapter 7

# Cox regression

## 7.1 Introduction

Cox proportional hazards (PH) regression model applies to survival data and is widely used in the medical literature. What characterizes a survival or lifetime variable compared to other variables is censoring. Censored data are values that have not been observed as the other ones but they are only known to be either above (right censoring) or below (left censoring) some given values. In this chapter we will basically focus on right censoring. We define the notions of lifetime variable and associated survival curve. We shall briefly describe how the survival curve can be estimated by the Kaplan-Meier method. Then we will introduce Cox PH regression model and outline the estimation of the model parameters. The problem of variable selection will be briefly alluded to and the chapter closes with some applications.

## 7.2 Lifetime variable

A *lifetime* variable also called *survival* or *time-to-event* variable, denoted by the capital letter $T$ in this chapter, is a continuous variable measuring the time elapsed between two events. Thus, the lifetime of a human being is the time interval between his/her birth and death. This is probably the most natural example of a survival variable. It can however apply also to animals, equipments, foods, plants and other living organisms in general. But a lifetime variable can also represent the length of a patient's hospital stay, the duration of unemployement of a worker, the duration of relationship between to individuals, the time interval between two asthma crisis, the time to cancer recurrence after recovery. Survival variables are generally expressed in time units (seconds, minutes, hours, weeks, months, years).

A lifetime variable $T$ has three distinct characteristcs :

1. $T \geq 0$ is a non negative variable
2. the distribtion of $T$ is usually positively skewed, i.e. asymmetric with a longer tail to the right, so that the mean of $T$ is greater than its median
3. $T$ is often right censored $(T > t)$

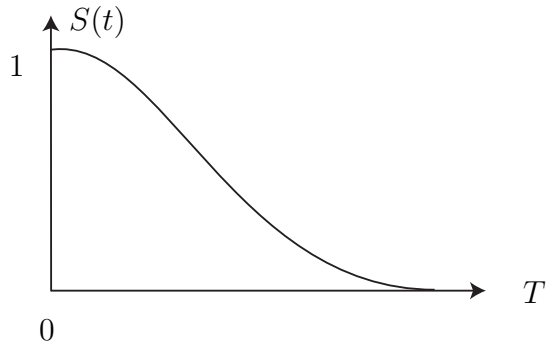
It follows that the observations of variable $T$ in a given sample of $n$ subjects (or objects) have to be separated into actually observed (uncensored) values and censored values. Censoring can be indicated in several ways (e.g., $> 25$ or $25^*$ or $25^+$). When entering lifetime data in a computer spreadsheet file for statistical analysis, this presentation is not optimal. Thus, it is preferred to use a censoring indicator variable $C$ such that $c = 0$ if the variable has been observed $(T = t)$ and $c = 1$ in case of censoring $(T > t)$. Note that in some statistical package, the reverse notation is used and $c = 1$ denotes an observed lifetime. In a time-to-event context, $c = 1$ denotes occurrence of the event (e.g. death, relapse, recurrence) and $c = 0$ non-occurrence. Thus, a sample of survival data will be represented by pairs of observations $(t_i, c_i), i = 1, \ldots, n$.

## 7.3 Kaplan-Meier survival curve

A lifetime variable $T$ is best described by its *survival curve or survival function*

$$S(t) = P[T > t] \qquad t \geq 0 \tag{7.1}$$

which displays the proportion of subjects still alive after time $t$. Thus, the curve represents the probability of surviving time $t$. By convention, at $t = 0, S(0) = 1$. Thereafter, as $t$ increases, $S(t)$ decreases. Thus, if $t_1 > t_2$, then $S(t_1) \leq S(t_2)$. Finally, as $t \to \infty, S(t) \to 0$.




In practice, the estimation of the survival curve from a sample of lifetime data $\{(t_i, c_i), i = 1, \ldots, n\}$, utilizes the *Kaplan-Meier (KM) method* as described below.

1. The $n$ survival data $t_i$ are sorted out in increasing order regardless of censoring. Cleary, in case of tie (equality) between a censored and a non-censored observation, the censored value should always be ranked after to the uncensored.

2. Then consider only the $k$ distinct non-censored survival data, say $t_1 < t_2 < \ldots < t_{k-1} < t_k$.

3. For each distinct value $t_i$, let $l_i$ be the number of subjects still alive just before that moment and $d_i$ the number of subjects who died at $T = t_i$. Be careful, any censored value before $t_i$ should not be accounted for because we ignore what happened to it.

4. The value of the survival curve at time $T = t_i$ can be estimated by the expression

$$\hat{S}(t_i) = \prod_{j=1}^{i} \frac{l_j - d_j}{l_j} \qquad (i = 1, \ldots, k) \tag{7.2}$$

or by the recurrence formula $(i = 1, \ldots, k)$

$$\hat{S}(t_i) = \hat{S}(t_{i-1}) \times \frac{l_i - d_i}{l_i} \tag{7.3}$$

where by convention if $i = 1, t_{i-1} = 0$ and $\hat{S}(0) = 1$.

Note that the Kaplan-Meier survival curve always starts at 1 but may not necessarily end at 0. This happens when the largest survival observation of the sample is censored. It is also seen that censored data preceding $t_1$ (the lowest distinct non-censored observation) are not taken into account in the KM process. Lastly, when two survival curves $\hat{S}_1(t)$ and $\hat{S}_2(t)$ of two groups are plotted on the same graph and $\hat{S}_2(t) > \hat{S}_1(t)$, subjects of group 2 have a longer lifetime than those of group 1.

## 7.4 Cox proportional hazards model

### 7.4.1 Problem definition

Consider $Y = T$ a "dependent" lifetime variable and $\underset{\sim}{X}^T = (X_1, \ldots, X_p)$ a vector of covariates as in all regression problems. No restriction is made on the number and type of the "independent" variables. Practically, the observation matrix has one additional column because of the (binary) censoring indicator $C$. Thus,

$$\underset{\sim}{X}_{n\times p} = \begin{pmatrix} (t_1, c_1) & x_{11} & \ldots & x_{1p} \\ (t_2, c_2) & x_{21} & \ldots & x_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ (t_n, c_n) & x_{n1} & \ldots & x_{np} \end{pmatrix} \tag{7.4}$$

where $y_i = (t_i, c_i)$ with $c_i = 0$ for an observed lifetime ($T = t_i$) and $c_i = 1$ for a censored lifetime ($T > t_i$) ($i = 1, \ldots, n$).

When "regressing" variable $T$ on vector $\underset{\sim}{X}$ to study the relationship between them, the classical multiple regression approach is hardly applicable because of the presence of censoring, unless there are no censored data in the sample. Indeed in this case, we could perform a multiple regression of $Y^* = \log T$ on $X_1, \ldots, X_p$, the log-transform being applied to eliminate right-skewness of $T$ and normalize as much as possible its distribution.

The regression of a lifetime variable $T$ on a vector of covariates $\underset{\sim}{X}$ is best be solved by the Cox proportional hazards model.

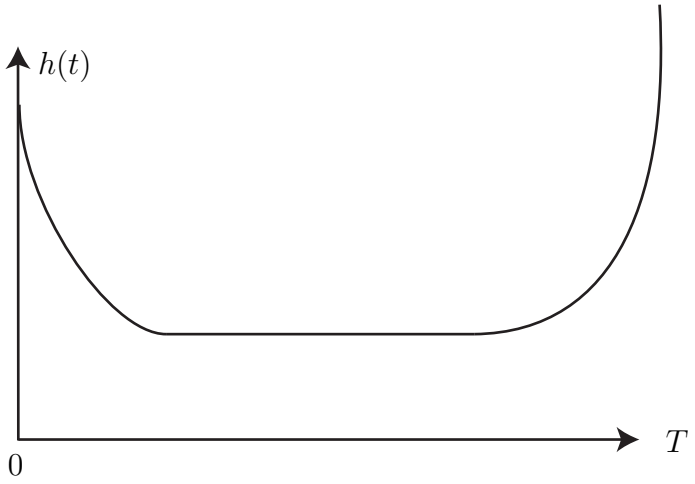
### 7.4.2 The hazard function

Let $S(t) = P[T > t]$ be the survival function of variable $T$. Then, the *hazard function*, noted $h(t)$, measures the immediate risk of death at time $t$. Clearly, this can only be envisaged as long as one lives until time $t$. By definition,

$$h(t) = \frac{f(t)}{S(t)} \tag{7.5}$$

where $f(t)$ is the density function of variable $T$. To some extent, $f(t)$ represents the overall risk of death at time $t$, i.e. among all subjects of the population not just those who lived until time $t$.

In human life, the function $h(t)$ has a bathtub shape. The immediate risk of death it high at birth, then decreases to level off (remain constant) during adulhood, and eventually rapidly increases with older ages to reach infinity.




Since $f(t) = -\dfrac{dS(t)}{dt}$, equation (7.5) can also be written

$$h(t) = -\frac{d}{dt}\log S(t) \tag{7.6}$$

Conversely, the survival function can be expressed in terms of the hazard function as follows

$$S(t) = e^{-\int_0^t h(u)du} \tag{7.7}$$

The integral part in equation (7.7) is often called the cumulative hazard function $H(t) = \int_0^t h(u)du$. Thus,

$$S(t) = e^{-H(t)} \tag{7.8}$$

These relations are essential to define Cox PH model.

### 7.4.3 Proportional hazards

The smart idea of Sir David R. Cox (1972) was to introduce the concept of *proportional hazards, PH*. Consider variable $T$ and two groups of subjects, and let $h_1(t)$ and $h_2(t)$ the corresponding hazard functions. Assume that these hazards are proportional.

$$h_1(t) = \lambda h_2(t)$$

or

$$\frac{h_1(t)}{h_2(t)} = \lambda \tag{7.9}$$

If $\lambda = 1$, the hazard functions are equal. If $\lambda > 1$, subjects of group 1 are at higher risk than those of group 2, and conversely if $\lambda < 1$.

By using equations (7.7) and (7.9) it is immediately seen that

$$S_1(t) = [S_2(t)]^\lambda \tag{7.10}$$

Thus, if $\lambda = 1$, the survival curves are equal. By contrast, if $\lambda > 1, S_1(t) < S_2(t)$ and if $\lambda < 1, S_1(t) > S_2(t)$, confirming what has been written above.

Equation (7.10) shows that, for a given survival curve $S_0(t)$, by letting $\lambda$ vary, one defines a family of (proportional or "parallel") survival functions below and above $S_0(t)$ depending on whether $\lambda$ is lower or greater than 1. This can be written as

$$S(t|\lambda) = [S_0(t)]^\lambda \tag{7.11}$$

Since $\lambda$ cannot be negative and given the importance of the pivotal threshold $\lambda = 1$, Cox's idea was to express $\lambda$ as the exponential of another quantity, say $\gamma$, i.e. $\lambda = e^\gamma$ where $\gamma$ can be negative, null or positive. Then, equation (7.11) writes

$$S(t|\gamma) = [S_0(t)]^{e^\gamma} \qquad \gamma \in R \tag{7.12}$$

### 7.4.4 Cox PH regression

*Cox proportional hazard model* allows to express the relation between a lifetime variable $T$ and a vector of covariates $\underset{\sim}{X}^T = (X_1, \ldots, X_p)$. Indeed, by letting $\underset{\sim}{\beta}^T \underset{\sim}{x} = \beta_1 x_1 + \ldots + \beta_p x_p$ as done previously in regression, and remembering that a weighted sum can take any value between $-\infty$ and $+\infty$, which is precisely what happens for $\gamma$ in expression (7.12), Cox proportional hazard model can be written as

$$S(t|\underset{\sim}{x}) = [S_0(t)]^{e^{\underset{\sim}{\beta}^T \underset{\sim}{x}}} \tag{7.13}$$

This quite complex model describes how the lifetime of subjects varies with respect to the covariates under study.

As before observe that if $\underset{\sim}{\beta} = \underset{\sim}{0}, e^{\underset{\sim}{\beta}^T \underset{\sim}{x}} = 1$, there is no association between $T$ and $\underset{\sim}{X}$. Likewise, we may test the null hypothesis that $\beta_i = 0$, that is test the usefulness of $X_i$ in the model. It is also seen that if the risk index $\underset{\sim}{\beta}^T \underset{\sim}{x}$ is negative, the corresponding survival curve is more favorable (above) than the baseline survival curve $S_0(t)$. By contrast, if $\underset{\sim}{\beta}^T \underset{\sim}{x}$ is positive, the corresponding survival curve is less favorable (below) $S_0(t)$. Hence, as $\underset{\sim}{\beta}^T \underset{\sim}{x}$ gets higher, survival continues to deteriorate.

Consider the simple case of a single binary covariate $X = 0, 1$. Then, the model (7.13) writes

$$S(t|x) = [S_0(t)]^{e^{\beta x}}$$

yielding the two survival curves $S_0(t)$ for $x = 0$ and $[S_0(t)]^\gamma$ where $\gamma = e^\beta$ for $x = 1$. This model is often used to compare two survival curves.

Remark that referring to the risk function, Cox model can be written as

$$h(t|\underset{\sim}{x}) = h_0(t) \cdot e^{\underset{\sim}{\beta}^T \underset{\sim}{x}} \tag{7.14}$$

The baseline risk is multiplied by an exponential function which incorporates the explanatory covariates $X_1, \ldots, X_p$.

## 7.5 Estimation of regression coefficients

The estimation of the regression coefficients $\underset{\sim}{\beta}^T = (\beta_1, \ldots, \beta_p)$ of Cox PH model from the observation matrix (7.4) is by no means an easy task. It was Cox ingeniosity to propose a solution to the problem by introducing a new kind of likelihood, called "partial likelihood". This is given by

$$L(\underset{\sim}{\beta}) = \prod_{j=1}^{k} \left\{ \frac{e^{\underset{\sim}{\beta}^T \underset{\sim}{x}_j}}{\sum_{m \in R_j} e^{\underset{\sim}{\beta}^T \underset{\sim}{x}_m}} \right\} \tag{7.15}$$

where $k$ is the number of actual death times, $R_j$ the subset of subjects alive just before time $t_j$.

The logarithm of the partial-likelihood function (7.15) writes

$$l(\underset{\sim}{\beta}) = \sum_{j=1}^{k} \left\{ \underset{\sim}{\beta}^T \underset{\sim}{x}_j - \log \sum_{m \in R_j} e^{\underset{\sim}{\beta}^T \underset{\sim}{x}_m} \right\} \tag{7.16}$$

To obtain the estimates $\hat{\underset{\sim}{\beta}} = \underset{\sim}{b}$, it suffices to maximize the log-likelihood function. This can be done by computer and the program also supplies the standard errors of the estimates, $SE(b_1), \ldots, SE(b_p)$, as in logistic regression.

Knowing vector $\underset{\sim}{b}$, equations (7.13) and (7.14) become

$$\hat{S}(t|\underset{\sim}{x}) = \left[\hat{S}_0(t)\right]^{e^{\underset{\sim}{b}^T \underset{\sim}{x}}} \tag{7.17}$$

and

$$\hat{h}(t|\underset{\sim}{x}) = \hat{h}_0(t) \cdot e^{\underset{\sim}{b}^T \underset{\sim}{x}}$$

To estimate $\hat{S}_0(t)$, we use the estimated cumulative hazard function

$$\hat{H}_0(t) = \sum_{t_j \leq t} \left[ \frac{d_j}{\sum_{m \in R_j} e^{\underset{\sim}{b}^T \underset{\sim}{x}_m}} \right] \tag{7.18}$$

and the fact that $\hat{S}_0(t) = e^{-\hat{H}_0(t)}$ for any value of $t$. Then, for a given individual $\underset{\sim}{X} = \underset{\sim}{x}$, The survival probability at time $T = t$ can be calculated by the formula

$$\begin{aligned} \hat{S}(t|\underset{\sim}{x}) &= \left[\hat{S}_0(t)\right]^{e^{\underset{\sim}{b}^T \underset{\sim}{x}}} \\ &= \left[e^{-\hat{H}_0(t)}\right]^{e^{\underset{\sim}{b}^T \underset{\sim}{x}}} \end{aligned} \tag{7.19}$$

where $\hat{H}_0(t)$ is given in (7.18).

## 7.6 Tests on the PH model

In general, one does not really want to compute the survival curve for a given individual but rather test null hypotheses on the PH model as done for logistic regression.

To test the null hypothesis $H_0 : \underset{\sim}{\beta} = \underset{\sim}{0}$, we use a likelihood ratio test distributed as a chi-square variable on $p$ degrees of freedom, as previously.

A Wald test may also be applied by computing the quadratic form (Mahalanobis distance) $\chi^2_p = \underset{\sim}{b}^T \cdot \underset{\sim}{V}(b)^{-1} \cdot \underset{\sim}{b}$, where $\underset{\sim}{V}(b)$ is the asymptotic estimated variance-covariance matrix of vector $\underset{\sim}{b}$. This quantity which is a chi-squared test on $p$ degrees of freedom can be compared to the critical threshold $Q_{\chi^2}(1-\alpha; p)$.

### 7.6.1 Utility of covariates

To test the hypotheses $H_0 : \beta_i = 0$ vs $H_1 : \beta_i \neq 0$ (utility of covariate $X_i$ in the model), one calculates the criterion

$$Z = \frac{b_i}{SE(b_i)} \quad (i = 1, \ldots, p) \tag{7.20}$$

or its square $Z^2$ and compare it to the upper 5% critical level of a chi-square distribution on 1 degree freedom equal to 3.84. The null hypothesis is rejected if $Z^2$ is greater than this threshold.

### 7.6.2 Hazard ratio

For each variable $X_i$, we have an estimate of the regression coefficient $b_i$ and of its standard error $SE(b_i)$. By definition, the *hazard ratio (HR)* associated with variable $X_i$ is given by

$$\hat{h}_i = e^{b_i} \tag{7.21}$$

A 95% confidence interval can be computed for the true (unknown) hazard ratio $h_i = e^{\beta_i}$ as follows : first we determine the confidence interval for $\beta_i$ by computing $b_i \pm 2\ SE(b_i)$. Next by taking the exponential of both values $b_{i1}$ and $b_{i2}$, we obtain the 95% confidence interval $h_{i1} \leq h_i \leq h_{i2}$. This approach allows to verify whether $h_i$ significantly differs from 1. It suffices to see whether the confidence interval contains the value 1. This is equivalent to test whether $\beta_i = 0$. If $\hat{h}_i$ is significantly greater (smaller) than 1, variable $X_i$ increases (decreases) the hazards (see (7.17)). This interpretation is particularly useful when all variables $X_1, \ldots, X_p$ are binary. It allows to discern which variables increase the risk of death (survival) and those which decrease it.

## 7.7 Variable selection methods

As in multiple or in logistic regression analysis, an ascending (forward), descending (backward) or stepwise variable selection method can be applied to

retain only useful variables in the model. The selection and stopping criterion are based on the maximum of the (partial) likelihood function and the chi-square test. In the forward approach, the process stops when none of the remaining variables significantly improves the model. In the backward approach, the process stops when none of the remaining variables significantly deteriorates the model.

## 7.8 Example

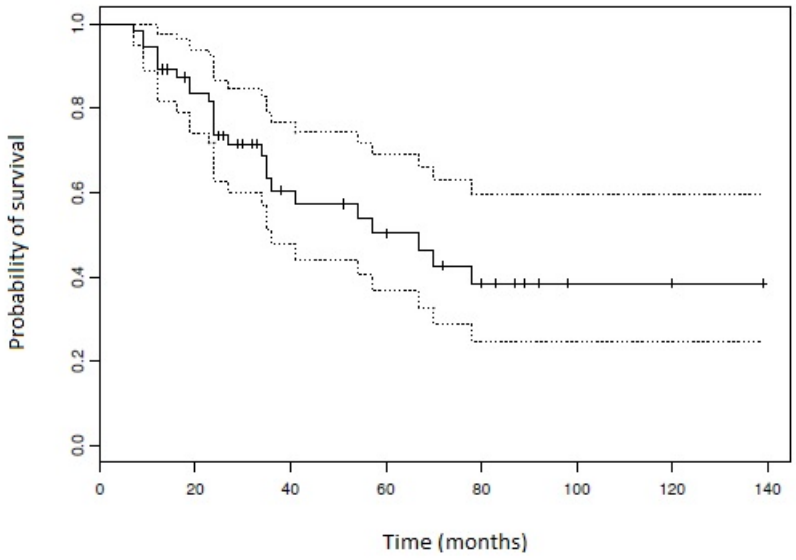
As an illustration, consider the postoperative survival time ($T$, months) of 56 patients with rectal cancer who received radiotherapy before surgery. The covariates included in the model were : $X_1$ = radiation dose received (0 if <5000 rads or 1 if $\geq$ 5000 rads), $X_2$ = age at diagnosis (years) and $X_3$ gender (0=female, 1=male). The complete dataset is given in Annex III.

### 7.8.1 Kaplan-Meier survival curve

When ranking the 56 survival times in increasing order, only $k = 17$ corresponded to actual (uncensored) distinct survival times. The survival table is given below.

| Time $t_i$ | Alive $l_i$ | Dead $d_i$ | Survival $\hat{S}(t_i)$ |
|---|---|---|---|
| 0 | 56 | 0 | 1.0 |
| 7 | 56 | 1 | 0.9821 |
| 9 | 53 | 2 | 0.9464 |
| 12 | 50 | 3 | 0.8929 |
| 16 | 46 | 1 | 0.8739 |
| 19 | 43 | 2 | 0.8350 |
| 23 | 42 | 1 | 0.8156 |
| 24 | 37 | 4 | 0.7360 |
| 27 | 33 | 1 | 0.7144 |
| 34 | 26 | 1 | 0.6879 |
| 35 | 24 | 2 | 0.6350 |
| 36 | 20 | 1 | 0.6048 |
| 41 | 18 | 1 | 0.5729 |
| 54 | 16 | 1 | 0.5392 |
| 57 | 14 | 1 | 0.5033 |
| 67 | 12 | 1 | 0.4646 |
| 70 | 11 | 1 | 0.4259 |
| 78 | 9 | 1 | 0.3833 |

The Kaplan-Meier survival curve with confidence region is displayed in the figure below.




### 7.8.2 Cox PH model

Of the 56 patients, 21 received a radiation dose <5000 rads and 35 a dose greater or equal to 5000 rads. There were 25 females and 31 males with a mean age of $62.1 \pm 12$ years. The median age was 62 years.

Cox PH regression analysis of survival time on the three covariates gave the following results :

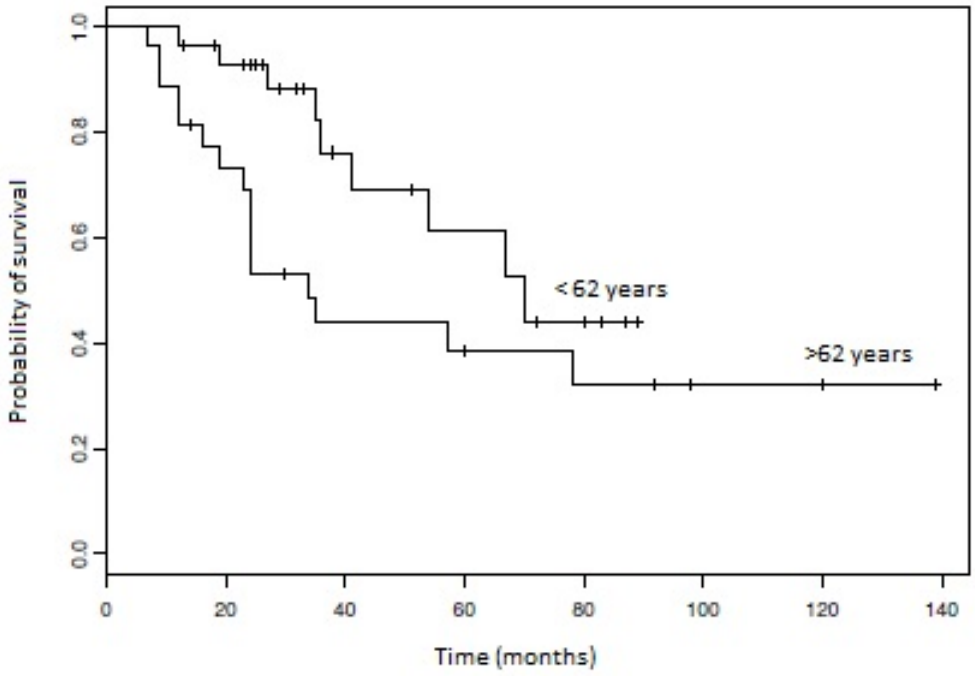
| Variable | Estimation | Standard error | $Z^2$ | p-value | $h$ |
|---|---|---|---|---|---|
| Dose | $b_1 = 0.08668$ | 0.4983 | 0.03 | 0.86 | 1.09 |
| Age | $b_2 = 0.04786$ | 0.0201 | 5.68 | 0.017 | 1.05 |
| Gender | $b_3 = -0.8987$ | 0.4718 | 3.63 | 0.057 | 0.41 |

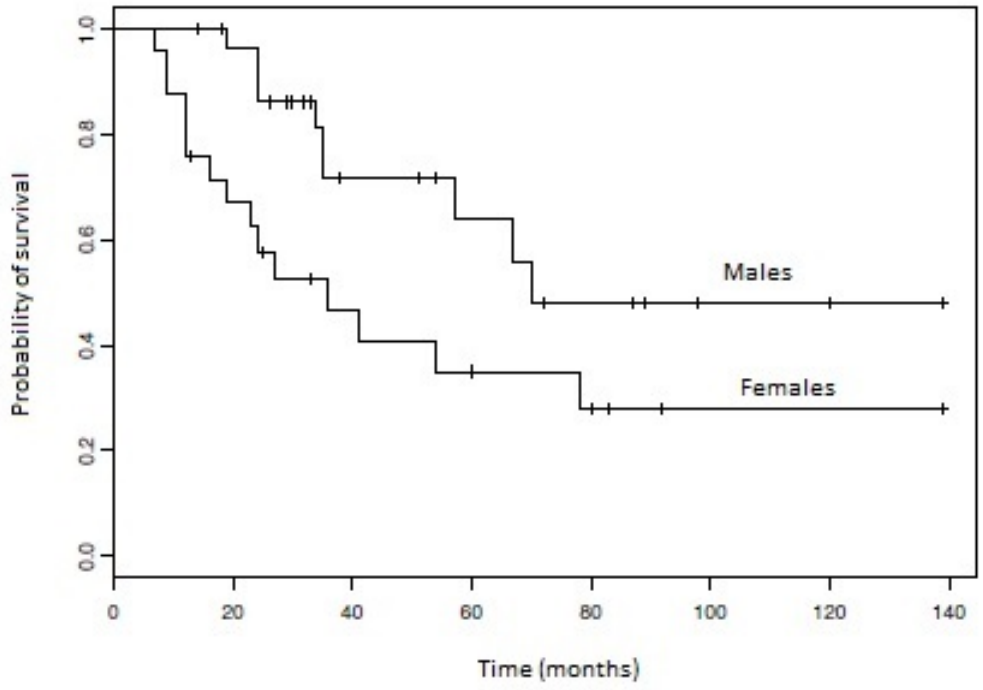
It is clear that radiation dose had no effect on survival (p=0.86). By contrast, age turned to be a significant factor ($p$=0.017) in the sense that the older the patient the higher the risk of death (positive coefficient). The hazard ratio was equal to 1.05 indicating a 5 percent risk increase each year after cancer surgery. As far as gender was concerned, an almost significant effect was noted ($p$=0.057). The coefficient being negative, hazards tended to be lower in male patients.

The likelihood ratio test showed that Cox model made sense since $LR = 11.17$ on 3 degrees of freedom ($p$=0.011). Likewise, Wald test was equal to 10.36 on 3 degrees of freedom ($p$=0.016).

Forward variable selection evidenced that both age ($p$=0.012) and gender ($p$=0.038) were jointly significant risk factors. Based on these factors, the risk index wrote $\underset{\sim}{b}^T\underset{\sim}{x} = 0.046 \times \text{ age } - 0.86 \times \text{gender}$ and the likelihood ratio test was $LR = 11.14$ on 2 degrees of freedom ($p$=0.0038). The hazard ratio was equal to 1.05 for age and 0.42 for gender; these values were close to those observed for the full model.

The figures below illustrate the effect of age on survival time by comparing patients with age below ($n$=29) and above ($n$=27) the median age, as well as the effect of gender (females verus males).




Probability of survival
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
120
140
Time (months)
Males
Females

# Chapter 8

# Discriminant analysis

## 8.1 Introduction

Discriminant analysis (DA) is one of the oldest and most important methods in multivariate statistics. In general, it concerns two or more populations of subjects (or objects) and a set of variables. The question is whether these variables "discriminate", that is differentiate or discern the populations. If this is the case, then the variables are discriminant. In medicine, populations are usually groups of patients suffering from various diseases and the variables may be clinical, biological, or radiological tests. Can these tests characterize and discern the diseases? This is a problem of differential diagnosis. In archeology, can measurements and observations made on excavated objects distinguish between periods when these objects were produced? In psychology, one often wants to differentiate anxiety profiles based on a specific questionnaire administered to the individuals? Examples of such kinds are numerous and can be found in all domains of life.

Today, discriminant analysis is predominantly used as a method to classify subjects (objects) among several populations. The main question states "How can we allocate a subject of unknown origin into one of several populations based on a single multivariate observation on this subject with a minimal risk of error?".

The chapter starts with discriminant analysis for two groups. We will introduce Fisher linear discriminant function, the notions of prior and posterior probabilities, error rate and correct classification rate. The relation between linear discriminant function and logistic regression will be discussed. We shall also envisage the variable selection problem by choosing the most discriminant variables. Then we shall consider discriminant analysis on more than two populations. This can be handled by a method called canonical

discriminant analysis, a method quite similar to that of principal component analysis. An alternative solution will consist in constructing several discriminant functions which will allow the calculation of posterior probabilities. Methods will be illustrated on various examples.

## 8.2 Discrimination between two groups

Consider two groups (populations), denoted $H_1$ and $H_2$, and a $p$-variate vector $\underset{\sim}{X}^T = (X_1, \ldots, X_p)$. Is it possible to discriminate between the two groups on the basis of $\underset{\sim}{X}$ ? Alternatively, can we classify a subject of unknown origin into $H_1$ or $H_2$ on the sole multivariate observation $X = \underset{\sim}{x}$ ? These are the salient questions arising in discriminant analysis. Let $\underset{\sim}{\mu}_1$ and $\underset{\sim}{\mu}_2$ the theoretical (unknown) mean vectors of the two populations and $\underset{\sim}{\sum}$ the variance-covariance (dispersion) matrix which we assume to be the same in the two populations (homoscedasticity assumption).

Now suppose we have a random sample of size $n_1$ from population $H_1$ and a sample of size $n_2$ from population $H_2$. This is a case of separate sampling ! The observation of vector $\underset{\sim}{X}$ in these two samples will generate two matrices of observations, respectively $\underset{\sim}{X}_{n_1 \times p}$ and $\underset{\sim}{X}_{n_2 \times p}$ as denoted before. These two matrices define in $\mathbb{R}^p$ two clusters (clouds) of points (see Chapter 2). From this perspective, the problem of discriminant analysis can be seen as examining whether the two clusters are more or less separated from each other or whether they overlap in a substantial way. Another approach would be find the hyperplane which separates them in the most satisfactory way.

## 8.3 Hotelling $T^2$ test

To determine whether vector $\underset{\sim}{X}$ discriminates the two groups $H_1$ and $H_2$, we need first to compare the mean vectors of $\underset{\sim}{X}$ in the two groups. If the mean vectors are not statistically different, then the problem is closed because the two clusters of points overlap markedly each other. Otherwise we can try to find a discriminant function which separates them.

To test the null hypothesis ($H_0$) of equality of the means of vector $\underset{\sim}{X}$ in the two groups $H_1$ and $H_2$, i.e. $\underset{\sim}{\mu}_1 = \underset{\sim}{\mu}_2$ against $\underset{\sim}{\mu}_1 \neq \underset{\sim}{\mu}_2$, we shall use Hotelling $T^2$ test. However this requires that vector $\underset{\sim}{X}$ consists only of normally distributed variables.

Using observation matrix $\underset{\sim}{X}_{n_1 \times p}$ from group $H_1$, we calculate (see Chapter 3) the sample mean vector $\underset{\sim}{\bar{x}}_1$ and the dispersion matrix $\underset{\sim}{S}_1$. We proceed in exactly the same way for observation matrix $\underset{\sim}{X}_{n_2 \times p}$ from group $H_2$, say $\underset{\sim}{\bar{x}}_2$ and $\underset{\sim}{S}_2$, respectively the mean vector and dispersion matrix.

Then we compute the pooled dispersion matrix,

$$\underset{\sim}{S} = \frac{(n_1 - 1)\underset{\sim}{S}_1 + (n_2 - 1)\underset{\sim}{S}_2}{n_1 + n_2 - 2} \tag{8.1}$$

and Hotelling's criterion

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2)^T \underset{\sim}{S}^{-1}(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2) \tag{8.2}$$

with on the right-hand side the Mahalanobis distance between the two means $\underset{\sim}{\bar{x}}_1$ and $\underset{\sim}{\bar{x}}_2$.

To test the equality of population means $\underset{\sim}{\mu}_1 = \underset{\sim}{\mu}_2$, it suffices to calculate the quantity

$$F = \frac{n_1 + n_2 - p - 1}{p(n_1 + n_2 - 2)} T^2 \tag{8.3}$$

distributed as a Snedecor $F$ test on $p$ and $n_1 + n_2 - p - 1$ degrees of freedom.

The null hypothesis $H_0$ is rejected if $F \geq Q_F(1 - \alpha; p, n_1 + n_2 - p - 1)$ the $(1 - \alpha)$-quantile of the Snedecor $F$ distribution on $p$ and $n_1 + n_2 - p - 1$ degrees of freedom, otherwise $H_0$ is not rejected. In the former case, vector $\underset{\sim}{X}$ discriminates $H_1$ and $H_2$ and a classification rule can be developed to allocate subjects into one or the other group.

## 8.4 Fisher linear discriminant function

When the difference between the two groups $H_1$ et $H_2$ is statistically significant, we can try to determine the direction which discriminates at best the two clusters of points in the $p$-dimensional space.

As we have seen, to switch from the $p$-dimensional space to the real line (1-dimensional space), it suffices to take a linear combination $\underset{\sim}{\beta}^T \underset{\sim}{X} = \beta_1 X_1 + \ldots + \beta_p X_p$ of vector $\underset{\sim}{X}$. Then, if we calculate $\underset{\sim}{\beta}^T \underset{\sim}{X}$ for all the observations of the data matrix $\underset{\sim}{X}_{n_1 \times p}$ of group 1 and those of $\underset{\sim}{X}_{n_2 \times p}$ of group 2, we obtain two univariate samples with means respectively equal to $\underset{\sim}{\beta}^T \underset{\sim}{\bar{x}}_1$ and $\underset{\sim}{\beta}^T \underset{\sim}{\bar{x}}_2$, whereas the pooled variance equals $\underset{\sim}{\beta}^T \underset{\sim}{S} \underset{\sim}{\beta}$ where $\underset{\sim}{S}$ is defined in (8.1).

Next we shall search for the direction (i.e. vector $\underset{\sim}{\beta}$) which maximizes the distance between the two means, that is $[\underset{\sim}{\beta}^T\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\beta}^T\underset{\sim}{\bar{x}}_2]^2 = [\underset{\sim}{\beta}^T(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2)]^2$ which accounts for the within group variance $\underset{\sim}{\beta}^T\underset{\sim}{S}\underset{\sim}{\beta}$. In other words, we look for the vector $\underset{\sim}{\beta}$ maximizing the "standardized" distance (ratio)

$$\lambda = \frac{[\underset{\sim}{\beta}^T(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2)]^2}{\underset{\sim}{\beta}^T\underset{\sim}{S}\underset{\sim}{\beta}} \tag{8.4}$$

It is easily shown that this direction is equal to $\underset{\sim}{\hat{\beta}} = \underset{\sim}{b} = \underset{\sim}{S}^{-1}(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2)$, an expression in which everything is known.

Then, the linear combination

$$\begin{aligned} L(\underset{\sim}{x}) &= \underset{\sim}{b}^T\underset{\sim}{x} \\ &= \left(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2\right)^T \underset{\sim}{S}^{-1}\underset{\sim}{x} \end{aligned} \tag{8.5}$$

is called "Fisher linear discriminant function". This function provides the best separation between the two groups (clusters of points) in the $p$-dimensional space.

## 8.5 Classification rule

To classify (allocate) a new individual (multivariate observation) $\underset{\sim}{x}$ in group $H_1$ or in group $H_2$ based on Fisher linear discriminant function $L(\underset{\sim}{x})$, we need to define a decision threshold (cut-off point). We may take the middle of the two mean values of $L(\underset{\sim}{x})$ in $H_1$ and $H_2$, that is

$$\begin{aligned} c &= \frac{L(\underset{\sim}{\bar{x}}_1) + L(\underset{\sim}{\bar{x}}_2)}{2} \\ &= \left(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2\right)^T \underset{\sim}{S}^{-1}\left(\frac{\underset{\sim}{\bar{x}}_1 + \underset{\sim}{\bar{x}}_2}{2}\right) \end{aligned} \tag{8.6}$$

Then, the subject $\underset{\sim}{x}$ would be classified in $H_1$ if $L(\underset{\sim}{x}) \geq c$ and in $H_2$ if $L(\underset{\sim}{x}) < c$. When subtracting the constant threshold value $c$ from $L(x)$, Fisher linear discriminant function becomes

$$L^*(\underset{\sim}{x}) = \left(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2\right)^T \underset{\sim}{S}^{-1}\left[\underset{\sim}{x} - \frac{1}{2}\left(\underset{\sim}{\bar{x}}_1 + \underset{\sim}{\bar{x}}_2\right)\right] \tag{8.7}$$

and the classification rule simplifies as follows:

$$\begin{array}{l} \text{“Classify } \underset{\sim}{x} \text{ in } H_1 \text{ if } L^*(\underset{\sim}{x}) \geq 0\text{”} \\ \text{“Classify } \underset{\sim}{x} \text{ in } H_2 \text{ if } L^*(\underset{\sim}{x}) < 0\text{”} \end{array} \tag{8.8}$$

The equation $L^*(\underset{\sim}{x}) = 0$ defines in the observation space $\mathbb{R}^P$ a hyperplane which separates the two clusters of points.

## 8.6 Error rate

To assess the quality of the allocation rule, or equivalently of Fisher linear discriminant function, one calculates for each observation $\underset{\sim}{x_i}$ of group $H_1$ and of group $H_2$ the value of $L^*(\underset{\sim}{x_i})$ defined in (8.7). In principle, for an observation of group $H_1$ we should expect $L^*(\underset{\sim}{x_i}) \geq 0$ whereas for an observation of group $H_2$ $L^*(\underset{\sim}{x_i}) < 0$. If this is not the case, then we can consider that the observation has been misclassified (error). Actually, two kinds of errors are possible : either observation $\underset{\sim}{x_i}$ comes from group $H_1$ but $L^*(\underset{\sim}{x_i}) < 0$ and therefore allocated to group $H_2$ (type 1 error) or observation $\underset{\sim}{x_i}$ comes from group $H_2$ but $L^*(\underset{\sim}{x_i}) \geq 0$ and therefore allocated to group $H_1$ (type 2 error). Let $r_1$ be the number of errors for group $H_1$ and $r_2$ the number of errors for group $H_2$. In these conditions, the “resubstitution” error rate of the classification rule (8.8) is

$$\varepsilon_R = \frac{r_1 + r_2}{n_1 + n_2} \tag{8.9}$$

## 8.7 Prior and posterior probabilities

When discriminant analysis is mainly considered as an allocation problem, there is a need to introduce the concepts of prior (a priori) and posterior (a posteriori) probabilities.

By definition, the prior probabilities $\pi_1 = P(H_1)$ and $\pi_2 = P(H_2)$ are the probabilities of belonging to groups $H_1$ and $H_2$ before any observation is made of vector $\underset{\sim}{X}$. They represent the proportions of the two groups in the mixture. Note that $\pi_1 + \pi_2 = 1$. For example, if $\pi_1 = 0.10$ and $\pi_2 = 0.90$, by taking a subject at random, there is only 10% chances the subject belongs to $H_1$. This element (piece of information) needs to be taken into account when classifying the subject after an observation has been made $\underset{\sim}{X} = \underset{\sim}{x}$.

The posterior probabilities are the probabilities of belonging to groups $H_1$ and $H_2$ after having observed vector $\underset{\sim}{X}$. They will be denoted

$$\begin{aligned} P_1(\underset{\sim}{x}) &= P[H_1|\underset{\sim}{x}] \\ \text{and} \\ P_2(\underset{\sim}{x}) &= P[H_2|\underset{\sim}{x}] \end{aligned} \tag{8.10}$$

and we still have $P_1(\underset{\sim}{x}) + P_2(\underset{\sim}{x}) = 1$.

Under specific normality assumptions, it can be shown that

$$\begin{aligned} P_1(\underset{\sim}{x}) &= \frac{e^{\log\frac{\pi_1}{\pi_2}+L^*(\underset{\sim}{x})}}{1+e^{\log\frac{\pi_1}{\pi_2}+L^*(\underset{\sim}{x})}} \\ \text{and} \\ P_2(\underset{\sim}{x}) &= \frac{1}{1+e^{\log\frac{\pi_1}{\pi_2}+L^*(\underset{\sim}{x})}} \end{aligned} \tag{8.11}$$

where $L^*(\underset{\sim}{x})$ is Fisher linear discriminant function defined in (8.7). We immediately recognize the logistic form of posterior probabilities.

Interestingly, the classification rule (8.8) can be stated as follows:

$$\begin{aligned} &\text{"Classify } \underset{\sim}{x} \text{ in } H_1 \text{ if } P_1(\underset{\sim}{x}) \geq P_2(\underset{\sim}{x}) \\ &\text{and in } H_2 \text{ otherwise."} \end{aligned} \tag{8.12}$$

Thus, the subject is allocated to the group with the highest posterior probability, which in fact sounds quite a natural and logical way of deciding. Posterior probabilities given in formulas (8.11) play a major role in discriminant analysis.

In terms of the Fisher linear discriminant function, the allocation rule (8.8) or (8.12) can also be written

$$\begin{aligned} &\text{"Classify } \underset{\sim}{x} \text{ in } H_1 \text{ if } L^*(\underset{\sim}{x}) \geq \log\frac{\pi_2}{\pi_1} \\ &\text{and in } H_2 \text{ otherwise."} \end{aligned} \tag{8.13}$$

When $\pi_1 = \pi_2$, we retrieve allocation rule (8.8). Further, when $\pi_2 > \pi_1$ (greater prior chances to belong to $H_2$), $\log\frac{\pi_2}{\pi_1} > 0$ and $L^*(\underset{\sim}{x})$ must exceed a positive threshold for $\underset{\sim}{x}$ being classified into $H_1$ (this threshold is more severe than the zero threshold). The converse holds when $\pi_2 < \pi_1$.

## 8.8 Additional remarks

As in regression analysis, there are techniques (ascending, descending or stepwise) to select the best discriminating variables without significant loss of information. They are based on a Snedecor $F$ test on 1 and $n-k-1$ degrees of freedom.

It has been shown that Fisher linear discriminant function could be obtained up to a constant factor by computing the multiple regression of a dummy binary variable $Y$ equal to 0 for $H_1$ and 1 for $H_2$. This is convenient as it suffices to run a classical multiple regression analysis on the multivariate dataset.

As far as the error rate ($\varepsilon$) is concerned, it can be estimated by other methods than just by resubstitution. First observe that, when $\underset{\sim}{X}$ is multinormal, the error rate can be expressed in an analytical form and then obtained by the "plug-in" method as follows:

$$\varepsilon_I = F_Z\left(-\frac{D}{2}\right) \tag{8.14}$$

where $F_Z(t)$ is the cumulative standard normal distribution, namely the function $P(Z \leq t)$ and $D$ the square root of Mahalanobis distance $(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2)^T \underset{\sim}{S}^{-1}(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2)$.

An other approach, called the "leaving-one-out" or "jackknife" method, consists in taking one observation out of the sample of $n = n_1 + n_2$ observations, calculating Fisher linear discriminant function based on the sample of the $n-1$ remaining observations and reallocating the observation left out. This process of leaving-one-out is repeated for all $n_1 + n_2$ observations. If $r'_1$ and $r'_2$ denote the number of misclassified observations, the error rate is given by the ratio

$$\varepsilon_J = \frac{r'_1 + r'_2}{n_1 + n_2} \tag{8.15}$$

In general, the error rate calculated by the jackknife method gives a more realistic estimation of the true error rate since it can be proven that $\varepsilon_J \geq \varepsilon_R$.

## 8.9 Application

To illustrate the discrimination between 2 groups and the use of Fisher linear discriminant function, we applied DA to Fisher iris versicolor (group $H_1$) and virginica (group $H_2$) datasets using the 4 traditional variables: sepal length ($X_1$), sepal width ($X_2$), petal length ($X_3$) and petal width ($X_4$). As we have

50 flowers in each group, $n = n_1 + n_2 = 100$. The table below reports the mean $\pm$ SD of each variable in the two groups.

| Variable | Iris versicolor $n_1 = 50$ | Iris virginica $n_2 = 50$ | $(\bar{x}_1 + \bar{x}_2)/2$ |
|---|---|---|---|
| Sepal length | $59.4 \pm 5.16$ | $65.9 \pm 6.36$ | 62.7 |
| Sepal width | $27.7 \pm 3.14$ | $29.7 \pm 3.23$ | 28.7 |
| Petal length | $42.6 \pm 4.70$ | $55.5 \pm 5.52$ | 49.1 |
| Petal width | $13.3 \pm 1.98$ | $20.3 \pm 2.75$ | 16.8 |

The dispersion matrices of the two groups $\underset{\sim}{S}_1$ and $\underset{\sim}{S}_2$ and the pooled dispersion matrix are respectively equal to

$$\underset{\sim}{S}_1 = \begin{pmatrix} 26.64 & & & \\ 8.518 & 9.847 & & \\ 18.29 & 8.265 & 22.08 & \\ 5.578 & 4.120 & 7.310 & 3.911 \end{pmatrix}$$

$$\underset{\sim}{S}_2 = \begin{pmatrix} 40.43 & & & \\ 9.376 & 10.40 & & \\ 30.33 & 7.138 & 30.46 & \\ 4.909 & 4.763 & 4.882 & 7.543 \end{pmatrix}$$

$$\underset{\sim}{S} = \begin{pmatrix} 33.54 & & & \\ 8.947 & 10.12 & & \\ 24.31 & 7.702 & 26.27 & \\ 5.244 & 4.442 & 6.096 & 5.727 \end{pmatrix}$$

After inverting matrix $\underset{\sim}{S}$, the Mahalanobis distance between the 2 groups writes

$$\begin{aligned} D^2 &= (\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2)^T \underset{\sim}{S}^{-1} (\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2) \\ &= 14.219 \end{aligned}$$

The Hotelling test is equal to

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} D^2 = \frac{2500}{100} \times 14.219 = 355.48$$

which yields a Snedecor $F$ value of

$$\begin{aligned} F &= \frac{50 + 50 - 4 - 1}{4 \times (100 - 2)} \times 355.48 \\ &= 86.15 \end{aligned}$$

The degrees of freedom are respectively equal to $\nu_1 = p = 4$ and $\nu_1 = n_1 + n_2 - p - 1 = 95$. Since the 95% percentile of the Snedecor F distribution is

$$Q_F(0.95; 4, 95) = 2.47$$

the null hypothesis $H_0 : \underset{\sim}{\mu}_1 = \underset{\sim}{\mu}_2$ is rejected. We conclude that the vector $\underset{\sim}{X}^T = (X_1, X_2, X_3, X_4)$ of sepals and petals discriminate the two iris species, versicolor and virginica. Note that the p-value associated with the Snedecor F test is highly significant ($p < 0.0001$).

Fisher linear discriminant function writes

$$\begin{aligned} L(\underset{\sim}{x}) &= (\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2)^T \underset{\sim}{S}^{-1} \underset{\sim}{x} \\ &= 0.3556X_1 + 0.5579X_2 - 0.6970X_3 - 1.239X_4 \end{aligned}$$

Since $c = -16.66$, the corrected Fisher linear discriminant function becomes

$$L^*(\underset{\sim}{x}) = 16.66 + 0.3556X_1 + 0.5579X_2 - 0.6970X_3 - 1.239X_4$$

Then, a flower will be classified in the iris versicolor species if $L^*(\underset{\sim}{x}) \geq 0$ and in the iris virginica species if $L^*(\underset{\sim}{x}) < 0$. As an illustration, by replacing vector $\underset{\sim}{X}$ by the mean values of the iris versicolor flowers (see table above), we naturally get a positive value. Indeed,

$$\begin{aligned} L^*(\underset{\sim}{x}) &= 16.66 + 0.3556 \times 59.4 + 0.5579 \times 27.7 - 0.6970 \times 42.6 - 1.239 \times 13.3 \\ &= 7.11 \end{aligned}$$

This value is nothing else than half of the Mahalanobis distance. In contrast, by substituting the mean values of iris virginica flowers, we obtain the same value $L^*(\underset{\sim}{x}) = -7.11$ but with an opposite sign.

When calculating the discriminant function $L^*(\underset{\sim}{x})$ for all the flowers of group 1, only one flower is misclassified ($r_1 = 1$). If we do the same for all the flowers of group 2, two of them are misclassified ($r_2 = 2$). Thus, the resubstitution error rate amounts

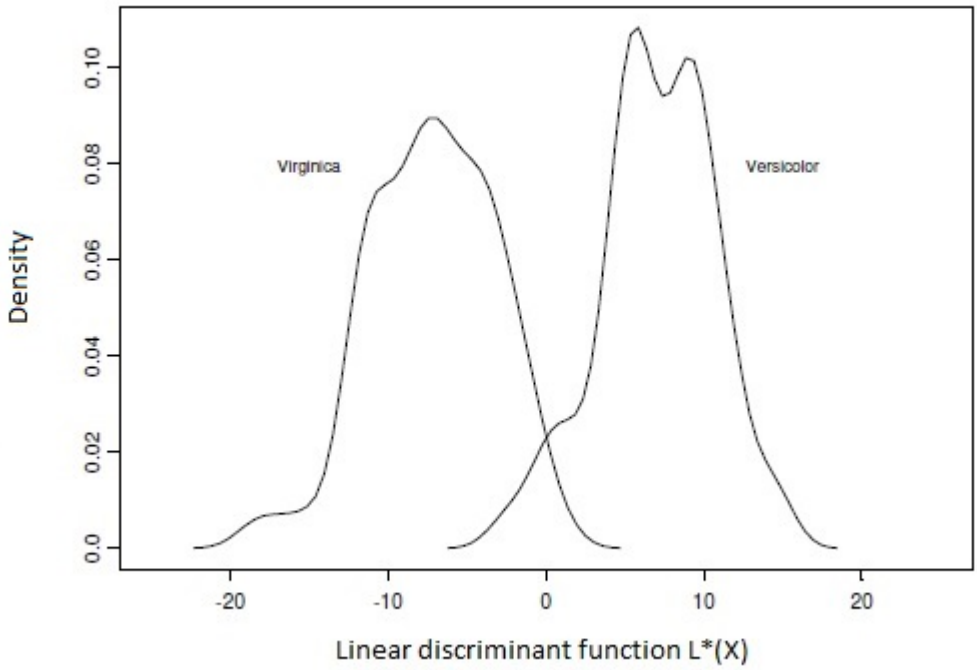
$$\varepsilon_R = \frac{1+2}{50+50} = 0.03$$

Since $D/2 = 1.8854$, the plug-in error rate

$$\varepsilon_I = F_G(-1.8854) = 0.0297,$$

is close to the value obtained by resubstitution. When using the jackknife method, the error rate is unchanged because $r' = 1$ and $r'_2 = 2$.

In all calculations above, the prior probabilities of the two groups were assumed to be the same, namely $\pi_1 = \pi_2 = 0.50$. The figure below which displays the density function of $L^*(\underset{\sim}{x})$ in each group, confirms the only slight degree of overlap between iris versicolor and iris virginica.




## 8.10 Logistic discrimination

In section 8.2.7 we mentioned the relationship between 2-group discriminant analysis and multiple regression. Today however most people prefer to solve the discriminant analysis problem by logistic regression. Nonetheless, it is essential to understand that in discriminant analysis the variable $Y$ is fixed and vector $\underset{\sim}{X}$ is observed, something we have not seen so far; actually it is the reverse situation of regression. In a remarkable proof, Anderson has shown that despite this particular (separate) sampling scheme a logistic regression can be performed as if the group variable $Y$ had been observed but

a correction should be applied to the intercept (constant term) of the linear discriminant function; specifically the correction concerns the decision threshold in order to account for the true prior probabilities. Indeed, when applying a logistic regression analysis, everything works as if the true prior probabilities were the sample proportions, i.e. $\pi_1 = n_1/n$ and $\pi_2 = n_2/n$, which is obviously not the case. The advantages of logistic regression result from the fact that it applies in all generality and that posterior probabilities are assumed to have the logistic form.

When performing a logistic regression of the binary variable $Y$(0 for $H_1$ and 1 for $H_2$) on vector $\underset{\sim}{X}$, we obtain the linear function

$$\Lambda(\underset{\sim}{x}) = b_0 + b_1 x_1 + \ldots + b_p x_p \tag{8.16}$$

which is not exactly the same as the (corrected) Fisher linear discriminant function derived by multiple regression.

If $\pi_1$ and $\pi_2$ are the true prior probabilities, a correction is applied to the constant term $b_0$ as follows

$$b_0^* = b_0 + \log \frac{n_2 \pi_1}{n_1 \pi_2} \tag{8.17}$$

Then, the logistic discriminant function

$$\Lambda^*(\underset{\sim}{x}) = b_0^* + \underset{\sim}{b}^T \underset{\sim}{x} \tag{8.18}$$

leads to the following classification rule :

$$\begin{array}{l} \text{“Classify } \underset{\sim}{x} \text{ in } H_1 \text{ if } \Lambda^*(\underset{\sim}{x}) \geq 0 \\ \text{and in } H_2 \text{ otherwise.”} \end{array} \tag{8.19}$$

Moreover, the posterior probabilities write

$$\begin{array}{l} P(H_1|\underset{\sim}{x}) = \dfrac{e^{\Lambda^*(\underset{\sim}{x})}}{1 + e^{\Lambda^*(\underset{\sim}{x})}} \\ \text{and} \\ P(H_2|\underset{\sim}{x}) = \dfrac{1}{1 + e^{\Lambda^*(\underset{\sim}{x})}} \end{array} \tag{8.20}$$

The resubstitution error rate is easily calculated but the recourse to the jackknife approach can quickly become very tedious. Further, variable selection methods as those described in Chapter 6 are basically also applicable.

We applied logistic regression to iris versicolor and virginica datasets and obtained the following results:

$$\Lambda(\underset{\sim}{x}) = 42.64 + 0.2465 X_1 + 0.6681 X_2 - 0.9429 X_3 - 1.829 X_4$$

Since $\pi_1 = \pi_2$ and $n_1 = n_2$, the correction term vanishes and $b_0^* = b_0$. Thus, $\Lambda^*(\underset{\sim}{x}) = \Lambda(\underset{\sim}{x})$. The likelihood ratio test is $LR = 126.7$ on 4 degrees of freedom; this evidences a highly significant difference bteween the two groups ($p<0.0001$). We also notice that the logistic discriminant function is not exactly equal to the Fisher linear discriminant function. The ratio between the coefficients of the two equations varies between 0.68 and 2.56 (mean value 1.46). It is theoretically proven that the relation between the logistic function and the standard normal (Gaussian) cumulative distribution function entails a factor equal to $\sqrt{8/\pi} = 1.60$. The mean value found above (1.46) is close to this multiplicative factor.

Using the logistic discriminant function above, a flower is allocated to the iris versicolor group when $\Lambda(\underset{\sim}{x}) \geq 0$, and to the iris virginica group otherwise.

## 8.11 Discrimination between several groups

### 8.11.1 Problem definition

Consider now the problem of discrimination (or equivalently of classification) between $g$ groups $H_1, \ldots, H_g$ based on vector $\underset{\sim}{X}^T = (X_1, \ldots, X_p)$.

To solve this problem, it is easy to show that $g-1$ discriminant functions are needed, say $L_1(\underset{\sim}{x}), \ldots, L_{g-1}(\underset{\sim}{x})$. Note that if $g = 2$ (case above), only one ($g - 1 = 1$) discriminant function is required.

Suppose that we have a random sample (matrix of observations) from each group. Let $\underset{\sim}{X}_{n_1 \times p}, \ldots, \underset{\sim}{X}_{n_g \times p}$ denote these matrices where the sample sizes $n_1, \ldots, n_g$ have been fixed by the investigators and $n = n_1 + \ldots + n_g$ is the total number of observations. For each data matrix we can calculate the mean vector $\underset{\sim}{\bar{x}}_i$ and the dispersion matrix $\underset{\sim}{S}_i$ ($i = 1, \ldots, g$). As before, we calculate the pooled matrix (under the assumption of homoscedasticity)

$$\underset{\sim}{S} = \frac{(n_1 - 1)\underset{\sim}{S}_1 + \ldots, +(n_g - 1)\underset{\sim}{S}_g}{n - g} \tag{8.21}$$

### 8.11.2 Canonical discriminant analysis

In the $p$-dimensional space, the $g$ data matrices correspond to $g$ clusters of points which overlap each other to various extents. Canonical discriminant analysis consists in obtaining an optimal graphical representation of these $p$-dimensional clusters on a plane (or in the 3-dimensional space). The axes

are chosen (as was the case for Fisher linear discriminant function) in such a way to maximize the between group variability (i.e. the differences between group mean values) with respect to the within group variability.

The between groups matrix of sums of squares and products is given by the expression

$$\underset{\sim}{H} = \sum_{j=1}^{g} \left(\bar{\underset{\sim}{x}}_j - \bar{\underset{\sim}{x}}\right) \left(\bar{\underset{\sim}{x}}_j - \bar{\underset{\sim}{x}}\right)^T \tag{8.22}$$

where $\bar{\underset{\sim}{x}} = (\bar{\underset{\sim}{x}}_1 + \ldots + \bar{\underset{\sim}{x}}_g)/g$, and the within groups matrix of sums of squares and products is equal to $\underset{\sim}{E} = (n-g)\underset{\sim}{S}$.

The problem resolution requires to determine the eigen values and eigen vectors of the matrix $\underset{\sim}{E}^{-1}\underset{\sim}{H}$, that is to find the roots of the equation

$$\left|\underset{\sim}{E}^{-1}\underset{\sim}{H} - \lambda\, \underset{\sim}{I}\right| = 0 \tag{8.23}$$

The rank of matrix $\underset{\sim}{E}^{-1}\underset{\sim}{H}$ is equal to the rank of $\underset{\sim}{H}$ because $\underset{\sim}{E}$, and hence $\underset{\sim}{E}^{-1}$, is of full rank. Given the definition (8.22), the rank of $\underset{\sim}{H}$ is less than or equal to $\min(g-1, p)$ and will be at most $g-1$ as long as $p \geq g$. Let $\lambda_1 > \ldots > \lambda_{g-1} \geq 0$, denote the $g-1$ eigen values in decreasing order.

The first canonical axis (variable) $Z_1 = \underset{\sim}{a}_1^T \underset{\sim}{X}$ corresponds to the largest eigen value $\lambda_1$ of matrix $\underset{\sim}{E}^{-1}\underset{\sim}{H}$ and $\underset{\sim}{a}_1$ is the corresponding eigen vector normalized to unity. In general, the first canonical variable is centered by subtracting the overall mean $\bar{\underset{\sim}{x}}$, soit $Z_1 = \underset{\sim}{a}_1^T(\underset{\sim}{x} - \bar{\underset{\sim}{x}})$.

The second canonical axis (variable) $Z_2 = \underset{\sim}{a}_2^T \underset{\sim}{X}$ corresponds to the second largest eigen value of $\underset{\sim}{E}^{-1}\underset{\sim}{H}$ and $\underset{\sim}{a}_2$ is the corresponding orthonormal eigen vector.

We can proceed the same way for the other canonical variables $Z_3, \ldots, Z_{g-1}$ but in general we restrict ourselves to the first two.

The quality of the discrimination will be assessed by the ratio

$$Q = \frac{\lambda_1 + \lambda_2}{\mathrm{tr}\left(\underset{\sim}{E}^{-1}\underset{\sim}{H}\right)} \tag{8.24}$$

as in principal component analysis.

The two canonical variables $Z_1 = \underset{\sim}{a}_1^T(\underset{\sim}{x} - \bar{\underset{\sim}{x}})$ and $Z_2 = \underset{\sim}{a}_2^T(\underset{\sim}{x} - \bar{\underset{\sim}{x}})$ will allow us to plot all the multivariate observations as well as the group means on

the plane thus defined. Further, groups will be easy to visualize and localize with each other.

If $\lambda_3 = 0$ then all subsequent eigen values $\lambda_4, \ldots, \lambda_{g-1}$ will also be 0 and we get an exact representation of the multivariate clusters on the plane. This will always be the case for $g = 3$ since there is always a plane through 3 points !

When $g = 2$, the first and only canonical variable will be Fisher linear discriminant function.

In canonical discriminant analysis, variable selection techniques (forward, backward, stepwise) are also feasible to keep only the most discriminating variables.

To illustrate the methodology, we applied canonical discriminant analysis on the entire Fisher iris dataset consisting 3 iris species ($H_1$=setosa, $H_2$=versicolor, $H_3$=virginica). The dataset has a total of 150 flowers and 4 variables as before (see Annex I). Since $g = 3$, we will have a perfect representation of the data on the canonical plane.

The table below reproduces the means ± SD of the 4 variables in the 3 groups and also globally:

| Variable | Iris setosa ($n_1 = 50$) | Iris versicolor ($n_2 = 50$) | Iris virginica ($n_3 = 50$) | Global ($n = 150$) |
|---|---|---|---|---|
| Sepal length | $50.1 \pm 3.52$ | $59.4 \pm 5.16$ | $65.9 \pm 6.36$ | $58.4 \pm 8.28$ |
| Sepal width | $34.3 \pm 3.79$ | $27.7 \pm 3.14$ | $29.7 \pm 3.23$ | $30.6 \pm 4.36$ |
| Petal length | $14.6 \pm 1.74$ | $42.6 \pm 4.70$ | $55.5 \pm 5.52$ | $37.6 \pm 17.7$ |
| Petal width | $2.46 \pm 1.05$ | $13.3 \pm 1.98$ | $20.3 \pm 2.75$ | $12.0 \pm 7.62$ |

The pooled variance-covariance matrix is

$$\underset{\sim}{S} = \begin{pmatrix} 26.50 & & & \\ 9.272 & 11.54 & & \\ 16.75 & 5.524 & 18.52 & \\ 3.84 & 3.271 & 4.267 & 4.188 \end{pmatrix}$$

A multivariate analysis of variance shows that the null hypothesis of equality of mean vectors of the three goups $H_0 : \underset{\sim}{\mu}_1 = \underset{\sim}{\mu}_2 = \underset{\sim}{\mu}_3$ is rejected ($p<0.0001$). Thus, vector $\underset{\sim}{X}$ discriminates the three iris groups.

The within groups matrix of sums of squares and products $\underset{\sim}{E} = (n - g)\underset{\sim}{S}$ writes

$$\underset{\sim}{E} = \begin{pmatrix} 3896 & & & \\ 1363 & 1696 & & \\ 2462 & 812.1 & 2722 & \\ 564.5 & 480.8 & 627.2 & 615.7 \end{pmatrix}$$

and the between groups matrix of sums of squares and products is

$$\underset{\sim}{H} = \begin{pmatrix} 6321 & & & \\ -1995 & 1135 & & \\ 16575 & -5724 & 43710 & \\ 7128 & -2293 & 18677 & 8041 \end{pmatrix}.$$

It follows that matrix $\underset{\sim}{E}^{-1}\underset{\sim}{H}$ writes

$$\underset{\sim}{E}^{-1}\underset{\sim}{H} = \begin{pmatrix} -3.052 & -5.565 & 8.075 & 10.493 \\ 1.079 & 2.180 & -2.942 & -3.418 \\ -8.096 & -14.975 & 21.505 & 27.538 \\ -3.452 & -6.312 & 9.140 & 11.840 \end{pmatrix}.$$

Note that this matrix is not symmetrical. Further, its rank is less than or equal to $min(g-1, p) = min(2, 4) = 2$.

The two non-zero eigen values of $\underset{\sim}{E}^{-1}\underset{\sim}{H}$ are respectively equal to $\lambda_1 = 32.19$ and $\lambda_2 = 0.2854$. Noting that the trace of the matrix is equal to 32.477, the first canonical variable (axis) represents alone 99.1% of the discrimination between the three groups.

The first and second canonical axes are respectively given by the expressions

$$Z_1 = -0.083X_1' - 0.153X_2' + 0.220X_3' + 0.281X_4'$$

and

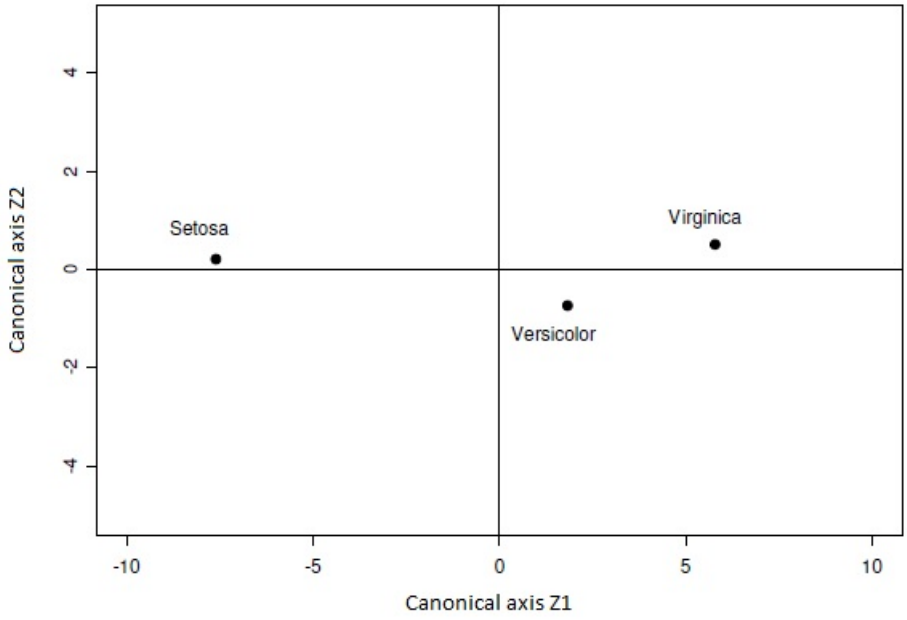
$$Z_2 = 0.002X_1' + 0.216X_2' - 0.093X_3' + 0.284X_4'$$

where the original variables were beforehand centered with respect to the overall mean, namely $X_1' = (X_1 - 58.4), X_2' = (X_2 - 30.6), X_3' = (X_3 - 37.6), X_4' = (X_4 - 12.0)$.

The mean vectors of the three groups projected on the 2-dimensional canonical space have the following components

$$\begin{aligned} &\text{Iris setosa} : \bar{Z}_1 = -7.61 \text{ et } \bar{Z}_2 = 0.215 \\ &\text{Iris versicolor} : \bar{Z}_1 = 1.83 \text{ et } \bar{Z}_2 = -0.728 \\ &\text{Iris virginica} : \bar{Z}_1 = 5.78 \text{ et } \bar{Z}_2 = 0.513 \end{aligned}$$

These points on the figure below illustrate the relative locations of the three groups in the 4-dimensional space.




## 8.12 Classification among several groups

Consider again $g$ groups $H_1, \ldots, H_g$ with prior probabilities $\pi_1, \ldots, \pi_g$ such that $(\pi_1 + \ldots + \pi_g = 1)$ and a vector $\underset{\sim}{X}^T = (X_1, \ldots, X_p)$. We also have a random sample (data matrix) $\underset{\sim}{X}_{n_i \times p}$ from each group $(i = 1, \ldots, g)$ with $n = n_1 + \ldots + n_g$.

Let $\bar{x}_1, \ldots, \bar{x}_g$ be the mean vectors and $\underset{\sim}{S}$ the pooled covariance matrix.

Then, to allocate a subject $\underset{\sim}{x}$ in one of the $g$ groups, $g-1$ discriminant functions are needed; choosing $\underset{\sim}{H}_g$ as the reference group, these functions are given by the expressions

$$
\begin{aligned}
L_1(\underset{\sim}{x}) &= (\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_g)^T \underset{\sim}{S}^{-1} \left[\underset{\sim}{x} - \tfrac{1}{2}\left(\bar{\underset{\sim}{x}}_1 + \bar{\underset{\sim}{x}}_g\right)\right] \\
L_2(\underset{\sim}{x}) &= (\bar{\underset{\sim}{x}}_2 - \bar{\underset{\sim}{x}}_g)^T \underset{\sim}{S}^{-1} \left[\underset{\sim}{x} - \tfrac{1}{2}\left(\bar{\underset{\sim}{x}}_2 + \bar{\underset{\sim}{x}}_g\right)\right] \\
&\ldots \\
L_{g-1}(\underset{\sim}{x}) &= (\bar{\underset{\sim}{x}}_{g-1} - \bar{\underset{\sim}{x}}_g)^T \underset{\sim}{S}^{-1} \left[\underset{\sim}{x} - \tfrac{1}{2}\left(\bar{\underset{\sim}{x}}_{g-1} + \bar{\underset{\sim}{x}}_g\right)\right]
\end{aligned}
\tag{8.25}
$$

When $g = 2$, we simply retrieve expression (8.7).

Next we calculate the posterior probabilities as follows

$$\begin{aligned} P_1(\underset{\sim}{x}) &= \frac{e^{\log\frac{\pi_1}{\pi_g}+L_1(\underset{\sim}{x})}}{1+\sum_{j=1}^{g-1} e^{\log\frac{\pi_j}{\pi_g}+L_j(\underset{\sim}{x})}} \\ &\dots \\ P_{g-1}(\underset{\sim}{x}) &= \frac{e^{\log\frac{\pi_{g-1}}{\pi_g}+L_{g-1}(\underset{\sim}{x})}}{1+\sum_{j=1}^{g-1} e^{\log\frac{\pi_j}{\pi_g}+L_j(\underset{\sim}{x})}} \\ P_g(\underset{\sim}{x}) &= 1 - P_1(\underset{\sim}{x}) - \ldots - P_{g-1}(\underset{\sim}{x}) \end{aligned} \tag{8.26}$$

When $g = 2$, we retrieve the equations (8.11).
At last, the classification rule can be stated as follows

$$\begin{aligned} &\text{“Classify the subject } \underset{\sim}{x} \text{ in the group with maximal} \\ &\text{posterior probability, i.e. classify in } H_j \text{ if} \\ &P_j(\underset{\sim}{x}) = \max\left\{P_1(\underset{\sim}{x}), \ldots, P_g(\underset{\sim}{x})\right\}\text{”} \end{aligned} \tag{8.27}$$

The performance of the allocation rule can be assessed by determining the classification matrix in which the true group of each observation is cross-classified with the allocated group as given by the decision rule.

| | True group | | | |
|---|---|---|---|---|
| Allocated group | $H_1$ | $H_2$ | $\ldots$ | $H_g$ |
| $H_1$ | $r_{11}$ | $r_{12}$ | $\ldots$ | $r_{1g}$ |
| $H_2$ | $r_{21}$ | $r_{22}$ | $\ldots$ | $r_{2g}$ |
| $\ldots$ | $\ldots$ | $\ldots$ | $\ldots$ | $\ldots$ |
| $H_g$ | $r_{g1}$ | $r_{g2}$ | $\ldots$ | $r_{gg}$ |
| Total | $n_1$ | $n_2$ | $\ldots$ | $n_g$ |

The “correct classification rate” is equal to

$$C = \frac{r_{11} + r_{22} + \ldots + r_{gg}}{n_1 + n_2 + \ldots + n_g} \tag{8.28}$$

and the total resubstitution error rate is given by $\varepsilon_R = 1 - C$.

Variable selection techniques can be applied as previously described in order to select the most discriminating variables.

Multiple group discriminant analysis can also be resolved by the “generalized” logistic regression method. This somewhat more complicated approach

yields $g-1$ discriminant functions

$$\begin{array}{l} \Lambda_1(\underset{\sim}{x}) = b_{01} + \underset{\sim}{b_1^T}\underset{\sim}{x} \\ \cdots \qquad \cdots \qquad \cdots \\ \Lambda_{g-1}(\underset{\sim}{x}) = b_{0,g-1} + \underset{\sim}{b_{g-1}^T}\underset{\sim}{x} \end{array} \tag{8.29}$$

but a correction has to be applied to the intercepts (constant terms) as follows

$$b^*_{0,i} = b_{0,i} + \log \frac{n_g \pi_i}{n_i \pi_g} \qquad (i = 1, \ldots, g-1) \tag{8.30}$$

as we did in the case of $g = 2$ groups (see (8.17)).

Then, posterior probabilities are given by the expressions

$$\begin{array}{l} P(H_i|\underset{\sim}{x}) = \dfrac{e^{\Lambda^*_i(\underset{\sim}{x})}}{1 + \sum_{j=1}^{g-1} e^{\Lambda^*_j(\underset{\sim}{x})}} \qquad (i = 1, \ldots, g-1) \\ \\ P(H_g|\underset{\sim}{x}) = 1 - \sum_{i=1}^{g-1} P(H_i|\underset{\sim}{x}) \end{array} \tag{8.31}$$

where $\Lambda^*_i(\underset{\sim}{x}) = b^*_{0,i} + \underset{\sim}{b_i^T}\underset{\sim}{x} \quad (i = 1, \ldots, g-1)$.

Remarks

1. Multiple group discriminant analysis can be seen as a regression problem where the "dependent" variable $Y$ is qualitative with $g$ modalities (categories). This is why it is also called "multinomial regression". We immediately notice that the problem cannot be solved with a single linear combination of the explanatory variables $X_1, \ldots, X_p$. Indeed the categories of $Y$ are not ordered and cannot be numbered in a unique way. As seen in Chapter 2, $g-1$ binary variables will be necessary. We now understand why $g-1$ discriminant functions are needed.

2. Most statistical packages do not immediately give the coefficients of the discriminant functions. Instead they provide the vectors of coefficients of so-called "discriminant scores", specifically the expressions

$$\underset{\sim}{d_i^T} = \underset{\sim}{\bar{x}_i^T}\underset{\sim}{S^{-1}} \quad (i = 1, \ldots, g)$$

as well as the intercepts (constants)

$$d_{0i} = \tfrac{1}{2} \underset{\sim}{d}_i^T \underset{\sim}{\bar{x}}_i = \tfrac{1}{2} \underset{\sim}{\bar{x}}_i^T \underset{\sim}{S}^{-1} \underset{\sim}{\bar{x}}_i \quad (i = 1, \ldots, g)$$

Hence, expressions (8.25) are easily derived as follows

$$L_i(\underset{\sim}{x}) = (\underset{\sim}{d}_i - \underset{\sim}{d}_g)^T \underset{\sim}{x} - (d_{0i} - d_{0g}) \quad (i = 1, \ldots, g)$$

Thus it suffices to calculate the differences betweeen discriminant scores and constants to obtain the discriminant functions. This way of presentation offers the avantage of freely choosing which group will be the reference group.

Based on the mean vectors of each iris species and the pooled dispersion matrix obtained previously, we can calculate the Mahalanobis distance between each two groups and the corresponding Hotelling $T^2$ tests. All distances (tests) are highly significant (p<0.0001), confirming the good discrimination between the three groups of flowers.

| | Setosa | Versicolor | Virginica |
|---|---|---|---|
| Setosa | 0 | | |
| Versicolor | 89.86 | 0 | |
| Virginica | 179.4 | 17.20 | 0 |

Assuming all prior probabilities equal $\pi_1 = \pi_2 = \pi_3$, the discriminant scores are given by the following expressions

| Variable | $d_{Setosa}$ | $d_{Versicolor}$ | $d_{Virginica}$ |
|---|---|---|---|
| Intercept | -85.21 | -71.75 | -103.27 |
| Sepal length | 2.354 | 1.570 | 1.245 |
| Sepal width | 2.359 | 0.707 | 0.369 |
| Petal length | -1.643 | 0.521 | 1.277 |
| Petal width | -1.734 | 0.643 | 2.108 |

Choosing group 3 as the reference group, the two discriminant functions write

$$\begin{aligned} L_1(\underset{\sim}{x}) &= 18.06 + 1.11X_1 + 1.99X_2 - 2.92X_3 - 3.85X_4 \\ L_2(\underset{\sim}{x}) &= 31.52 + 0.325X_1 + 0.339X_2 - 0.756X_3 - 1.465X_4. \end{aligned}$$

A flower will be classified in $H_1$ (iris setosa) if $L_1(\underset{\sim}{x}) \geq L_2(\underset{\sim}{x})$ and $L_1(\underset{\sim}{x}) \geq 0$, in $H_2$ (iris versicolor) if $L_2(\underset{\sim}{x}) > L_1(\underset{\sim}{x})$ and $L_2(\underset{\sim}{x}) \geq 0$, in $H_3$ (iris virginica) otherwise, i.e. if $L_1(\underset{\sim}{x}) < 0$ and $L_2(\underset{\sim}{x}) < 0$.

The classification matrix obtained by the resubstitution method is given below

| | True group | | | |
|---|---|---|---|---|
| Allocated group | Setosa | Versicolor | Virginica | Total |
| Setosa | 50 | 0 | 0 | 50 |
| Versicolor | 0 | 48 | 1 | 49 |
| Virginica | 0 | 2 | 49 | 51 |
| Total | 50 | 50 | 50 | 150 |

It is seen that the iris setosa group is completely separated from the two other groups which in turn only slightly overlap as we have seen before. The total correct classification rate amounts $C = (50 + 48 + 49)/150 = 0.98$ so that the total error is only $\varepsilon_r = 0.02$ (2%). We found that the jackknife method gave exactly the same value.

# Annex I - Fisher iris dataset

| | Setosa | | | | | | Versicolor | | | | | | Virginica | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| #. | Y | $X_1$ | $X_2$ | $X_3$ | $X_4$ | #. | Y | $X_1$ | $X_2$ | $X_3$ | $X_4$ | #. | Y | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
| 1 | 1 | 50 | 33 | 14 | 2 | 1 | 2 | 65 | 28 | 46 | 15 | 1 | 3 | 64 | 28 | 56 | 22 |
| 2 | 1 | 46 | 34 | 14 | 3 | 2 | 2 | 62 | 22 | 45 | 15 | 2 | 3 | 67 | 31 | 56 | 24 |
| 3 | 1 | 46 | 36 | 10 | 2 | 3 | 2 | 59 | 32 | 48 | 18 | 3 | 3 | 63 | 28 | 51 | 15 |
| 4 | 1 | 51 | 33 | 17 | 5 | 4 | 2 | 61 | 30 | 46 | 14 | 4 | 3 | 69 | 31 | 51 | 23 |
| 5 | 1 | 55 | 35 | 13 | 2 | 5 | 2 | 60 | 27 | 51 | 16 | 5 | 3 | 65 | 30 | 52 | 20 |
| 6 | 1 | 48 | 31 | 16 | 2 | 6 | 2 | 56 | 25 | 39 | 11 | 6 | 3 | 65 | 30 | 55 | 18 |
| 7 | 1 | 52 | 34 | 14 | 2 | 7 | 2 | 57 | 28 | 45 | 13 | 7 | 3 | 58 | 27 | 51 | 19 |
| 8 | 1 | 49 | 36 | 14 | 1 | 8 | 2 | 63 | 33 | 47 | 16 | 8 | 3 | 68 | 32 | 59 | 23 |
| 9 | 1 | 44 | 32 | 13 | 2 | 9 | 2 | 70 | 32 | 47 | 14 | 9 | 3 | 62 | 34 | 54 | 23 |
| 10 | 1 | 50 | 35 | 16 | 6 | 10 | 2 | 64 | 32 | 45 | 15 | 10 | 3 | 77 | 38 | 67 | 22 |
| 11 | 1 | 44 | 30 | 13 | 2 | 11 | 2 | 61 | 28 | 40 | 13 | 11 | 3 | 67 | 33 | 57 | 25 |
| 12 | 1 | 47 | 32 | 16 | 2 | 12 | 2 | 55 | 24 | 38 | 11 | 12 | 3 | 76 | 30 | 66 | 21 |
| 13 | 1 | 48 | 30 | 14 | 3 | 13 | 2 | 54 | 30 | 45 | 15 | 13 | 3 | 49 | 25 | 45 | 17 |
| 14 | 1 | 51 | 38 | 16 | 2 | 14 | 2 | 58 | 26 | 40 | 12 | 14 | 3 | 67 | 30 | 52 | 23 |
| 15 | 1 | 48 | 34 | 19 | 2 | 15 | 2 | 55 | 26 | 44 | 12 | 15 | 3 | 59 | 30 | 51 | 18 |
| 16 | 1 | 50 | 30 | 16 | 2 | 16 | 2 | 50 | 23 | 33 | 10 | 16 | 3 | 63 | 25 | 50 | 19 |
| 17 | 1 | 50 | 32 | 12 | 2 | 17 | 2 | 67 | 31 | 44 | 14 | 17 | 3 | 64 | 32 | 53 | 23 |
| 18 | 1 | 43 | 30 | 11 | 1 | 18 | 2 | 56 | 30 | 45 | 15 | 18 | 3 | 79 | 38 | 64 | 20 |
| 19 | 1 | 58 | 40 | 12 | 2 | 19 | 2 | 58 | 27 | 41 | 10 | 19 | 3 | 67 | 33 | 57 | 21 |
| 20 | 1 | 51 | 38 | 19 | 4 | 20 | 2 | 60 | 29 | 45 | 15 | 20 | 3 | 77 | 28 | 67 | 20 |
| 21 | 1 | 49 | 30 | 14 | 2 | 21 | 2 | 57 | 26 | 35 | 10 | 21 | 3 | 63 | 27 | 49 | 18 |
| 22 | 1 | 51 | 35 | 14 | 2 | 22 | 2 | 57 | 29 | 42 | 13 | 22 | 3 | 72 | 32 | 60 | 18 |
| 23 | 1 | 50 | 34 | 16 | 4 | 23 | 2 | 49 | 24 | 33 | 10 | 23 | 3 | 61 | 30 | 49 | 18 |
| 24 | 1 | 46 | 32 | 14 | 2 | 24 | 2 | 56 | 27 | 42 | 13 | 24 | 3 | 61 | 26 | 56 | 14 |
| 25 | 1 | 57 | 44 | 15 | 4 | 25 | 2 | 57 | 30 | 42 | 12 | 25 | 3 | 64 | 28 | 56 | 21 |
| 26 | 1 | 50 | 36 | 14 | 2 | 26 | 2 | 66 | 29 | 46 | 13 | 26 | 3 | 62 | 28 | 48 | 18 |
| 27 | 1 | 54 | 34 | 15 | 4 | 27 | 2 | 52 | 27 | 39 | 14 | 27 | 3 | 77 | 30 | 61 | 23 |
| 28 | 1 | 52 | 41 | 15 | 1 | 28 | 2 | 60 | 34 | 45 | 16 | 28 | 3 | 63 | 34 | 56 | 24 |
| 29 | 1 | 55 | 42 | 14 | 2 | 29 | 2 | 50 | 20 | 35 | 10 | 29 | 3 | 58 | 27 | 51 | 19 |
| 30 | 1 | 49 | 31 | 15 | 2 | 30 | 2 | 55 | 24 | 37 | 10 | 30 | 3 | 72 | 30 | 58 | 16 |
| 31 | 1 | 54 | 39 | 17 | 4 | 31 | 2 | 58 | 27 | 39 | 12 | 31 | 3 | 71 | 30 | 59 | 21 |
| 32 | 1 | 50 | 34 | 15 | 2 | 32 | 2 | 62 | 29 | 43 | 13 | 32 | 3 | 64 | 31 | 55 | 18 |
| 33 | 1 | 44 | 29 | 14 | 2 | 33 | 2 | 59 | 30 | 42 | 15 | 33 | 3 | 60 | 30 | 48 | 18 |
| 34 | 1 | 47 | 32 | 13 | 2 | 34 | 2 | 60 | 22 | 40 | 10 | 34 | 3 | 63 | 29 | 56 | 18 |
| 35 | 1 | 46 | 31 | 15 | 2 | 35 | 2 | 67 | 31 | 47 | 15 | 35 | 3 | 77 | 26 | 69 | 23 |
| 36 | 1 | 51 | 34 | 15 | 2 | 36 | 2 | 63 | 23 | 44 | 13 | 36 | 3 | 60 | 22 | 50 | 15 |
| 37 | 1 | 50 | 35 | 13 | 3 | 37 | 2 | 56 | 30 | 41 | 13 | 37 | 3 | 69 | 32 | 57 | 23 |
| 38 | 1 | 49 | 31 | 15 | 1 | 38 | 2 | 63 | 25 | 49 | 15 | 38 | 3 | 74 | 28 | 61 | 19 |
| 39 | 1 | 54 | 37 | 15 | 2 | 39 | 2 | 61 | 28 | 47 | 12 | 39 | 3 | 56 | 28 | 49 | 20 |
| 40 | 1 | 54 | 39 | 13 | 4 | 40 | 2 | 64 | 29 | 43 | 13 | 40 | 3 | 73 | 29 | 63 | 18 |
| 41 | 1 | 51 | 35 | 14 | 3 | 41 | 2 | 51 | 25 | 30 | 11 | 41 | 3 | 67 | 25 | 58 | 18 |
| 42 | 1 | 48 | 34 | 16 | 2 | 42 | 2 | 57 | 28 | 41 | 13 | 42 | 3 | 65 | 30 | 58 | 22 |
| 43 | 1 | 48 | 30 | 14 | 1 | 43 | 2 | 61 | 29 | 47 | 14 | 43 | 3 | 69 | 31 | 54 | 21 |
| 44 | 1 | 45 | 23 | 13 | 3 | 44 | 2 | 56 | 29 | 36 | 13 | 44 | 3 | 72 | 36 | 61 | 25 |
| 45 | 1 | 57 | 38 | 17 | 3 | 45 | 2 | 69 | 31 | 49 | 15 | 45 | 3 | 65 | 32 | 51 | 20 |
| 46 | 1 | 51 | 38 | 15 | 3 | 46 | 2 | 55 | 25 | 40 | 13 | 46 | 3 | 64 | 27 | 53 | 19 |
| 47 | 1 | 54 | 34 | 17 | 2 | 47 | 2 | 55 | 23 | 40 | 13 | 47 | 3 | 68 | 30 | 55 | 21 |
| 48 | 1 | 51 | 37 | 15 | 4 | 48 | 2 | 66 | 30 | 44 | 14 | 48 | 3 | 57 | 25 | 50 | 20 |
| 49 | 1 | 52 | 35 | 15 | 2 | 49 | 2 | 68 | 28 | 48 | 14 | 49 | 3 | 58 | 28 | 51 | 24 |
| 50 | 1 | 53 | 37 | 15 | 2 | 50 | 2 | 67 | 30 | 50 | 17 | 50 | 3 | 63 | 33 | 60 | 25 |

$X_1$ = Sepal length (mm)
$X_2$ = Sepal width (mm)
$X_3$ = Petal length (mm)
$X_4$ = Petal width (mm)
Y= Group (1=iris setosa 2=iris versicolor, 3=iris virginica)

# Annex II - Severe head injury dataset

Outcome at 6 months (1=low or moderate disability, 2=severe disability or persistent vegetative state, 3=death), age (years), cerebrospinal fluid (CSF) CK-BB isoenzyme (UI/l) in 60 patients with severe head injury. (Source: Hans et al., 1985)

| Patient | Outcome | Age | CK-BB | Patient | Outcome | Age | CK-BB |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 19 | 100 | 31 | 3 | 59 | 76 |
| 2 | 1 | 11 | 220 | 32 | 3 | 61 | 303 |
| 3 | 1 | 38 | 6 | 33 | 3 | 45 | 1560 |
| 4 | 1 | 7 | 281 | 34 | 3 | 61 | 353 |
| 5 | 1 | 17 | 17 | 35 | 3 | 20 | 1370 |
| 6 | 1 | 19 | 27 | 36 | 3 | 24 | 671 |
| 7 | 1 | 12 | 96 | 37 | 3 | 22 | 60 |
| 8 | 1 | 8 | 23 | 38 | 3 | 30 | 356 |
| 9 | 1 | 24 | 253 | 39 | 3 | 20 | 543 |
| 10 | 1 | 16 | 60 | 40 | 3 | 45 | 120 |
| 11 | 1 | 18 | 126 | 41 | 3 | 16 | 700 |
| 12 | 1 | 12 | 100 | 42 | 3 | 16 | 16 |
| 13 | 1 | 8 | 200 | 43 | 3 | 50 | 216 |
| 14 | 1 | 28 | 70 | 44 | 3 | 16 | 800 |
| 15 | 1 | 10 | 146 | 45 | 3 | 19 | 90 |
| 16 | 1 | 46 | 46 | 46 | 3 | 19 | 303 |
| 17 | 1 | 35 | 40 | 47 | 3 | 11 | 183 |
| 18 | 1 | 6 | 136 | 48 | 3 | 18 | 740 |
| 19 | 1 | 6 | 286 | 49 | 3 | 15 | 1256 |
| 20 | 2 | 8 | 230 | 50 | 3 | 56 | 523 |
| 21 | 2 | 23 | 509 | 51 | 3 | 40 | 350 |
| 22 | 2 | 19 | 283 | 52 | 3 | 18 | 126 |
| 23 | 2 | 4 | 140 | 53 | 3 | 18 | 153 |
| 24 | 2 | 29 | 80 | 54 | 3 | 20 | 913 |
| 25 | 2 | 23 | 576 | 55 | 3 | 19 | 193 |
| 26 | 2 | 20 | 76 | 56 | 3 | 41 | 323 |
| 27 | 2 | 62 | 206 | 57 | 3 | 51 | 443 |
| 28 | 3 | 17 | 253 | 58 | 3 | 29 | 156 |
| 29 | 3 | 29 | 490 | 59 | 3 | 21 | 463 |
| 30 | 3 | 7 | 1087 | 60 | 3 | 20 | 230 |

# Annex III - Rectal cancer patients

Lifetime (months), preoperative dose of radiotherapy (0 = < 5000 rads, 1 = ≥ 5000 rads), age (years), gender (0 = *female*, 1 = *male*) in 56 patients with rectal cancer. (Source : Harris et Albert, 1991)

| Patient | Lifetime | Dose | Age | Gender | Patient | Lifetime | Dose | Age | Gender |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 7 | 0 | 68 | 0 | 29 | 19 | 1 | 60 | 0 |
| 2 | 9 | 0 | 69 | 0 | 30 | 23* | 1 | 54 | 0 |
| 3 | 12 | 0 | 68 | 0 | 31 | 24* | 1 | 62 | 1 |
| 4 | 12 | 0 | 71 | 0 | 32 | 25* | 1 | 55 | 0 |
| 5 | 19 | 0 | 77 | 1 | 33 | 26* | 1 | 39 | 1 |
| 6 | 23 | 0 | 70 | 0 | 34 | 27 | 1 | 50 | 0 |
| 7 | 24 | 0 | 67 | 0 | 35 | 29* | 1 | 58 | 1 |
| 8 | 24 | 0 | 68 | 1 | 36 | 30* | 1 | 75 | 1 |
| 9 | 24 | 0 | 88 | 1 | 37 | 32* | 1 | 61 | 1 |
| 10 | 24 | 0 | 89 | 1 | 38 | 33* | 1 | 53 | 0 |
| 11 | 29* | 0 | 28 | 1 | 39 | 33* | 1 | 57 | 1 |
| 12 | 34 | 0 | 73 | 1 | 40 | 35 | 1 | 50 | 1 |
| 13 | 41 | 0 | 60 | 0 | 41 | 35 | 1 | 78 | 1 |
| 14 | 54 | 0 | 60 | 0 | 42 | 35* | 1 | 55 | 1 |
| 15 | 72* | 0 | 44 | 1 | 43 | 35* | 1 | 65 | 1 |
| 16 | 78 | 0 | 82 | 0 | 44 | 35* | 1 | 73 | 1 |
| 17 | 80* | 0 | 62 | 0 | 45 | 36 | 1 | 53 | 0 |
| 18 | 83* | 0 | 53 | 0 | 46 | 38* | 1 | 47 | 1 |
| 19 | 92* | 0 | 66 | 0 | 47 | 51* | 1 | 60 | 1 |
| 20 | 139* | 0 | 63 | 0 | 48 | 54* | 1 | 54 | 1 |
| 21 | 139* | 0 | 68 | 1 | 49 | 57 | 1 | 66 | 1 |
| 22 | 9 | 1 | 77 | 0 | 50 | 60* | 1 | 64 | 0 |
| 23 | 12 | 1 | 55 | 0 | 51 | 67 | 1 | 60 | 1 |
| 24 | 12* | 1 | 78 | 0 | 52 | 70 | 1 | 41 | 1 |
| 25 | 13* | 1 | 47 | 0 | 53 | 87* | 1 | 58 | 1 |
| 26 | 14* | 1 | 69 | 1 | 54 | 89* | 1 | 45 | 1 |
| 27 | 16 | 1 | 68 | 0 | 55 | 98* | 1 | 73 | 1 |
| 28 | 18* | 1 | 62 | 1 | 56 | 120* | 1 | 63 | 1 |

* Censored lifetime

# Annex IV - Statistical tables

Table A Gaussian (Normal) distribution $Z \sim N(0,1)$.
Upper probabilities $P[Z > z]$, $z \geq 0$

Table B Chi-squared distribution on $\nu$ degrees of freedom $\chi^2_\nu$
Upper percentiles $Q_{\chi^2}(1-\alpha;\nu)$

Table C Student $t$ distribution on $\nu$ degrees of freedom $t_\nu$
Upper percentiles $Q_t(1-\frac{\alpha}{2};\nu)$
Note : row $\nu = \infty$ corresponds to Gaussian upper percentiles $Q_Z(1-\frac{\alpha}{2})$

Table D Snedecor F distribution on $\nu_1$ and $\nu_2$ degrees of freedom $F_{\nu_1,\nu_2}$
Upper percentiles $Q_F(1-\alpha;\nu_1,\nu_2)$ for $\alpha = 0.05$ and $\alpha = 0.01$

Table A Gaussian distribution $Z \sim N(0,1)$
Upper probabilities $P[Z > z]$, $z \geq 0$

| z | *0.00* | *0.01* | *0.02* | *0.03* | *0.04* | *0.05* | *0.06* | *0.07* | *0.08* | *0.09* |
|---|---|---|---|---|---|---|---|---|---|---|
| *0.0* | 0.5000 | 0.4960 | 0.4920 | 0.4880 | 0.4840 | 0.4801 | 0.4761 | 0.4721 | 0.4681 | 0.4641 |
| *0.1* | 0.4602 | 0.4562 | 0.4522 | 0.4483 | 0.4443 | 0.4404 | 0.4364 | 0.4325 | 0.4286 | 0.4247 |
| *0.2* | 0.4207 | 0.4168 | 0.4129 | 0.4090 | 0.4052 | 0.4013 | 0.3974 | 0.3936 | 0.3897 | 0.3859 |
| *0.3* | 0.3821 | 0.3783 | 0.3745 | 0.3707 | 0.3669 | 0.3632 | 0.3594 | 0.3557 | 0.3520 | 0.3483 |
| *0.4* | 0.3446 | 0.3409 | 0.3372 | 0.3336 | 0.3300 | 0.3264 | 0.3228 | 0.3192 | 0.3156 | 0.3121 |
| *0.5* | 0.3085 | 0.3050 | 0.3015 | 0.2981 | 0.2946 | 0.2912 | 0.2877 | 0.2843 | 0.2810 | 0.2776 |
| *0.6* | 0.2743 | 0.2709 | 0.2676 | 0.2643 | 0.2611 | 0.2578 | 0.2546 | 0.2514 | 0.2483 | 0.2451 |
| *0.7* | 0.2420 | 0.2389 | 0.2358 | 0.2327 | 0.2296 | 0.2266 | 0.2236 | 0.2206 | 0.2177 | 0.2148 |
| *0.8* | 0.2119 | 0.2090 | 0.2061 | 0.2033 | 0.2005 | 0.1977 | 0.1949 | 0.1922 | 0.1894 | 0.1867 |
| *0.9* | 0.1841 | 0.1814 | 0.1788 | 0.1762 | 0.1736 | 0.1711 | 0.1685 | 0.1660 | 0.1635 | 0.1611 |
| *1.0* | 0.1587 | 0.1562 | 0.1539 | 0.1515 | 0.1492 | 0.1469 | 0.1446 | 0.1423 | 0.1401 | 0.1379 |
| *1.1* | 0.1357 | 0.1335 | 0.1314 | 0.1292 | 0.1271 | 0.1251 | 0.1230 | 0.1210 | 0.1190 | 0.1170 |
| *1.2* | 0.1151 | 0.1131 | 0.1112 | 0.1093 | 0.1075 | 0.1056 | 0.1038 | 0.1020 | 0.1003 | 0.0985 |
| *1.3* | 0.0968 | 0.0951 | 0.0934 | 0.0918 | 0.0901 | 0.0885 | 0.0869 | 0.0853 | 0.0838 | 0.0823 |
| *1.4* | 0.0808 | 0.0793 | 0.0778 | 0.0764 | 0.0749 | 0.0735 | 0.0721 | 0.0708 | 0.0694 | 0.0681 |
| *1.5* | 0.0668 | 0.0655 | 0.0643 | 0.0630 | 0.0618 | 0.0606 | 0.0594 | 0.0582 | 0.0571 | 0.0559 |
| *1.6* | 0.0548 | 0.0537 | 0.0526 | 0.0516 | 0.0505 | 0.0495 | 0.0485 | 0.0475 | 0.0465 | 0.0455 |
| *1.7* | 0.0446 | 0.0436 | 0.0427 | 0.0418 | 0.0409 | 0.0401 | 0.0392 | 0.0384 | 0.0375 | 0.0367 |
| *1.8* | 0.0359 | 0.0351 | 0.0344 | 0.0336 | 0.0329 | 0.0322 | 0.0314 | 0.0307 | 0.0301 | 0.0294 |
| *1.9* | 0.0287 | 0.0281 | 0.0274 | 0.0268 | 0.0262 | 0.0256 | 0.0250 | 0.0244 | 0.0239 | 0.0233 |
| *2.0* | 0.0228 | 0.0222 | 0.0217 | 0.0212 | 0.0207 | 0.0202 | 0.0197 | 0.0192 | 0.0188 | 0.0183 |
| *2.1* | 0.0179 | 0.0174 | 0.0170 | 0.0166 | 0.0162 | 0.0158 | 0.0154 | 0.0150 | 0.0146 | 0.0143 |
| *2.2* | 0.0139 | 0.0136 | 0.0132 | 0.0129 | 0.0125 | 0.0122 | 0.0119 | 0.0116 | 0.0113 | 0.0110 |
| *2.3* | 0.0107 | 0.0104 | 0.0102 | 0.0099 | 0.0096 | 0.0094 | 0.0091 | 0.0089 | 0.0087 | 0.0084 |
| *2.4* | 0.0082 | 0.0080 | 0.0078 | 0.0075 | 0.0073 | 0.0071 | 0.0069 | 0.0068 | 0.0066 | 0.0064 |
| *2.5* | 0.0062 | 0.0060 | 0.0059 | 0.0057 | 0.0055 | 0.0054 | 0.0052 | 0.0051 | 0.0049 | 0.0048 |
| *2.6* | 0.0047 | 0.0045 | 0.0044 | 0.0043 | 0.0041 | 0.0040 | 0.0039 | 0.0038 | 0.0037 | 0.0036 |
| *2.7* | 0.0035 | 0.0034 | 0.0033 | 0.0032 | 0.0031 | 0.0030 | 0.0029 | 0.0028 | 0.0027 | 0.0026 |
| *2.8* | 0.0026 | 0.0025 | 0.0024 | 0.0023 | 0.0023 | 0.0022 | 0.0021 | 0.0021 | 0.0020 | 0.0019 |
| *2.9* | 0.0019 | 0.0018 | 0.0018 | 0.0017 | 0.0016 | 0.0016 | 0.0015 | 0.0015 | 0.0014 | 0.0014 |
| *3.0* | 0.0013 | 0.0013 | 0.0013 | 0.0012 | 0.0012 | 0.0011 | 0.0011 | 0.0011 | 0.0010 | 0.0010 |
| *3.1* | 0.0010 | 0.0009 | 0.0009 | 0.0009 | 0.0008 | 0.0008 | 0.0008 | 0.0008 | 0.0007 | 0.0007 |
| *3.2* | 0.0007 | 0.0007 | 0.0006 | 0.0006 | 0.0006 | 0.0006 | 0.0006 | 0.0005 | 0.0005 | 0.0005 |
| *3.3* | 0.0005 | 0.0005 | 0.0005 | 0.0004 | 0.0004 | 0.0004 | 0.0004 | 0.0004 | 0.0004 | 0.0003 |
| *3.4* | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0002 |

Table B Chi-squared distribution on $\nu$ degrees of freedom $\chi^2_\nu$
Upper percentiles $Q_{\chi^2}(1-\alpha;\nu)$

| | $1-\alpha$ | | | | | |
|---|---|---|---|---|---|---|
| $\nu$ | 0.90 | 0.95 | 0.975 | 0.99 | 0.995 | 0.999 |
| *1* | 2.706 | 3.841 | 5.024 | 6.635 | 7.879 | 10.83 |
| *2* | 4.605 | 5.991 | 7.378 | 9.210 | 10.60 | 13.82 |
| *3* | 6.251 | 7.815 | 9.348 | 11.34 | 12.84 | 16.27 |
| *4* | 7.779 | 9.488 | 11.14 | 13.28 | 14.86 | 18.47 |
| *5* | 9.236 | 11.07 | 12.83 | 15.09 | 16.75 | 20.52 |
| *6* | 10.64 | 12.59 | 14.45 | 16.81 | 18.55 | 22.46 |
| *7* | 12.02 | 14.07 | 16.01 | 18.48 | 20.28 | 24.32 |
| *8* | 13.36 | 15.51 | 17.53 | 20.09 | 21.95 | 26.12 |
| *9* | 14.68 | 16.92 | 19.02 | 21.67 | 23.59 | 27.88 |
| *10* | 15.99 | 18.31 | 20.48 | 23.21 | 25.19 | 29.59 |
| *11* | 17.28 | 19.68 | 21.92 | 24.72 | 26.76 | 31.26 |
| *12* | 18.55 | 21.03 | 23.34 | 26.22 | 28.30 | 32.91 |
| *13* | 19.81 | 22.36 | 24.74 | 27.69 | 29.82 | 34.53 |
| *14* | 21.06 | 23.68 | 26.12 | 29.14 | 31.32 | 36.12 |
| *15* | 22.31 | 25.00 | 27.49 | 30.58 | 32.80 | 37.70 |
| *16* | 23.54 | 26.30 | 28.85 | 32.00 | 34.27 | 39.25 |
| *17* | 24.77 | 27.59 | 30.19 | 33.41 | 35.72 | 40.79 |
| *18* | 25.99 | 28.87 | 31.53 | 34.81 | 37.16 | 42.31 |
| *19* | 27.20 | 30.14 | 32.85 | 36.19 | 38.58 | 43.82 |
| *20* | 28.41 | 31.41 | 34.17 | 37.57 | 40.00 | 45.31 |
| *21* | 29.62 | 32.67 | 35.48 | 38.93 | 41.40 | 46.80 |
| *22* | 30.81 | 33.92 | 36.78 | 40.29 | 42.80 | 48.27 |
| *23* | 32.01 | 35.17 | 38.08 | 41.64 | 44.18 | 49.73 |
| *24* | 33.20 | 36.42 | 39.36 | 42.98 | 45.56 | 51.18 |
| *25* | 34.38 | 37.65 | 40.65 | 44.31 | 46.93 | 52.62 |
| *26* | 35.56 | 38.89 | 41.92 | 45.64 | 48.29 | 54.05 |
| *27* | 36.74 | 40.11 | 43.19 | 46.96 | 49.64 | 55.48 |
| *28* | 37.92 | 41.34 | 44.46 | 48.28 | 50.99 | 56.89 |
| *29* | 39.09 | 42.56 | 45.72 | 49.59 | 52.34 | 58.30 |
| *30* | 40.26 | 43.77 | 46.98 | 50.89 | 53.67 | 59.70 |
| *35* | 46.06 | 49.80 | 53.20 | 57.34 | 60.27 | 66.62 |
| *40* | 51.81 | 55.76 | 59.34 | 63.69 | 66.77 | 73.40 |
| *45* | 57.51 | 61.66 | 65.41 | 69.96 | 73.17 | 80.08 |
| *50* | 63.17 | 67.50 | 71.42 | 76.15 | 79.49 | 86.66 |
| *60* | 74.40 | 79.08 | 83.30 | 88.38 | 91.95 | 99.61 |
| *70* | 85.53 | 90.53 | 95.02 | 100.4 | 104.2 | 112.3 |
| *80* | 96.58 | 101.9 | 106.6 | 112.3 | 116.3 | 124.8 |
| *90* | 107.6 | 113.1 | 118.1 | 124.1 | 128.3 | 137.2 |
| *100* | 118.5 | 124.3 | 129.6 | 135.8 | 140.2 | 149.4 |

Table C Student $t$ distribution on $\nu$ degrees of freedom $t_\nu$
Upper percentiles $Q_t(1-\frac{\alpha}{2};\nu)$

| | $1-\frac{\alpha}{2}$ | | | | | |
|---|---|---|---|---|---|---|
| $\nu$ | 0.90 | 0.95 | 0.975 | 0.99 | 0.995 | 0.999 |
| *1* | 3.0777 | 6.3138 | 12.706 | 31.821 | 63.657 | 318.31 |
| *2* | 1.8856 | 2.9200 | 4.3027 | 6.9646 | 9.9248 | 22.327 |
| *3* | 1.6377 | 2.3534 | 3.1824 | 4.5407 | 5.8409 | 10.215 |
| *4* | 1.5332 | 2.1318 | 2.7764 | 3.7469 | 4.6041 | 7.1732 |
| *5* | 1.4759 | 2.0150 | 2.5706 | 3.3649 | 4.0321 | 5.8934 |
| *6* | 1.4398 | 1.9432 | 2.4469 | 3.1427 | 3.7074 | 5.2076 |
| *7* | 1.4149 | 1.8946 | 2.3646 | 2.9980 | 3.4995 | 4.7853 |
| *8* | 1.3968 | 1.8595 | 2.3060 | 2.8965 | 3.3554 | 4.5008 |
| *9* | 1.3830 | 1.8331 | 2.2622 | 2.8214 | 3.2498 | 4.2968 |
| *10* | 1.3722 | 1.8125 | 2.2281 | 2.7638 | 3.1693 | 4.1437 |
| *11* | 1.3634 | 1.7959 | 2.2010 | 2.7181 | 3.1058 | 4.0247 |
| *12* | 1.3562 | 1.7823 | 2.1788 | 2.6810 | 3.0545 | 3.9296 |
| *13* | 1.3502 | 1.7709 | 2.1604 | 2.6503 | 3.0123 | 3.8520 |
| *14* | 1.3450 | 1.7613 | 2.1448 | 2.6245 | 2.9768 | 3.7874 |
| *15* | 1.3406 | 1.7531 | 2.1314 | 2.6025 | 2.9467 | 3.7328 |
| *16* | 1.3368 | 1.7459 | 2.1199 | 2.5835 | 2.9208 | 3.6862 |
| *17* | 1.3334 | 1.7396 | 2.1098 | 2.5669 | 2.8982 | 3.6458 |
| *18* | 1.3304 | 1.7341 | 2.1009 | 2.5524 | 2.8784 | 3.6105 |
| *19* | 1.3277 | 1.7291 | 2.0930 | 2.5395 | 2.8609 | 3.5794 |
| *20* | 1.3253 | 1.7247 | 2.0860 | 2.5280 | 2.8453 | 3.5518 |
| *21* | 1.3232 | 1.7207 | 2.0796 | 2.5176 | 2.8314 | 3.5272 |
| *22* | 1.3212 | 1.7171 | 2.0739 | 2.5083 | 2.8188 | 3.5050 |
| *23* | 1.3195 | 1.7139 | 2.0687 | 2.4999 | 2.8073 | 3.4850 |
| *24* | 1.3178 | 1.7109 | 2.0639 | 2.4922 | 2.7969 | 3.4668 |
| *25* | 1.3163 | 1.7081 | 2.0595 | 2.4851 | 2.7874 | 3.4502 |
| *26* | 1.3150 | 1.7056 | 2.0555 | 2.4786 | 2.7787 | 3.4350 |
| *27* | 1.3137 | 1.7033 | 2.0518 | 2.4727 | 2.7707 | 3.4210 |
| *28* | 1.3125 | 1.7011 | 2.0484 | 2.4671 | 2.7633 | 3.4082 |
| *29* | 1.3114 | 1.6991 | 2.0452 | 2.4620 | 2.7564 | 3.3962 |
| *30* | 1.3104 | 1.6973 | 2.0423 | 2.4573 | 2.7500 | 3.3852 |
| *35* | 1.3062 | 1.6896 | 2.0301 | 2.4377 | 2.7238 | 3.3400 |
| *40* | 1.3031 | 1.6839 | 2.0211 | 2.4233 | 2.7045 | 3.3069 |
| *45* | 1.3006 | 1.6794 | 2.0141 | 2.4121 | 2.6896 | 3.2815 |
| *50* | 1.2987 | 1.6759 | 2.0086 | 2.4033 | 2.6778 | 3.2614 |
| *60* | 1.2958 | 1.6706 | 2.0003 | 2.3901 | 2.6603 | 3.2317 |
| *70* | 1.2938 | 1.6669 | 1.9944 | 2.3808 | 2.6479 | 3.2108 |
| *80* | 1.2922 | 1.6641 | 1.9901 | 2.3739 | 2.6387 | 3.1953 |
| *90* | 1.2910 | 1.6620 | 1.9867 | 2.3685 | 2.6316 | 3.1833 |
| *100* | 1.2901 | 1.6602 | 1.9840 | 2.3642 | 2.6259 | 3.1737 |
| *120* | 1.2886 | 1.6577 | 1.9799 | 2.3578 | 2.6174 | 3.1595 |
| *200* | 1.2858 | 1.6525 | 1.9719 | 2.3451 | 2.6006 | 3.1315 |
| $\infty$ | 1.2816 | 1.6449 | 1.9600 | 2.3263 | 2.5758 | 3.0902 |

Table D Snedecor F distribution on $\nu_1$ and $\nu_2$ degrees of freedom
Upper percentiles $Q_F(1-\alpha;\nu_1,\nu_2)$ for $1-\alpha = 0.95$

| | $\nu_1$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\nu_2$ | *1* | *2* | *3* | *4* | *5* | *6* | *8* | *12* | *24* | $\infty$ |
| *1* | 161.45 | 199.50 | 215.71 | 224.58 | 230.16 | 233.99 | 238.88 | 243.91 | 249.05 | 254.3 |
| *2* | 18.51 | 19.00 | 19.16 | 19.25 | 19.30 | 19.33 | 19.37 | 19.41 | 19.45 | 19.50 |
| *3* | 10.13 | 9.55 | 9.28 | 9.12 | 9.01 | 8.94 | 8.85 | 8.74 | 8.64 | 8.53 |
| *4* | 7.71 | 6.94 | 6.59 | 6.39 | 6.26 | 6.16 | 6.04 | 5.91 | 5.77 | 5.63 |
| *5* | 6.61 | 5.79 | 5.41 | 5.19 | 5.05 | 4.95 | 4.82 | 4.68 | 4.53 | 4.37 |
| *6* | 5.99 | 5.14 | 4.76 | 4.53 | 4.39 | 4.28 | 4.15 | 4.00 | 3.84 | 3.67 |
| *7* | 5.59 | 4.74 | 4.35 | 4.12 | 3.97 | 3.87 | 3.73 | 3.57 | 3.41 | 3.23 |
| *8* | 5.32 | 4.46 | 4.07 | 3.84 | 3.69 | 3.58 | 3.44 | 3.28 | 3.12 | 2.93 |
| *9* | 5.12 | 4.26 | 3.86 | 3.63 | 3.48 | 3.37 | 3.23 | 3.07 | 2.90 | 2.71 |
| *10* | 4.96 | 4.10 | 3.71 | 3.48 | 3.33 | 3.22 | 3.07 | 2.91 | 2.74 | 2.54 |
| *11* | 4.84 | 3.98 | 3.59 | 3.36 | 3.20 | 3.09 | 2.95 | 2.79 | 2.61 | 2.40 |
| *12* | 4.75 | 3.89 | 3.49 | 3.26 | 3.11 | 3.00 | 2.85 | 2.69 | 2.51 | 2.30 |
| *13* | 4.67 | 3.81 | 3.41 | 3.18 | 3.03 | 2.92 | 2.77 | 2.60 | 2.42 | 2.21 |
| *14* | 4.60 | 3.74 | 3.34 | 3.11 | 2.96 | 2.85 | 2.70 | 2.53 | 2.35 | 2.13 |
| *15* | 4.54 | 3.68 | 3.29 | 3.06 | 2.90 | 2.79 | 2.64 | 2.48 | 2.29 | 2.07 |
| *16* | 4.49 | 3.63 | 3.24 | 3.01 | 2.85 | 2.74 | 2.59 | 2.42 | 2.24 | 2.01 |
| *17* | 4.45 | 3.59 | 3.20 | 2.96 | 2.81 | 2.70 | 2.55 | 2.38 | 2.19 | 1.96 |
| *18* | 4.41 | 3.55 | 3.16 | 2.93 | 2.77 | 2.66 | 2.51 | 2.34 | 2.15 | 1.92 |
| *19* | 4.38 | 3.52 | 3.13 | 2.90 | 2.74 | 2.63 | 2.48 | 2.31 | 2.11 | 1.88 |
| *20* | 4.35 | 3.49 | 3.10 | 2.87 | 2.71 | 2.60 | 2.45 | 2.28 | 2.08 | 1.84 |
| *21* | 4.32 | 3.47 | 3.07 | 2.84 | 2.68 | 2.57 | 2.42 | 2.25 | 2.05 | 1.81 |
| *22* | 4.30 | 3.44 | 3.05 | 2.82 | 2.66 | 2.55 | 2.40 | 2.23 | 2.03 | 1.78 |
| *23* | 4.28 | 3.42 | 3.03 | 2.80 | 2.64 | 2.53 | 2.37 | 2.20 | 2.01 | 1.76 |
| *24* | 4.26 | 3.40 | 3.01 | 2.78 | 2.62 | 2.51 | 2.36 | 2.18 | 1.98 | 1.73 |
| *25* | 4.24 | 3.39 | 2.99 | 2.76 | 2.60 | 2.49 | 2.34 | 2.16 | 1.96 | 1.71 |
| *26* | 4.23 | 3.37 | 2.98 | 2.74 | 2.59 | 2.47 | 2.32 | 2.15 | 1.95 | 1.69 |
| *27* | 4.21 | 3.35 | 2.96 | 2.73 | 2.57 | 2.46 | 2.31 | 2.13 | 1.93 | 1.67 |
| *28* | 4.20 | 3.34 | 2.95 | 2.71 | 2.56 | 2.45 | 2.29 | 2.12 | 1.91 | 1.65 |
| *29* | 4.18 | 3.33 | 2.93 | 2.70 | 2.55 | 2.43 | 2.28 | 2.10 | 1.90 | 1.64 |
| *30* | 4.17 | 3.32 | 2.92 | 2.69 | 2.53 | 2.42 | 2.27 | 2.09 | 1.89 | 1.62 |
| *40* | 4.08 | 3.23 | 2.84 | 2.61 | 2.45 | 2.34 | 2.18 | 2.00 | 1.79 | 1.51 |
| *60* | 4.00 | 3.15 | 2.76 | 2.53 | 2.37 | 2.25 | 2.10 | 1.92 | 1.70 | 1.39 |
| *120* | 3.92 | 3.07 | 2.68 | 2.45 | 2.29 | 2.18 | 2.02 | 1.83 | 1.61 | 1.25 |
| $\infty$ | 3.84 | 3.00 | 2.60 | 2.37 | 2.21 | 2.10 | 1.94 | 1.75 | 1.52 | 1.00 |

Table D Snedecor F distribution on $\nu_1$ and $\nu_2$ degrees of freedom
Upper percentiles $Q_F(1-\alpha;\nu_1,\nu_2)$ for $1-\alpha = 0.99$

| | $\nu_1$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\nu_2$ | *1* | *2* | *3* | *4* | *5* | *6* | *8* | *12* | *24* | $\infty$ |
| *1* | 4052 | 4999 | 5403 | 5625 | 5764 | 5859 | 5982 | 6106 | 6234 | 6366 |
| *2* | 98.50 | 99.00 | 99.17 | 99.25 | 99.30 | 99.33 | 99.37 | 99.42 | 99.46 | 99.50 |
| *3* | 34.12 | 30.82 | 29.46 | 28.71 | 28.24 | 27.91 | 27.49 | 27.05 | 26.60 | 26.13 |
| *4* | 21.20 | 18.00 | 16.69 | 15.98 | 15.52 | 15.21 | 14.80 | 14.37 | 13.93 | 13.46 |
| *5* | 16.26 | 13.27 | 12.06 | 11.39 | 10.97 | 10.67 | 10.29 | 9.89 | 9.47 | 9.02 |
| *6* | 13.75 | 10.92 | 9.78 | 9.15 | 8.75 | 8.47 | 8.10 | 7.72 | 7.31 | 6.88 |
| *7* | 12.25 | 9.55 | 8.45 | 7.85 | 7.46 | 7.19 | 6.84 | 6.47 | 6.07 | 5.65 |
| *8* | 11.26 | 8.65 | 7.59 | 7.01 | 6.63 | 6.37 | 6.03 | 5.67 | 5.28 | 4.86 |
| *9* | 10.56 | 8.02 | 6.99 | 6.42 | 6.06 | 5.80 | 5.47 | 5.11 | 4.73 | 4.31 |
| *10* | 10.04 | 7.56 | 6.55 | 5.99 | 5.64 | 5.39 | 5.06 | 4.71 | 4.33 | 3.91 |
| *11* | 9.65 | 7.21 | 6.22 | 5.67 | 5.32 | 5.07 | 4.74 | 4.40 | 4.02 | 3.61 |
| *12* | 9.33 | 6.93 | 5.95 | 5.41 | 5.06 | 4.82 | 4.50 | 4.16 | 3.78 | 3.36 |
| *13* | 9.07 | 6.70 | 5.74 | 5.21 | 4.86 | 4.62 | 4.30 | 3.96 | 3.59 | 3.17 |
| *14* | 8.86 | 6.51 | 5.56 | 5.04 | 4.69 | 4.46 | 4.14 | 3.80 | 3.43 | 3.01 |
| *15* | 8.68 | 6.36 | 5.42 | 4.89 | 4.56 | 4.32 | 4.00 | 3.67 | 3.29 | 2.87 |
| *16* | 8.53 | 6.23 | 5.29 | 4.77 | 4.44 | 4.20 | 3.89 | 3.55 | 3.18 | 2.76 |
| *17* | 8.40 | 6.11 | 5.18 | 4.67 | 4.34 | 4.10 | 3.79 | 3.46 | 3.08 | 2.66 |
| *18* | 8.29 | 6.01 | 5.09 | 4.58 | 4.25 | 4.01 | 3.71 | 3.37 | 3.00 | 2.57 |
| *19* | 8.18 | 5.93 | 5.01 | 4.50 | 4.17 | 3.94 | 3.63 | 3.30 | 2.92 | 2.49 |
| *20* | 8.10 | 5.85 | 4.94 | 4.43 | 4.10 | 3.87 | 3.56 | 3.23 | 2.86 | 2.42 |
| *21* | 8.02 | 5.78 | 4.87 | 4.37 | 4.04 | 3.81 | 3.51 | 3.17 | 2.80 | 2.36 |
| *22* | 7.95 | 5.72 | 4.82 | 4.31 | 3.99 | 3.76 | 3.45 | 3.12 | 2.75 | 2.31 |
| *23* | 7.88 | 5.66 | 4.76 | 4.26 | 3.94 | 3.71 | 3.41 | 3.07 | 2.70 | 2.26 |
| *24* | 7.82 | 5.61 | 4.72 | 4.22 | 3.90 | 3.67 | 3.36 | 3.03 | 2.66 | 2.21 |
| *25* | 7.77 | 5.57 | 4.68 | 4.18 | 3.85 | 3.63 | 3.32 | 2.99 | 2.62 | 2.17 |
| *26* | 7.72 | 5.53 | 4.64 | 4.14 | 3.82 | 3.59 | 3.29 | 2.96 | 2.58 | 2.13 |
| *27* | 7.68 | 5.49 | 4.60 | 4.11 | 3.78 | 3.56 | 3.26 | 2.93 | 2.55 | 2.10 |
| *28* | 7.64 | 5.45 | 4.57 | 4.07 | 3.75 | 3.53 | 3.23 | 2.90 | 2.52 | 2.06 |
| *29* | 7.60 | 5.42 | 4.54 | 4.04 | 3.73 | 3.50 | 3.20 | 2.87 | 2.49 | 2.03 |
| *30* | 7.56 | 5.39 | 4.51 | 4.02 | 3.70 | 3.47 | 3.17 | 2.84 | 2.47 | 2.01 |
| *40* | 7.31 | 5.18 | 4.31 | 3.83 | 3.51 | 3.29 | 2.99 | 2.66 | 2.29 | 1.80 |
| *60* | 7.08 | 4.98 | 4.13 | 3.65 | 3.34 | 3.12 | 2.82 | 2.50 | 2.12 | 1.60 |
| *120* | 6.85 | 4.79 | 3.95 | 3.48 | 3.17 | 2.96 | 2.66 | 2.34 | 1.95 | 1.38 |
| $\infty$ | 6.63 | 4.61 | 3.78 | 3.32 | 3.02 | 2.80 | 2.51 | 2.18 | 1.79 | 1.00 |

# Contents