Current Natural Sciences

Lang WU and Jin QIU

Applied Multivariate Statistical Analysis and Related Topics with R

SCIENCE PRESS edp sciences

This book was originally published by Science Press, © Science Press, 2014.

Printed in France

EDP Sciences – ISBN(print): 978-2-7598-2601-8 – ISBN(ebook): 978-2-7598-2602-5

Preface

Multivariate analysis models and methods are very useful in data analysis, since in practice data are often collected on more than one variables and these variables are often associated or correlated. A main consideration in multivariate analysis methods is to incorporate the associations or correlations between variables, so multivariate analysis methods are usually more efficient than univariate analysis methods which ignore the associations between variables. However, many multivariate analysis methods are mathematically intractable, so students often get lost in the complicated mathematical expressions, which prevents them to truly understand the basic ideas behind many multivariate analysis models and methods. In teaching a multivariate analysis course for undergraduate or Master level graduate students in statistics, we believe that the main goal should let students to understand the basic ideas of the models and methods and then use these models and methods in real data analysis. Theoretical proofs should be treated as secondary and should be left as exercises which may help students to better understand the methods and to prepare them for further studies or research.

A key feature of this textbook is that it focuses on detailed explanations of the basic ideas of common multivariate analysis models and methods using simple language and illustrations of the methods using software R. Tedious mathematical derivations are omitted from the main text and are left in the exercises. With this approach, students can focus on understanding of the basic ideas without being distracted by tedious mathematical arguments, as well as applications of the models and methods in data analysis using software R. Students with strong mathematical background and with strong interest in further study or research in any topics are encouraged to work on the theoretical exercises available at the end of each chapter. Moreover, many classic books on multivariate analysis contain detailed theoretical results, which are listed in the references, so interested readers can easily assess these materials.

Many books on multivariate analysis focus on inference for multivariate normal populations, such as parameter estimation and inference for multivariate normal distributions and normal regression models. This is restrictive since in practice there are

often discrete or categorical data or skewed data which do not follow normal distributions. In this textbook, we fill the gaps with chapters on multivariate discrete data, copula models, and generalized linear models, in addition to standard topics based on multivariate normal or continuous data. Categorical or discrete data are common in practice, but this topic often receives little attention in many multivariate textbooks and undergraduate and graduate curriculums. In this textbook, we provide an overview of multivariate categorical data analysis, including analysis of contingency tables and loglinear models. Copula models have received much attention in recent years, especially in finance. Copula models allow us to build multivariate distributions from any univariate distributions. Thus they are powerful tools for multivariate analysis. Generalized linear models allow non-normal response variables so they greatly extend the applicability of linear regression models.

Results of data analysis can be greatly influenced by noises in the data, such as missing data, measurement errors, and outliers. For example, in the presence of missing data, analysis results may be biased or less efficient if the missing data are simply discarded, and a few outliers in the data may completely change conclusions from data analysis so the conclusions do not represent population characteristics. These problems are especially common and bad for multivariate data. However, these problems often receive little attention in many books on multivariate analysis, so students do not know how to handle these problems in data analysis. In this textbook, we provide comprehensive discussions of these issues on separate chapters and offer practical suggestions for data analysts.

This book may be used as a textbook for a multivariate analysis course for undergraduate and Master-level graduate students in statistics, as well as students or researchers in other fields who wish to learn basics of multivariate analysis methods and apply these methods in data analysis. The English language is simpler than most English textbooks on multivariate analysis, so it is ideal for students whose first language is not English. Exercises are available at the end of each chapter. These exercises contain both theoretical problems and data analysis problems. Some of the theoretical derivations of the methods in the main text are left as exercise problems. We strongly suggest students to practice these theoretical problems, since such practice will help students to better understand the models and methods. For the applied exercise problems, students can follow the procedures explained in the examples and the R code in the examples. Some materials in certain chapters are challenging, so they are optional for undergraduate students. At the end of Chapter 1 we provide

some general advice on good data analysis practice. The datasets and R code in the book are available at www.stat.ubc.ca/∼lang/text.

We thank many colleagues and students who have provided us useful suggestions and help. We also thank Zhejiang University of Finance and Economics and School of Mathematics and Statistics for their supports and encouragements.

Lang Wu and Jin Qiu
Feb. 2014

Contents

Chapter 1
Introduction

1.1 Goal of Statistics

A main goal of statistics is to analyze data in order to obtain useful information and make important decisions. In other words, statisticians analyze data to extract useful information from the data in a sample and draw important conclusions about the population. In modern world, many important decisions are based on information from data. With the developments of modern computers and internet, massive data are available and can be easily obtained, but important information in the data may not be easily obtained without using modern statistical methods. Therefore, statistics is becoming one of the most important subjects in the 21 century, and statistical methods are among the most widely used tools in almost every area, including banks, insurance industries, economics, finance, medicine, and engineering. As a *New York Times* article says (August 5, 2009, **For Today's Graduate, Just One Word: Statistics**): "For many different jobs in today's world, mostly what you do is data analysis (statistics), even for jobs which seem unrelated to statistics ... Many today's decisions in industry and government are based on data analysis results. Statisticians are thus in high demand ...".

Data can be collected in many ways, such as survey, internet, company records and designed experiments. Our goal is to analyze the data to extract as much information as possible and then draw some conclusions about the whole population. For example, if a new drug is found to be effective on 20 randomly selected patients (sample) based on statistical analysis, will this drug also be effective for all patients (population)? If exam scores are found to be related to students' IQ scores as well as students' attitude on 50 randomly selected students (sample), is this also true for all students (population)? Such a generalization from sample to population is called *statistical inference*. Sometimes, however, we may just wish to obtain useful information from the sample, without necessarily making inference about the population, especially if the sample is not a random and representative sample.

In practice, data analysis often consists of two stages:

- exploratory data analysis.

- formal (or confirmatory) data analysis.

In exploratory data analyses, data are simply summarized using common statistics (e.g., means, standard deviations, correlations) and are displayed using common graphical tools (e.g., histograms, boxplots, scatterplots). In this stage, we simply present and summarize the data, without trying to generalize the conclusions obtained from summary statistics and graphs to the whole population. In this stage, we do not need to make any distributional assumptions for the data, i.e., we do not need to *assume* that data follow certain distributions such as the normal distributions. Thus, the conclusions obtained from exploratory analysis do not depend on the validity of any assumptions. Exploratory analysis can reveal important features of the data, which may lead to preliminary conclusions. **Exploratory data analysis is an important step in any data analysis and should not be skipped**. However, exploratory data analysis is usually followed by a formal or confirmatory analysis, which is used to confirm the preliminary conclusions from the exploratory analysis.

In formal (or confirmatory) data analysis, we *assume* models or distributions for the data or population, estimate unknown parameters in the models or distributions, and attempt to make statistical inference so that we may generalize the results based on the sample to the whole population. For example, we may assume that the population follows a normal distribution, and then we use data to estimate the parameters in the normal distribution (mean and variance). Note that the models or distributions are only assumptions, so they may not be true. In other words, the assumed models and distributions should be checked for their validity based on the data. **Since the assumed models or distributions rarely hold exactly, formal analysis results should only be viewed as approximate, and it is desirable to use different statistical models or methods to further validate the results.**

In practice, data are often collected on many variables, such as age, income, and education. When we analyze data on each variable separately, the data on each variable are called *univariate data*, and an analysis of univariate data is called *univariate analysis*. **In univariate analysis, the association or correlation between different variables are ignored.** In other words, when we analyze data on one variable in univariate analysis, we cannot borrow information from data of other variables. For example, suppose that we have data on income from a sample survey. In univariate analysis, we can compute the average income and the standard deviations. But if an income is missing or not reported, we cannot estimate it without information from other variables such as age and job title. Many statistical methods in introductory statistics courses, such as data summaries and hypothesis testing, are designed for univariate analysis.

Since different variables or data may be associated or correlated, analyzing data on several correlated variables *simultaneously* should be more desirable than univari-

ate analysis since additional information from the association can be used. Statistical analysis on data from more than one variables is called *multivariate analysis*. Multivariate analysis is usually more efficient than univariate analysis, although it is more complicated. In addition, multivariate analysis allows us to study the association or correlation between different variables. Thus, **multivariate analysis allows us to study the relationship between different variables and uses more information than univariate analysis.** Regression analysis can also be viewed as multivariate analysis since it studies relationship between variables.

In summary, the advantages of multivariate analysis over univariate analysis are: (i) a multivariate analysis incorporates the association between variables so it uses more information, (ii) a multivariate analysis may provide more efficient inference than a univariate analysis, (iii) a multivariate analysis may avoid biased results in a univariate analysis, and (iv) a multivariate analysis allows us to understand the association between variables. Although multivariate analysis offers many advantages over univariate analysis, multivariate analysis also has some limitations. For example, a multivariate analysis is usually more complicated than a univariate analysis, and many useful statistical tools for univariate analysis may not be available for multivariate analysis. Moreover, graphical displays of data, which are very important for statistical analysis, can be difficult for multivariate data.

In the following sections, we provide a brief overview of some simple univariate and multivariate analysis methods.

1.2 Univariate Analysis

The first step in data analysis should be exploratory analysis, which summarize the data using simple statistics and display the data using graphs. Univariate continuous data are often summarized by simple statistics such as the *mean* and *standard deviation*: the mean measures the center of the data while the standard deviation measures the variation of the data. Univariate data can be displayed by graphical tools such as *histogram* and *boxplot*: a histogram show the distribution (frequencies) of the data while the boxplot shows five number summaries of the data. It is important to display data using graphs since graphs may show unusual patterns in the data and outliers. These unusual patterns and outliers may make the mean and standard deviation misleading and unreliable. **A picture is worth a thousand words!** Graphical tools are important components of statistical analysis.

In statistical analysis, choice of statistical methods depends on the types of data in hand. Generally, there are two types of data (or variables):

- *continuous data (variables)*, which take continuous values.
- *categorical or discrete data (variables)*, which take discrete values.

Examples of continuous data include weight, age, income, blood pressure, etc. Examples of categorical data include gender, location, yes/no answers, blood type, etc. Continuous data may be transformed into categorical data by grouping the data based on some threshold values, such as subjects with ages greater than 60 and subjects with ages less than or equal to 60. Such categorizations of continuous data may simplify analysis but it may also lead to some loss of information. Note that categorical data cannot be converted to continuous data. For a categorical variable, its value may represent a category rather than a numerical value, such as gender (male/female). **Statistical methods for analyzing continuous data and categorical data are quite different.** In this textbook, we will mostly focus on continuous data, but we will also discuss categorical data.

Continuous data are probably most common in practice and thus will be our main focus. For univariate continuous data, the two most important features are

- the center of the data, measured by the sample *mean*.
- the variation of the data, measured by the sample *standard deviation*.

The mean and the standard deviation should always be reported in statistical analysis. They give us some idea about the average of the data and the variation in the data- the two most important features of continuous data. Note that the mean and the standard deviation can be very sensitive to outliers and the data distributions, i.e., a few outliers in the data or a severely skewed data distribution can greatly affect the values of the mean and standard deviation and thus may lead to misleading conclusions. Graphical tools such as histograms may reveal outliers and data distributions, so they should be used in data analysis.

Let $x_1, x_2, \cdots, x_n$ be a random sample on a continuous random variable X. Then the sample *mean* $\bar{x}$ and the sample *standard deviation* s are given respectively by

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \quad s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}.$$

Other summary statistics sometimes may also be useful in data analysis, such as the percentiles, quartiles, medians, etc. These summary statistics are useful when there may be outliers or when the data distributions may not be symmetric (so not normally distributed). For example, a 5th *percentile* is the value such that 5 % data are smaller than it and 95% data are larger than it. A *median* is the 50th percentile, i.e., half the data are smaller than the median and half the data are larger than the median. The *first quartile* is the 25th percentile, and the *third quartile* is the 75th percentile. Note that both the mean and the median measure the center of the data, but the mean can be greatly influenced by an outlier while the median is robust against outliers.

Pictures can reveal important features of the data which may not be seen from numerical summaries. For example, outliers and data distributions can be best detected by pictures. Common graphical methods for summarizing univariate data from a continuous variable include histograms and boxplots. A *histogram* shows the frequencies of the data points or the distribution of the data (see Figure 1.1). A *boxplot* shows five number summaries of the data: minimum, first quartile, median, third quartile, and maximum (see Figure 1.2).

Discrete or categorical data are summarized in different ways than continuous data. We usually use *counts* or *percentages* to summarize discrete data. For example, we can say that there are 25 (or 50%) male students and 25 (or 50%) female students in the class, or in a sample survey 20 (or 15%) participants are teachers, 40 (or 30%) participants are first year students, 20 (or 15%) participants are second year students, etc. In other words, we do not use means and standard deviations to summarize discrete or categorical data, but instead we use counts or percentages. Graphical tools for discrete data include bar-charts and pie-charts, but they are less commonly used than graphical tools for continuous data.

1.2.1 Statistical software

There are many statistical software, such as R, SAS, SPSS, Stata, etc. Any software will produce same or similar results when the same statistical model or method are used, but some software may be preferred over others. We will use software R in this book. R is free and open source. It typically includes latest statistical methods, and it produces nice pictures. R can be downloaded from the R home page: *http://cran.r-project.org/* (select a mirror and go to "Download and Install R"). Readers can go to the R webpage for information on many R packages, including some tutorials. R compiles and runs on a wide variety of UNIX platforms, Windows, and MacOS. R provides a wide variety of statistical and graphical techniques, and is highly extensible.

1.2.2 An Example in R

In the following, we show some simple R examples using a dataset which contains five quiz scores of 53 students in a semester (called *"quiz dataset"*).

```
# Import data from the data file "class.dat1"
> class.dat <- read.table("class.dat1",head=T)  # a dataframe
> attach(class.dat)  # make this data a priority in this R session.
# part of the data
> class.dat
   ID gender quiz1 quiz2 quiz3 quiz4 quiz5
1   1      M    90    79    90    90    93
2   2      F    55    60    58    70    79
3   3      F    60    72    75    80    77
```

```
4   4       F     66     48     89     70     72
..........
# save the following figures to a pdf file "histogram1.pdf"
> pdf("histogram1.pdf")
> par(mfrow=c(2,2)) # 2 by 2 figures on one page
# Show histograms of the first 4 quiz scores
> hist(quiz1, main="Scores for quiz 1",xlab="score", col="grey")
> hist(quiz2, main="Scores for quiz 2",xlab="score", col="grey")
> hist(quiz3, main="Scores for quiz 3",xlab="score", col="grey")
> hist(quiz4, main="Scores for quiz 4",xlab="score", col="grey")
> dev.off()   # save the figure
```

From the histograms in Figure 1.1, we can learn something about the *distributions* of the quiz scores. For example, in Quiz 2 the quiz scores seem roughly symmetric, while in Quiz 3 the quiz scores seem to skewed to the left and there may be outliers. Thus, a normal distribution assumption may be reasonable for Quiz 2 scores but may be inappropriate for Quiz 3 scores. Moreover, the mean and standard deviation for Quiz 3 scores may be misleading because they may be greatly affected by the outliers.

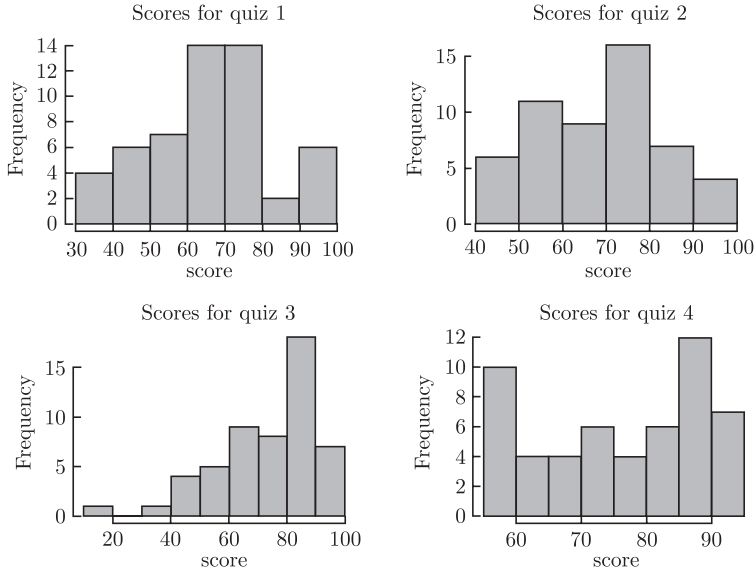


Figure 1.1 Four histograms for the first four quiz scores.

Next, let's also summarize the univariate data for each quiz using boxplots.

```
# Show boxplots of the first 4 quiz scores
> pdf("boxplot1.pdf")
> par(mfrow=c(1,1)) # one figure a page
> boxplot(quiz1,quiz2,quiz3,quiz4, main="First 4 Quiz Scores",
          names=c(1,2,3,4), col="blue")
> dev.off()   # save the figure
```

From the boxplots in Figure 1.2, we see that in Quiz 1 and Quiz 3 there are outliers (a very low score in each quiz). The median quiz scores have an increasing trend. Scores in Quiz 1 and Quiz 3 seem to have more variations than scores in the other two quizes. In Quiz 1 and Quiz 3, one student did much worse than the rest of students, but overall students' performances improve over time.

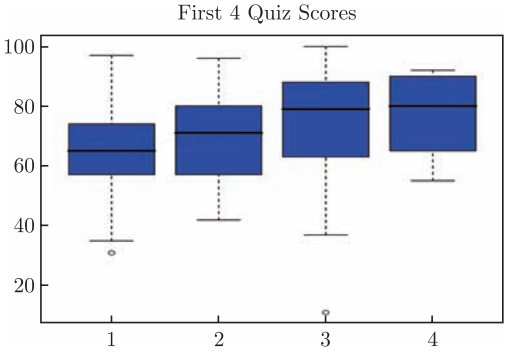


Figure 1.2 Four boxplots for the first four quiz scores.

We can also summarize the quiz data numerically. For example, the scores in Quiz 1 can be summarized as follows: the mean score is 65.7, the standard deviation is 15.65, the minimum is 31, the maximum is 97, the first quartile is 57, the median is 65, and the third quartile is 74.

```
# Summary statistics of the first quiz scores
> summary(quiz1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   31.0    57.0    65.0    65.7    74.0    97.0
> sd1 <- sqrt(var(quiz1))  # standard deviation
 15.65
```

In this example, we have seen the most common ways to summarize univariate data both graphically and numerically. These exploratory analyses give us rough ideas how students perform over time.

1.3 Multivariate Analysis

In practice, data are often collected on more than one variables, and these variables may be associated or correlated, i.e., change in one variable corresponds to changes in other variables. For example, income and education may be correlated, and gender and smoking status may be associated. When variables are associated or correlated, it is desirable to incorporate the association or correlation and analyze the data on these variables simultaneously. In other words, when several variables are associated,

it may be more desirable to use multivariate analysis than univariate analysis, since multivariate analysis uses extra information from the correlation or association. In other words, when analyzing data on one variable from a multivariate dataset, we can borrow information from data on other variables. Note that the term "correlation" is usually used for continuous data while the term "association" may be used for both continuous and discrete data. Data on two or more variables are called *multivariate data*, and statistical analysis of multivariate data is called *multivariate analysis*. **In multivariate analysis, a key consideration is to incorporate the correlation or association between the variables.**

Methods to measure the correlation among continuous data and methods to measure the association among discrete data are quite different. For continuous multivariate data, we usually use correlation matrices while for discrete multivariate data we usually use odds ratio or other measurements, which will be described in details in later chapters. We first focus on multivariate continuous data, where each variable is a continuous variable.

As an example for multivariate continuous data, suppose that a bank or a credit card company wish to classify all customers into two groups: individuals with good credit risks and individuals with bad credit risks. The classification can be based on customers' education (x_1), income (x_2), age (x_3), and past credit history (x_4). It would be easy to do the classification if we just consider one variable, say income. For example, individuals with high income may be classified as good credit risk, while those with low income may be classified as bad credit risk. However, since these variables (x_1, x_2, x_3, x_4) may be associated, we need to consider all variables simultaneously. The classification based on all variables can be challenging. For example, an individual with high income may have low education and young age. Therefore, special multivariate analysis methods are required to do the classification.

When there are many variables to be considered simultaneously, statistical analysis is challenging because: (i) it may be difficult to assume a joint distribution for all variables for statistical inference; (ii) there may be too many parameters in an assumed model for all variables; (iii) the correlation matrix may be of high dimension so it may be singular; and (iv) graphical displays of the multivariate data can be difficult. Therefore, an important question for multivariate analysis is to see if we can reduce the number of variables without much loss of information in the data. This is called *dimension reduction*, and it will greatly simplify statistical analysis. Such a dimension reduction method is often possible since some variables may be highly correlated, so they may be combined in some ways. For example, scores on algebra, geometry, and calculus may be replaced by one variable called mathematical ability. Special multivariate methods are available to do such dimension reduction.

In data analysis, the first step is to summarize the data. continuous multivariate

data are usually summarized by (i) the *mean* of each variable, (ii) the *standard deviation* or *variance* of each variable, and (iii) the *correlation* or *covariance* between the variables. These quantities reflect the center, variation, and association in multivariate data, which are the most important characteristics of multivariate data. In multivariate analysis, the covariance matrix or the correlation matrix play an important role since it contains information about the association among the variables.

Graphical displays of multivariate data can be difficult if the number of variables is more than two. A simple approach is to check the pairwise correlations for any two variables using *scatterplots*, which allow us to visually check the correlation between any two variables. Some other graphical methods such as *star-plots* can also be useful in displaying multivariate data in some cases, but their values are often limited.

1.3.1 Covariance matrix and correlation matrix

Since a main focus in multivariate analysis is to incorporate the correlation between variables, the covariance matrix or the correlation matrix play an important role in multivariate analysis for continuous data or variables. Suppose that X and Y are two continuous random variables, and suppose that $\{x_1, x_2, \cdots, x_n\}$ and $\{y_1, y_2, \cdots, y_n\}$ are data collected on X and Y respectively. The correlation between X and Y can be measured by the *covariance* $Cov(X, Y)$ or the *correlation coefficient* r:

$$Cov(X, Y) = E[(X - E(X))(Y - E(Y))],$$
$$r = \frac{Cov(X, Y)}{\sigma_X \sigma_Y},$$

where $\sigma_X = \sqrt{Var(X)}$ and $\sigma_Y = \sqrt{Var(Y)}$ are the standard deviations of X and Y respectively. The correlation coefficient r is a number between -1 and 1, and it measures the *linear correlation* between two variables. If $r > 0$, X and Y are *positively correlated*. If $r < 0$, X and Y are *negatively correlated*. The larger the value $|r|$, the stronger the correlation. When $r = 0$, X and Y are not linearly correlated.

Note that the correlation coefficient r only measures a linear relationship, not other relationships, i.e., the observed values of X and Y roughly fall on a straight line. For example, when $r = 0$, X and Y may still have a nonlinear relationship.

More generally, let $\mathbf{X} = (X_1, X_2, \cdots, X_p)^{\mathrm{T}}$ be a random vector with p continuous random variables. The mean vector of $\mathbf{X}$ is

$$E(\mathbf{X}) = \boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_p)^{\mathrm{T}}.$$

The covariance matrix of $\mathbf{X}$ is given by

$$\Sigma = Cov(\mathbf{X}) = (\sigma_{ij})_{p \times p}, \quad \text{where } \sigma_{ij} = E(X_i - E(X_i))(X_j - E(X_j)),$$

and the correlation between X_i and X_j is

$$r_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}, \quad i, j = 1, 2, \cdots, p.$$

For multivariate data, the mean vector $\boldsymbol{\mu}$ represents the *center* of the data, and the covariance matrix Σ represents the *variation* (variance) and *correlation* of the data.

The covariance matrix Σ or the correlation matrix $R = (r_{ij})_{p\times p}$ play a key role in multivariate analysis since they measure the strength of the association between continuous variables. **Much information about the distribution of multivariate continuous data is contained in the *eigenvalues* and *eigenvectors* of the covariance matrix Σ. The eigenvalues describe the *shape* of the data region, while the eigenvectors show the *directions* of the axes of the data region.** Specifically, let $\{(\lambda_i, \mathbf{a}_i),\ i = 1, \cdots, p\}$ be the eigenvalues-eigenvectors of the covariance matrix Σ. Then, the variance of data points in the $\mathbf{a}_i$ direction is $\lambda_i,\ i = 1, 2, \cdots, p$. See Figure 6 on page 23 for an example in the case of $p = 2$. This is a very important fact to keep in mind in order to understand many of the multivariate analysis methods.

Since eigenvalues represent the variation (variance) of data in certain directions, if a few eigenvalues are much larger than the rest eigenvalues of a covariance matrix, the multivariate data essentially fall onto the sub-space spanned by the eigenvectors corresponding to the few large eigenvalues, so the dimension of the multivariate data can be reduced to that subspace. This is the essential idea behind dimension-reduction methods in multivariate analysis, such as the principal components method. We will see many examples when describing the principal components method.

In practice, the parameters in the population mean vector $\boldsymbol{\mu}$, the covariance matrix Σ, and the correlation matrix $\boldsymbol{R}$ are all unknown. However, they can be estimated from data. Specifically, let $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ be an i.i.d. sample obtained from a population with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ, denoted by $\mathbf{x}_i \sim (\boldsymbol{\mu}, \Sigma)$ (note that here the population does not have to be normally distributed), where

$$\mathbf{x}_j = (x_{j1}, \cdots, x_{jp})^{\mathrm{T}}, \quad \boldsymbol{\mu} = (\mu_1, \cdots, \mu_p)^{\mathrm{T}}, \quad \Sigma = (\sigma_{kl})_{p\times p}, \qquad j = 1, 2, \cdots, n.$$

Then, $\boldsymbol{\mu}$ and Σ can be estimated by the *sample mean* $\bar{\mathbf{x}}$ and the *sample covariance matrix* S as follows

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i, \quad \boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{T}} = (\hat{\sigma}_{lk})_{p\times p}.$$

In data analysis, we can use these estimates to replace the unknown population parameters. As we can imagine, the larger the sample size n, the more accurate the estimates.

Sometimes statistical analysis results may depend on the units or scales of the measurements. For example, a person's height may be measured in meters or centimetres, leading to different values, which may affect analysis results. Therefore, it is sometimes desirable to use scale-free statistics or data, which do not depend on the units or scales of measurements. Two commonly used scale-free measurements are the standardized data and correlation matrix. The *standardized data* are defined as

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{\hat{\sigma}_{jj}}}, \quad i = 1, \cdots, n;\ j = 1, \cdots, p.$$

which has a mean 0 and variance 1. The correlation between two random variables X_i and X_j can be estimated by the following *sample correlation coefficient*

$$\hat{r}_{ij} = \frac{\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)^2 \sum_{k=1}^{n}(x_{kj} - \bar{x}_j)^2}}.$$

The sample correlation matrix $\boldsymbol{R}$ is given by $\boldsymbol{R} = (\hat{r}_{ij})_{p\times p}$, which has the diagonal elements all being 1 and off-diagonal elements all being values between –1 and 1.

Outliers may be present in multivariate data. *Outliers* are unusual observations which are distinctly different from rest of the data. For example, a score of 0 in an exam may be viewed as an outlier, if most students did very well and only very few students did poorly. Outliers may lead to misleading statistical analysis results if they are not handled appropriately. For example, the sample mean and the sample covariance can be greatly influenced by a few outliers in the data. For univariate data, we often can visually detect a possible outlier from a picture. Unfortunately, for multivariate data, it is difficult to use graphical tools to detect outliers since it is difficult to graphically display multivariate data. In statistical analysis, we should always pay attention to possible outliers in the data.

Note that so far we have not assumed any distributions for the data. In other words, we do not need to assume any distributions for the random variables or the data in order to define the foregoing quantities. The only assumption we have made is that the data are continuous.

1.3.2 Examples in R

In the following, we show some simple R examples to summarize and display multivariate continuous data. These examples are for illustrations only.

Example 1. Consider the quiz score dataset described in the previous section again. We can view this dataset as a multivariate dataset if we view each quiz as a random

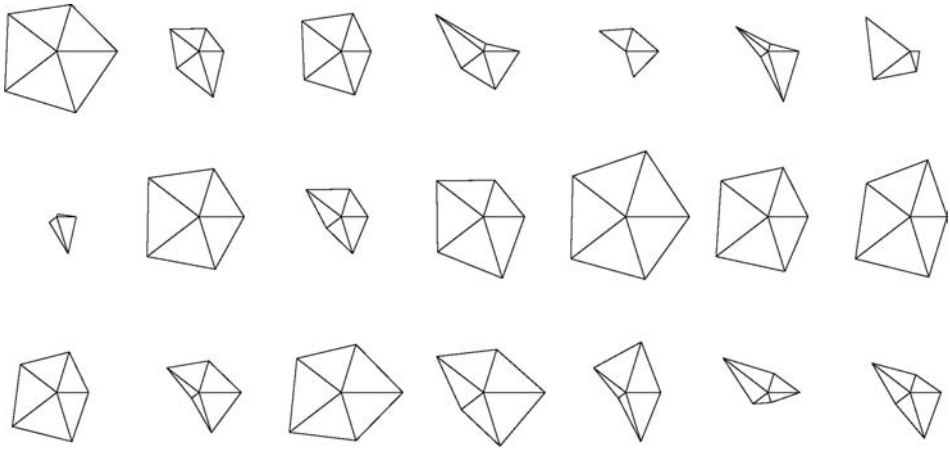
variable. The quiz five scores for each student are clearly correlated, so a multivariate analysis would be desirable.

```
# Read data from the data file "class.dat1"
>  class.dat <- read.table("class.dat1",head=T)
>  attach(class.dat)
>  library(lattice)  # package for some R graphical functions
>  options(digits=2)   # just need 2 decimal points
>  class.dat2 <- class.dat[,-c(1,2)] # remove the first 2 columns (ID
    and gender)

# starplot to check possible outliers
> pdf("star.pdf")
> stars(as.matrix(class.dat2))
> dev.off()

# pairwise scatterplots to show correlations between two quizes
> pdf("pairwise.pdf")
> pairs(class.dat2)
> dev.off()
```

Figure 1.3 shows the *star-plot* of the multivariate quiz data. Each "star" represents one observation (i.e., the five quiz scores from one student), with the edges representing the values of the components of the observation. Such a picture allows us to find potential outliers (the stars with unusual shapes). For example, the first observation on the second row is potentially an outlier. Figure 1.4 shows pairwise scatterplots. Each plot shows a scatterplot between two variables (quizes). Such a picture allows us to exam possible relationship between any two variables. For example, quiz 1 scores and quiz 2 scores seem highly correlated, but quiz 2 scores and quiz 3 scores seem less correlated.



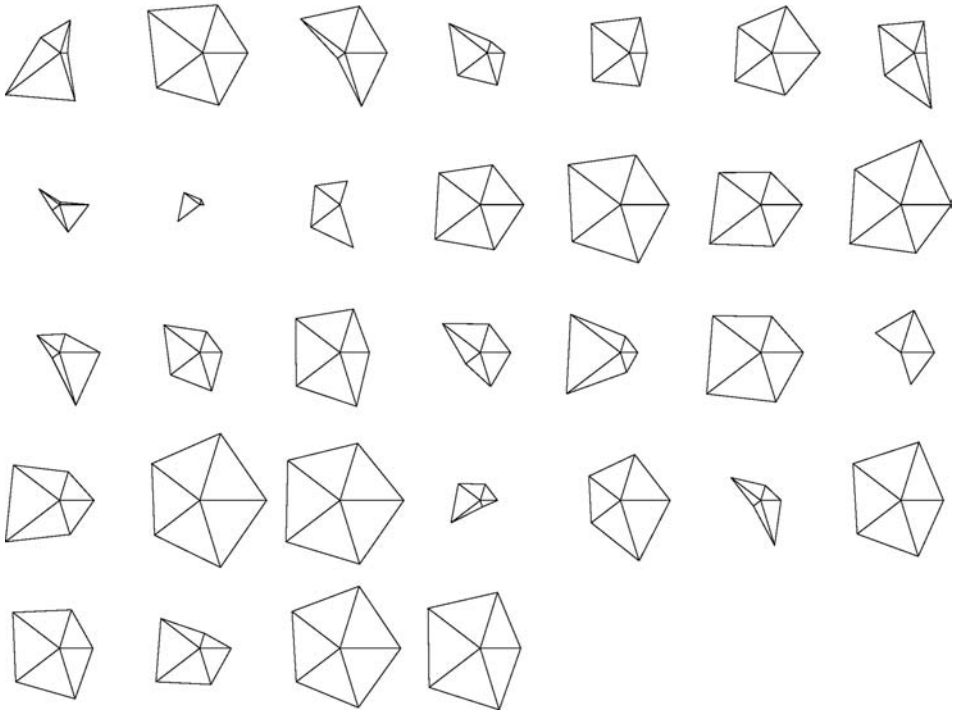
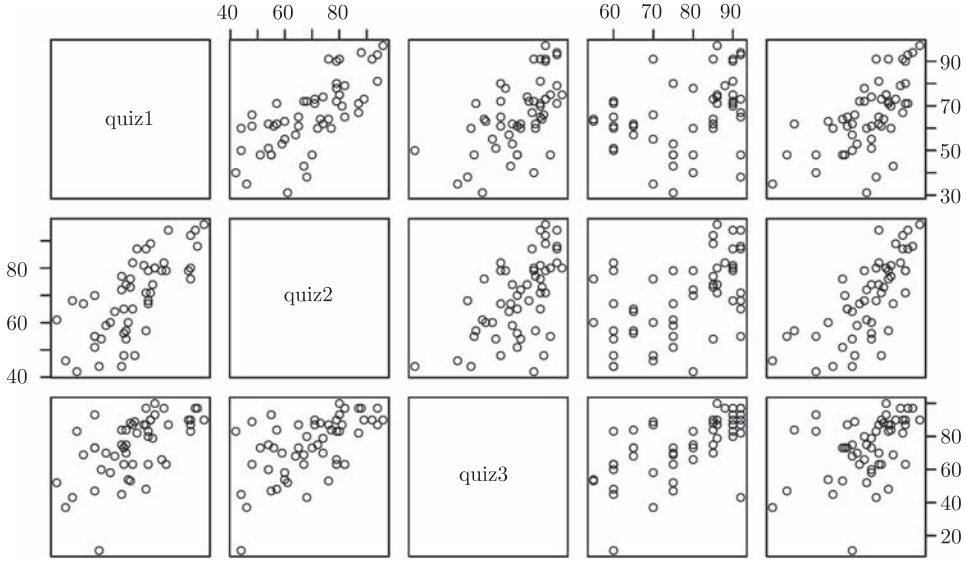


Figure 1.3 Starplot for the quiz scores. Each figure represent five quiz scores of one student. Figures with unusual shapes may be potential outliers (e.g., the first one on second row).



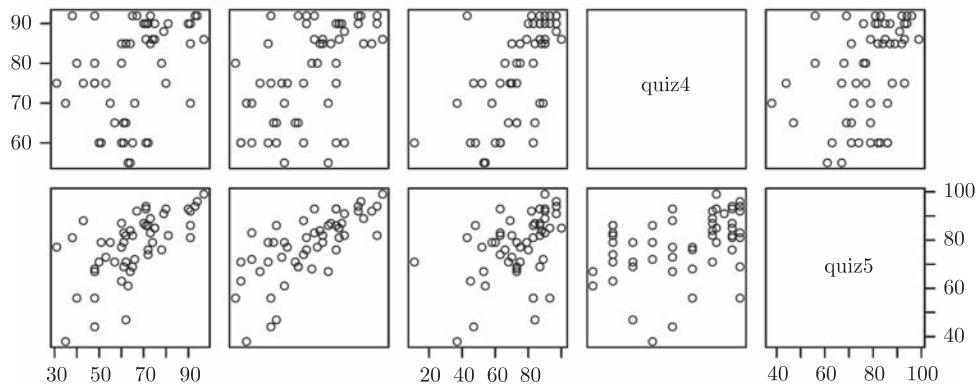


Figure 1.4 Pairwise scatterplots for the quiz scores. Each plot shows the scatterplot of the scores from two quizzes.

The covariance matrix and the correlation matrix can be easily obtained using R function *"cov"* and *"cor"* respectively. The diagonal elements of the covariance matrix are the sample variances of the variables, and the off-diagonal elements of the correlation matrix are the correlation coefficients between variables. We see that the variance of quiz3 is 338 and the variance of quiz4 is 148, so the scores of quiz3 are more spreadout than quiz4. We also see that the correlation between quiz1 and quiz2 is 0.72 and the correlation between quiz1 and quiz4 is 0.32, so quiz1 and quiz2 are highly correlated while quiz1 and quiz4 are less correlated.

```
# covariance matrix Sigma
> cov(class.dat2)
      quiz1 quiz2 quiz3 quiz4 quiz5
quiz1   245   163   168    61   132
quiz2   163   208   155    93   139
quiz3   168   155   338   142   113
quiz4    61    93   142   148    63
quiz5   132   139   113    63   178

# correlation matrix R
> cor(class.dat2)
      quiz1 quiz2 quiz3 quiz4 quiz5
quiz1  1.00  0.72  0.58  0.32  0.63
quiz2  0.72  1.00  0.58  0.53  0.72
quiz3  0.58  0.58  1.00  0.64  0.46
quiz4  0.32  0.53  0.64  1.00  0.39
quiz5  0.63  0.72  0.46  0.39  1.00
```

Example 2. The *Chinese consumption dataset* consists of per capita annual consumption expenditures of urban households in 31 regions in China in year 2007 (courtesy of the Chinese Statistics Bureau). The variables include the consumption expenditures (in RMB yuan) on eight categories: food (Food), clothing (Cloth), residence (Resid),

household facilities and articles and services (HousF), health care and medical services (Health), transport and communication (TranC), education, culture and recreation (Educ), and miscellaneous goods (Miscel). We perform an exploratory multivariate analysis on this dataset.

```
# Read data from the data file "consum2007.txt"
> consum.1 <- read.table("consum2007.dat",head=T)
> consum<-consum.1[,1:8]  # consider only the first 8 columns
# part of the data
> consum[1:3,]
        Food Cloth Resid HousF Health TranC Educ Miscel
Beijing 4934  1513  1246   981   1294  2329 2384    650
Tianjin 4249  1024  1417   761   1164  1310 1640    464
Hebei   2790   976   917   547    834  1011  895    266
.........
# starplot
> stars(consum,key.loc=c(14,2), main="consumption expenditure:stars(*,
    full=F)", full=FALSE)
# the segment plots occupy the (upper) semicircle
```

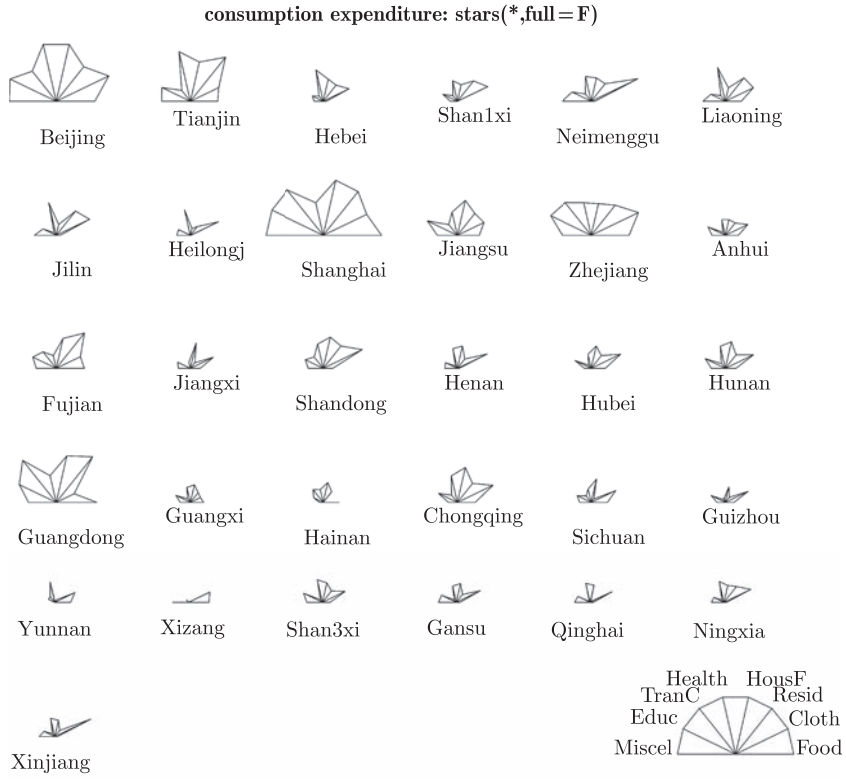


Figure 1.5 Starplot for the Chinese Consumption Data(not include Hong Kong, Macao and Taiwan)

Figure 1.5 shows the star-plot of the data. From Figure 1.5, we can roughly classify all the regions into 2 groups. One group contains all the regions with relatively high consumption levels, such as Beijing, Shanghai, Zhejiang, Guangdong, Tianjin, etc., which are the more developed regions in China. The other group consists of regions with median to low consumption levels. This is only a rough classification based on the picture.

For formal multivariate analysis, matrix algebra plays an important role since it greatly simplifies many mathematical expressions. In the following, we introduce some basic matrix operations in R. Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be matrices with appropriate dimensions.

```
At <- t(A)   # function t() returns the transpose of matrix A
AB <- A %*% B   # matrix A times matrix B
A.inv <- solve(A)   # the inverse of matrix A
sum(diag(A))  # sums the diagonal elements of matrix A; same as trace(A)
det(A)   # determinant of matrix A
eigen(A)    # eigenvalues and eigenvectors of matrix A
```

1.4 Multivariate Normal Distribution

In the previous sections, we do not assume any parametric distributions for the random variables or data. When we perform statistical inference such as hypothesis testing, however, we need to assume that the multivariate data follow some parametric distributions. The most common distribution for continuous multivariate data is the *multivariate normal distribution*. There are several reasons. First, by the Central Limit Theorem, many statistics (such as sample means) asymptotically follow normal distributions even if the original data do not follow normal distributions. In other words, when the sample size is large, a normality assumption may be reasonable for these statistics. Second, the normal distributions have many attractive properties. For example, a normal distribution is completely determined by its mean and variance (covariance), which are the two most important characteristics of data. Third, many continuous data, even if they may not be normally distributed, may be transformed into data which are roughly normally distributed. For example, we may consider a log-transformation for data which are positive (such as age) or skewed (such as survival time). Lastly, for multivariate continuous data, there are not many reasonable multivariate distributions to choose. Therefore, for multivariate continuous data, if a distributional assumption is required for statistical inference, we often assume that the data follow a multivariate normal distribution. However, note that this is only an *assumption*, so it needs to be checked based on the data for its validity.

The random vector $\mathbf{x} = (x_1, \cdots, x_p)^{\mathrm{T}}$ follows a p-dimensional *multivariate normal distribution*, denoted by $N_p(\boldsymbol{\mu}, \Sigma)$, if any linear combination of the components of the random vector $\mathbf{x}$ follows a univariate normal distribution. That is, for any constant

vector $\mathbf{a} = (a_1, a_2, \cdots, a_p)^{\mathrm{T}}$, the univariate random variable $y = \mathbf{a}^{\mathrm{T}}\mathbf{x} = \sum_{i=1}^{p} a_i x_i$ follows a univariate normal distribution

$$\mathbf{a}^{\mathrm{T}}\mathbf{x} \sim N\left(\sum_{i=1}^{p} a_i E(x_i), \quad var(\sum_{i=1}^{p} a_i x_i)\right).$$

Thus, if $\mathbf{x}$ follows $N_p(\boldsymbol{\mu}, \Sigma)$, each component x_k follows $N(\mu_k, \sigma_{kk})$, $k = 1, \cdots, p$. Note that the reverse may not be true: if each x_k follows $N(\mu_k, \sigma_{kk})$, $\mathbf{x}$ may or may not follow a normal distribution. The probability density function of $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma)$ can be written as

$$\begin{aligned} f(\mathbf{x}) \quad = \quad & \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right], \\ & -\infty < x_j < \infty, \qquad j = 1, 2, \cdots, p, \end{aligned}$$

which reduces to a univariate normal density when $p = 1$.

As an example, consider a bivariate normal random vector $\mathbf{x} = (x_1, x_2)^{\mathrm{T}}$ with the mean vector and covariance matrix given by

$$\boldsymbol{\mu} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} 4 & 3 \\ 3 & 3 \end{pmatrix}.$$

Then, we have $x_1 \sim N(2, 4)$, $x_2 \sim N(1, 3)$, and the correlation between x_1 and x_2 is

$$r = \frac{Cov(x_1, x_2)}{\sqrt{Var(x_1)Var(x_2)}} = \frac{3}{\sqrt{4 \times 3}} = 0.866.$$

The eigenvalues and eigenvectors of matrix Σ are given by

$$\lambda_1 = 6.54, \quad \mathbf{a}_1 = (-0.76, -0.65)^{\mathrm{T}}; \qquad \lambda_2 = 0.46, \quad \mathbf{a}_2 = (0.65, -0.76)^{\mathrm{T}},$$

respectively. Figure 1.6 shows a picture where the sample points from this population may fall. Note that the data in the $\mathbf{a}_1 = (-0.76, -0.65)^{\mathrm{T}}$ direction have a variance of $\lambda_1 = 6.54$, and the data in the $\mathbf{a}_2 = (0.65, -0.76)^{\mathrm{T}}$ direction have a variance of $\lambda_2 = 0.46$. Since λ_1 is much larger than λ_2, most variation in the data is in the $\mathbf{a}_1$ direction. So we may reduce the dimension from 2 to 1 while retain most of the information (variation) in the data.

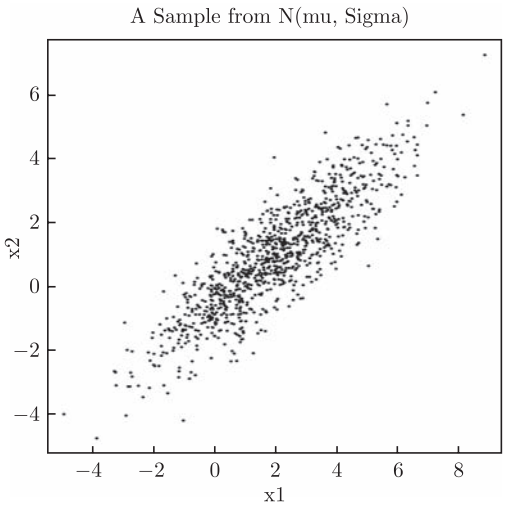


Figure 1.6 Sample points from the bivariate normal distribution $N(\boldsymbol{\mu}, \Sigma)$ in the example.

1.4.1 Properties of multivariate normal distributions

Since the multivariate normal distribution is the most important distribution for multivariate continuous data, we list some of its important properties, which are useful in multivariate analysis. Note that properties (i) and (ii) below do not require that the distribution is normal.

Let $\mathbf{x}$ and $\mathbf{y}$ be two random vectors, and let $\boldsymbol{B}$ and $\mathbf{b}$ be a constant matrix and a constant vector respectively. Let $\mathbf{y} = \boldsymbol{B}\mathbf{x} + \mathbf{b}$ be a linear transformation. Then, we have the following properties:

(i) $E(\mathbf{y}) = \boldsymbol{B}E(\mathbf{x}) + \mathbf{b}$;

(ii) $Cov(\mathbf{y}) = \boldsymbol{B}Cov(\mathbf{x})\boldsymbol{B}^{\mathrm{T}}$;

(iii) If $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma)$, then

$$\mathbf{y} = \boldsymbol{B}\mathbf{x} + \mathbf{b} \sim N(\boldsymbol{B}\boldsymbol{\mu} + \mathbf{b}, \boldsymbol{B}\Sigma\boldsymbol{B}^{\mathrm{T}}),$$

i.e., a linear transformation of a normal random vector is still normally distributed;

(iv) If $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma)$ and

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

then

$$\mathbf{x}_1 \sim N(\mu_1, \Sigma_{11}), \quad \mathbf{x}_2 \sim N(\mu_2, \Sigma_{22}),$$

and the conditional distribution of $\mathbf{x}_1$ given $\mathbf{x}_2$ is still normally distributed and is given by

$$\mathbf{x}_1 | \mathbf{x}_2 \sim N(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \mu_2),\ \Sigma_{11.2}),$$

where $\Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. In other words, for a multivariate normal distribution, its components' distributions and conditional distributions are still normal.

1.4.2 Check for Multivariate Normality

When we assume a multivariate normal distribution for data analysis, we should check to see if this assumption is supported by the data. Unlike univariate normal distribution, it is not straightforward to check multivariate normal distributional assumption. In the following, we discuss methods that may be used to check multivariate normality.

First, as a simple and naive approach, we may consider methods for checking univariate normality, which may also be useful for checking multivariate normality. Note that, if observations were generated from a multivariate normal distribution, then each univariate distribution will also be normal. In other words, if the univariate data are not normally distributed, then the multivariate data will not be normally distributed either. Boxplots and histograms can be used to check if the univariate data are symmetric or not (note that all normal data are symmetric), but symmetric data are not necessarily normal. If the univariate data are not symmetric, then the normality cannot hold. A more formal method to check univariate normality is normal quantile-quantile (Q-Q) plot. A *normal Q-Q plots* shows the theoretical quantiles from a normal distribution and the quantiles computed from the data. If a Q-Q plot shows a straight line ($y = x$), then the univariate data may be considered as normally distributed. A more formal method, called the Shapiro-Wilk test, may also be used to check normality.

A formal method for checking multivariate normality is the *Chi-Squared plot*. It is a generalization of a Q-Q plot based on the squared *Mahalanobis distance*

$$d_j^2 = (\mathbf{x}_j - \bar{\mathbf{x}})^{\mathrm{T}} \boldsymbol{S}^{-1} (\mathbf{x}_j - \bar{\mathbf{x}}), \quad j = 1, 2, \cdots, n$$

where $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ are the sample observations, $\bar{\mathbf{x}}$ is the sample mean vector, and $\boldsymbol{S}$ is the sample covariance matrix. If the population is multivariate normal and $n - p$ is large, each of the squared distances $d_1^2, d_2^2, \cdots, d_n^2$ should behave as a chi-square random variable. We can order the squared distances as $d_{(1)}^2 \leqslant d_{(2)}^2 \leqslant \cdots \leqslant d_{(n)}^2$, and then graph the pairs $(q((j - 1/2)/n),\ d_{(j)}^2)$, where $q((j - 1/2)/n)$ is the $(j - 1/2)/n$ quantile of the chi-square distribution with p degrees of freedom. Under multivariate normality, the plot should resemble a straight line through the origin with slope 1. A systematic curved pattern indicates that normality may not hold. A few points far above the line suggests outliers. We give an example below.

To illustrate the idea of the Chi-Squared plot, we simulate a sample of size 100 from the bivariate normal distribution $(x_1, x_2) \sim N_2(\boldsymbol{\mu}, \Sigma)$, with the mean vector and the covariance matrix given by

$$\boldsymbol{\mu} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} 10 & -3 \\ -3 & 2 \end{pmatrix}.$$

Then we draw Q-Q plots to verify univariate normalities of the component random variables x_1 and x_2 respectively, and draw a Chi-Squared plot to verify the bivariate normality of the random vector (x_1, x_2). See Figure 1.7. The R code is given below.

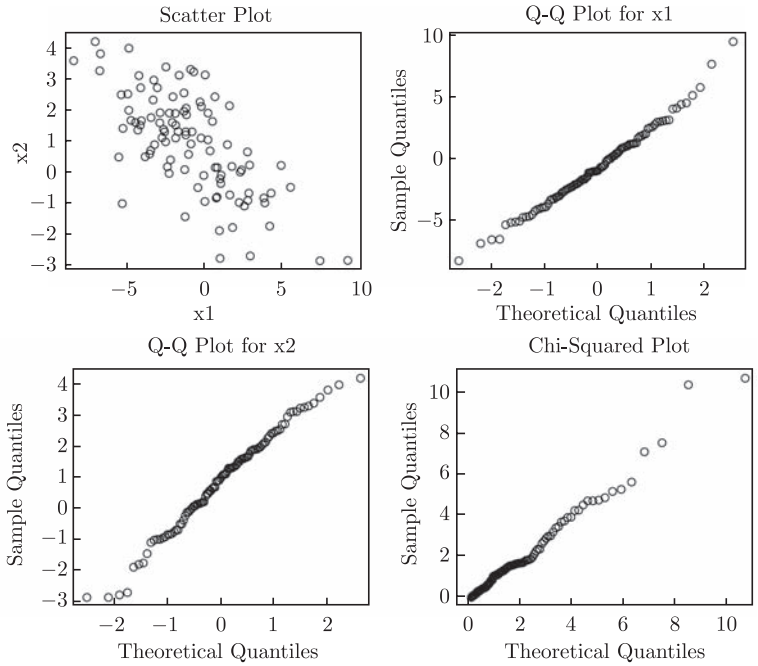


Figure 1.7 Checking Multivariate Normality.

```
> library(mvtnorm)
> library(MASS)
> Sigma<-matrix(c(10,-3,-3,2),2,2)
> X<-mvrnorm(n=100, c(-1, 1), Sigma)
> x1<-X[,1]
> x2<-X[,2]
> mX<-colMeans(X) # compute the mean vector
> mX
[1] -0.96  1.04
> S<-var(X) # compute the sample covariance matrix
> S
     [,1] [,2]
[1,]  9.5 -3.2
[2,] -3.2  2.3
# compute the Mahalanobis distances and sort them
> d<-apply(X,1,function(z) t(z-mX)%*%solve(S)%*%(z-mX))
> d.sort<-sort(d)
```

```
> q<-qchisq((1:nrow(X)-0.5)/nrow(X),2) # compute the quantile of the
    Chi-Squared distribution
> par(mfrow=c(2,2))
> plot(x1,x2, xlab="x1", ylab="x2", main="Scatter Plot")
> qqnorm(x1, main="Q-Q Plot for x1" )
> qqnorm(x2, main="Q-Q Plot for x2")
> plot(q,d.sort, xlab="Theoretical Quantiles", ylab="Sample Quantiles",
 main="Chi-Squared Plot") # draw the Chi-Squared plot
```

1.5 Unsupervised Learning and Supervised Learning

Statistics is about learning from data. In practice, data are often collected on more than one variables, so many statistical methods may be viewed as multivariate analysis in a broad sense. In general, we may classify these methods into two general approaches:

- *Unsupervised learning*: we treat all variables symmetrically or equally, with the goal of understanding the underlying association structures between these variables. Examples of unsupervised learning include principal components analysis, factor analysis, and cluster analysis.
- *Supervised learning*: we treat one or more variables as responses and other variables as predictors which are used to partially explain the variations in the responses. Regression models are examples of supervised learning.

These two approaches are used to answer different questions, so the choice of methods depends on the study objectives. There is a wide variety of statistical methods available for each approach. If our goal is to understand the relationship among all the variables or if we want to reduce the number of variables, we should consider unsupervised learning methods. If our objective is to predict one or more variables using the other variables or to explain the variations in one or more variables using other variables, we should use regression models.

In unsupervised learning, all variables are treated equally and the goal is to understand the covariance structures in these variables or to reduce the dimension of the data space. Commonly used unsupervised learning methods for multivariate continuous data include principal components analysis (PCA), factor analysis, discriminant analysis, and cluster analysis. For example, in PCA and factor analysis, the original set of variables can be replaced by a smaller set of new variables which may explain most of the variation in the original data. These new variables are usually special linear combinations of the original variables, and they allow us to use graphical tools to display the data and to interpret the data more easily than the original sets of variables. In discriminant analysis and cluster analysis, we classify multivariate data into different clusters based on the "distances" between the observations. In these proce-

dures, distributional assumptions for the data may not be needed. The covariance matrices or correlation matrices play the key role.

Regression models are among the most useful statistical methods. There are many types of regression models. The types of regression models are determined based on the types of the response variables, not the types of predictors. For example, if the response is a continuous variable, we may consider a linear regression model, but if the response is a binary (discrete) variable, we may consider a logistic regression model. The following regression models are commonly used in practice: linear models, nonlinear models, generalized linear models, survival models, and models for longitudinal data or clustered data. Linear regression models are often considered when the response variables are continuous and roughly normally distributed. Analysis of variance (ANOVA) models are special linear models in which all predictors are categorical or discrete. Nonlinear regression models may be used when the response variables are continuous and roughly normally distributed, and there is a good understanding of the mechanisms that generate the data. Generalized linear models (GLMs) are often used when the response variables are binary or count or follow distributions in the exponential family. Survival models are used when the response variables are the times to some events of interest, such as times to death or times to accidents. The foregoing regression models are used for independent data. When the data are correlated or clustered, we should use models for clustered data.

In a regression model, if there are more than one response variables, the model is called a *multivariate regression model*. For example, if there are two or more responses in an ANOVA model, the model is called a *multivariate ANOVA (MANOVA)* model.

1.5.1 Statistical Inference

A main goal of statistical inference is to generalize the results obtained from a sample to the general population. To achieve this, the sample has to be representative of the population, and the data are assumed to follow some parametric distributions such as normal distributions. Such an assumption allows us to do probability calculation required in inference (e.g., p-values in hypothesis testing).

Multivariate continuous data are often assumed to follow multivariate normal distributions. Under this distributional assumption, we can perform usual statistical inference, such as confidence regions and hypothesis testing for the unknown population mean vectors or the covariance matrices. For example, for univariate continuous data, the most well-known test is perhaps the t-test for the population mean, while for multivariate continuous data, the most well-known test is perhaps the Hotelling's T^2-test for the mean vector. A major consideration in multivariate analysis is to incorporate the correlation between the variables. This allows for more efficient inference

than univariate analysis, which ignores the correlation between variables.

Multivariate discrete data are often assumed to follow *multinomial distributions*. Under this distributional assumption, statistical inference for the unknown parameters can be done using standard methods such as the maximum likelihood method. The simplest and also the most common multivariate discrete data are often summarized by 2×2 tables. For example, we may want to compare two methods, with the response being either positive or negative. The results can then be summarized by a 2×2 table. Many statistical methods are available to analyze such 2×2 tables. More general multivariate discrete data may be summarized by $k \times m$ contingency tables.

1.6 Data Analysis Strategies and Statistical Thinking

The main goal of statistics is to learn from data in order for us to make good decisions and to understand real world problems. In data analysis, the most important skill for a statistician is to develop the ability of *statistical thinking*: how to obtain good data, how to choose appropriate methods to analyze the data, and how to interpret analysis results and draw reliable conclusions. It takes time to develop such skills since real understanding of statistical methods is harder than one may imagine. Statistical thinking is different from mathematical thinking, since mathematics often involves either black or white (i.e., right or wrong) while statistics may involve many grey areas which may not be as simple as either right or wrong. Therefore, in statistics sometimes it may be more important to understand the concepts, models, and methods than to do mathematical derivations or proofs. Statistics is becoming one of the most important subjects in modern world since many important decisions in almost all fields in modern world are based on information from data, obtained via data analysis. As Samuel Wells wrote "Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write".

In data analysis, it is desirable to follow certain procedures. These procedures are reflections of statistical thinking. Specifically, a good statistical analysis should consist of the following steps:

1. objectives.
2. data collection.
3. exploratory analysis.
4. confirmatory analysis.
5. interpretation of results.
6. conclusions.

Before collecting data, we should be clear about the study objectives, which allow us to decide how to collect data. Once the objectives are clear, the next step is to decide how to collect data. Getting good data is an important step, since there is not much

statistics can do if the data is poorly collected. There are generally two ways to collect data: designed experiments or observation studies (e.g., sample surveys). Designed experiments often involve randomization which allows us to make causal inference. Observational studies such as sample surveys allow us to find associations. Nowadays, there are many other ways that massive data are automatically generated, such as data from internet and records from business transactions. A good understanding of how the data are generated can help us to make reliable conclusions from data analysis.

Once data are obtained, the next step is to analyze the data to get as much information from the data as possible. Statistical models and methods are used in this step for data analysis. This step is also our main focus. Here, it is important to have a good understanding of what statistical models and methods are really about. Statistical models usually do not represent the truth – they may be viewed as approximations to the truth, but they can help us to quantify the evidence in the data to answer our questions. Statistical models and methods require *assumptions* about the data, such as distributional assumptions and model assumptions. These assumptions almost never hold exactly, but they may hold approximately or may be reasonable so analysis based on these assumptions may be reasonable. Therefore, when interpreting analysis results, we should consider how sensitive the analysis conclusions are to these assumptions. For example, we may assume that the data follow normal distributions, but this assumption may only hold approximately, so we should consider whether the conclusions will change much if this assumption is violated. In other words, conclusions based on assumed models should never be interpreted as exactly true. That is why alternative models and methods should also be used to validate the conclusions.

In data analysis, we should first summarize the data *without any assumptions*. This step is called *exploratory data analysis (EDA)*. In EDA, we can use simple summary statistics such as the means, standard deviations (or variance), and correlations to summarize the data. We can also use graphical tools such as histograms, boxplots, and scatterplots to display the data. EDA can reveal important features in the data which allow us to draw preliminary conclusions. We then confirm these preliminary conclusions based on statistical models. Since conclusions based on EDA do not rely on any assumptions, they are usually reliable. However, EDA is not sufficient, since EDA only provides preliminary and rough results. Statistical models are required to confirm the results from EDA and to obtain more definite conclusions. Statistical analysis based on models may be called *confirmatory analysis*. Therefore, data analysis should consist of two steps:

1. exploratory data analysis (EDA).
2. confirmatory analysis.

Both steps are important parts of data analysis. If the results from both steps are consistent, we are reasonably confident about our conclusions. If the results are different, we should do a further exploration to find out why the results are different: is it because the model assumptions are violated or the statistical methods are inappropriate?

Statistical models *assume* that the data are generated from some mathematical models with random errors. Since a model is only an assumption, it should always be checked to see if it is consistent with the data. In fact, all models are wrong since the truth is unknown, but some models are closer to the truth and are more useful than other models. Because all models are assumed, it is a good idea to fit different models to the same dataset to see if similar conclusions can be obtained. Note that complex models are not necessary better than simpler models. Moreover, complex models are harder to check for their validity. On the other hand, if a model is too simple, it may not be of much value. Thus, some kind of balance is required, and scientific consideration is important (i.e., whether the models make sense in the particular application of real world problems).

Once models are assumed, statistical methods are used to estimate the unknown parameters in the models. Many methods are available to estimate model parameters. The most common method is probably the maximum likelihood method, which chooses parameter estimates to maximize the likelihood of the data under the assumed models. The accuracy of an parameter estimate can be measured by its standard error: the smaller the standard error, the more accurate the estimate. So standard errors should always be reported with parameter estimates. Usually, the larger the sample size, the more accurate the parameter estimates. Note that many formulas to compute standard errors are based on asymptotic results, i.e., these formulas hold only when the sample size is very large (actually infinite). In practice, when the sample size is not large, the results are only approximate, and we often do not know how accurate the approximation is (we only know that the larger the sample size, the more accurate the formula). When a model is too complicated, which suggests that it may contain too many parameters, the estimates may be less accurate. Note that a parameter estimate may be viewed as a guess of the true parameter value (which is never known). The guess may be close or far from the true value, even if it is unbiased. Statistical inference methods, such as confidence intervals and hypothesis testing, are used to provide some degree of certainty about where the true parameter values are likely to be.

Standard statistical models and methods assume that the data are "clean". In practice, however, the observed data may be "dirty" in the sense that there may be outliers, missing values, measurement errors, and other issues. In this case, blindly applying standard statistical models and methods without addressing these data com-

plications may lead to severely biased results. For example, a few outliers can have a great impact on the analysis results. The conclusions we obtain from data analysis should be valid for most individuals, so they should not be sensitive to a few outliers or influential observations. When missing data are informative (i.e., not missing at random), ignoring the missing data in statistical analysis may also lead to biased results. In order to obtain reliable conclusions, a comprehensive data analysis should address all issues in the data or warn the readers how reliable the conclusions may be if some of these issues are not addressed due to their complexities.

All data analysis results should be viewed as approximate, since the assumed models and distributions and other assumptions usually only hold approximately, not exactly. For example, a p-value of 0.045 is only an approximate answer, so claiming that the results is significant at 5% level should not be viewed as absolute – it should be confirmed with alternative approaches. This is another reason why different models and methods should be used to analyze the same dataset so that we can be reasonably confident about the conclusions. The conclusions should not be just based on results from a single model.

For a given dataset, when analyses based on different models and methods give similar conclusions, we can be reasonably confident about the conclusions. Whether these conclusions can be generalized to a larger population would depend on the nature of the sample (i.e., how the data are obtained). If the sample is truly random and representative, we may generalize the conclusions to the population. However, a true random sample is not common. Most samples or data in practice are probably not random and representative samples, such as convenience samples. Therefore, we may need to be careful to make certain general statements, such as the new drug *is* effective for all patients, because such statements should be based on a true simple random sample.

1.7 Outline

The topics for multivariate analysis can be quite extensive, since many statistical models and methods involving more than one variables may be viewed as multivariate analysis in a general sense. In some classic textbooks, multivariate analysis focuses mostly on models and methods for multivariate normal distributions with i.i.d. data. Such a focus allows theoretical developments since multivariate normal distributions have many nice properties and have elegant mathematical expressions. In practice, however, real data are highly complex. The multivariate normal distribution assumption may not hold for some real world problems. This book focuses more on how to analyze data from real world problems. In practice, the data may not follow normal distributions, may be discrete, and may not be independent. We select the topics

which are among the most commonly used in practice. Due to space limitation, some topics, which may also be important in practice, have to be omitted.

Chapters 2 to 4 may be viewed as exploratory multivariate analysis for continuous data. The goal is to reduce the dimension of multivariate data or to classify multivariate observations. Distributional assumptions may not be required, although in some cases we may assume multivariate normal distributions for inference. Chapter 5 considers statistical inference for multivariate normal distributions, including hypothesis testing for the mean vector and covariance matrix, which is often a major focus in classical textbooks. Chapter 6 reviews methods for multivariate discrete data, including inference for contingency tables. Chapter 7 briefly introduces Copula models, which are useful models for non-normal multivariate data. Chapters 8 to 10 briefly review linear and generalized linear regression models, with MANOVA models being viewed as special linear models. Chapter 11 shows models for dependent data, especial models for repeated measurements and longitudinal data. In Chapter 12, we discuss how to handle missing data in multivariate datasets since missing data are very common in practice and they can have substantial effects on analysis results. Chapter 13 briefly describes robust methods for multivariate analysis with outliers, which is an important topic since outliers are not easily detected in multivariate datasets but can have serious impact on analysis results. In Chapter 14, we briefly describe some general methods which are useful in multivariate analysis.

The focus of this book is on conceptual understanding of the models and methods for multivariate data, rather than tedious mathematical derivations or proofs. Extensive real data examples are presented in R. Students completing this course should be ready to perform statistical analysis of multivariate data from real world problems.

Exercises 1

1.1. Prove that the covariance matrix for any random vector is positive semidefinite.

1.2. Let random variable $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma)$. Show that the random variable

$$\mathbf{z} = (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

has a $\chi^2(p)$ distribution.

1.3. Let $(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)$ be a random sample from population $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma)$. Show that the squared Mahalanobis distances

$$d_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{T}} \boldsymbol{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})$$

distributes approximately as $\chi^2(p)$.

1.4. Draw boxplots and pairwise scatterplots for the Chinese consumption data in Section 2.

1.5. Generate a random sample from a three-dimensional multivariate normal population. Then verify the following properties for multivariate normality using R.

(1) All of the marginal distributions are normal.

(2) Any linear combination of the components is normal.

1.6. Check the multivariate normality for the Chinese consumption data in Section 2.

Chapter 2

Principal Components Analysis

2.1 The Basic Idea

In practice, a dataset may contain many variables, such as exam scores on algebra, calculus, geometry, violin, piano, and guitar. Some of these variables may be highly correlated, such as scores on violin, piano, and guitar. When analyzing such multivariate data, it is difficult to graphically display the data. Moreover, if we have to assume parametric distributions for these variables for statistical inference, there may be too many parameters to be estimated or there may be multi-collinearity problems. For example, if we assume a multivariate normal distribution for 6 variables, the covariance matrix will contain 21 parameters. That is, the covariance matrix will be of high dimension, so it is likely to be singular or ill behaved. When the sample size is not large, these parameters may be poorly estimated. Therefore, it is important to *reduce* the number of variables if the loss of information is not much. This dimension reduction is possible because, if some variables are highly correlated, they may be replaced by fewer new variables without much loss of information.

For example, if we have exam scores on six courses (variables): algebra, calculus, geometry, violin, piano, and guitar, we can use a new variable called *mathematics skills* to represent the first three variables and use another new variable called *music skills* to represent the last three variables. These two new variables, mathematics skills and music skills, retain most information in the original six variables. Moreover, these two new variables are uncorrelated and can be obtained by linear combinations of the original six variables. Therefore, we have reduced the number of variables from 6 to 2, without much loss of information. The scores of the original six variables can be converted scores of the two new variables, and we can plot the scores of the two new variables to check the normal assumption and possible outliers. That is the basic idea of *dimension reduction* (the dimension of the original data space is 6, while the dimension of the new data space is 2), and the idea behind principal components analysis. The two new variables are called *principal components*.

Much information in the data or variables can be measured by the *variability* (or *variance*) of the data or variables. In other words, if the values of the data are

all the same, there will be little information in the data. The basic idea of principal components analysis (PCA) is to explain the variability in the original set of correlated variables through a smaller set of uncorrelated new variables. These new variables are obtained by certain linear combinations of the original variables, and they are called the *principal components (PCs)*. That is, the goal of PCA is to reduce the number of original variables while maintain most of the information (variation) in the original data, i.e., it is a dimension-reduction method. Since the PCs are linear combinations of the original variables, normality and outliers in the original data should still be present in the "new data" of the PCs (called *PC scores*), so we can check multivariate normality of the original data or check outliers in the original data based on the PC scores, which is easier since the new data have a lower dimension.

In the example described above, the original data have 6 variables (i.e., 6 dimensions), so it is difficult to graphically display the data on 6 variables and to check their joint distributions or outliers. However, if we reduce the 6 variables to 2 new variables, it is easy to graphically display the data on the two new variables, which allows us to visually check the possible joint distribution of the two variables and to check if there are potential outliers.

We consider another example. In regression analysis, when there are too many predictors in the regression model, some predictors may be highly correlated. This may lead to the so-called "multi-collinearity problem" in regression analysis, which may lead to many problems such as poor parameter estimates and unreliable results. By using PCA, however, we may be able to reduce the number of predictors, replacing the original set of predictors by a new set of predictors (PCs) to avoid the multi-collinearity problem. This approach can greatly improve regression analysis. PCA is also widely used in many other problems, such as statistical genome where there are typically many variables but sample size is small so a PCA is quite valuable.

In summary, PCA is useful for screening multivariate continuous data. It can be used as a first step in data analysis. In the following, we describe the PCA method in details.

2.2 The Principal Components

Let $\mathbf{x} = (x_1, x_2, \cdots, x_p)^{\mathrm{T}}$ be a set of p continuous variables. The basic idea of a PCA method is to transform the set of variables $(x_1, x_2, \cdots, x_p)$ into a *smaller* set of *uncorrelated new* variables and try to explain most variability in the original variables $\mathbf{x}$ through these new variables. Specifically, let

$$\boldsymbol{\mu} = E(\mathbf{x}) = (\mu_1, \mu_2, \cdots, \mu_p)^{\mathrm{T}}, \quad \Sigma = Cov(\mathbf{x}) = (\sigma_{ij})_{p\times p}$$

be the mean vector and the covariance matrix of $\mathbf{x}$ respectively. Note that the mean vector $\boldsymbol{\mu}$ represents the center of $\mathbf{x}$, and the covariance matrix Σ represents the vari-

ations (the diagonal elements of Σ) and correlations (the off-diagonal elements of Σ) of the random vector $\mathbf{x}$.

For ease interpretation, we usually replace $\mathbf{x}$ by its centered version $\mathbf{x} - \boldsymbol{\mu}$. We consider the following linear combinations of the components of vector $\mathbf{x} - \boldsymbol{\mu}$

$$\begin{aligned} y_1 &= \mathbf{a}_1^{\mathrm{T}}(\mathbf{x} - \boldsymbol{\mu}) = a_{11}(x_1 - \mu_1) + a_{12}(x_2 - \mu_2) + \cdots + a_{1p}(x_p - \mu_p), \\ y_2 &= \mathbf{a}_2^{\mathrm{T}}(\mathbf{x} - \boldsymbol{\mu}) = a_{21}(x_1 - \mu_1) + a_{22}(x_2 - \mu_2) + \cdots + a_{2p}(x_p - \mu_p), \\ &\cdots\cdots \\ y_p &= \mathbf{a}_p^{\mathrm{T}}(\mathbf{x} - \boldsymbol{\mu}) = a_{p1}(x_1 - \mu_1) + a_{p2}(x_2 - \mu_2) + \cdots + a_{pp}(x_p - \mu_p), \end{aligned}$$

where $\mathbf{a}_k = (a_{k1}, a_{k2}, \cdots, a_{kp})^{\mathrm{T}}$ is a vector of constants. That is, each y_k is a linear combination of the original random vector $\mathbf{x}$. Then, we can show that

$$Var(y_i) = \mathbf{a}_i^{\mathrm{T}}\Sigma\mathbf{a}_i, \quad Cov(y_i, y_j) = \mathbf{a}_i^{\mathrm{T}}\Sigma\mathbf{a}_j, \quad i, j = 1, 2, \cdots, p.$$

Note that the variance of y_i increases as the length of $\mathbf{a}_i$, denoted by $||\mathbf{a}_i||$, increases. To ensure uniqueness of $\mathbf{a}_i$, we can assume that $\mathbf{a}_i$ is a unit vector, i.e., $||\mathbf{a}_i|| = 1$, for $i = 1, 2, \cdots, p$. This can be easily done since a vector can always be rescaled to have length of one.

The *first principal component* (PC) is defined as

$$y_1 = \mathbf{a}_1^{\mathrm{T}}\mathbf{x}$$

where $\mathbf{a}_1$ is chosen to maximize the variance $Var(\mathbf{a}_1^{\mathrm{T}}\mathbf{x})$ over all constant vectors $\mathbf{a}_1$ subject to the restriction $||\mathbf{a}_1|| = 1$. The *second principal component* (PC) is defined as

$$y_2 = \mathbf{a}_2^{\mathrm{T}}\mathbf{x}$$

where $\mathbf{a}_2$ is chosen to maximize the variance $Var(\mathbf{a}_2^{\mathrm{T}}\mathbf{x})$ over all constant vectors $\mathbf{a}_2$ subject to the restrictions

$$||\mathbf{a}_2|| = 1, \qquad Cov(\mathbf{a}_2^{\mathrm{T}}\mathbf{x},\ \mathbf{a}_1^{\mathrm{T}}\mathbf{x}) = 0.$$

Thus, the first PC and the second PC are uncorrelated. Similarly, the *k-th principal component* (PC) is the linear combination

$$y_k = \mathbf{a}_k^{\mathrm{T}}\mathbf{x}$$

where $\mathbf{a}_k$ is chosen to maximize the variance $Var(\mathbf{a}_k^{\mathrm{T}}\mathbf{x})$, subject to the restrictions

$$||\mathbf{a}_k|| = 1, \qquad Cov(\mathbf{a}_k^{\mathrm{T}}\mathbf{x},\ \mathbf{a}_j^{\mathrm{T}}\mathbf{x}) = 0 \quad \text{for all } j < k,$$

$k = 1, 2, \cdots, p$. Thus, the k-th PC is uncorrelated with the first $k - 1$ PCs.

Based on the above constructions of the principal components, the first PC has the largest variation, the second PC has the second largest variation, and so on. Moreover, all the PCs are uncorrelated. Thus, The first few principal components

may explain most of the variation in the original $\mathbf{x}$. In the following, we show how to find the $\mathbf{a}_j$'s which satisfy the foregoing constructions of the principal components.

Let $\{(\lambda_j, \mathbf{a}_j), j = 1, \cdots, p\}$ be the eigenvalues and the corresponding eigenvectors of the covariance matrix Σ respectively, where $\mathbf{a}_j = (a_{j1}, a_{j2}, \cdots, a_{jp})^{\mathrm{T}}$ is an eigenvector. Suppose that the eigenvalues λ_j's are arranged in a *decreasing* order and that the corresponding eigenvectors are *normalized* (so that they all have lengths of 1), i.e.,

$$\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p, \quad ||\mathbf{a}_j|| = 1 \quad \text{for all } j.$$

Then, by the method of Lagrange multipliers for constrained maximization, it can be shown that the *first principal component* (PC), denoted by y_1, is given by

$$y_1 = \mathbf{a}_1^{\mathrm{T}}(\mathbf{x} - \boldsymbol{\mu}) = a_{11}(x_1 - \mu_1) + a_{12}(x_2 - \mu_2) + \cdots + a_{1p}(x_p - \mu_p),$$

with

$$Var(y_1) = \lambda_1.$$

That is, the first PC is a linear combination of the centered variables $\{x_j - \mu_j,\ j = 1, 2, \cdots, p\}$ (which all have mean 0), with the coefficients of the linear combination given by the components of the first eigenvector $\mathbf{a}_1$. Moreover, the variance (variation) of the first PC y_1 is the largest eigenvalue λ_1 of the covariance matrix Σ.

Similarly, the *second principal component* (PC), denoted by y_2, is given by

$$y_2 = \mathbf{a}_2^{\mathrm{T}}(\mathbf{x} - \boldsymbol{\mu}),$$

with

$$Var(y_2) = \lambda_2.$$

That is, the second PC is a linear combination of the centered variables $\{x_j - \mu_j,\ j = 1, 2, \cdots, p\}$, with the coefficients of the linear combination given by the components of the second eigenvector $\mathbf{a}_2$. Moreover, the variance (variation) of the second PC y_2 is the *second largest* eigenvalue λ_2 of the covariance matrix Σ. It can be shown that the first PC y_1 and the second PC y_2 are uncorrelated, i.e., $cov(y_1, y_2) = 0$.

In general, the *k-th principal component* (PC), denoted by y_k, is given by

$$y_k = \mathbf{a}_k^{\mathrm{T}}(\mathbf{x} - \boldsymbol{\mu}), \qquad \text{with} \qquad Var(y_k) = \lambda_k, \qquad k = 1, \cdots, p,$$

and the k-th PC y_k is uncorrelated with the first $(k-1)$ PCs $y_1, y_2, \cdots, y_{k-1}$. In other words, the variance of y_k is the maximum among the variances of all linear combinations $\mathbf{c}^{\mathrm{T}}(\mathbf{x} - \boldsymbol{\mu})$ that are *uncorrelated* with the first $k-1$ PCs, and this maximum variance is equal to the k-th largest eigenvalue $\lambda_k,\ k = 1, 2, \cdots, p$.

Therefore, the covariance matrix Σ plays the most important role in PCA, while the mean vector $\boldsymbol{\mu}$ only represents the center location of the data points. This is

because the covariance matrix contains all the information about the variabilities and correlations of the variables, so it contains all the information for dimension reduction. On the other hand, the mean vector only shows where the data points might fall. The eigenvalues and the eigenvectors of the covariance matrix Σ describe the key characteristics of the covariance matrix and the random vector $\mathbf{x}$: **the eigenvalues represent the variations and the eigenvectors present the directions of the random vector x**. See Figure 2.1 for an example. Although the data in Figure 2.1 are assumed to follow a bivariate normal distribution, the basic idea remains essentially the same for other continuous data. Note that, in PCA, the data or variables are assumed to be continuous, not necessarily follow normal distributions. That is, in the above descriptions, we only assume a mean vector and a covariance matrix, without any distributional assumption.

From the foregoing description, we see that the principal components y_j's are special linear combinations of the original variables x_j's. Geometrically, these represent the selection of a new coordinate system obtained by *rotating* the original system. The new axes represent the directions with maximum variabilities and provide a simpler and more parsimonious description of the covariance structure. See Figure 2.1 for an illustration. The first PC explains as much of the variability in the data as possible. Each succeeding PC explains as much of the *remaining* variability as is possible. **If the first few PCs can explain most variability in the data, we can just use the first few PCs to replace the original set of variables.** This is the essential idea behind principal components analysis. In the following, we will present a simple example to illustrate the points.

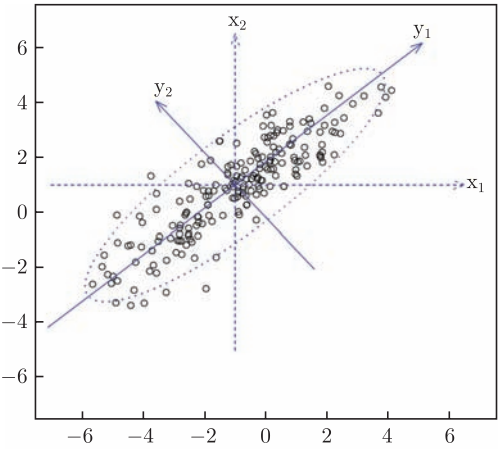


Figure 2.1 Principal Components for the bivariate normal Data.

Example

The data points in Figure 2.1 come from a bivariate normal population $\mathbf{x} = (x_1, x_2)^{\mathrm{T}} \sim N_2(\boldsymbol{\mu}, \Sigma)$ with

$$\boldsymbol{\mu} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} 4 & 3 \\ 3 & 3 \end{pmatrix}.$$

The correlation between x_1 and x_2 is $r = 0.866$, so the scatter plot shows that x_1 and x_2 are highly correlated. The eigenvalues and eigenvectors of covariance matrix Σ are given by

$$\lambda_1 = 6.54, \quad \mathbf{a}_1 = (0.76, 0.65)^{\mathrm{T}}; \qquad \lambda_2 = 0.46, \quad \mathbf{a}_2 = (-0.65, 0.76)^{\mathrm{T}},$$

respectively. Thus, the first PC is

$$y_1 = 0.76(x_1 + 1) + 0.65(x_2 - 1)$$

whose variance is $\lambda_1 = 6.54$. The second PC is

$$y_2 = -0.65(x_1 + 1) + 0.76(x_2 - 1)$$

whose variance is $\lambda_2 = 0.46$.

In Figure 2.1, the two PCs constitute a new coordinate system by rotating the original system of x_1 and x_2 centered at the means. The ellipse is a 95% contour for the bivariate normal distribution $N_2(\boldsymbol{\mu}, \Sigma)$. The major and minor axes of the ellipse lie along the first PC and the second PC respectively. Clearly, the variation in the y_1 direction is much larger than that in the y_2 direction. So we can reduce the dimension from 2 to 1 while retain most of the information (variation) in the data.

2.3 Choose Number of Principal Components

The purpose of PCA is to reduce dimension, i.e., reduce the number of variables. In practice, we need to decide how many dimensions we can reduce without much loss of information. In other words, we should decide how many principal components should be retained. This question can be answered by the amount of variation that can be explained through the first few principal components.

Note that the *total variation* (variance) in the data is

$$tr(\Sigma) = \sigma_{11} + \cdots + \sigma_{pp} = \lambda_1 + \cdots + \lambda_p.$$

Thus, the importance of the j-th PC can be measured by the ratio

$$\frac{\lambda_j}{tr(\Sigma)}, \qquad j = 1, 2, \cdots, p;$$

i.e., the proportion of the total variability explained by the j-th PC. For example, the importance of the first two PCs can be measured by the ratio

$$\frac{\lambda_1 + \lambda_2}{tr(\Sigma)}.$$

If the first few PCs can explain most (e.g., 70%∼80%) of the total variability, then these first few PCs can replace all the original p variables without much loss of information, where the information is measured by the variability. For example, if the first two PCs (y_1 and y_2) can explain 70% variation in the original $p = 10$ variables $(x_1, \cdots, x_{10})$, i.e., if $(\lambda_1 + \lambda_2)/tr(\Sigma) = 0.7$, we can just use the two new variables (i.e., the first two PCs y_1 and y_2) instead of the original 10 variables $(x_1, \cdots, x_{10})$ in data analysis, so the dimension of the data space is reduced from 10 to 2 (a big reduction in dimension!). Then, we can use graphical tools to display the " new data" on the two new variables. Although we loss some information by using the two new variables instead of the original ten variables, we gain a lot in data analysis, such as better parameter estimates and better use of graphical tools.

There have been some suggestions in the literature on choosing the number of principal components. For example, some authors suggest that, if we do PCA on the correlation matrix (not the covariance matrix), then the eigenvalues greater than 1 should be retained, which means that the PCs with variance larger than 1 are retained. A scree plot (see Figure 2.4) is also a useful visual aid for deciding the number of principal components. We will illustrate these methods in the R examples later. These methods are rules of thumb and should be treated as a guideline only. In real applications, however, we do not need to follow these guidelines strictly. The decision for choosing the number of principal components should be based on subject-matter interpretation, i.e., whether the chosen principal components make good sense in the particular problem under consideration and whether the chosen number of principal components can help us in data analysis. For example, if we choose two principal components, we will be able to use graphical tools, but if we choose three or more principal components, we are unable to use graphical tools. On the other hand, we usually hope that the chosen number of principal components can explain most of the variation in the data, such as at least 70% of the total variation. In summary, choosing the number of principal components should be guided by data analysis rather than certain strict rules.

2.4 Considerations in Data Analysis

In practice, the true population mean vector $\boldsymbol{\mu}$ and covariance matrix Σ are unknown. Thus, in data analysis, we should use the sample estimates of the mean vector and the covariance matrix to replace the unknown population mean vector and covariance

matrix. Specifically, given a sample of data $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$, where $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})^{\mathrm{T}}$, we can use the following sample mean vector $\bar{\mathbf{x}} = \hat{\boldsymbol{\mu}}$ and sample covariance matrix $\boldsymbol{S} = \hat{\Sigma}$ for PCA:

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i, \quad \boldsymbol{S} = (\hat{\sigma}_{ij})_{p\times p} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{T}}.$$

The accuracies of these estimates depend on the sample size n. The larger the sample size, the closer the sample estimates to the population parameters.

Note that PCA results may depend on the *scales* or *units* of the variables. For example, a distance x_1 can be measured in centermeter or in meter, and their values can differ by 100 times. The PCA results may depend on the scale (or unit) of x_1. Usually, it is desirable that all the variables in the original data have similar scales, i.e., the magnitudes of the values are comparable (e.g., not some values are around 0.0001 while other values are around 10000). To address this issue, it is generally desirable to perform PCA on the correlation matrix $\boldsymbol{R}$ rather than the original covariance matrix Σ, or perform PCA on the *standardized data*:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{\hat{\sigma}_{jj}}}, \quad i = 1, 2, \cdots, n; \quad k = 1, 2, \cdots, p,$$

where $\bar{x}_j = \sum_{i=1}^{n} x_{ij}/n$, which are transformations of the original data, with mean 0 and variance 1.

Once we find the PCs, i.e., the new variables y_j's, we can convert the original data x_{ij} into "new data" of the PCs y_j's. For example, for individual i, let

$$\hat{y}_{ik} = \hat{\mathbf{a}}_k^{\mathrm{T}}(\mathbf{x}_i - \bar{\mathbf{x}}), \qquad i = 1, 2, \cdots, n; \quad k = 1, 2, \cdots, p,$$

where $\hat{\mathbf{a}}_k$'s are the eigenvectors of the sample covariance matrix $\boldsymbol{S} = \hat{\Sigma}$ or the sample correlation matrix $\boldsymbol{R}$. These "new data"

$$\{\hat{y}_{ik} : \ i = 1, 2, \cdots, n; \ k = 1, 2, \cdots, p\}$$

are called *PC scores*, and they can be used for further analysis. For example, we may proceed with "new data" on the first two PCs $\{\hat{y}_{ik} : \ i = 1, 2, \cdots, n; \ k = 1, 2\}$. in data analysis. In other words, data analysis is performed on the new data with two variables rather than the original data with p variables.

Principal components are linear combinations of the original variables. Original variables have practical meanings, but the principal components do not always have practical meanings. However, sometimes we may be able to interpret interesting practical meanings for some principal components, as illustrated in Examples 2 and 3 in next section.

PCA can also be used in initial screening of choosing number of predictors in regression analysis. For example, in a regression model, if there are too many predictors, there may be difficulties in data analysis, such as poor estimates of the parameters, multi-collinearity, and possibly singular design matrices. The problem is that many predictors may be highly correlated, so some predictors may be unnecessary and should be removed. In this case, we may perform a PCA to see how many predictors may be removed. For example, if there are originally 10 predictors, and a PCA shows that the first two PCs explains about 70% variation in the 10 predictors, a regression models with only two predictors may be reasonable.

2.5 Examples in R

In this section, we consider several real data examples for PCA and illustrate the use of software R for PCA.

Example 1. In this study (Johnson and Wichern, 2007), 48 individuals who had applied for a job with a large firm were interviewed and rated on 15 criteria. Individuals were rated on the form of their letter of application (FL), their appearance (APP), academic ability (AA), likability (LA), self-confidence (SC), lucidity (LC), honesty(HON), salesmanship (SMS), experience (EXP), drive (DRV), ambition (AMB), ability to grasp concepts (GSP), potential (POT), keenness to join (KJ), and their suitability (SUIT). Each criterion was evaluated on a scale ranging from 0 to 10, with 0 being a very low and very unsatisfactory rating, and 10 being a very high rating. In this example, there are 15 variables, so the dimension of the data is 15. So it is difficult to graphically display the data, and to check any outliers and multivariate normality of the data. We consider a PCA to reduce the dimension.

```
> applican.dat0 <- read.table("applicant.dat", head=T)  # import data
> applican.dat1 <- applican.dat0[,-1] # remove ID (the first column)
> attach(applican.dat1)  # put this data as a priority
> options(digits=2)   # just need 2 decimal points

# See part of the data
> applican.dat1
   FL APP AA LA SC LC HON SMS EXP DRV AMB GSP POT KJ SUIT
1   6   7  2  5  8  7   8   8   3   8   9   7   5  7   10
2   9  10  5  8 10  9   9  10   5   9   9   8   8  8   10
3   7   8  3  6  9  8   9   7   4   9   9   8   6  8   10
4   5   6  8  5  6  5   9   2   8   4   5   8   7  6    5
.........................

# Covariance matrix Sigma
> Sigma.ap <- var(applican.dat1)
# Correlation matrix P
> P.ap <- cor(applican.dat1)
# Check pairwise correlations between the 15 variables
> P.ap
        FL APP    AA    LA    SC    LC   HON   SMS   EXP DRV  AMB GSP POT   KJ
FL    1.00 0.2  0.04  0.30  0.09  0.23 -0.10  0.27  0.55 0.3 0.28 0.3 0.4  0.5
```

```
APP    0.24 1.0   0.12   0.38   0.43   0.37   0.35   0.49   0.14 0.3 0.55 0.5 0.5   0.3
AA     0.04 0.1   1.00   0.00   0.001 0.08 -0.03   0.05   0.27 0.1 0.04 0.2 0.3 -0.3
LA     0.31 0.4   0.00   1.00   0.30   0.48   0.64   0.36   0.14 0.4 0.35 0.5 0.6   0.7
SC     0.09 0.4   0.00   0.30   1.00   0.81   0.41   0.80   0.02 0.7 0.84 0.7 0.7   0.5
LC     0.23 0.4   0.07   0.48   0.80   1.00   0.35   0.82   0.15 0.7 0.76 0.9 0.8   0.5
HON   -0.11 0.4 -0.03   0.64   0.41   0.36   1.00   0.24 -0.16 0.3 0.21 0.4 0.4   0.4
SMS    0.27 0.5   0.05   0.36   0.80   0.82   0.24   1.00   0.26 0.8 0.86 0.8 0.8   0.6
EXP    0.55 0.1   0.26   0.14   0.01   0.15 -0.15   0.26   1.00 0.3 0.20 0.3 0.3   0.2
DRV    0.35 0.3   0.09   0.39   0.70   0.70   0.28   0.81   0.34 1.0 0.78 0.7 0.8   0.6
AMB    0.28 0.5   0.04   0.34   0.84   0.76   0.21   0.86   0.20 0.8 1.00 0.8 0.8   0.5
GSP    0.34 0.5   0.19   0.50   0.72   0.88   0.38   0.78   0.30 0.7 0.78 1.0 0.9   0.5
POT    0.37 0.5   0.29   0.60   0.67   0.78   0.41   0.75   0.35 0.8 0.77 0.9 1.0   0.5
KJ     0.47 0.3 -0.32   0.68   0.48   0.53   0.44   0.56   0.21 0.6 0.55 0.5 0.5   1.0
SUIT   0.59 0.4   0.14   0.32   0.25   0.42   0.003 0.56   0.69 0.6 0.43 0.5 0.6   0.4
.........
```

We see that all 15 variables are correlated and some are highly correlated (e.g., POT and GSP have a correlation of 0.9, LC and GSP also have a correlation of 0.9, etc). Since all variables here have comparable scales, we may perform PCA on the sample covariance matrix $S = \hat{\Sigma}$.

```
# Eigenvalues and eigenvectors of Sigma
> eigen.ap <- eigen(Sigma.ap)
> eigen.ap
$values:
 [1] 66.54 18.18 10.59  6.77  3.99  3.63  2.92  2.84  1.96  1.61  1.14  0.87
[13]  0.71  0.51  0.30
$vectors:
       [,1]  [,2]  [,3]  [,4]  [,5]  [,6]   [,7]  [,8]  [,9]  [,10]  [,11]
 [1,] -0.15  0.37  0.20 -0.28  0.64  0.144  0.05 -0.189 -0.38  0.224  0.09
 [2,] -0.13 -0.03  0.04  0.13  0.04  0.757  0.12  0.269  0.03 -0.028 -0.09
 [3,] -0.03  0.10 -0.13  0.60  0.17 -0.002  0.12 -0.365  0.06  0.450 -0.20
 [4,] -0.20 -0.09  0.62  0.13  0.05 -0.019 -0.13 -0.046  0.52  0.135  0.32
 [5,] -0.23 -0.24 -0.19 -0.07 -0.03 -0.012  0.25  0.103 -0.26  0.150  0.48
 [6,] -0.34 -0.20 -0.12  0.05  0.23 -0.367 -0.35  0.318 -0.07  0.161  0.21
 [7,] -0.12 -0.30  0.45  0.26 -0.33  0.059  0.18  0.095 -0.54  0.138 -0.04
 [8,] -0.38 -0.09 -0.28 -0.17 -0.18  0.094 -0.06  0.087  0.24  0.543 -0.30
 [9,] -0.16  0.64  0.03  0.17 -0.19 -0.297  0.48  0.388  0.09  0.007  0.07
[10,] -0.32  0.01 -0.11 -0.13 -0.34 -0.119  0.08 -0.578 -0.16 -0.080  0.02
[11,] -0.31 -0.12 -0.24 -0.15  0.11  0.226  0.34 -0.038  0.24 -0.238  0.20
[12,] -0.34 -0.07 -0.05  0.21  0.26 -0.130 -0.12  0.237 -0.17 -0.357 -0.47
[13,] -0.36 -0.02  0.04  0.32  0.11 -0.029  0.02 -0.291  0.10 -0.414  0.06
[14,] -0.23 -0.04  0.39 -0.46 -0.03 -0.149  0.19 -0.031  0.10  0.011 -0.44
[15,] -0.27  0.47  0.02 -0.02 -0.35  0.255 -0.57 -0.004 -0.13 -0.062  0.12
 .........

######## PCA on the covariance matrix Sigma #########
> princomp.ap <- princomp(applican.dat1, cor=F)  # PCA function "princomp"
> summary(princomp.ap)
Importance of components:
                          Comp. 1 Comp. 2 Comp. 3 Comp. 4 Comp. 5 Comp. 6 Comp. 7
    Standard deviation      8.07    4.22   3.220   2.574   1.976    1.88   1.690
Proportion of Variance      0.54    0.15   0.086   0.055   0.033    0.03   0.024
 Cumulative Proportion      0.54    0.69   0.778   0.833   0.866    0.90   0.919
                          Comp. 8 Comp. 9 Comp. 10 Comp. 11 Comp. 12 Comp. 13
    Standard deviation      1.666   1.384    1.257   1.0549   0.9245   0.8318
Proportion of Variance      0.023   0.016    0.013   0.0093   0.0071   0.0058
 Cumulative Proportion      0.942   0.958    0.971   0.9805   0.9876   0.9934
                          Comp. 14 Comp. 15
```

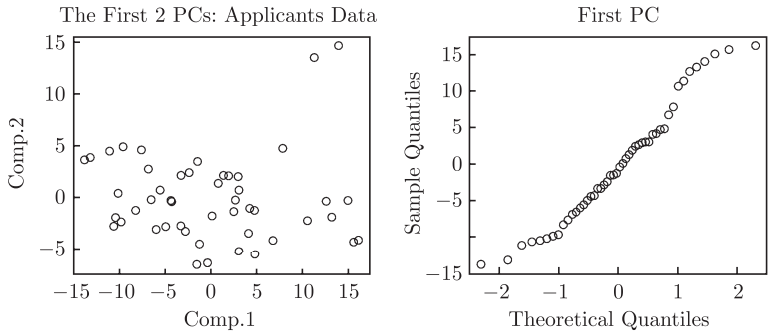
```
    Standard deviation    0.7056    0.5432
Proportion of Variance    0.0042    0.0025
 Cumulative Proportion    0.9975    1.0000
```

In the above results, the "Cumulative Proportion" row gives the cumulative proportions of variation explained by the first PC, the first two PCs, the first three PCs, etc. We see that the first 2 PCs explain 69% variations (the first two eigenvalues of $\hat{\Sigma}$ are 66.54 and 18.18 respectively): the first PC explains 54% variation and the second PC explains 15% variation. The first 3 PC's explain about 78% of total variations. Thus, the 15-dimensional original data may be reduced to 2 or 3 dimensions!

The "new data" (i.e., the PC scores) can be obtained from the original data. To show graphical displays of the new data, we consider PC scores from the first two PCs and then we graphically display the PC scores to check normality of the new data and any possible outliers.

```
> pdf("pca1.pdf")    # save figure to file "pca1.pdf"
> par(mfrow=c(2,2))    # draw 2 by 2 figures
# get PC scores of the first two PCs
> pc.ap2 <- princomp.ap$scores[,1:2]
# scatterplot of first two PC scores
> plot(pc.ap2, main="The First 2 PCs: Applicants Data")
> identify(pc.ap2)  # identify outliers
[1] 42 41
    # Cases 41 and 42 may be outliers
# QQ plot for checking normality of the first 2 PCs
> qqnorm(pc.ap2[,1], main="First PC")
> qqnorm(pc.ap2[,2], main="Second PC")
> dev.off()
```

Figure 2.2 displays the results. We see that there are two potential outliers (cases 41 and 42) in the first two PC scores, suggesting that these two cases may also be outliers in the original data. The normality of the PC scores seems reasonable, except the two possible outliers, suggesting that a normality assumption for the original data may also be reasonable.



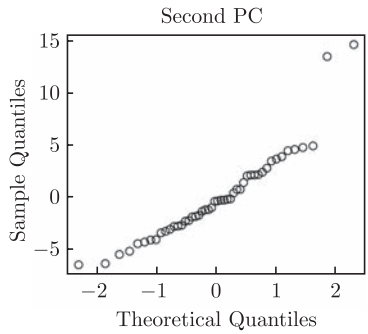


Figure 2.2 Scatterplot and QQ plots of the first two PC scores for the Job Applicants data.

Example 2. In this example, we consider a PCA for the Chinese consumption data using R (see Chapter 1 for a detailed description of the dataset).

```
> consum.1 <- read.table("consum2007.txt",head=T)
> consum<-consum.1[,1:8]
> Sigma.con<-var(consum)
> Cor.con<-cor(consum)
        Food Cloth Resid HousF Health TranC Educ Miscel
Food    1.00  0.26  0.71  0.72   0.39  0.90 0.83   0.72
Cloth   0.26  1.00  0.40  0.45   0.58  0.36 0.54   0.63
Resid   0.71  0.40  1.00  0.77   0.69  0.79 0.81   0.72
HousF   0.72  0.45  0.77  1.00   0.58  0.78 0.89   0.72
Health  0.39  0.58  0.69  0.58   1.00  0.47 0.63   0.63
TranC   0.90  0.36  0.79  0.78   0.47  1.00 0.88   0.75
Educ    0.83  0.54  0.81  0.89   0.63  0.88 1.00   0.84
Miscel  0.72  0.63  0.72  0.72   0.63  0.75 0.84   1.00
```

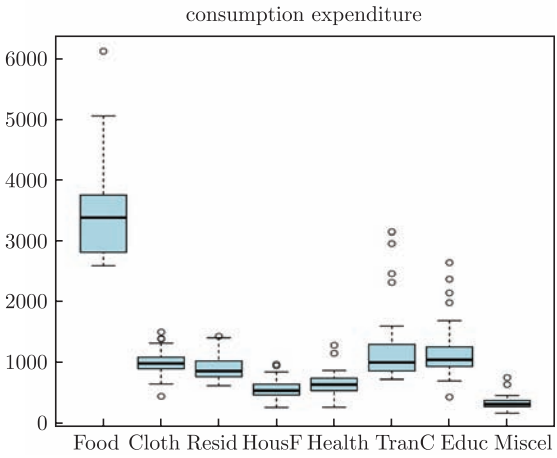


Figure 2.3 Boxplots for the Consumption Data.

The correlation matrix shows that all variables are moderately or highly correlated, so PCA may be conducted to reduce the number of variables. Figure 2.3 shows the boxplots of the variables, which allow us to see the rough distributions of the data and their magnitudes. We see that the magnitudes of the variables differ substantially, such as the values of Food and Cloth. Moreover, the spreads (variations) are also quite different. For example, there are lots of variation in the Food data, but not much variations in the Cloth data. There are also outliers in the data of each variable. Therefore, we decide to perform a PCA on sample correlation matrix rather than the covariance matrix.

```
> princomp.con2<-princomp(consum,cor =T)
> summary(princomp.con2)
Importance of components:
                       Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
Standard deviation       2.39   1.01  0.710  0.522  0.431   0.40  0.295
Proportion of Variance   0.71   0.13  0.063  0.034  0.023   0.02  0.011
Cumulative Proportion    0.71   0.84  0.904  0.938  0.962   0.98  0.993
                       Comp.8
Standard deviation     0.2416
Proportion of Variance 0.0073
Cumulative Proportion  1.0000

> princomp.con2$loadings

Loadings:
       Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
Food    0.353 -0.429  0.175  0.299         -0.377  0.651
Cloth   0.250  0.677  0.521         -0.398 -0.133  0.134
Resid   0.371        -0.442         -0.589  0.530  0.167
HousF   0.374                -0.789  0.260         0.117  0.372
Health  0.302  0.472 -0.628  0.226  0.253 -0.413
TranC   0.376 -0.324  0.123  0.127 -0.279 -0.271 -0.695  0.298
Educ    0.404                -0.200  0.132        -0.156 -0.857
Miscel  0.374  0.118  0.283  0.408  0.518  0.551         0.142

               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
SS loadings      1.00   1.00   1.00   1.00   1.00   1.00   1.00   1.00
Proportion Var   0.12   0.12   0.12   0.12   0.12   0.12   0.12   0.12
Cumulative Var   0.12   0.25   0.38   0.50   0.62   0.75   0.88   1.00

> screeplot(princomp.con2,type="lines")
```

From the above results, we see that the first PC explains 71% of total variation, and the first two PC's explain 84% of total variation. This can also be seen in the scree plot in Figure 2.4. Note that only two eigenvalues are greater than 1. Thus, a reasonable choice of the number of principal components is 2 (or even 1!). So the dimension of the data may be reduced to 2. The values in the eigenvectors show that

the first PC appears to be essentially an average of the 8 original variables, which may be viewed as a composite consumption component. The second PC appears to be a contrast between a weighted average of consumptions on cloth, health care, and medical services and a weighted average of consumptions on food, transport, and communication. Therefore, in this example, we are able to not only reduce the data dimension but also interpret the PCs with interesting practical meanings! This interpretation also justifies the use of the two PCs to replace the original 8 variables. We will do a more elaborate analysis in the next chapter.

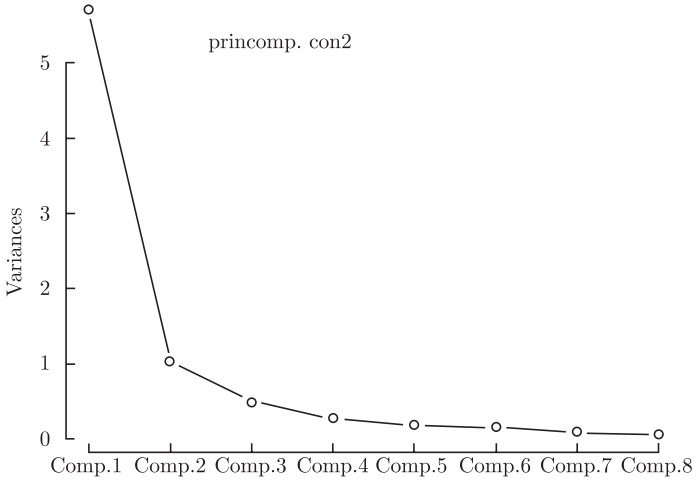


Figure 2.4 Scree Plot for the Consumption Data.

Example 3. The dataset consists of weekly rate returns of 2 stocks listed at Shenzhen Stock Exchange and 4 stocks listed at Shanghai Stock Exchange. The six stocks are VANKE (China Vanke Co., Ltd., code 000002), CMPD (China Merchants Property Development Co., Ltd., code 000024), PRE (Poly Real Estate Group Co., Ltd., code 600048), SINOPEC (China Petroleum & Chemical Corporation, code 600028), SHER (Shanghai Datun Energy Resources Co., Ltd., code 600508), CSEC (China Shenhua Energy Co., Ltd., code 601088). Data from 149 successive weeks in the period of August 2009 to June 2012 are recorded (the data were obtained from www.resset.cn).

We first check the correlation matrix of the weekly rate returns of these 6 stocks

$$\begin{array}{l} \text{VANKE} \\ \text{CMPD} \\ \text{PRE} \\ \text{SINOPEC} \\ \text{SHER} \\ \text{CSEC} \end{array} \begin{pmatrix} 1.00 & . & . & . & . & . \\ 0.88 & 1.00 & . & . & . & . \\ 0.86 & 0.87 & 1.00 & . & . & . \\ 0.49 & 0.51 & 0.56 & 1.00 & . & . \\ 0.50 & 0.51 & 0.48 & 0.55 & 1.00 & . \\ 0.61 & 0.58 & 0.61 & 0.68 & 0.72 & 1.00 \end{pmatrix}.$$

We find that the weekly rate returns of these 6 stocks are highly correlated, which suggests that a PCA may be useful. The results of a PCA on the correlation matrix of the 6 stocks are shown in Table 2.1.

Table 2.1 Principal components analysis of the 6 stocks

Stock	PC1	PC2	PC3	PC4	PC5	PC6
VANKE	0.43	−0.39	0.104	0.082	−0.487	0.638
CMPD	0.43	−0.38	0.049	−0.180	−0.293	−0.737
PRE	0.44	−0.36	−0.102	−0.011	0.809	0.120
SINOPEC	0.36	0.41	−0.773	−0.283	−0.125	0.068
SHER	0.36	0.50	0.608	−0.487	0.083	0.082
CSEC	0.41	0.40	0.095	0.802	0.001	−0.156
Cumulative percentage of total variance	0.72	0.869	0.922	0.95	0.980	1.000

We see that the first PC explains 72% of total variation and the first two PC's explain 86.9% of the total variation. So the dimension of the data space can be reduced from 6 to 2 (or even 1!). From Table 2.1, we also see that the first PC appears to be essentially an average of the 6 stocks, which represents a common market component, and the second PC appears to a contrast between the first 3 stocks and the remaining 3 stocks. Actually, the first 3 stocks belong to the real estate industry, while the remaining 3 stocks belong to the energy industry. Thus, the second PC distinguishes the real estate industry from the energy industry. Again, in this example the first two PCs have very interesting and meaningful practical interpretations! We will do a factor analysis to obtain more elaborate results in the next chapter.

Exercises 2

2.1. Use the method of Lagrange multipliers for constrained maximization to derive the principal components.

2.2. Show that the principal components $y_1, y_2, \cdots, y_p$ are uncorrelated.

2.3. Show that the variance of the k-th principal component is the k-th largest eigenvalue λ_k, $k = 1, 2, \cdots, p$.

2.4. Compute the correlation coefficient between the i-th PC and the j-th variable, $i, j = 1, 2, \cdots, p$.

2.5. Take a logarithm transformation of the data in the dataset "consum2007". Then carry out a principal component analysis. Do you find any differences in the results with those based on the original data?

2.6. The data set "police.dat"consist of anthropometric and physical fitness measurements that were taken on 50 white male applicants to the police department of a major metropolitan city. The variables include reaction time in seconds to a visual stimulus (REACT), the applicant's height in centimeters (HEIGHT), the applicant's weight in kilograms (WEIGHT), the applicant's shoulder width in centimeters (SHLDR), the applicant's pelvic width in centimeters (PELVIC), the applicant's minimum chest circumference in centimeters

(CHEST), the applicant's thigh skinfold thickness in millimeters (THIGH), the applicant's resting pulse rate (PULSE), the applicant's diastolic blood pressure (DIAST), the number of chin-ups the applicant was able to complete (CHNUP), the applicant's maximum breathing capacity in liters (BREATH), the applicant's pulse rate after 5 minutes of recovery from treadmill running (RECVR), the applicant's maximum treadmill speed (SPEED), the applicant's treadmill endurance time in minutes (ENDUR), and the applicant's total body fat measurement (FAT). Thus, in this study there are 14 variables. Conduct a PCA to reduce the dimensionality.

2.7. The decathlon is a combined event in athletics consisting of ten track and field events held over two consecutive days. The events are 100 meters, long jump, shot put, high jump, 400 meters, 110 meters hurdles, discus throw, pole vault, javelin throw, and 1500 meters. Performance is judged on a points system in each event, and the winners are determined by the combined performance in all events. The following correlation coefficients matrix is computed based on the performance of the Olympic athletes on each events. Find the principal components of the decathlon and interpret the results.

$$
\begin{array}{l}
\text{100m} \\
\text{long jump} \\
\text{shot put} \\
\text{high jump} \\
\text{400m} \\
\text{110m hurdles} \\
\text{discus throw} \\
\text{pole vault} \\
\text{javelin throw} \\
\text{1500m}
\end{array}
\begin{pmatrix}
1.00 & . & . & . & . & . & . & . & . & . \\
0.59 & 1.00 & . & . & . & . & . & . & . & . \\
0.35 & 0.42 & 1.00 & . & . & . & . & . & . & . \\
0.34 & 0.51 & 0.38 & 1.00 & . & . & . & . & . & . \\
0.63 & 0.49 & 0.19 & 0.29 & 1.00 & . & . & . & . & . \\
0.40 & 0.52 & 0.36 & 0.46 & 0.34 & 1.00 & . & . & . & . \\
0.28 & 0.31 & 0.73 & 0.27 & 0.17 & 0.32 & 1.00 & . & . & . \\
0.20 & 0.36 & 0.24 & 0.39 & 0.23 & 0.33 & 0.24 & 1.00 & . & . \\
0.11 & 0.21 & 0.44 & 0.17 & 0.13 & 0.18 & 0.34 & 0.24 & 1.00 & . \\
-0.07 & 0.09 & -0.08 & 0.18 & 0.39 & 0.00 & -0.02 & 0.17 & -0.00 & 1.00
\end{pmatrix}.
$$

Chapter 3

Factor Analysis

3.1 The Basic Idea

In principal components analysis (PCA), we try to reduce the dimension of multivariate data to simplify multivariate analysis. A PCA gives us some idea about the minimal number of variables which contain most information in the original set of variables. A disadvantage of PCA is that the principal components may not always have practical interpretations. That is, sometimes the principal components do not have meaningful interpretations in practice, which is a disadvantage in real data analysis. In this section, we try to determine the minimal set of variables which also have meaningful interpretations in practice. Such analysis is called *factor analysis*, and the factors usually have practical interpretations.

Factor analysis (FA) tries to describe the variance-covariance relationship among variables in terms of a smaller set of unobservable and uncorrelated new random variables called *factors*. These factors cannot be directly observed, but they have practical meanings and can be used for data analysis. **The essential idea of factor analysis is to group the original variables so that all variables in the same group are highly correlated. Each group then represents a factor (a new variable) that explains the variation and correlation in the original variables in the group.** The original set of variables may then be replaced by these factors in data analysis. Thus, factor analysis is closely related to PCA. A main difference is that the factors have practical meanings while the principal components may not have practical meanings.

For example, suppose that a multivariate dataset contains exam scores on mathematics, physics, computer science, English, Chinese, French, income, education, and professionals. We wish to perform a multivariate analysis on this dataset. We see that the dimension of the original data is 9, and we hope to reduce the dimension in data analysis. Note that exam scores on mathematics (x_1), physics (x_2), and computer science (x_3) may be represented by an unobservable factor called *intelligence* (f_1). Exam scores on English (x_4), Chinese (x_5), and French (x_6) may be represented by an unobservable factor called *verbal ability* (f_2). Data on income (x_7), education

(x_8), and professionals (x_9)may be represented by an unobservable factor called *social status* (f_3). Thus, we have grouped the original 9 variables that are highly correlated, and obtain three factors (f_1, f_2, f_3): intelligence, verbal ability, and social status. These factors are unobservable and uncorrelated, and they have practical meanings. Therefore, the original 9 variables may be represented by three factors, which is a big reduction of data dimension. Although these three factors may not completely represent the original 9 variables, the three factors should contain most information in the original variables or explain most variability in the original variables. Compared to PCA, an attractive feature of factor analysis is that the factors have practical interpretation, since intelligence, verbal ability, and social status are meaningful variables. This is the idea behind factor analysis.

3.2 The Factor Analysis Model

The idea of factor analysis can be formally stated as follows. Let $\mathbf{x} = (x_1, \cdots, x_p)^{\mathrm{T}} \sim (\boldsymbol{\mu}, \Sigma)$ be the original set of variables with mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_p)^{\mathrm{T}}$ and covariance matrix $\Sigma = (\sigma_{ij})_{p\times p}$. Note that $\mathbf{x} \sim (\boldsymbol{\mu}, \Sigma)$ means that the random vector $\mathbf{x}$ has a mean vector $\boldsymbol{\mu}$ and covariance matrix Σ, without a distributional assumption. The *factor analysis* (FA) *model* can be written as

$$x_i = \mu_i + \lambda_{i1} f_1 + \cdots + \lambda_{im} f_m + \eta_i, \qquad i = 1, 2, \cdots, p, \quad m \leqslant p, \tag{3.1}$$

where the f_j's are random variables called *factors* or *common factors*, the quantities λ_{ij}'s are called *loadings*, m is an positive integer smaller than the original number of variables p, and η_i's are random errors.

In the factor analysis model (3.1), the original set of variables x_j's are written as linear combinations of the common factors f_j's plus random errors. In other words, the variation in the original set of variables can be partially explained by the variation in the common factors. Typically, the number of factors is less than the number of original variables, i.e., $m < p$. The factors are unobservable. The loading λ_{ij}'s represent the *contribution* (importance) of factor f_j to variable x_i. The random errors η_i's represent variations that cannot be explained by the factors.

Note that the FA model (3.1) is different from a regression model. In a regression model, both the responses and the predictors are *observed* (or known). In the FA model, however, the common factors f_j's are *not observed* (or unknown). Thus, statistical methods for FA models are different from those for regression models.

To estimate the unknown parameters and factors based on given data, we must make some assumptions for the FA model. The common assumptions for the FA model (3.1) are

- the factors f_j's are i.i.d.$\sim (0,1)$, i.e., the factors are independently and identically distributed with mean 0 and variance 1;
- the random errors η_j's $\sim (0, \psi_j)$, and are independent, i.e., the random errors are independent with mean 0 and variances ψ_j's;
- the factor f_k and the random error η_j are independent for any k, j.

These assumptions are needed for a standard factor analysis.

The FA model and its assumptions can be written in the following compact matrix form:

$$\begin{aligned} \mathbf{x} &= \boldsymbol{\mu} + \Lambda \mathbf{f} + \eta, \\ \mathbf{f} &\sim (0, I), \quad \eta \sim (0, \Psi), \quad \mathbf{f} \text{ and } \eta \text{ are independent}, \end{aligned} \tag{3.2}$$

where $\Lambda = (\lambda_{ij})_{p\times m}$ is the *loading matrix*, $\mathbf{f} = (f_1, \cdots, f_m)^{\mathrm{T}}$, $\eta = (\eta_1, \cdots, \eta_p)^{\mathrm{T}}$, and $\Psi = diag(\psi_1, \cdots, \psi_p)$ is a diagonal matrix. This matrix form is convenient for presentation and mathematical arguments. Like PCA, the covariance matrix Σ of $\mathbf{x}$ plays a key role in factor analysis, since the covariance matrix measures the variation and correlation in the data. The mean vector $\boldsymbol{\mu}$ simply measures the location of the data, so it does not contain any information about the variation and correlation in the data.

In both PCA and FA, the variation in the original set of variables are partially explained by the variation in a smaller set of new and uncorrelated variables (principal components and factors). However, PCA mostly focuses on dimension reduction and the principal components may or may not have meaningful practical interpretation, while in factor analysis the factors typically have meaningful practical interpretation. Often, a PCA can be used to roughly determine the number m of factors needed in factor analysis.

3.3 Methods for Estimation

In this section, we describe several methods to estimate the parameters in a FA model. In FA model (3.2), the loading matrix Λ and the factors $\mathbf{f}$ and error η are all unobservable. Since $\mathbf{f}$ and η are assumed to be independent, based on FA model (3.2) and its assumptions, we can obtain the following *factor analysis* (FA) *equation*

$$\Sigma = \Lambda\Lambda^{\mathrm{T}} + \Psi. \tag{3.3}$$

The above FA equation leads to the following partition of variance for variable x_j:

$$\sigma_{jj} = \sum_{k=1}^{m} \lambda_{jk}^2 + \psi_j, \qquad j = 1, \cdots, p,$$

i.e., the total variance (variation) of the original variable x_j can be partitioned into the contributions from the common factors and the random variation. The *communality* of variable x_j, defined as

$$c_j = \sum_{k=1}^{m} \lambda_{jk}^2 / \sigma_{jj}, \qquad j = 1, \cdots, p$$

is the proportion of the variance of x_j that is explained by the common factors $(f_1, \cdots, f_m)$. Therefore, the communality c_j indicates the importance of the common factors to variable x_j, $j = 1, 2, \cdots, p$. In other words, the communality c_j indicates how much variation in the original variable x_j can be explained by the common factors $(f_1, \cdots, f_m)$. The variance of random error ψ_i is called *uniqueness* or *specificity*.

The FA equations can be solved using different methods, such as the the maximum likelihood method and the *principal factor method*. For the maximum likelihood method, we assume that the error η follows the multivariate normal distribution $N(0, \Phi)$, and then we maximize the likelihood to obtain the maximum likelihood estimates of Γ and $\mathbf{f}$. The principal factor method uses a different approach and does not require the normality assumption for η. We omit the technical details here. Interested readers can find the technical details in Johnson and Wichern (2007). Although different methods are available for solving the FA equations, each method has its own advantages and limitations. Thus, in data analysis, a good strategy is to use different methods to analyze the same dataset and then compare the results. If all the results are similar, the conclusions may be reliable. If the results based on different methods lead to different conclusions, we should do a further investigation and try to gain some insights as to why the results differ.

Note that the factor loading matrix Λ is not unique, i.e., different loading matrices may satisfy the same FA equations. In fact, for any orthogonal matrix $\boldsymbol{Q}$, the new matrix Λ^* obtained by the following orthogonal transformation

$$\Lambda^* = \Lambda \boldsymbol{Q}$$

is also a loading matrix, since $\Lambda^* \Lambda^{*\mathrm{T}} = \Lambda \boldsymbol{Q} \boldsymbol{Q}^{\mathrm{T}} \Lambda = \Lambda \Lambda^{\mathrm{T}}$. The matrix Λ^* is called a *rotation* of the loading matrix Λ, since Λ^* is an orthogonal transformation of Λ. Thus, any rotation of a loading matrix is also a loading matrix. This is in fact an advantage of factor analysis, since in real data analysis we can rotate the factors so that the new factors are easy to interpret. **This easy interpretability is a major advantage of factor analysis, compared to PCA**. In other words, in data analysis, we should choose a rotation Λ^* so that the factors are practically meaningful.

There are also many methods for rotating the loading matrix, such as the *varimax method* and the *quartimax method*. That is, there are many methods to choose the

orthogonal matrix $\boldsymbol{Q}$ (Johnson and Wichern, 2007). Although different methods are available, the basic idea behind these methods is as follows: **we should choose a rotation to make as many factor loadings as possible near zero and in the mean time to maximize as many of the other factor loadings as possible**. Such a rotation makes the interpretation of the factors relatively easy in practice. In other words, when trying to interpret the results in practice, we can focus on the factors with large loadings and ignore the factors with small loadings, as illustrated in the examples given below. Again, each method has its own advantages and limitations. In data analysis, we may try different methods and choose the one that makes the interpretation easiest and most meaningful for the given real world problem.

An illustrating example

In the stock example (Example 3) of Chapter 2, we find that the dimension of the six stock data can be reduced to two. This suggests that, in factor analysis, we may be able to use just two factors to analyze the stock data. The results of a factor analysis are displayed in Table 3.1, which contains the factor loadings before rotation and the factor loadings after rotation using the varimax method. Table 3.1 also shows that the first two factors explain 77% of the total variation in the data.

Table 3.1 Factor analysis for the six stocks

Stock name	Loadings before rotation		Loadings after rotation		Communality
	f_1	f_2	f_1	f_2	
VANKE	0.90	−0.24	0.86	0.36	0.87
CMPD	0.90	−0.28	0.88	0.33	0.88
PRE	0.90	−0.23	0.85	0.36	0.86
SINOPEC	0.66	0.29	0.34	0.63	0.52
SHER	0.66	0.38	0.30	0.70	0.58
CSEC	0.81	0.50	0.33	0.89	0.90
Cumulative % of total variance	0.66	0.77	0.43	0.77	

Note that, for each factor, the corresponding values of the factor loadings represent the importance or contribution of that factor to the corresponding original variables. For example, factor f_1 is important for the first three stocks VANKE, CMPD, and PRE (with loadings being 0.90 respectively), and factor f_2 is important for the stock CSEC (with loading being 0.50). Based on the factor loadings before rotation in Table 3.1, however, we do not see a very clear pattern to separate the stocks. For example, all stocks seem important for factor f_1, although some are more important than others. Thus, we try to do a rotation and hope to get a better interpretation. After rotation, the loadings in Table 3.1 clearly indicate that factor f_1 may represent the first three stocks (VANKE, CMPD, and PRE), while factor 2

may represent the remaining stocks (SINOPEC, SHER, and CSEC). This also has an interesting practical interpretation: the first factor represents the real estate industry and the second factor represents the energy industry. The rotation is illustrated in Figure 3.1.

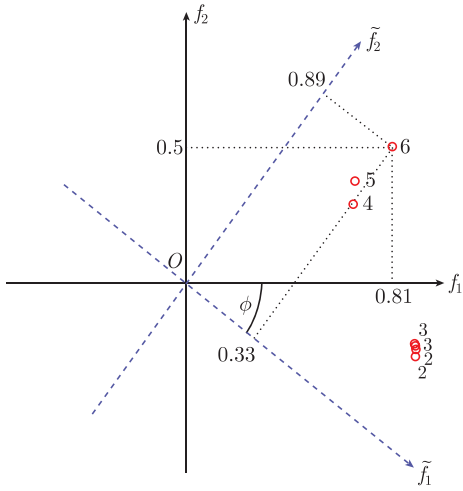


Figure 3.1 Factor Rotation for the six Stocks.

In practice, the number of factors m can be chosen based on the PCA method. For example, if the PCA shows that at least 70% of the total variation can be explained by two principal components, we may simply choose two factors ($m = 2$) in factor analysis.

In factor analysis, although the factors are unobservable, they can be estimated from the data. The estimated values of the factors f_j's are called *factor scores*. These factors scores may be viewed as "data" of the factors, and they can be used for further analysis.

3.4 Examples in R

In this section, we present several examples to show how to use R to do factor analysis.

Example 1. The weekly rates of the return of the following 5 stocks were determined (Johnson and Wichern, 2007): Applied Chemical (x_1), duPont (x_2), Union Carbide (x_3), Exxon (x_4), Texaco (x_5). Observations in 100 successive weeks were obtained. Based on a PCA, the first two PCs explain about 73% variation. The first two PCs are given by

$$y_1 = 0.46x_1 + 0.46x_2 + 0.47x_3 + 0.42x_4 + 0.42x_5,$$
$$y_2 = 0.24x_1 + 0.51x_2 + 0.26x_3 - 0.53x_4 - 0.58x_5.$$

We see that the first PC y_1 is an equally weighted sum of individual stocks, so it may be interpreted as representing *market component*. The second PC y_2 is a contrast between the chemical stocks (the first three stocks x_1, x_2, x_3), and the oil stocks (the last two stocks x_4, x_5), so it may be interpreted as representing the *industry component*. Thus, most of the variation in these five stock returns is due to market activity and uncorrelated industry activity.

We can also do a factor analysis, which may allows us to obtain better interpretations. Based on the above PCA, we consider 2 factors. The following FA results are obtained: one without rotation and one with rotation of the loadings

variable	factor loadings		rotated factor loadings	
	f1	f2	f1*	f2*
Applied Chemical	0.68	0.19	0.60	0.38
du Pont	0.69	0.52	0.85	0.16
Union Carbide	0.68	0.24	0.64	0.33
Exxon	0.62	-0.07	0.36	0.51
Texaco	0.79	-0.44	0.20	0.88

We see that the rotated factors f_1^* and f_2^* are easier to interpret: the chemical companies Applied Chemical, du Pont, and Union Carbide contribute most to the first factor f_1^* (the corresponding loadings are high), while the oil companies Exxon and Texaco contribute most to the second factor f_2^* (the corresponding loadings are high). Thus, we can interpret the two factors as follows: factor f_1^* represents chemical stocks and factor f_2^* represents oil stocks. Such a meaningful and practical interpretation is unavailable in PCA, and it is an advantage of factor analysis.

Example 2. We return to the job applicant data described in Chapter 2. Here we do a factor analysis on this dataset for comparison. In PCA, it was shown that the first two PCs can explain most variation in the data. Thus, in the following we consider a factor analysis using two factors ($m = 2$). We first repeat the PCA, and then do a factor analysis.

```
> applican.dat0 <- read.table("applicant.dat", head=T)
> applican.dat1 <- applican.dat0[,-1] # remove ID
> attach(applican.dat1)
> options(digits=2)    # only show 2 decimal pts

# Sample correlation matrix
> R.app <- cor(applican.dat1)
# Determinant of the correlation matrix
> det(R.app)
[1] 1.6e-07
# The determinant is very small, so there is definitely a need
# for factor analysis, i.e., some variables may be highly correlated.

# PCA on the correlation matrix R
```

```
> pc.app <- princomp(applican.dat1, cor=T)
> summary(pc.app)
Importance of components:
                       Comp. 1 Comp. 2 Comp. 3 Comp. 4 Comp. 5 Comp. 6 Comp. 7
    Standard deviation     2.7    1.43   1.207    1.09   0.860   0.703   0.593
Proportion of Variance     0.5    0.14   0.097    0.08   0.049   0.033   0.023
 Cumulative Proportion     0.5    0.64   0.735    0.81   0.864   0.897   0.921
                       Comp. 8 Comp. 9 Comp. 10 Comp. 11 Comp. 12 Comp. 13
    Standard deviation   0.557   0.507    0.430     0.39   0.3124   0.2980
Proportion of Variance   0.021   0.017    0.012     0.01   0.0065   0.0059
 Cumulative Proportion   0.941   0.958    0.971     0.98   0.9874   0.9933
                       Comp. 14 Comp. 15
    Standard deviation   0.2542   0.1890
Proportion of Variance   0.0043   0.0024
 Cumulative Proportion   0.9976   1.0000
```

We see that the first 2 PCs explain 64% of total variation in the original data, and the first 3 PCs explain 73.5% of total variation. Let's try a FA with 2 factors. Note that in the previous chapter we did a PCA on the covariance matrix Σ rather than the correlation matrix $\boldsymbol{R}$ here. The results were slightly different. In the following, we use the maximum likelihood method for factor analysis.

```
# FA using the maximum likelihood method
> fact3.app <- factanal(applican.dat1, factors=2,method="mle",rotation="
    none")
> fact3.app$loadings
     Factor1 Factor2
  FL   0.372   0.560
 APP   0.533
  AA   0.116   0.187
  LA   0.506
  SC   0.836  -0.389
  LC   0.882  -0.177
 HON   0.359  -0.287
 SMS   0.911
 EXP   0.324   0.730
 DRV   0.861   0.101
 AMB   0.894  -0.145
 GSP   0.903
 POT   0.888
  KJ   0.631
SUIT   0.608   0.666

# No need to interpret the results now, since interpretation
# is ususally easier after a rotation.

# Try rotation using the Varimax method
> fact4.app <- factanal(applican.dat1, factors=2,method="mle",rotation="
    varimax")
# Loading matrix after rotation
> fact4.app$loadings
     Factor1 Factor2
  FL   0.144   0.656
 APP   0.486   0.222
```

```
  AA           0.216
  LA   0.445   0.253
  SC   0.920
  LC   0.886   0.154
 HON   0.439  -0.138
 SMS   0.867   0.284
 EXP           0.798
 DRV   0.767   0.405
 AMB   0.885   0.188
 GSP   0.843   0.324
 POT   0.796   0.404
  KJ   0.566   0.284
SUIT   0.326   0.841
# Variables that are highly correlated with the 2 rotated factors are:
# Factor 1: APP, LA, SC, LC, HON, SMS, DRV, AMB, GSP, POT, KJ
# Factor 2: FL, EXP, DRV, POT, SUIT
```

After rotation, we see that factor 1 explains much variation in variables APP, LA, SC, LC, HON, SMS, DRV, AMB, GSP, POT, and KJ, while factor 2 explains much variation in variables FL, EXP, and SUIT. Therefore, we may interpret factor 1 as *personal characteristics* such as appearance, self confidence, honesty, and potential, while we may interpret factor 2 as *working experience and suitability* for the position. These interpretations seem meaningful, which are unavailable in PCA.

Example 3. We perform a factor analysis for the Chinese consumption data described in the previous chapter for comparison. Based on the PCA in the previous chapter, we find that the first two PCs can explain most of the variation in the original data, so we consider a factor analysis using two factors.

```
> consum.1 <- read.table("consum2007.txt",head=T)
> consum<-consum.1[,1:8]
> fact.con1<-factanal(consum,factors=2, rotation="none") # no rotation
> fact.con1$loadings
Loadings:
       Factor1 Factor2
Food     0.89   -0.32
Cloth    0.50    0.56
Resid    0.84    0.10
HousF    0.88    0.13
Health   0.61    0.49
TranC    0.93   -0.20
Educ     0.97    0.10
Miscel   0.85    0.20

                Factor1 Factor2
SS loadings       5.43   0.786
Proportion Var    0.68   0.098
Cumulative Var    0.68   0.778

# compute the communalities
```

```
> apply((fact.con1$loadings)^2,1,sum)
  Food   Cloth   Resid   HousF Health   TranC    Educ Miscel
  0.90    0.56    0.72    0.79    0.62    0.91    0.95    0.76
```

We see that the cumulative proportion of variance explained by the two factors is 77.8%. At the same time, the communalities show that the two factors can explain most variations in most of the eight variables. We may try to do a factor analysis with three factors and compare the results.

```
# Analysis with three factors and rotation
> fact.con2<-factanal(consum,factors=3, rotation="varimax")
> fact.con2$loadings
Loadings:
       Factor1 Factor2 Factor3
Food   0.92    0.12    0.21
Cloth  0.16    0.76    0.12
Resid  0.57    0.29    0.77
HousF  0.68    0.47    0.33
Health 0.22    0.61    0.50
TranC  0.88    0.21    0.29
Educ   0.78    0.52    0.29
Miscel 0.63    0.56    0.26

               Factor1 Factor2 Factor3
SS loadings       3.48    1.90    1.24
Proportion Var    0.43    0.24    0.15
Cumulative Var    0.43    0.67    0.83

> apply((fact.con2$loadings)^2,1,sum)
  Food   Cloth   Resid   HousF Health   TranC    Educ Miscel
  0.89    0.63    1.00    0.79    0.68    0.91    0.95    0.78
```

The three factors explain 83% of total variation. The communalities for Cloth and Health somewhat increase. After rotation for the three factors, we find that Food, HousF, TranC, Educ and Miscel are highly correlated with Factor 1, Cloth and Health are highly correlated with Factor 2, and Resid is highly correlated with Factor 3. These three factors seem meaningful in practice.

Exercises 3

3.1. Based on the FA model (3.2), derive the FA equation (3.3).

3.2. Based on the FA model (3.2), solve the FA equation using the maximum likelihood method.

3.3. Based on the FA model (3.1), compute the correlation coefficient between the ith variable and the jth factor.

3.4. Will the communalities change after factor rotation? Will the cumulative variance explained by the factors change after rotation? Why? Use Example 3 to verify your conclusions.

3.5. Perform a factor analysis for the decathlon performance data in Exercise 2.6. Interpret the meanings of the factors.

3.6. Perform factor analyses for data “consum2000” and “consum2010” respectively. Is there any difference in the results between these two years?

3.7. Analyze the police data described in the previous chapter (Exercise 2.6) using a factor analysis method.

Chapter 4

Discriminant Analysis and Cluster Analysis

4.1 Introduction

In practice, we often wish to separate individuals into different groups based on observed multivariate data. For example, a bank may wish to separate good customers from bad customers based on their credit history, education, and income, or a teacher may wish to separate good students from bad students based on their grades, motivation, attitude, and other performances. In other words, given observed multivariate data, we wish to decide which group (or population) an individual in the sample belongs to. This type of analysis is called *discriminant analysis* or *cluster analysis*. Such analysis would be easy if the separation is only based on one variable, such as students' grades or customers' income. However, when we have multivariate data, the analysis may not be easy. For example, a student may have an average grade but excellent attitude, or a customer may have high income but poor credit history and low education, then it is not clear which group the student or customer belong to.

Discriminant analysis and cluster analysis are important in many areas. The difference between discriminant analysis and cluster analysis depends on whether the groups (or populations) are known in advance or not. If the groups are known in advance, e.g., we already know that there are good and bad students in a class or there are good customers and bad customers in a bank and we just want to separate them, then the analysis is called discriminant analysis. If the groups are not known in advance, e.g., we don't know whether there are good students or bad students in a class and we wish to determine if all students in a class can be separated into two groups (sometimes maybe all students in a class are good students), then the analysis is called cluster analysis.

In the following section, we first introduce the basic idea of discriminant analysis using simple examples, and then we describe cluster analysis and discuss the differences and similarities between these two analyses. We first consider continuous data, and then we discuss how to treat discrete or categorical data.

4.2 Discriminant Analysis

In discriminant analysis, we wish to decide which population an observation in a sample comes from, if the possible populations are known in advance. For example, suppose that we know in advance that some customers are "good" and some customers are "bad" (so the two possible populations, good customers and bad customers, are known in advance). Given a randomly selected customer, we wish to determine if he/she is a good customer or not based on his/her records (data) such as credit history (x_1), education (x_2), and income (x_3). Such a discriminant analysis may be very useful for (say) credit card applicants so that good applicants will receive credit cards while bad applicants do not receive credit cards. For the convenience of statistical analysis, sometimes we may assume that the sample (x_1, x_2, x_3) (or its transformations, e.g., $(\log(x_1), \log(x_2), \log(x_3))$), follows a multivariate normal distribution, but some discriminant analysis methods do not require this assumption.

As another example, suppose that a company reviews job applicants based on their academic records (x_1), education (x_2), working experience (x_3), self confidence (x_4), and motivation (x_5). All job applicants can be classified as either "suitable" or "not suitable" based on the given information. So the company may perform a discriminant analysis to separate suitable applicants from unsuitable applicants based on the data all applicants provide.

More generally, suppose that there are two multivariate normally distributed populations, denoted by

$$\text{population } \pi_1 : N_p(\boldsymbol{\mu}_1, \Sigma_1), \qquad \text{population } \pi_2 : N_p(\boldsymbol{\mu}_2, \Sigma_2).$$

Given an observation $\mathbf{x} = (x_1, x_2, \cdots, x_p)^{\mathrm{T}}$ from a sample, we want to find out whether $\mathbf{x}$ is from population π_1 or population π_2. This is a discriminant analysis. Note that, if $p = 1$ (i.e., if there is only one variable of interest), then it is very easy to separate the observations. All we need to do is to decide a threshold value, say K, so that we can do the separation based on whether $x \leqslant K$ or $x > K$. For example, if we just wish to separate students based on their grades, it is easy to see which students are good and which are not, such as the ones with grades over 80% and the ones with grades less than 80% (so $K = 80$). However, when $p \geqslant 2$, it is less straightforward to separate the observations. For example, if we wish to separate students based on their grades and their music skills, then it may be hard to do the separation since a student may have very good grades but poor music skills. In this case, we need more advance statistical methods to do the separation.

There are many methods available for discriminant analysis. For example, the following two methods are simple and useful ones:

- *Likelihood method*: we may choose population π_1 if the likelihood for π_1 is larger

than the likelihood for π_2, or vice versa. This method requires distributional assumption, such as multivariate normal distributions.

- *Mahalanobis distance method*: we may consider the Mahalanobis distances between an observation $\mathbf{x}$ and the population mean $\boldsymbol{\mu}_i$:

$$d_i = \sqrt{(\mathbf{x} - \boldsymbol{\mu}_i)'\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)}, \qquad i = 1, 2,$$

assuming $\Sigma_1 = \Sigma_2 = \Sigma$. We can then choose population π_1 if $d_1 < d_2$, or vice versa. This method does not require distributional assumption.

Both methods are intuitive, simple, and useful in practice.

For the likelihood method, given an observation $\mathbf{x}$, if the likelihood for population π_1 is larger than the likelihood for population π_2, the observation $\mathbf{x}$ is more likely from population π_1 than from population π_2, or vice versa. **Mahalanobis distance is commonly used in multivariate data to measure the "distance" between two multivariate observations, incorporating the variances and correlations of the variables.** That is, a variable with a larger variance receives less weight than a variable with a smaller variance. Moreover, the correlation between two variables affects the "distance" of the two variables. Thus, if an observation $\mathbf{x}$ is closer to the center of population π_1 than to the center of population π_2, the observation may be more likely from population π_1.

Note that the likelihood method requires distributional assumption while the Mahalanobis distance method does not. However, the Mahalanobis distance method assumes that the two populations have the same covariance matrices while the likelihood method does not require such assumption. Thus, each method has its advantages and limitations. In addition, the likelihood method is based on the "likelihood" or "probability" of an observation belonging to a population, while the Mahalanobis distance method is based on the "distance" of an observation to a population center, the basic ideas behind the two methods are different. Certainly, no method can give 100% correct results or correct classifications. In other words, for any methods, mis-classification is inevitable. The performance of a discriminant method can be evaluated by the probability of correct classification, i.e., the probability that an observation is correctly classified into the right group/population.

4.2.1 Canonical discriminant analysis (Fisher's method)

Another commonly used method for discriminant analysis is called the *canonical discriminant analysis* (*Fisher's method*). The basic idea of canonical discriminant analysis is to create new variables, called *canonical variables*, by taking special linear combinations of the original variables so that the new variables contain all useful information. This idea is similar to principal components analysis and factor analysis.

Specifically, suppose that we have m populations denoted by

$$\pi_i : N_p(\boldsymbol{\mu}_i, \Sigma), \qquad i = 1, 2, \cdots, m,$$

with the same covariance matrix Σ. Suppose also that a sample $\mathbf{x}_i = (\mathbf{x}_{i1}, \cdots, \mathbf{x}_{in_i})$ of size n_i is obtained from population π_i, where $\mathbf{x}_{ij} = (x_{ij1}, \cdots, x_{ijp})^{\mathrm{T}}$ is the j-th observation of random vector $\mathbf{x}_i$ in the i-th sample and each observation has p measurements (i.e., p variables). Let

$$\hat{\boldsymbol{\mu}}_i = \bar{\mathbf{x}}_i = \sum_{j=1}^{n_i} \mathbf{x}_{ij}/n_i, \quad \bar{\mathbf{x}} = \sum_{i=1}^{m} n_i \hat{\mu}_i / n$$

be the sample mean for sample $\mathbf{x}_i$ and the total mean respectively, where $n = \sum_{i=1}^{n} n_i$ is the total sample size. The *between-sample mean sum of squares* and the *within-sample sum of squares* are given respectively by

$$B = \sum_{i=1}^{m} n_i (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^{\mathrm{T}}, \qquad W = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)^{\mathrm{T}},$$

which measure the between-sample variation and the within-sample variation respectively. The canonical discriminant analysis proceeds as follows.

Consider the matrix $(\boldsymbol{B}+\boldsymbol{W})^{-1}\boldsymbol{B}$, which measures the proportion of the between-sample variation to the total variation. Let l_1 be the *largest* eigenvalue of matrix $(\boldsymbol{B} + \boldsymbol{W})^{-1}\boldsymbol{B}$, with the corresponding eigenvector $\mathbf{b}_1$. Then, for a given observation $\mathbf{x} = (x_1, x_2, \cdots, x_p)^{\mathrm{T}}$, the *first canonical function*, defined as

$$y_1 = \mathbf{b}_1^{\mathrm{T}} \mathbf{x},$$

is the linear discriminant function that provides the maximum separation between the mean vectors $\hat{\boldsymbol{\mu}}_1, \cdots, \hat{\boldsymbol{\mu}}_m$. If all means lie on a straight line, we may compute the m distances

$$d_i = |\mathbf{b}_1^{\mathrm{T}} \mathbf{x} - \mathbf{b}_1^{\mathrm{T}} \hat{\boldsymbol{\mu}}_i|, \qquad i = 1, \cdots, m,$$

and assign observation $\mathbf{x}$ to population π_k such that the distance d_k is the *smallest* among all m distances $\{d_1, d_2, \cdots, d_m\}$.

If the mean vectors do not lie on a straight line but on a plane (i.e., a two-dimensional space), we need the first two mutually orthogonal canonical discriminant functions

$$y_1 = \mathbf{b}_1^{\mathrm{T}} \mathbf{x}, \qquad y_2 = \mathbf{b}_2^{\mathrm{T}} \mathbf{x},$$

where $\mathbf{b}_2$ is the eigenvector of matrix $(\boldsymbol{B}+\boldsymbol{W})^{-1}\boldsymbol{B}$ corresponding to its second largest eigenvalue. Then we compute the m distances in the 2-dimensional space

$$d_i^2 = (\mathbf{b}_1^{\mathrm{T}}\mathbf{x} - \mathbf{b}_1^{\mathrm{T}}\hat{\boldsymbol{\mu}}_i)^2 + (\mathbf{b}_2^{\mathrm{T}}\mathbf{x} - \mathbf{b}_2^{\mathrm{T}}\hat{\boldsymbol{\mu}}_i)^2, \qquad i = 1, 2, \cdots, m,$$

and assign observation $\mathbf{x}$ to population π_k such that d_k^2 is the smallest among all m distances $\{d_1^2, d_2^2, \cdots, d_m^2\}$.

If necessary, we can compute the 3rd or 4th canonical functions, although in practice usually the first or the first two canonical functions are sufficient. Note that, to determine the dimensionality of the canonical space, we can check the eigenvalues of matrix $(\boldsymbol{B}+\boldsymbol{W})^{-1}\boldsymbol{B}$ and then use a method similar to PCA to decide the appropriate dimension of the canonical space. Often, one or two dimensions are preferable because we can visualize them, as illustrated in the examples later.

The foregoing canonical discriminant analysis uses ideas similar to principal component analysis and factor analysis in the sense that it is also based on linear combinations of the original variables. The matrix $(\boldsymbol{B}+\boldsymbol{W})^{-1}\boldsymbol{B}$ may be viewed as the ratio of the between-sample variation to the total variation, similar to the idea used in analysis of variance (ANOVA).

4.2.2 Discriminant analysis for categorical data

The discriminant analysis methods discussed so far assume that all the variables or data are *continuous*. When some variables or data are *categorical* or *discrete*, the above methods cannot be used since the means and covariance matrices are no longer meaningful for categorical variables or data. When some variables are categorical, a simple approach is to use logistic regression models for discriminant analysis, as illustrated below. Note that a logistic regression model is a generalized linear model, which will be described in details in Chapter 9.

As an example, consider the case of two populations. Let $y = 1$ if an observation $\mathbf{x} = (x_1, x_2, \cdots, x_p)^{\mathrm{T}}$ is from population 1 and $y = 0$ if the observation is from population 2. Then, we can consider the following logistic regression model

$$\log \frac{P(y=1)}{1-P(y=1)} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p,$$

where some x_j's may be categorical and some may be continuous. Given data, the above logistic regression model can be used to fit the data. Then, we can estimate the probability $P(y = 1)$ based on the fitted logistic regression model. For observation $\mathbf{x}_i$, if the estimated probability $\widehat{P}(y_i = 1) > 0.5$, observation $\mathbf{x}_i$ is more likely from population 1; otherwise it is more likely from population 2. This method can be extended to more than two populations.

4.3 Cluster Analysis

The goal of a cluster analysis is to identify *homogeneous groups* in the data by grouping observations based on their similarities or dis-similarities, i.e., similar observations are assigned to the same group. In other words, the idea is to partition all observations into subgroups or *clusters* (or populations), so that observations in the same cluster have similar characteristics.

For example, a marketing professional may want to partition all consumers into subgroups or clusters so that consumers in the same subgroup or cluster have similar buying habits. The partition can be based on age (x_1), education (x_2), income (x_3), and monthly payments (x_4). This is an example of cluster analysis based on four variables. The marketing professional may then design special advertisement strategies for different groups/clusters of consumers. Note that here the number of subgroups or clusters is not known before the cluster analysis.

Cluster analysis is similar to discriminant analysis in the sense that both methods try to separate observations into different groups. However, in discriminant analysis, the number of clusters (or subgroups or populations) are *known* in advance, and the objective is to determine which cluster (or population) an observation is likely to come from. In cluster analysis, on the other hand, the number of clusters (or subgroups or populations) is *not* known in advance, and the objective is find out distinct clusters and determine which cluster an observation is likely to come from. In other words, in cluster analysis one needs to find out how many clusters there may be and determine the cluster membership of an observation. Therefore, statistical methods for discriminant analysis may not be directly used in cluster analysis.

In cluster analysis, our objective is to devise a classification scheme, i.e., we need to find a rule to measure the similarity or dissimilarity between any two observations so that similar observations are grouped together to form clusters. Specifically, let $\mathbf{x} = (x_1, x_2, \cdots, x_p)^{\mathrm{T}}$ and $\mathbf{x}^* = (x_1^*, x_2^*, \cdots, x_p^*)^{\mathrm{T}}$ be two observations. The similarity between them can be measured by the "distance" between them, so that the two observations are similar if the distance between them is small. For example, the *Euclidean distance* between $\mathbf{x}$ and $\mathbf{x}^*$ is defined as

$$d_0(\mathbf{x}, \mathbf{x}^*) = \sqrt{(\mathbf{x} - \mathbf{x}^*)^{\mathrm{T}}(\mathbf{x} - \mathbf{x}^*)} = \sqrt{\sum_{j=1}^{p} (x_j - x_j^*)^2}.$$

However, the Euclidean distance does not take into account the *variations* and the *correlations* of the component variables. For multivariate data, each individual component variable has its own variance and the component variables may be correlated. Therefore, a better measure of the "distance" between two multivariate observations $\mathbf{x}$ and $\mathbf{x}^*$ is the *Mahalanobis distance*:

$$d(\mathbf{x}, \mathbf{x}^*) = \sqrt{(\mathbf{x} - \mathbf{x}^*)^{\mathrm{T}} \Sigma^{-1} (\mathbf{x} - \mathbf{x}^*)},$$

where $\Sigma = Cov(\mathbf{x}) = Cov(\mathbf{x}^*)$ is the covariance matrix, assuming the two observations have the same covariance matrices. Thus, if the distance $d(\mathbf{x}, \mathbf{x}^*)$ is small, we can consider that $\mathbf{x}$ and $\mathbf{x}^*$ are "close" and put them in the same cluster/group. Otherwise, we can put them in different clusters/groups. For a sample of n observations, $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$, each observation can be assigned into one of the clusters, based on a clustering method.

There are many cluster analysis methods available. In the following, we briefly discuss a few commonly used methods: nearest neighbour method, k-means algorithm, and two hierarchical cluster methods. Each method has its advantages and disadvantages, and different methods may lead to different results. In practice, it is always desirable to try at least two methods to analyze a dataset to see if the results agree or how much the results differ (which may provide some insights about the data).

4.3.1 Nearest neighbor method

A commonly used clustering method is called the *nearest neighbour method*. It begins with n clusters where each cluster contains only one observation, and then it keeps combining clusters based on the "distance" between them. Specifically, the nearest neighbour method is as follows:

1. Start with n clusters where each cluster contains only one point/observation.
2. Combine two closest points based on the Mahalanobis distance between them.
3. Define the distance between a point and a cluster to be the minimum distance between this point and all points in the cluster.
4. Continue combining clusters that are closest to one another until all observations are assigned a cluster.

Thus, the nearest neighbor method is quite intuitive, but it is not always clear how many final clusters there should be since sometimes the distinction between clusters may not be obvious. Moreover, the Mahalanobis distance may not always be the best measure of distance between observations in some applications.

4.3.2 The k-means algorithm

Another commonly used method is called the *k-means algorithm*. The k-means algorithm assigns each observation to the cluster whose center is nearest, where the *center* of a cluster is defined as the average of all the points in the cluster. For example, consider a cluster with two observations $\mathbf{x} = (x_1, x_2, x_3)$ and $\mathbf{y} = (y_1, y_2, y_3)$. Then the center of the cluster is given by

$$\mathbf{z} = ((x_1 + y_1)/2,\ (x_2 + y_2)/2,\ (x_3 + y_3)/2).$$

The k-means algorithm proceeds as follows:

1. Choose the number of clusters k ($1 \leqslant k \leqslant n$).

2. Randomly generate k clusters and determine the cluster centers, or directly generate k random points as cluster centers.

3. Assign each observation to the nearest cluster center.

4. Recompute the new cluster centers.

5. Repeat the two previous steps, until some convergence criterion is met.

A main advantage of the k-means algorithm is its simplicity and speed, which allow the algorithm to run on large datasets. Its disadvantage is that it does not yield the same result with each run, since the resulting clusters depend on the initial random assignments.

4.3.3 Hierarchical cluster methods

Hierarchical cluster analyses are also commonly used, which consist of either a series of successive mergers or a series of successive divisions. In the following we briefly describe a few of hierarchical methods.

An *agglomerative hierarchical method* starts with the individual observations. It keeps combining the most similar groups based on their similarities. Eventually, as the similarity decreases, all subgroups are merged into a single cluster. Specifically, an agglomerative hierarchical clustering algorithm can be described as follows:

1. Start with n clusters, with each cluster containing only one observation, and an $n \times n$ symmetric matrix of distances (or similarities) between any two observations.

2. Merge the nearest pair of clusters based on the distance matrix.

3. Update the distance matrix by computing the distances between the newly formed clusters and the remaining clusters based on a given criteria.

4. Repeat steps 2 and 3 for $n - 1$ times. Record the identities of clusters that are merged and the levels at which the mergers take place.

The distance between clusters can be defined using three methods: *single linkage, complete linkage*, and *average linkage*. The single linkage method defines the distance between two clusters as the minimum distance between two observations, one from each cluster. The complete linkage method defines the distance between two clusters as the distance between the two observations, one from each cluster, that are most distant. The distance in the average linkage method is defined as the average distance between all pairs of observations with one observation from a cluster.

Division hierarchical methods work in the opposite direction: an initial single group including all points is divided into two subgroups such that the points in one subgroup are "far" from the points in the other subgroup. These subgroups are then further divided into dissimilar subgroups; the process continues until there are as many clusters as points.

The *Ward's hierarchical clustering method* is another popular technique whose main idea lies in merging the groups at the minimum loss of information.

The results of hierarchical clustering methods may be displayed in a dendrogram, which is flexible in applications. We will demonstrate these methods in Example 3.

There are other cluster analysis methods. Different methods may give different results. Even for the same method, different run of the method may lead to different results, as the k-mean algorithm. Like other statistical methods, each cluster analysis method has its advantages and limitations. There is no single method that works the best all the times. In practice, it is suggested that we try different methods to analyze the same dataset. If the results from different methods are the same, we have a high confidence in the conclusions and the conclusions are likely to be reliable. If the results from different methods are different, they may give us some insights about the dataset and suggest that we should perform a further analysis.

4.4 Examples in R

Example 1. We consider a discriminant analysis for the famous Fisher's iris dataset. These data consist of 50 samples from each of three varieties of iris plants: Setosa (IS), Versicolor (IC), Virginica (IV). The variables measured on each plant are sepal length (SL), sepal width (SW), petal length (PL), and petal width (PW). Based on the data from these measurements, we do a discriminant analysis to separate these three plants.

```
> library(MASS)
> data(iris)  # it's a R built-in dataset
> attach(iris)
# part of the data
> iris[1:3,]
        Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1          3.5          1.4          0.2  setosa
2          4.9          3.0          1.4          0.2  setosa
3          4.7          3.2          1.3          0.2  setosa
.......
# Pairwise scatterplots
> library(car)
> scatterplotMatrix(~ Sepal.Length + Sepal.Width + Petal.Length
    + Petal.Width | Species,
      data=iris, smooth=FALSE, reg.line=FALSE, ellipse=TRUE,
      levels=0.95, by.groups=TRUE, diagonal="none")
> iris.Sp<-rep(c("s","c","v"), rep(50,3)) # label the three species
# linear discriminant analysis
> iris.lda<-lda(iris[,-5],factor(iris$Species))
> iris.lda
Call:
lda(iris[, -5], factor(iris$Species))
```

```
Prior probabilities of groups:
    setosa versicolor  virginica
 0.3333333  0.3333333  0.3333333

Group means:
           Sepal.Length Sepal.Width Petal.Length Petal.Width
setosa            5.006       3.428        1.462       0.246
versicolor        5.936       2.770        4.260       1.326
virginica         6.588       2.974        5.552       2.026

Coefficients of linear discriminants:
                    LD1         LD2
Sepal.Length  0.8293776  0.02410215
Sepal.Width   1.5344731  2.16452123
Petal.Length -2.2012117 -0.93192121
Petal.Width  -2.8104603  2.83918785

Proportion of trace:
   LD1    LD2
0.9912 0.0088
> z1<-predict(iris.lda,dim=2)$x  #prediction based on two-dimension
    discriminants
> eqscplot(z1, type="n", xlab="first linear discriminant",
    ylab="second linear discriminant", main="Iris Data") # figure to
        visualize
> text(z1,labels=as.character(iris.Sp))
> z2<-predict(iris.lda,dim=2)$class
# Compute the misclassification numbers
> table(z2,Species)
            Species
  z2           setosa versicolor virginica
  setosa         50          0         0
  versicolor      0         48         1
  virginica       0          2        49
```

Figure 4.1 shows pairwise scatterplots of the four measurements classified by species, with 95% concentration ellipses. Figure 4.2 shows results for linear discriminant analysis based on all four measurements. From Figure 4.1, we see that two-dimensional classifications may not be very accurate. For example, the classifications based on Sepal length and Sepal width may not be accurate since the ellipses for versicolor and virginica overlap, but the classifications based on Petal length and Petal width may be better. The linear discriminant analysis gives better classifications. From Figure 4.2, we see that the first linear discriminant provides the most separation between the three species. The first two linear discriminants separate most species correctly. The misclassification rates are quite low: Setosa's are all correctly classified, while 2 of Versicolor is classified as virginica and 1 of Virginica is classified as Versicolor. Overall, the discriminant method performs quite well.

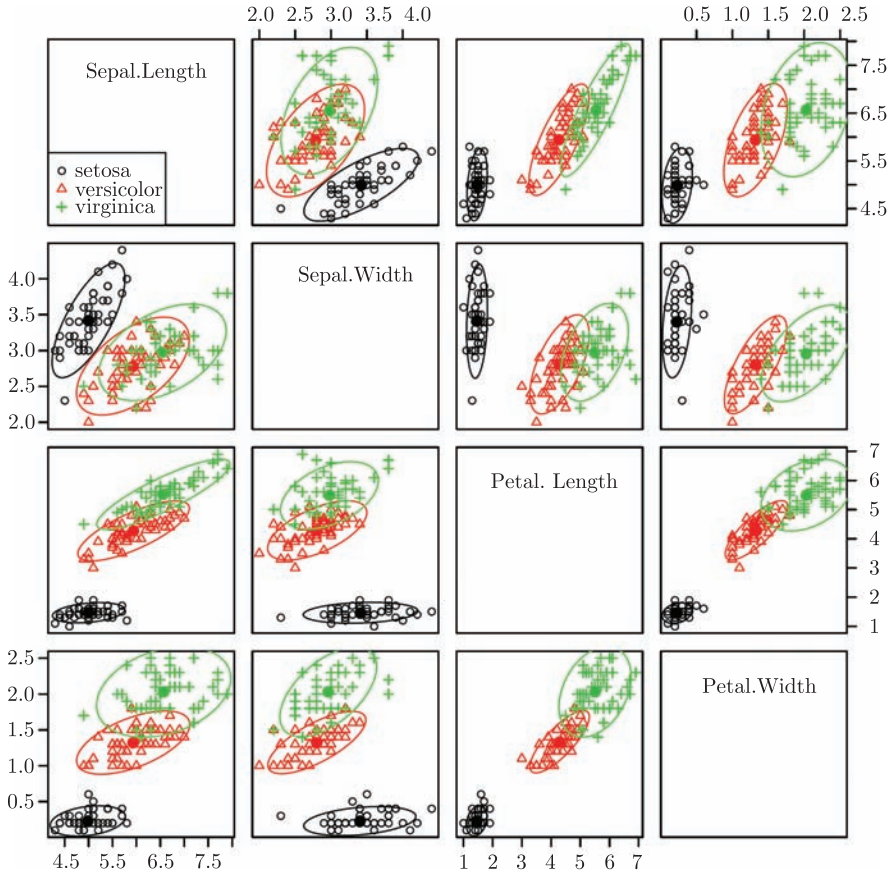


Figure 4.1 Pairwise scatterplots for the Iris data, with 95% concentration ellipses.

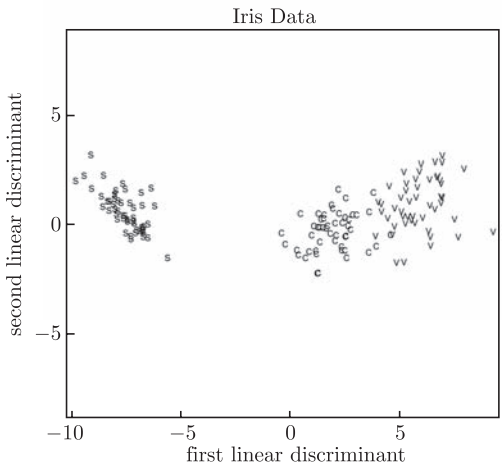


Figure 4.2 Discriminant analysis for the Iris dataset.

Example 2. We consider a cluster analysis for the job applicants data described in previous chapters to separate all the applicants. We may wish to separate the good applicants from the rest applicants. The rest applicants may be called bad ones (two clusters) or reasonable and bad ones (three clusters). Then, we may invite the good applicants for a second interview and reject the bad ones immediately. The corresponding cluster analysis is shown below.

```
> applican.dat0 <- read.table("applicant.dat", head=T)  # import data
> applican.dat1 <- applican.dat0[,-1] # remove ID
> attach(applican.dat1)
> options(digits=2)
# Let's look at part of the data
> applican.dat1
   FL APP AA LA SC LC HON SMS EXP DRV AMB GSP POT KJ SUIT
1   6   7  2  5  8  7   8   8   3   8   9   7   5  7   10
2   9  10  5  8 10  9   9  10   5   9   9   8   8  8   10
3   7   8  3  6  9  8   9   7   4   9   9   8   6  8   10
4   5   6  8  5  6  5   9   2   8   4   5   8   7  6    5
5   6   8  8  8  4  4   9   5   8   5   5   8   8  7    7
6   7   7  7  6  8  7  10   5   9   6   5   8   6  6    6
7   9   9  8  8  8  8   8   8  10   8  10   8   9  8   10
........

# First, we use the k-means method for cluster analysis with 2 clusters.
> app.km1 <- kmeans(applican.dat1, 2)
> app.km1
K-means clustering with 2 clusters of sizes 23, 25

Cluster means:
   FL APP  AA  LA  SC  LC HON SMS EXP DRV AMB GSP POT  KJ SUIT
1 5.3   6 7.0 5.2 5.2 4.0 7.4 1.9 3.2 3.0 3.5 4.0 3.3 4.3  4.1
2 6.6   8 7.2 7.0 8.6 8.4 8.6 7.6 5.2 7.5 8.3 8.4 7.9 6.7  7.7

Clustering vector:
 [1] 2 2 2 1 2 2 2 2 2 2 2 2 1 1 1 2 2 1 1 2 2 2 2 2 1 1 2 1 1 1 1 1 1 1
     1 1 2 2
[39] 2 2 1 1 1 2 1 2 1 1
```

We see that 23 applicants are "bad" ones, and 25 applicants are "good" ones. For example, the first three applicants are "good", and the 4th person is "bad", the fifth to the 12th applicants are "good", etc.

Next, we do cluster analysis with 3 clusters: "good", "reasonable", and "bad".

```
> app.km2 <- kmeans(applican.dat1, 3)
> app.km2
K-means clustering with 3 clusters of sizes 21, 9, 18

Cluster means:
```

```
   FL APP  AA  LA  SC  LC HON  SMS EXP DRV AMB GSP  POT  KJ SUIT
1 5.5 7.3 7.0 7.0 6.9 6.4 8.9 3.81 3.4 3.9 5.3 6.4 5.57 5.3  4.5
2 4.7 4.9 7.1 2.8 3.2 1.3 5.4 0.44 3.1 2.0 1.9 1.3 0.67 2.6  3.7
3 7.3 7.9 7.2 6.9 8.9 8.7 8.4 8.28 5.7 8.6 8.8 8.6 8.33 7.4  8.8

Clustering vector:
 [1] 3 3 3 1 1 1 3 3 3 3 3 3 1 1 1 3 3 1 1 3 1 3 3 3 1 1 1 2 2 1 1 1 1 2
     2 1 1 1
[39] 3 3 2 2 2 3 1 1 2 2
```

We see that 21 are "reasonable", 9 are "bad", and 18 are "good". For example, the first three are "good", the second three are "reasonable", and the last two are "bad". Thus, depending on how we would like to separate the applications, the results may differ.

Example 3. We consider a cluster analysis for the Chinese consumption data to separate 31 regions into groups with similar consumption levels.

```
> consum.1<-read.table("consum2007.txt",head =T)
> consum<-consum.1[,1:8]
> km.con1<-kmeans(consum, 2)
> km.con1
K-means clustering with 2 clusters of sizes 25, 6

Cluster means:
  Food Cloth Resid HousF Health TranC Educ Miscel
1 3171   983   843   507    619   976 1036    309
2 4926  1172  1325   811    905  2306 2043    529

Clustering vector:
  Beijing   Tianjin     Hebei   Shan1xi Neimenggu  Liaoning     Jilin
        2         2         1         1         1         1         1
 Heilongj  Shanghai   Jiangsu  Zhejiang     Anhui    Fujian   Jiangxi
        1         2         1         2         1         2         1
 Shandong     Henan     Hubei     Hunan Guangdong   Guangxi    Hainan
        1         1         1         1         2         1         1
Chongqing   Sichuan   Guizhou    Yunnan    Xizang   Shan3xi     Gansu
        1         1         1         1         1         1         1
  Qinghai   Ningxia  Xinjiang
        1         1         1
```

The above cluster analysis Separates the 31 regions to 2 clusters. Cluster 2 includes Beijing, Tianjin, Shanghai, Zhejiang, Fujian, and Guangdong, which are the regions with relatively higher consumption levels. The other regions belong to Cluster 1. This result is similar to the visual classification using a star-plot in Chapter 1.

Next, we use a hierarchical clustering method to see if we get the same results or not.

```
> hc.con1<-hclust(dist(consum),"ward") # use agglomeration method "ward"
```

```
> plot(hc.con1)
```

A dendrogram can help us to identify clusters easily. From Figure 4.3, we can separate the regions into three meaningful groups, which are defined as "high", "low", and "medium" groups. Each group can be further divided into even smaller groups. We can choose two groups, three groups, or even more groups if necessary.

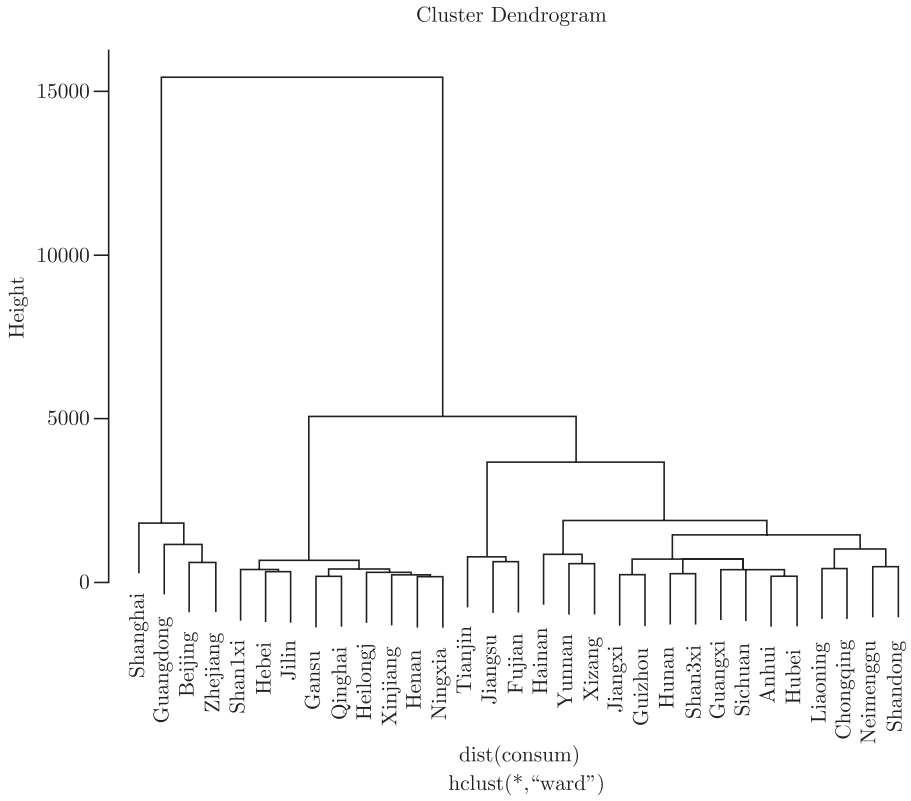


Figure 4.3 Cluster Dendrogram for Consumption Data.

Exercises 4

4.1. Derive the Fisher's canonical discriminant.

4.2. Consider the iris data.

(a) Draw boxplots of the four variables for different species to check the differences between groups.

(b) Carry out a discrimination analysis and a clustering analysis based only on the sizes of the sepal, i.e., Sepal.Length and Sepal.Width. Compare the results to those obtained in Examples 1.

(c) Darroch and Mosimann (1985) think that the iris can be characterized by its shape. They suggested using Y_1 = Sepal.length/Sepal.width and Y_2 = Petal.length/Petal.width to describe the shape of the sepal and petal respectively. Do a discrimination analysis and a cluster analysis based on $\log Y_1$ and $\log Y_2$. Compare the results to those obtained in Examples 1.

4.3. For the Chinese consumption data, perform a cluster analysis for data "consum2007" based on correlations.

4.4. For the Chinese consumption data, perform a cluster analysis for the 31 regions based on data "consum2000" and "consum2010" respectively. Compare the results to those obtained based on data "consum2007".

4.5. Perform a cluster analysis for the 6 stocks in Example 3 in Chapter 2.

4.6. Perform a cluster analysis for the 10 track and field events in Exercise 2.6.

4.7. Perform a cluster analysis for the 31 regions in China based on the pollution data described in Exercise 8.2.

Chapter 5

Inference for a Multivariate Normal Population

5.1 Introduction

In previous chapters, we have mostly focused on multivariate exploratory analyses, without statistical inference. That is, our focus is mostly on finding important descriptive features of data, such as data dimensions and data classifications, without extending the results or conclusions to the whole population. Distributional assumptions for the data are often not required for such exploratory analysis. Exploratory data analysis is an important component of statistical analysis. However, exploratory analysis is not sufficient. Often we wish to extend the conclusions from exploratory analysis to the whole population. That is, we also wish to make statistical inference, such as hypothesis testing or confidence intervals.

In this chapter, we consider statistical inference for multivariate continuous data, i.e., we try to extend the results from the data to the whole population. In order to make inference, the following *assumptions* are often required: (i) the sample is a random and representative subset of the population, e.g., an *i.i.d.* (independent and identically distributed) sample from the population, and (ii) the population is assumed to follow a parametric distribution, such as a multivariate normal distribution. In this chapter, we focus on the most important distribution for multivariate continuous data, i.e., the multivariate normal distribution. A multivariate normal distribution, denoted by $N(\boldsymbol{\mu}, \Sigma)$, is completely determined by its mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_p)^{\mathrm{T}}$ and its covariance matrix $\Sigma = (\sigma_{ij})_{p\times p}$.

Inference for a multivariate normal distribution is often the main focus of many classic multivariate analysis books. Many nice features of the multivariate normal distribution allow us to do theoretical arguments and derive many elegant results. In practice, a multivariate normal distributional assumption may be reasonable for continuous data in many cases (perhaps after some transformations of the original data), especially when the sample size is large so the central limit theorems can be used to justify the normality assumption for the sample means. In some other cases in practice, such as small sample sizes or skewed data or discrete data, the multivariate

normal assumption may be too strong or unreasonable. In real data analysis, when we use a method or model which require normality assumption, we should check to see if this assumption is indeed reasonable.

In the following sections, we consider inference for both the mean vector and the covariance matrix of a multivariate normal distribution. Since theoretical derivations of these results are available in many classic books and our focus is to use these methods in practice, we skip the theoretical derivations and focus on explaining the methods and their applicability in data analysis. Readers interested in mathematical derivations are referred to many classic multivariate analysis books (e.g., Johnson and Wichern 2007).

5.2 Inference for Multivariate Means

We first review the well-known t-test for inference of a univariate mean. In *univariate* analysis, the t-test is widely used to test the mean parameter assumed for a continuous variable. Suppose that $\{x_1, x_2, \cdots, x_n\}$ is a random sample from a univariate normal population $N(\mu, \sigma^2)$. Consider a two-sided test for the mean parameter μ:

$$H_0 : \mu = \mu_0 \qquad \text{versus} \qquad H_1 : \mu \neq \mu_0,$$

where μ_0 is known. Let the sample mean and the sample standard deviation be $\bar{x}$ and s respectively. The t test statistic is given by

$$t = \frac{\sqrt{n}(\bar{x} - \mu_0)}{s},$$

which is in fact a standardize version of the sample mean $\bar{x}$ under H_0. Under the null hypothesis H_0, the test statistic t follows a t-distribution with $n - 1$ degrees of freedom. An alternative and equivalent test statistic is given by

$$t^2 = \frac{n(\bar{x} - \mu_0)^2}{s^2},$$

which follows a $F(1, n-1)$-distribution. These tests are relatively robust against small to moderate departure from the assumed normality of the population, especially when the sample size is large. In fact, when the sample size is large, the sample mean $\bar{x}$ will approximately follow a normal distribution for any population distribution, based on the central limit theorem. In other words, t-tests can be used as long as the sample size is large, even if the population is not normally distributed.

The above univariate t-test can be extended to the *multivariate* case. For multivariate data, a key consideration is to incorporate the correlation or covariance between the variables. The most well-known extension is the so-called *Hotelling's T test*, which is described as follows. Let $\mathbf{x}_1, \cdots, \mathbf{x}_n$ be a random sample from the

multivariate normal population $N_p(\boldsymbol{\mu}, \Sigma)$, where $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})^{\mathrm{T}}$ and both $\boldsymbol{\mu}$ and Σ are unknown. Let

$$\bar{\mathbf{x}} = \frac{\sum_{i=1}^{n} \mathbf{x}_i}{n}, \quad \boldsymbol{S} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{T}}$$

be the sample mean vector and sample covariance matrix respectively. Suppose that we wish to test the following two-sided multivariate hypotheses

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \qquad \text{versus} \qquad H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0,$$

where $\boldsymbol{\mu}_0$ is a known vector. The Hotelling's T^2 test statistic is given by

$$T^2 = n(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^{\mathrm{T}} \boldsymbol{S}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}_0).$$

Under H_0, we have

$$T^{*2} = \frac{n-p}{p(n-1)} T^2 \sim F(p, n-p).$$

Thus, we can reject H_0 when $T^{*2} > F_\alpha(p, n-p)$, where $F_\alpha(p, n-p)$ is the upper α-th percentile of the $F(p, n-p)$ distribution, or a p-value can be computed based on this null distribution.

When the number of variables $p = 1$ (i.e., the univariate case), the Hotelling's T^2 statistic reduces to the above univariate test statistic t^2, i.e., $T^2 = t^2$ when $p = 1$. In other words, the Hotelling's T^2 test is equivalent to the univariate t-test when there is only one variable, or the Hotelling's T^2 test is an extension of the univariate t-test to multivariate data.

A *confidence region* for the mean vector $\boldsymbol{\mu}$ is an extension of a confidence interval for a scaler parameter μ to the multivariate case. It is a region in which the true parameter vector $\boldsymbol{\mu}$ is likely to fall. In general, confidence intervals/regions and hypothesis tests are closely related: **a $(1-\alpha)\times 100\%$ confidence interval/region can be obtained by inverting the acceptance region of a level α test**. Therefore, a confidence region for the mean vector $\boldsymbol{\mu}$ can be obtained by inverting the acceptance region of the Hotelling's T^2 test. Specifically, a $(1-\alpha) \times 100\%$ confidence region for $\boldsymbol{\mu}$ is given by

$$R(\mathbf{x}) = \left\{ \boldsymbol{\mu} : \; n(\bar{\mathbf{x}} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{S}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}) \leqslant \frac{(n-1)p}{n-p} F_\alpha(p, n-p) \right\}.$$

The confidence region $R(\mathbf{x})$ is an ellipsoid centered at the sample mean $\bar{\mathbf{x}}$. The axes of the confidence ellipsoid and their relative lengths are determined from the eigenvalues and eigenvectors of the matrix $\boldsymbol{S}$.

Figure 5.1 shows a confidence region for the mean vector $\boldsymbol{\mu}$ of a bivariate normal population $N_2(\boldsymbol{\mu}, \Sigma)$, with both $\boldsymbol{\mu}$ and Σ unknown. Suppose that we have obtained the following sample mean vector and sample covariance matrix

$$\bar{\mathbf{x}} = \begin{pmatrix} -1.8 \\ 1.2 \end{pmatrix}, \quad \boldsymbol{S} = \begin{pmatrix} 8.0 & -2.5 \\ -2.5 & 1.6 \end{pmatrix}$$

based on a sample of size 100 from the population $N_2(\boldsymbol{\mu}, \Sigma)$. The eigenvalues and the corresponding eigenvectors of the sample covariance matrix $\boldsymbol{S}$ are

$$\lambda_1 = 8.87,\ e_1 = (-0.94, 0.33)^{\mathrm{T}}, \quad \lambda_2 = 0.72,\ e_2 = (0.33, 0.94)^{\mathrm{T}},$$

respectively. In Figure 5.1, the larger ellipse is the 95% confidence region of $\boldsymbol{\mu}$, while the smaller one is the 50% confidence region of $\boldsymbol{\mu}$. The axes of the ellipses lie along the directions determined by the eigenvectors, passing through the center at $\bar{\mathbf{x}}$, with their lengths proportional to the square root of the corresponding eigenvalues. The interpretation of the (say) 95% confidence region is as follows: if we take many samples from the population $N_2(\boldsymbol{\mu}, \Sigma)$ and construct a 95% confidence region of $\boldsymbol{\mu}$ based on the above formula for each sample, roughly 95% of these confidence regions will cover the (unknown) true mean vector $\boldsymbol{\mu}$. Equivalently, if we test the hypotheses (say)

$$H_0 : \boldsymbol{\mu} = (-2, 1)^{\mathrm{T}} \quad \text{versus} \quad H_1 : \boldsymbol{\mu} \neq (-2, 1)^{\mathrm{T}}.$$

At 5% level, we would fail to reject H_0 (since the vector $(-2, 1)$ is covered by the 95% confidence region).

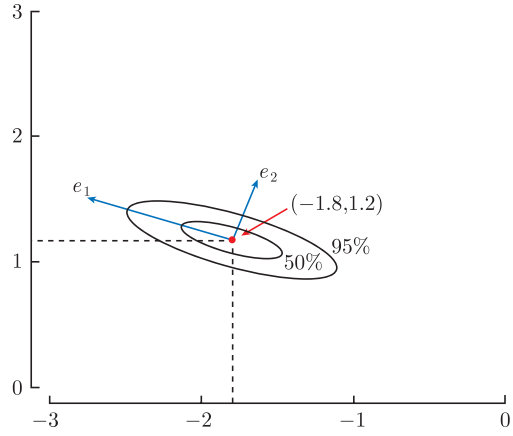


Figure 5.1 A 95% and 50% confidence regions for the mean vector $\boldsymbol{\mu}$.

The Hotelling's T^2 test and the corresponding confidence region can be extended for inference about *two* multivariate normal population mean vectors. Specifically, let

$\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ be a random sample from multivariate normal population $N_p(\boldsymbol{\mu}_1, \Sigma_1)$, where $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})^{\mathrm{T}}$, and let $\{\mathbf{y}_1, \cdots, \mathbf{y}_n\}$ be a random sample from the multivariate normal population $N_p(\boldsymbol{\mu}_2, \Sigma_2)$, where $\mathbf{y}_i = (y_{i1}, \cdots, y_{ip})^{\mathrm{T}}$, and both $\boldsymbol{\mu}_j$ and Σ_j are unknown, $j = 1, 2$. We may consider testing the equality of two multivariate mean vectors

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2, \qquad \text{versus} \qquad H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2.$$

Conceptually, the basic idea for testing the above hypotheses is similar to the one for the univariate case, where we assume equal variances (i.e., in the multivariate case we also assume equal covariance matrices $\Sigma_1 = \Sigma_2$). However, the mathematical expressions of the test statistic and its null distribution are somewhat more complicated (so they are omitted here). In the case of equal sample sizes of the two samples (as above), an alternative approach is to define a new variable $\mathbf{z}_i = \mathbf{x}_i - \mathbf{y}_i$, and then the two-sample test problem is reduced to the one-sample testing problem as described above.

The Hotelling's T test is used to test multivariate two-sided hypotheses. It can *not* be used to test multivariate one-sided hypothesis such as $H_0 : \boldsymbol{\mu} \geqslant \boldsymbol{\mu}_0$ versus $H_1 : \boldsymbol{\mu} < \boldsymbol{\mu}_0$. Multivariate one-sided tests are more complicated than the two-sided tests described above. For multivariate one-sided hypotheses, a main difficulty is that the null distribution of a test statistic is difficult to obtain due to the shape of the null parameter space. There is a large literature on multivariate one-sided or order-restricted testing problems. For space reason, we do not discuss them here. Interested readers are referred to Silvapulle and Sen (2004) for a detailed description.

5.3 Inference for Covariance Matrices

For a multivariate normal population $N_p(\boldsymbol{\mu}, \Sigma)$, inference for the covariance matrix Σ can also be performed. However, the computation associated with the test can be tedious since closed-form null distributions of test statistics are often unavailable, so computer software is needed for computation.

In practice, it is common to test the equality of two covariance matrices. For example, when testing two multivariate mean vectors, it is assumed that the two unknown covariance matrices are equal (as noted in the previous section). This assumption can be tested. Specifically, we can test the equality of two population covariance matrices

$$H_0 : \Sigma_1 = \Sigma_2 \quad \text{versus} \quad \Sigma_1 \neq \Sigma_2.$$

Let $\{\mathbf{x}_1, \cdots, \mathbf{x}_{n_1}\}$ be a random sample from population $N_p(\boldsymbol{\mu}_1, \Sigma_1)$, and let $\{\mathbf{y}_1, \cdots, \mathbf{y}_{n_2}\}$ be an independent random sample from population $N_p(\boldsymbol{\mu}_2, \Sigma_2)$. The test statistic is given by

$$\Lambda = \frac{|\hat{\Sigma}_1|^{(n_1-1)/2} |\hat{\Sigma}_2|^{(n_2-1)/2}}{|\hat{\Sigma}|^{(n-2)/2}},$$

where $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ are sample covariance matrices of Σ_1 and Σ_2 respectively,

$$\hat{\Sigma} = ((n_1 - 1)\hat{\Sigma}_1 + (n_2 - 1)\hat{\Sigma}_2)/(n - 2)$$

is the pooled estimate of the covariance matrix, and $n = n_1 + n_2$ is the total sample size. The null distribution of Λ is complicated, but computer software can be used to obtain p-values of the test.

5.4 Large Sample Inferences about a Population Mean Vector

When the sample size is large, statistical inference for multivariate means can be done without the assumption of normality, based on asymptotic theory. Specifically, let $\mathbf{x}_1, \cdots, \mathbf{x}_n$ be a random sample from a multivariate population with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ (the population does not need to follow a multivariate normal distribution). When $n - p$ is large, we have

$$P\{n(\bar{\mathbf{x}} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{S}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}) \leqslant \chi^2_\alpha(p)\} \approx 1 - \alpha.$$

Therefore, an approximately $(1 - \alpha) \times 100\%$ confidence region for $\boldsymbol{\mu}$ is given by

$$R(\mathbf{x}) = \left\{\boldsymbol{\mu} : \ n(\bar{\mathbf{x}} - \boldsymbol{\mu})' \boldsymbol{S}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}) \leqslant \chi^2_\alpha(p)\right\},$$

where $\chi^2_\alpha(p)$ is the upper αth percentile of the χ^2 distribution with p degrees of freedom. Similarly, for hypothesis testing, the null hypothesis $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ is rejected at level α if

$$n(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)' S^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}_0) > \chi^2_\alpha(p).$$

For multivariate data, the multivariate normality assumption often does not hold, even approximately. Therefore, it is desirable to have large samples so that asymptotic methods may be used for inference. When the sample size is small and the multivariate normality does not hold, we can use bootstrap methods for inference.

5.5 Examples in R

Example 1. Return to the quiz score dataset described in Chapter 1. Suppose that Quiz 1 and Quiz 2 were given before the midterm exam and Quiz 3 and Quiz 4 were given after the midterm exam. Suppose also that the instructor wishes to check if scores in Quiz 1 and Quiz 2 roughly have the same averages, and if scores in Quiz 3 and Quiz 4 roughly have the same averages, i.e., if students' performances in Quiz 1 and Quiz 2 are similar and students' performances in Quiz 3 and Quiz 4 are similar. Equivalently, we can check if the differences between Quiz 1 and Quiz 2 scores are different from zero and if the differences between Quiz 3 and Quiz 4 scores are different from zeros. More specifically, let x_{ij} be the score of student i in quiz

$j,\ i = 1, 2, \ldots, n = 53;\ j = 1, 2, 3, 4$. Define new variables $y_{i1} = x_{i1} - x_{i2},\ y_{i2} = x_{i3} - x_{i4}$, with $E(y_{i1}) = \mu_1$ and $E(y_{i2}) = \mu_2,\ i = 1, 2, \cdots, 53$. We assume that $\{(y_{i1}, y_{i2}), \cdots, (y_{in}, y_{in})\}$ is an i.i.d. sample from $N_2(\boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu} = (\mu_1, \mu_2)^{\mathrm{T}}$ and Σ are unknown. The hypotheses to be tested are

$$H_0 : \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \mathbf{0}, \qquad \text{versus} \qquad H_1 : \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \neq \mathbf{0}.$$

We can then consider a Hotelling's T^2 test, as illustrated in R below.

```
> dat <- read.table("class.dat2", head=T)[,-c(1,2,3)]  # use only quiz
    scores
## compute the differences between quiz1 and quiz2 scores, quiz3 and
    quiz4 scores
> y <- cbind(dat$quiz1-dat$quiz2, dat$quiz3-dat$quiz4)
> n <- dim(y)[1]   # sample size n
> p <- dim(y)[2]   # number of variables p
> y.bar <- colMeans(y)  # sample means of y=(y1,y2)
# The sample mean vector is given by
> y.bar
[1] -3.442308 -3.538462
> S.y <- var(y)  # sample covariance
# The sample covariance matrix is given by
> S.y
          [,1]       [,2]
[1,] 125.74170  47.87481 [2,]  47.87481 202.88084

# Compute Hotelling's T  statistic
> T.sq <- n*t(y.bar)%*%solve(S.y)%*%y.bar  # T statistic
> T.sq2 <- ((n-p)/(p*(n-1)))*T.sq    # T* statistic
# The value of the test statistic is given by
> T.sq2
        [,1]
[1,] 3.08726
> p-value <- 1-pf(T.sq2, p,n-p)  # p-value for the test
# The p-value for the Hotelling's T test is given by
> p-value
     [,1]
[1,] 0.054
```

The p-value for the Hotelling's T test is about 0.054, implying that there are some differences between quiz 1 and quiz 2 scores, and quiz 3 and quiz 4 scores. However, the differences are not statistically significant at 5% level, although they are statistically significant at 10% level.

Example 2. Based on some economic theory, the consumption patterns tend to change with the development of economics. For example, Engel's law states that as incomes rise, the proportions of incomes spent on food fall, even if actual expenditures on food rise. On the other hand, as incomes rise, consumers likely move to needs of a lower priority, such as education, health, etc. In this example, to check if consumption

patterns change as a country becomes richer, we compare the consumption structures in China between year 2000 and year 2010. The data come from China Statistical Yearbook.

To get a rough idea of the consumption patterns in these two years, we show boxplots in Figure 5.2, where the numbers 1 to 8 on the horizontal axis represent eight consumption categories: food (Food), clothing (Cloth), residence (Resid), household facilities, articles and services (HousF), health care and medical services (Health), transport and communication (TranC), education, culture and recreation (Educ), miscellaneous goods (Miscel). The boxplots show some patterns of change in consumption. Although consumption expenditure on food is still the majority, there is a systematic decrease in year 2010 compared to year 2000. The consumption expenditure on transport and communication increases substantially in year 2010. Although we see some differences in the values of the variables between 2000 and 2010, the boxplot in Figure 5.2 cannot tell us whether these differences are statistically significant or not. Thus, we need to perform a formal hypothesis testing to test these differences.

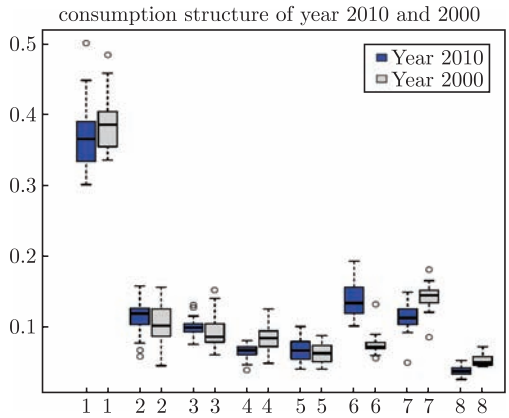


Figure 5.2 Boxplots for the Consumption Data. The numbers 1 to 8 on the horizontal axis represent eight consumption categories: food (Food), clothing (Cloth), residence (Resid), household facilities, articles and services (HousF), health care and medical services (Health), transport and communication (TranC), education, culture and recreation (Educ), miscellaneous goods (Miscel).

As an illustration, we select the five crucial variables: Food, Colth, Resid, TranC, and Educ, and perform a multivariate hypothesis test to test the differences between year 2000 and year 2010, using the Hotelling's T test. That is, we can test

$$H_0 : \boldsymbol{\mu} = 0 \quad \text{versus} \quad \boldsymbol{\mu} \neq 0,$$

where $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_5)^{\mathrm{T}}$ is the population means of the 5 variables, with the population covariance matrix Σ unknown.

```
> consum.2<-read.table("consum2010.txt",head =T)
> consum.3<-read.table("consum2000.txt",head =T)
> X<-consum.2[,1:8]/rowSums(consum.2[,1:8]) # compute the expenditure
    share
> Y<-consum.3[,1:8]/rowSums(consum.3[,1:8]) # compute the expenditure
    share
> XY.d<-X[,c(1,2,3,6,7)]-Y[,c(1,2,3,6,7)] # differences of 5 variables
    between 2 years
> d.mean<-colMeans(XY.d)  # sample means of the 5 variables
   Food    Cloth    Resid    TranC     Educ
-0.0213   0.0103   0.0055   0.0635 -0.0299
> d.S<-var(XY.d)  # sample covariance matrix
> n<-dim(XY.d)[1] # sample size
> p<-dim(XY.d)[2] # number of variables p=5
> T2<-n*t(d.mean)%*%solve(d.S)%*%d.mean   # Hotelling's T statistics
> pvalue<-1-pf((n-p)*T2/((n-1)*p),p,n-p)  # p-value
> pvalue
         [,1]
[1,] 7.5e-14
```

The p-value is near 0, indicating that there are highly significant differences in the consumption structures in China between year 2000 and 2010 for the five variables. In fact, we could perform a hypothesis test for each variable separately, i.e., testing $H_0 : \mu_j = 0$ versus $\mu_j \neq 0,\ j = 1, 2, 3, 4, 5$, separately, using the usual t-test. However, such univariate tests will be less powerful than the above multivariate test since the multivariate test incorporates the correlation between the 5 variables while the univariate tests do not.

Exercises 5

5.1. Show that, when $p = 1$, the Hotelling's T test reduces to the usual t-test in the univariate case.

5.2. When $p = 1$, write down the form of the confidence region $R(\mathbf{x})$.

5.3. Let $\mathbf{x}_1, \cdots, \mathbf{x}_n$ be a random sample from the multivariate population with mean vector $\boldsymbol{\mu}$ and positive definite covariance matrix Σ. Prove that, when $n - p$ is large,

$$P\{n(\bar{\mathbf{x}} - \boldsymbol{\mu})' \boldsymbol{S}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}) \leqslant \chi^2_\alpha(p)\} \approx 1 - \alpha,$$

where $\bar{\mathbf{x}}$ is the sample mean vector and S is the sample covariance matrix.

5.4. In Example 2, based on the 5 selected crucial consumption variables, we have checked that there exist significant differences in the consumption structures between year 2010 and 2000. Use all the consumption variables to do the Hotelling's T test again. Report what you find.

5.5. Use data consum2000 and consum2007 to check if there are any differences in the consumption structure between year 2000 and year 2007.

5.6. In Example 2, test data on each variable separately using univariate t tests. Report what you find.

Chapter 6

Discrete or Categorical Multivariate Data

6.1 Discrete or Categorical Data

In previous chapters, we have mostly focused on multivariate continuous data, such as quiz scores and stock prices. The exploratory analysis methods presented in previous chapters, such as the principal component analysis, factor analysis, and cluster analysis, are designed for multivariate continuous data. For multivariate continuous data, we have seen that (i) the most important summary statistics are means, variances, and covariances/correlations; (ii) the most common graphical tools are histograms, boxplots, and scatterplots; and (iii) the most popular distributional assumption is the multivariate normal distribution. In practice, however, many variables or data are discrete or categorical, such as gender and locations. Discrete or categorical variables/data are very different from continuous variables/data. For a discrete variable, its values represent different categories which may not have particular orders. For example, the value of variable "gender" can be "male" (1) or "female" (0). So the means, variances, and correlations cannot be defined for discrete data. Therefore, the standard summary statistics, graphical tools, and distributional assumption for continuous data cannot be used for discrete data since they are meaningless for discrete data.

For discrete data, the most important summary statistics are counts and percentages or proportions. For example, consider gender in a class of 50 students. We can summarize the data as follows: 30 (60%) students are male and 20 (40%) students are female. When summarizing discrete data, it is desirable to show both counts and percentages or proportions. We can measure the association between two discrete variables by cross-tabulating the data in a two-way contingency table. Similarly, we can assess the association between p discrete variables by a p-way contingency table. The χ^2 statistic computed from a contingency table can give some indication of the strength of the association. We can also use correlation-like measures such as the ϕ-coefficient and Cramer's V to describe the strength of relationship between several discrete variables. In the simplest case, data from two binary variables can be summarized by a 2×2 contingency table, and their association can be measured by *odds*

ratio.

In the analysis of categorical data, often it may be desirable to combine some categories in order to reduce the number of categories of a discrete variable. For example, the variable "smoking status" may have three categories: heavy smokers, light smokers, and non-smokers. In data analysis, sometimes it may be desirable to combine the first two categories so that the new variable "smoking status" only have two categories: smokers and non-smokers. Such approach may not only simplify analysis but may also make analysis results more reliable. This is because many statistical methods for discrete data, such as the χ^2 test for contingency tables, require that the counts in a contingency table to be reasonably large (say, at least 5). Combining categories may increase the counts, although some information may be lost.

In regression analysis, when the response is a binary variable, logistic regression models can be used. Logistic regression models belong to the class of generalized linear models, and they are widely used in practice. The standard logistic regression model can also be extended to response variables with more than two categories. Regression analysis for contingency tables can be based on log-linear models (Agresti, 2012).

Note that continuous data may be converted into discrete data by grouping the observations. For example, age is usually treated as a continuous variable, but it can be converted into a discrete variable, such as "old" and "young" (say, based on whether age is larger than 50 or not). Similarly, income may be converted into "high" or "low" based on some threshold value. Such a conversion can be convenient in data analysis, but usually lead to some loss of information. This type of discrete variable is often called ordinal categorical variables since there is an order of the values of the variables. In data analysis, sometimes it may be appropriate to treat ordinal discrete data as continuous data, although such an approach may not be ideal.

In this chapter, we briefly discuss statistical tools for analyzing multivariate discrete or categorical data. For a thorough discussion of categorical data analysis, readers are referred to Agresti (2012) or other similar books.

6.2 The Multinomial Distribution

For continuous multivariate data, the basic distributional assumption is the multivariate normal distribution. For discrete or categorical multivariate data, the basic distributional assumption is the *multinomial distribution*. The multinomial distribution is a generalization of the well-known binomial distribution. Note that, for either continuous data or discrete data, distributional assumptions are usually only required for statistical inference, but they may not be required for descriptive analysis or exploratory analysis.

The *binomial distribution*, denoted by $B(n, p)$, is usually assumed for a binary variable in statistical inference. A binary variable takes only two possible values, such as "success" or "failure", "yes" or "no" answers, usually coded as 1 or 0. Let the probability of "success" (or "1") be a constant p. For an i.i.d. sample of size n, the number of "successes" (or "1"'s) x follows a binomial distribution, with

$$P(x = j) = \left(\begin{array}{c} n \\ j \end{array} \right) p^j (1-p)^{n-j}, \qquad j = 0, 1, 2, \cdots, n,$$

where

$$\left(\begin{array}{c} n \\ j \end{array} \right) = \frac{n!}{j!(n-j)!},$$

and $n! = n(n-1)\cdots 2 \times 1$. For example, the number of heads in n tosses of a fair coin follows a binomial distribution $B(n, 0.5)$. The multinomial distribution is an extension of a binomial distribution in which each trial has k possible outcomes ($k \geqslant 2$). The distribution of n independent replicates on such trials follows a multinomial distribution. For example, the number of possible outcomes in n tosses of a "die" (which has $k = 6$ faces) follows a multinomial distribution.

Specifically, consider n independent trials in which each trial has k possible outcomes, with probability p_j for observing outcome j in each trial, $j = 1, 2, \cdots, k$. Let the random variable x_j be the number of times outcome j is observed in the n trials. Then, the random vector $\mathbf{x} = (x_1, \cdots, x_k)^{\mathrm{T}}$ follows a *multinomial distribution*, denoted by $M(n, \mathbf{p})$, with probability distribution given by

$$\begin{aligned} P(x_1 = n_1, x_2 = n_2, \cdots, x_k = n_k) &= \frac{n!}{n_1! n_2! \cdots n_k!} p_1^{n_1} \cdots p_k^{n_k}, \qquad (6.1) \\ 0 \leqslant n_j \leqslant n,\ 0 < p_j < 1,\ j = 1, \cdots, k; & \\ \sum_{j=1}^{k} n_j = n; \qquad \sum_{j=1}^{k} p_j = 1, & \end{aligned}$$

where $\mathbf{p} = (p_1, \cdots, p_k)^{\mathrm{T}}$. When $k = 2$, the multinomial distribution reduces to a binomial distribution. Note also that each x_j separately follows a binomial distribution $B(n, p_j)$.

For the multinomial distribution $M(n, \mathbf{p})$, the means, variances, and covariances are given by

$$\begin{aligned} E(x_j) = np_j, \qquad & var(x_j) = np_j(1-p_j), \\ cov(x_i, x_j) = -np_i p_j, \qquad & i \neq j, \ \ i, j = 1, \cdots, k. \end{aligned}$$

The covariance matrix of $\mathbf{x}$, whose (i, j)-th component is $cov(x_i, x_j)$ for $i \neq j$ and whose diagonal elements are $var(x_j)$'s, is a $k \times k$ positive semidefinite matrix of rank

$k-1$. The covariance matrix does not have full rank because there is a constrain $\sum_{j=1}^{k} p_j = 1$, which is not the case for multivariate normal distributions. Moreover, the variances and covariances share the same parameters as the means (i.e., parameters p_j's). The correlation between x_i and x_j is given by

$$\rho(x_i, x_j) = -\sqrt{\frac{p_i p_j}{(1-p_i)(1-p_j)}}, \qquad i, j = 1, \cdots, k.$$

The correlation between x_i and x_j is always negative because an increase in x_i leads to a decrease in x_j for any $i \neq j$.

6.3 Contingency Tables

Discrete or categorical multivariate data are often summarized using *contingency tables*. A contingency table contains counts or number of observations that fall into each cell of the table, where each cell corresponds to a category of each variable. So the cell counts show the frequencies of the categories of the variables. The simplest contingency table is a 2×2 *table*, which is used to summarize data from two binary variables (i.e., two categorical variables with each variable having only two categories).

To illustrate, suppose that we wish to compare two groups, such as a treatment group and a control group or a group for male and a group for female. Let y be a binary variable taking only two possible values, say 1 or 0 (e.g., cancer or no cancer, death or alive, improvement or no improvement, success or failure). Let x be the group indicator, say group 1 or group 2. For example, we may wish to study if getting a cancer is associated with smoking or not (so y is a cancer indicator and x is a smoking indicator), or if more female students find jobs than male students (so y is a job status and x is a gender indicator). Suppose that there are n_1 and n_2 individuals in the two groups respectively. Let c_{11} and c_{12} be the numbers of individuals having values $y = 1$ in the two groups respectively, and let c_{21} and c_{22} be the numbers of individuals having values $y = 0$ in the two groups respectively. The data may then be summarized in the following 2×2 table (also called a 2-way table):

		x	
		1	2
y	1	c_{11}	c_{12}
	0	c_{21}	c_{22}
	Total	n_1	n_2

For example, in a sample of 100 individuals, with 50 smokers and 50 non-smokers. Suppose that 20 of the smokers get cancer, while only 12 of the non-smokers get cancer, in a follow-up of 10 years. The data may then be summarized in the following table

		Smoking	
		yes	no
Cancer	yes	20	12
	no	30	38
	Total	50	50

Based on this table, we may wish to know if smoking and cancer are significantly associated or if smoking has a higher risk of cancer than non-smoking.

Thus, a 2×2 table summarizes data for the two binary variables (x, y). The objective may be to test if x and y are associated. In a 2×2 table, the column totals n_1 and n_2 are not necessarily fixed, but the total sample size $n = n_1 + n_2$ is usually fixed. The test for association is equivalent to tests for equality of two proportions (i.e., if the proportions of $y = 1$ are the same in the two groups) or tests for equal risks (e.g., if the risk of cancer is the same in the two groups). These 2×2 tables are simple but are quite useful in practice, so they are widely used especially in biostatistics. Some useful statistics, such as odds ratio and relative risk, can be easily computed from these tables.

The 2×2 tables can be extended to more general cases. First, each of the two variables can have more than two categories. If x has p categories and y has q categories, data for (x, y) can be summarized in a $p \times q$ table. For example, a 2×3 table can be written as the following 2-way table

		x		
		1	2	3
y	1	c_{11}	c_{12}	c_{13}
	0	c_{21}	c_{22}	c_{23}
	Total	n_1	n_2	n_3

Second, contingency tables can be used to summarize data from more than two categorical variables. For example, data from three binary variables (x, y, z) can be summarized in the following $2 \times 2 \times 2$ table (also called a 3-way table):

$x = 1$

		z	
		1	2
y	1	c_{11}	c_{12}
	2	c_{21}	c_{22}
	Total	n_1	n_2

$x = 2$

		z	
		1	2
y	1	d_{11}	d_{12}
	2	d_{21}	d_{22}
	Total	m_1	m_2

Similar, data from k categorical variables can be summarized in a k-way contingency table, but the presentation becomes more tedious. The k-way tables can be used to study the association between k categorical variables.

6.4 Associations Between Discrete or Categorical Variables

The association between two continuous variables is usually measured using the correlation coefficient or covariance. However, for two categorical variables, the correlation coefficient is no longer appropriate. In this section, we discuss a few methods to measure the association among categorical variables.

For two binary variables or a 2×2 table, the association is often measured using the odds ratio, as illustrated below. Let p_i be the probability (or *risk*) of variable y taking value 1 in group i, $i = 1, 2$. Then, the ratio p_1/p_2 is called the *relative risk*, and the ratio

$$q = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

is called the *odds ratio*, which can be used to measure the *association* between two binary variables x and y. An odds ratio of 1 indicates the independence of two binary variables, while an odds ratio moving away from 1 (in either direction) indicates increasing association between the two binary variables. Testing the association between x and y is equivalent to testing the equal risks hypothesis $H_0 : p_1 = p_2$ (i.e., there is no risk difference between the two groups, or variable x and y are independent), versus an alternative (say) $H_1 : p_1 \neq p_2$ (i.e., there is a risk difference between the two groups, or variable x and y are associated). Alternatively, we can test the odds ratio $H_0 : q = 1$ (independence) versus $H_1 : q \neq 1$ (association).

For the smoking/cancer example in the previous section, the estimated proportions are $\hat{p}_1 = 20/50 = 0.4$ and $\hat{p}_2 = 12/50 = 0.24$. The estimated relative risk is $\hat{p}_1/\hat{p}_2 = 0.4/0.24 = 1.67$, and the estimated odds ratio is

$$\hat{q} = \frac{0.4/0.6}{0.24/0.76} = 2.11.$$

So the estimated odds ratio is far from 1, suggesting a high association between smoking and cancer. However, such an association may or may not be statistically significant. To verify this, we can perform a hypothesis testing of association for a 2×2 table, as described below.

When the sample size is small, a commonly used method for testing the hypothesis $H_0 : p_1 = p_2$ versus $H_1 : p_1 \neq p_2$ (or testing the independence between the two variables in a 2×2 table) is called the *Fisher's exact test*. The Fisher's exact test is obtained by conditioning on the marginal total $m_1 = c_{11} + c_{12}$ (i.e., treating the marginal totals as fixed). Conditional on m_1, the probability (likelihood) under the

null hypothesis H_0 is given by the following hypergeometric distribution

$$P(x_1 = j \mid x_1 + x_2 = m_1) = \frac{\begin{pmatrix} n_1 \\ j \end{pmatrix}\begin{pmatrix} n - n_1 \\ m_1 - j \end{pmatrix}}{\begin{pmatrix} n \\ m_1 \end{pmatrix}}, \qquad j = l, \cdots, u,$$

where $n = n_1 + n_2$, $l = \max(0, m_1 - n_2)$, and $u = \min(m_1, n_1)$. Based on the above hypergeometric distribution, a p-value for testing H_0 can be obtained according to Fisher's suggestion that computing the total probability of getting the observed data, and all data sets with more extreme deviations when the marginal totals are the same as in the observed table. The Fisher's exact test is based on exact computation, not on asymptotic results, so it can be used when the sample size is not large.

When the sample size is large, several large-sample tests can be considered. These tests are unconditional tests. For example, the following χ^2-*test* can be used to test for association of two categorical variables (x, y) in a $r \times c$ table (i.e., a 2-way contingency table for y having r categories and x having c categories). Specifically, let p_{ij} be the probability of an observation falling in cell (i, j) (i.e., the i-th row and j-th column), $i = 1, \cdots, r; j = 1, \cdots, c$. Let $p_{i.} = \sum_j p_{ij}$ and $p_{.j} = \sum_i p_{ij}$. Consider testing the *independence* between x and y, which is equivalent to testing the hypothesis

$$H_0 : p_{ij} = p_{i.}p_{.j} \qquad \text{versus} \qquad H_1 : p_{ij} \neq p_{i.}p_{.j}.$$

The χ^2 test statistic is given by

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(c_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}, \tag{6.2}$$

where c_{ij} is the observed count in cell (i, j), $\hat{e}_{ij}$ is the estimated expected count in cell (i, j) under H_0, i.e.,

$$\hat{e}_{ij} = \left(\sum_{j=1}^{c} c_{ij}\right)\left(\sum_{i=1}^{r} c_{ij}\right)/n,$$

and n is the total sample size. Under the null hypothesis H_0, the test statistic

$$X^2 \sim \chi^2((r-1)(c-1))$$

asymptotically. So an approximate p-value can be obtained based on this asymptotic null distribution. Other large-sample tests include the Mantel-Haenszel test and the likelihood ratio test. All these large-sample tests are asymptotically equivalent, although their finite-sample performances may differ.

A 2×2 table may be used in different settings. For example, suppose that x and y are two binary variables. A single sample of size n is taken, and data on (x, y) are

obtained and summarized in a 2×2 table. In this case, the total sample size n is fixed but the column totals n_1 and n_2 are not fixed, and the data of (x, y) may be assumed to follow a multinomial distribution $M(n, \mathbf{p})$, with $\mathbf{p} = (p_{11}, p_{12}, p_{21}, p_{22})^{\mathrm{T}}$.

More generally, suppose that $\mathbf{x} = (x_1, \cdots, x_k)^{\mathrm{T}}$ is a multivariate discrete random vector consisting of k categorical variables, with each x_j having d_j categories. Then, data on $\mathbf{x}$ can be summarized in a *k-way contingency table*. The χ^2 test in (6.2) may be used to test the association between the k categorical variables. In other words, the χ^2 test in (6.2) is a general method which can be used to test association in any contingency tables, not just 2×2 tables. However, when use the χ^2 test, the cell counts should not be too small. Usually, it is recommended that each cell count should be at least 5 in order for the χ^2 test to perform well. This is because the χ^2 test is an asymptotic test (i.e., a large sample test). When cell counts are small, it may be desirable to use a bootstrap test or to combine some categories in order to make the cell counts larger. For example, if x_1 has 3 categories, say "non-smoker", "light smoker", and "heavy smoker". To increase cell counts, we may combine the last two categories so that the new x_1 has only two categories "non-smoker" and "smoker". This is a usual approach to avoid sparse tables.

If there is an association between categorical variables, it may be desirable to describe the strength of the association. We can use correlation-like measures such as the *Phi coefficient* and *Cramer's V* to describe the strength of relationship between categorical variables. The values of these coefficients range from 0 to 1, since we cannot have a negative relationship between categorical variables. The Phi coefficient (ϕ) is a measure of nominal association applicable only to 2×2 tables. It is calculated as:

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

where χ^2 is the value of Pearson's χ^2 given above, and n is the total sample size. For contingency tables that are larger than 2×2 tables, Cramer's V is the choice of nominal association measure. The formula for Cramer's V is given by:

$$V = \sqrt{\frac{\chi^2}{n(k-1)}},$$

where n is the total sample size and k is the lesser of the number of rows or columns. Since in 2×2 tables $k = 2$, Cramer's V equals the Phi coefficient for 2×2 tables. More details of these measures may be found in Agresti (2012).

6.5 Logit Models for Multinomial Variables

In regression analysis, when the response is a binary variable with only two possible categories, a logistic regression model is a natural choice. When the response is a

multinomial variable with more than two categories, a multinomial regression model may be considered, which is a generalization of the logistic regression model. There are different approaches to model multinomial responses. In this section, we briefly describe a simple approach based on logistic regression models. The basic idea is to use a separate logistic regression model for *each pair* of the response categories, leading to a set of related logistic regression models.

Let y be a discrete or categorical response variable with m categories, and let the vector $\mathbf{x} = (x_1, x_2, \cdots, x_p)$ contain a set of predictors or covariates. Let $p_j(\mathbf{x}) = P(y = j|\mathbf{x}),\ j = 1, 2, \cdots, m$, with $\sum_{j=1}^{m} p_j(\mathbf{x}) = 1$. Given a dataset, we can assume that the counts in the m categories of y follow a multinomial distribution with probabilities $\{p_1(\mathbf{x}), \cdots, p_m(\mathbf{x})\}$. To build a regression model for the multinomial response y, we may consider the following $m - 1$ logistic regression models (called *multinomial logit model*)

$$\log \frac{p_j(\mathbf{x})}{p_m(\mathbf{x})} = \beta_{0j} + \boldsymbol{\beta}_j^{\mathrm{T}} \mathbf{x}, \qquad j = 1, 2, \cdots, m, \tag{6.3}$$

which simultaneously describe the effects of predictors $\mathbf{x}$ on the $m - 1$ logits $p_j(\mathbf{x})/p_m(\mathbf{x}), j = 1, 2, \cdots, m - 1$. Thus, the above multicategory (polytomous) logit models simultaneously describe log odds for all $m(m - 1)/2$ pairs of categories, since

$$\log \frac{p_i(\mathbf{x})}{p_j(\mathbf{x})} = \log \frac{p_i(\mathbf{x})}{p_m(\mathbf{x})} - \log \frac{p_j(\mathbf{x})}{p_m(\mathbf{x})}, \qquad \text{for } i \neq j.$$

Note that, in the $m - 1$ logistic models (6.3), we pair each response category with a baseline category, which is chosen as the last category m (but it can be other categories). The $m - 1$ logistic regression models are related since each model contains $p_m(\mathbf{x})$ and $\sum_{j=1}^{m} p_j(\mathbf{x}) = 1$. The effects of the predictors vary according to the response paired with the baseline, i.e., the regression coefficient $\boldsymbol{\beta}_j$ are specific to the corresponding logistic model. The multinomial probabilities may be obtained as

$$p_j(\mathbf{x}) = \frac{\exp(\beta_{0j} + \boldsymbol{\beta}_j^{\mathrm{T}} \mathbf{x})}{1 + \sum_{k=1}^{m-1} \exp(\beta_{0k} + \boldsymbol{\beta}_k^{\mathrm{T}} \mathbf{x})}, \qquad j = 1, 2, \cdots, m,$$

with $\beta_{0m} = 0$ and $\boldsymbol{\beta}_m = 0$.

Estimates of all parameters in models (6.3) can be obtained *simultaneously* using the maximum likelihood method, assuming the response data follow a multinomial distribution. Most statistical software can provide the fitting results. When one only knows how to fit logistic regression models, an alternative fitting approach is to fit

logistic regression models *separately* for the $m-1$ logistic regression models in (6.3). This approach is simple, but the resulting estimates may be less efficient, with larger standard errors, than the overall maximum likelihood method.

There are also other approaches for building regression models for multinomial responses. One such approach is to consider a multivariate generalized linear model (GLM), assuming the multinomial distribution for the response. Specifically, let $\mathbf{y}_i = (y_{i1}, \cdots, y_{im})$, where $y_{ij} = 1$ if the response of individual i is in category j and $y_{ij} = 0$ otherwise (so $\sum_j y_{ij} = 1$), $i = 1, 2, \cdots, n;\ j = 1, 2, \cdots, m$. Then, a multivariate GLM can be written as

$$\mathbf{g}[E(\mathbf{y}_i)] = X_i\boldsymbol{\beta},$$

where $\mathbf{g}$ is a vector of link functions and X_i is a design matrix containing predictors. When the categorical responses are ordinal (i.e., the categories have orders, such as "small", "average", and "large"), an alternative approach is to use logits of cumulative response probabilities. For example, the following *proportional odds model* may be considered

$$\text{logit}[P(y \leqslant j|\mathbf{x})] = \beta_{0j} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}, \quad j = 1, 2, \cdots, m-1,$$

which simultaneously uses all cumulative logits.

6.6 Loglinear Models for Contingency Tables

For multivariate categorical variables, if one variable is treated as a response and other variables as predictors, we can use logit models to study their associations, as described in the previous section. When two or more categorical variables are treated as responses, we may use contingency tables to summarize data on these variables, and then use loglinear models to study their associations and interactions between the responses. In other words, loglinear models treat categorical response variables symmetrically, and they focus on associations and interactions in their joint distribution. Logit models, on the other hand, describe how a single categorical response depends on predictors. In this section, we briefly discuss these loglinear models.

First we consider loglinear models for two-way contingency tables. Let X be a categorical variable with I categories, let Y be a categorical variable with J categories, and let $m = IJ$ be the total number of cells in a contingency table (i.e., number of all combinations of different categories). A sample of n observations on the random vector (X, Y) can be summarized in an $I \times J$ contingency table. Let p_{ij} be the probability that an observation falls in cell (i, j) in the table, and let $p_{i+} = \sum_j p_{ij}$ and $p_{+j} = \sum_i p_{ij}$ be the marginal probabilities. The observed counts in cell (i, j) are

denoted by n_{ij}, and the expected cell counts are $\mu_{ij} = np_{ij}$. We assume a multinomial distribution with cell probabilities $\{p_{ij}, i = 1, \cdots, I, j = 1, \cdots, J\}$. If X and Y are *independent*, we have

$$p_{ij} = p_{i+}p_{+j}$$

for all i, j. This is equivalent to the following *loglinear model* of independence

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y, \tag{6.4}$$

where λ_i^X is the effect of X and λ_j^Y is the effect of Y. Note that identifiability requires that $\lambda_I^X = \lambda_J^Y = 0$. The *saturated model*, which is the most general model, can be written as

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}, \tag{6.5}$$

where λ_{ij}^{XY} is an association term that reflects deviation from independence. This is similar to a two-way ANOVA model with interaction.

Next, consider a three-way $I \times J \times K$ contingency table on three categorical variable X, Y, and Z. We assume a multinomial distribution with cell probabilities p_{ijk}. We use notation similar to that for two-way tables. The three variables are *mutually independent* if $p_{ijk} = p_{i++}p_{+j+}p_{++k}$ for all i, j, k. This independence model is equivalent to the following loglinear model (denoted by (X, Y, Z))

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z. \tag{6.6}$$

If X *and* Y *are conditionally independent given* Z, we have the following loglinear model (denoted by (XZ, YZ))

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}, \tag{6.7}$$

which is equivalent to the following equation

$$p_{ijk} = \frac{p_{i+k}p_{+jk}}{p_{++k}}$$

for all i, j, k. Note that conditional independence does not imply marginal independence. The following loglinear model allows for all three pairs of conditional independence (denoted by (XZ, YZ, XY))

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ik}^{XY}, \tag{6.8}$$

The most general model (saturated model) is (denoted by (XYZ))

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ik}^{XY} + \lambda_{ik}^{XYZ}. \tag{6.9}$$

Setting certain terms in the saturated model to zero leads to various reduced models, including various conditional independence models. In data analysis, reduced models are preferred. The strength of (conditional) association may be measured by odds ratio.

In a similar way, we may consider loglinear models for four-way contingency tables or higher order tables. Loglinear models for three-way tables are more complex than loglinear models for two-way tables, because there are different types of associations among the variables. Similarly, loglinear models for higher order tables are even more complex, due to the variety of association structures. From a practical point of view, interactions higher than order of two may become harder to interpret.

In data analysis, we prefer parsimonious models which fit the data reasonably well. The usual model selection methods, such as the χ^2 test for contingency tables and AIC criterion, can be used to select a reasonable model. The likelihood ratio test can also be used to compare different loglinear models. Note that many tests are based on asymptotic theory, which requires that the cell counts are not too small (say, larger than 5). For tables with small cell counts, Fisher's exact test or the bootstrap method may be used.

6.7 Example in R

Example 1. A researcher wishes to study if male drivers are more likely to get car accidents than female drivers. A sample of 100 male and 100 female drivers is obtained and the number of car accidents in a 5-year period is shown in the following 2×2 table:

	Male	Female
Accident	53	40
No accident	47	60
	100	100

The *risk* (probability) of accident in 5 years for male subjects is estimated to be $\hat{p}_1 = 0.53$, with a 95% confidence interval $(0.43, 0.63)$, and the risk for female subjects is estimated to be $\hat{p}_2 = 0.40$, with a 95% confidence interval $(0.30, 0.50)$. Note that the confidence intervals may be obtained using the normal approximation to a binomial distribution. The *odds ratio* is estimated to be 1.69, with a 95% confidence interval $(0.93, 3.08)$. The goal is to check if there is a significant difference between the two risks p_1 and p_2 (i.e., if male subjects are more likely to get accidents than female subjects). These questions can be answered by testing the hypothesis of equal risks versus two-sided alternative (a risk difference) or one-sided alternative (one risk is higher than the other).

For testing the hypothesis of equal risks $H_0 : p_1 = p_2$, a two-sided Fisher's exact test produces a p-value of 0.09, and a one-sided Fisher's exact test produces a p-value of 0.04. A two-sided large-sample χ^2 test produces a p-value of 0.09. Therefore, there are some evidence that the two risks are different and that male drivers are more likely to get car accidents than female drivers. In other words, there is some association between car accident and gender. In this example, the sample sizes are large, so large-sample tests may be appropriate. Fisher's exact tests may not be necessary here.

The R code for the above analysis is shown below.

```
# The data in a 2 by 2 table
> car <- matrix(c(53, 40, 47, 60), byrow=T, nrow=2,
           dimnames=list(c("Accident", "No"),c("Male","Female")))
> car
           Male   Female
Accident    53      40
No          47      60

# Fisher's exact test
> fisher.test(car)# two-sided test
        Fisher's Exact Test for Count Data

data:  car
p-value = 0.08865
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.929964 3.080613
sample estimates:
odds ratio
  1.687032

> fisher.test(car, alternative="greater")  # one-sided test
        Fisher's Exact Test for Count Data

data:  car
p-value = 0.04432
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 1.01642     Inf
sample estimates:
odds ratio
  1.687032

# Large sample chi-square test
> chisq.test(car)
        Pearson's Chi-squared test with Yates' continuity correction

data:  car
X-squared = 2.8942, df = 1, p-value = 0.0889
```

Example 2. The following dataset contains numbers of people using alcohol and cigarette in two age groups (ages 20–29 and ages 30–39) in a sample of 2271 people. The data can be summarized in a $2 \times 2 \times 2$ three-way contingency table as shown below. So this is a three-variable multivariate discrete dataset. The three categorical variables are denoted by A (alcohol use), C (cigarette use), and G (group indicator). We consider a loglinear model to study the association structures among the three variables.

Alcohol	Cigarette	Group	
		I	II
Yes	Yes	909	537
	No	45	454
No	Yes	4	44
	No	2	276

The R library "MASS" contains many functions for categorical data analysis. We fit four loglinear models (6.6) – (6.9) and then choose the best model.

```
> library(MASS)
# data presented in 3-way table
> table3 <- data.frame(expand.grid(group=factor(c("1","2")),
     cigarette=factor(c("Yes","No")),
     alcohol=factor(c("Yes","No"))),
     count=c(909,537,45,454,4,44,2,276))
> table3    # display of data
  group cigarette alcohol count
1     1       Yes     Yes   909
2     2       Yes     Yes   537
3     1        No     Yes    45
4     2        No     Yes   454
5     1       Yes      No     4
6     2       Yes      No    44
7     1        No      No     2
8     2        No      No   276

# fit a saturated loglinear model
> fitACG <- loglm(count~alcohol*cigarette*group, data=table3,param=T,fit
    =T)
# fit a pairwise conditional independence loglinear model
> fitAC.AG.CG <- update(fitACG, .~. - alcohol:cigarette:group)
# fit a loglinear model with A conditional independent of C given G
> fitAG.CG <- loglm(count~alcohol+cigarette+group+alcohol:group+
  cigarette:group,data=table3,param=T,fit=T)
# fit a loglinear model of independence
> fitA.C.G <- loglm(count~alcohol+cigarette+group, data=table3,param=T,
    fit=T)

# Comparison of nested models using the anova method
```

```
> anova(fitAC.AG.CG, fitAG.CG, fitA.C.G)

LR tests for hierarchical log-linear models

Model 1:
 count ~ alcohol + cigarette + group
Model 2:
 count ~ alcohol + cigarette + group + alcohol:group + cigarette:group
Model 3:
 count ~ alcohol + cigarette + group + alcohol:cigarette + alcohol:group
 + cigarette:group

           Deviance      df   Delta(Dev)   Delta(df) P(> Delta(Dev))
Model 1   1262.866       4
Model 2    182.244       2     1080.622           2           0.000
Model 3      0.114       1      182.129           1           0.000
Saturated    0.000       0        0.114           1           0.735

#Likelihood ratio chi-squared test statistics
> summary(fitAC.AG.CG)
.......
statistics:
                      X^2   df  P(> X^2)
Likelihood Ratio    0.114        1 0.734
Pearson             0.119        1 0.729
```

From the above results, we see that "Model 3" (fitAC.AG.CG) seems the best since it fits much better than Models 1 and 2 (very small p-values) while it is not significantly different from the saturated model (so it is preferred since it is simpler than the saturated model). Model 3 shows that the three variables are pairwise conditionally independent given the remaining variables, i.e., variables A and C are conditionally independent given G, variables A and G are conditionally independent given C, and variables G and C are conditionally independent given A. The independence model (Model 1), although the simplest one, fits the data quite poorly, indicating that the three variables are not independent (i.e., alcohol, cigarette, and age group are correlated).

Finally, we can get estimates of the parameters in "Model 3" using the "glm" function with Poisson distributional assumption.

```
# Parameter estimates using glm
fit.glm <-
glm(count~alcohol+cigarette+group+alcohol:group+cigarette:group+
alcohol:cigarette,data=table3, family=poisson)
> summary(fit.glm)

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)              5.62172    0.06003  93.646  < 2e-16 ***
alcoholYes               0.49557    0.07606   6.516 7.24e-11 ***
cigaretteYes            -1.84586    0.16046 -11.503  < 2e-16 ***
```

```
group1                    -5.13007     0.43824 -11.706   < 2e-16 ***
alcoholYes:group1          2.82751     0.42696   6.622 3.53e-11 ***
cigaretteYes:group1        2.82782     0.16242  17.410   < 2e-16 ***
alcoholYes:cigaretteYes    2.01525     0.17191  11.722   < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

All parameter estimates are highly significant. The interaction terms are also highly significant, indicating that the numbers of people using alcohol or cigarette depend on age.

Exercises 6

6.1. Derive the means, variances, covariances for the multinomial distribution $M(n, \mathbf{p})$.

6.2. (a) Show that, for a 2×2 table with cell counts c_{ij}, $i, j = 1, 2$, the odds ratio is invariant to: (i) interchange of rows and columns, and (ii) multiplication of cell counts within rows or columns by a nonzero constant. Also, show that the difference of proportions do not have these invariant properties.

(b) Derive an approximate formula for the standard error of the odds ratio, and then construct an approximate 95% confidence interval for odds ratio, assuming the sample size is large.

6.3. In the following contingency table, there are 13 males between the ages of 11 and 30 who received an operation for knee injures using a arthroscopic surgery. The patients were classified based on the type of injury: direct blow (D), or both twisted knee and direct blow (B). The results of the surgery were also classified as excellent (E) or good (G).

Type of injury	Surgery result		Total
	Excellent	Good	
Direct (D)	3	2	5
Both (B)	7	1	8
Total	10	3	13

Since the cell counts are small, use the Fisher's exact test to test whether the result of a patient's surgery is independent of the type of injury.

6.4. The following table contains results of a study by Correa et al. (1983) to investigate the effect of passive smoking on lung cancer. Each of the 155 non-smokers is classified according to whether his/her spouse smoke or not.

Spouse smoke	Lung cancer		Total
	Yes	No	
Yes	14	61	75
No	8	72	80
Total	22	133	155

(a) Conduct the Fisher's exact test to test the following hypotheses separately:

$$H_0 : \theta = 1$$

against the alternatives (i) $H_1 : \theta > 1$ or (ii) $H_1 : \theta \neq 1$, where θ is the odds ratio.

(b) Repeat the tests based on a large-sample method and compare the results to those in (a). Are the results consistent?

(c) Estimate the strength of the association between spouse smoke status and lung cancer.

6.5. The following table gives information for the numbers of applicants admitted to graduate programs in the six largest major programs at the University of California, Berkeley, in the fall of 1973.

Major	Men		Women	
	No. of applicants	No. admitted	No. of applicants	No. admitted
A	825	512	108	89
B	560	353	25	17
C	325	120	593	202
D	417	138	375	202
E	191	53	393	94
F	373	22	341	24

(a) Is there any indication of biasedness in graduate admissions based on the sex of the applicants?

(b) Construct separate 2×2 tables for both men and women for each major and conduct individual tests for each sub-table. What do you find? Are the conclusions the same as that in (a)? Why?

6.6. Fit loglinear models to the $2 \times 2 \times 6$ contingency table in Exercise 6.5 and choose the best model.

Chapter 7

Copula Models

7.1 Introduction

The most commonly used multivariate distributions are probably the multivariate normal distributions for continuous data and the multinomial distributions for discrete data. For a multivariate normal distribution, all the marginal distributions are normal distributions. Similarly, for a multinomial distribution, all the marginal distributions are binomial distributions, which are special multinomial distributions. Thus, for either a multivariate normal distribution or a multinomial distribution, the univariate distributions of the component random variables all follow the same distributions. These distributions can be restrictive in some applications. For example, in financial risk analysis, the market risk has portfolio value distributions which can be approximated by a normal distribution, but the credit and operational risks are often approximated by skewed distributions because of occasional extreme events, and the different risks are usually associated or correlated. Also, for lifetimes and long-tail claims in actuarial science, the normal distributions usually do not provide adequate approximations to the data distributions.

In addition, in practice we sometimes may want to use a multivariate distribution whose marginal distributions are of different types. For example, in some applications we may wish to assume that X follows a normal distribution but Y follows a binomial distribution. Then, if X and Y are associated and a distributional assumption is desirable for the multivariate random vector (X, Y), it is not clear what the multivariate joint distribution for (X, Y) should be. As another example, suppose that X follows a normal distribution and Y follows a exponential distribution, then the joint distribution for (X, Y) is not clear, since there seems no existing multivariate distributions whose marginal distributions are normal and exponential respectively.

More specifically, consider the multivariate normal distribution, which is a standard distributional assumption in multivariate analysis. Although the multivariate normal distribution is very useful and is a reasonable assumption in many applications, it can also be restrictive in some applications, such as applications in finance

or insurance data. This is because i) the tails of its marginal distributions (i.e., normal distributions) may be too thin for some data, so it may not incorporate extreme values which may arise in practice; ii) it fails to capture the phenomenon of joint extreme movements, so simultaneous large values may be infrequent, and it is generally believed to lack tail dependence; and iii) it is too symmetric so it lacks skewness. In other words, the multivariate normal distribution may be inappropriate for modelling some financial or insurance data, since these data may exhibit long tails and extreme values, such as un-usual stock price changes and financial crisis. Therefore, we need a more flexible approach to specify a multivariate distribution for these kinds of data.

In practice, it is easier to make reasonable distributional assumptions for univariate data or data on one variable, since we can use graphical tools to display univariate data and check the distributional assumption using Q-Q plots. To specify a multivariate distribution, a main consideration is to incorporate the association between different variables. Thus, an appealing approach for constructing a multivariate distribution is i) assume a univariate distribution for each variable; and then ii) incorporate the association between different variables separately. **That is, we separate the marginal distributions from the association structure. This is the key idea behind a Copula function.** A *Copula* is a useful tool which allows us to build multivariate distributions from univariate distributions, and it allows us to model different types of dependence for multivariate data. In other words, to specify a multivariate distribution, we can assume a distribution for each variable separately, where the distributions for different variables can be different, and then we construct a multivariate distribution from these univariate distributions by incorporating the association between different variables.

In multivariate analysis, the association or correlation between different variables play a key role. One of the main goals in multivariate analysis is to study the dependency between random variables. For example, in the financial world, the market risk and credit risk may be associated since both are related to the interest rate. If an investor has a portfolio with two loans to two companies, he may reduce the portfolio risk by holding assets that are not highly correlated with each other. A commonly used dependence measure is the correlation or correlation coefficient. However, a correlation coefficient only measures a *linear* relationship. It is a reasonable measure of association when the random variables are multivariate normally distributed or elliptically distributed. Correlation coefficient may not be a good measure for dependence when the data distributions are skewed or heavy-tailed or the association is not linear. Moreover, a correlation is not invariant under monotone transformations of the random variables, such as log-transformation. In these cases, the Copula provides an alternative measure of association and may be better

than the correlation coefficient.

In the following sections, we briefly describe the Copula function and its properties. Then, we review some applications of Copula functions, especially in finance and actuarial sciences where Copulas are widely used.

7.2 Copula Models

Copula is a useful way to model dependence between variables, especially for non-normal data. It allows one to combine univariate distributions to form a joint multivariate distribution with a particular dependence structure. In this section, we illustrate the basic idea of the copula method in a bivariate setting. Extensions to multivariate settings are straightforward.

Let $F_x(x)$ and $F_y(y)$ be the cumulative distribution functions (cdf's) of random variables X and Y respectively. It can be shown that there is a function $C(u, v) : [0, 1]^2 \to [0, 1]$ such that $C(F_x(x), F_y(y))$ is a cdf of the bivariate random vector (X, Y). In other words, we can build a multivariate cdf from the univariate cdf's $F_x(x)$ and $F_y(y)$ through the copula function $C(\cdot, \cdot)$. Sklar's theorem, which is the fundamental result on copulas theory, states that, for a given joint distribution $F_{x,y}(x, y)$ and its corresponding marginal distributions $F_x(x)$ and $F_y(y)$, there exists a *copula function* $C(u, v)$ such that

$$F_{x,y}(x, y) = C(F_x(x), F_y(y)),$$

and the function $C(u, v)$ is unique if X and Y are continuous. Conversely, if $C(u, v)$ is a copula function and if $F_x(x)$ and $F_y(y)$ are univariate distribution functions, then the function $F_{x,y}(x, y) = C(F_x(x), F_y(y))$ is a bivariate distribution function with marginal distributions given by $F_x(x)$ and $F_y(y)$ respectively. Therefore, a copula function $C(u, v)$ allows one to build a multivariate distribution $F_{x,y}(x, y)$ from given univariate distributions $F_x(x)$ and $F_y(y)$. The dependence between X and Y is often incorporated by one or more parameters in the copula function $C(u, v)$, as illustrated below.

There are many common choices for copula functions. We first provide a formal definition of a copula function in the bivariate setting. A bivariate *copula* is a joint cumulative distribution function defined on the two dimension region $[0, 1]^2$ with uniform marginals. That is, the function

$$C(u, v) : [0, 1]^2 \to [0, 1]$$

is a *copula* if it satisfies the following conditions

(i) $C(0, v) = C(u, 0) = 0, \qquad 0 \leqslant u, v \leqslant 1,$

(ii) $C(1, u) = C(u, 1) = u, \qquad 0 \leqslant u \leqslant 1,$

(iii) For all $u_1 \leqslant u_2, v_1 \leqslant v_2$,

$$\int_{v_1}^{v_2}\int_{u_1}^{u_2}\frac{\partial^2 C(x,y)}{\partial x \partial y}\mathrm{d}x\mathrm{d}y = C(u_2,v_2) - C(u_1,v_2) - C(u_2,v_1) + C(u_1,v_1) \geqslant 0.$$

Note that the above three conditions are properties of a cdf. Extensions to more than two variables are similar.

The foregoing results suggest that we can use copula to build multivariate distributions based on marginal distributions. This approach is flexible and powerful since it separates the choice of the dependence structure from the choice of marginal distributions. There are many parametric copula families available, and these parametric copulas often contain parameters that control the strengths of dependence. For example, we can choose the Gaussian copula for linear correlations, the Gumbel copula for extreme distributions, and the Archimedean copula for dependence in the tails. These commonly used copula families are described below.

The *Gaussian copula* with correlation matrix Σ can be written as

$$C_\Sigma(u_1,u_2) = \Phi_\Sigma(\Phi^{-1}(u_1), \Phi^{-1}(u_2)),$$

where Φ is the cdf of the univariate standard normal distribution $N(0,1)$ and Φ_Σ is the joint cdf of the bivariate normal distribution $N_2(\mathbf{0}, \Sigma)$.

The *Archimedean copula* has the following form

$$C(u_1,u_2) = \psi(\psi^{-1}(u_1) + \psi^{-1}(u_2)),$$

where ψ is called a *generator*. Examples of commonly used generators are the *Clayton generator*

$$\psi(t) = (1+\theta t)^{-1/\theta}, \qquad 0 < \theta < \infty,$$

the *Gumbel generator*

$$\psi(t) = \exp(-t^{1/\theta}), \qquad 1 \leqslant \theta < \infty,$$

the *Frank generator*

$$\psi(t) = -\frac{1}{\theta}\log(1-(1-\exp(-\theta))\exp(-t)), \qquad 0 < \theta < \infty,$$

the *independence generator*

$$\psi(t) = \exp(-t),$$

and the *Joe generator*

$$\psi(t) = 1-(1-\exp(-t))^{1/\theta}, \qquad 1 \leqslant \theta < \infty.$$

In the above generators, the parameter θ controls the strength of the dependence between the variables X and Y. In practice, Archimedean copulas are popular since they allow modelling dependence in arbitrarily high dimensions with only one parameter governing the strength of dependence. In other words, the Archimedean representation allows us to reduce the study of a multivariate copula to a single univariate function. Moreover, Archimedean copulas are not difficult to construct and have many different forms.

Copula functions have the invariance property. Specifically, suppose that the random vector $\mathbf{x} = (x_1, \cdots, x_n)$ has a copula function C, and suppose that $g_1, \cdots, g_n$ are non-decreasing continuous functions of $x_1, \cdots, x_n$ respectively. Then, the transformed random vector $(g_1(x_1), \cdots, g_n(x_n))$ has the same copula function C. In this sense, the copula accounts for all the dependence between the random variables. This invariance property can be useful in many applications. For example, suppose that we have a copula describing a joint distribution of insurance losses of different types, and we decide that the quantity of interest is a transformation (e.g. logarithm) of these losses, then the multivariate distribution structure does not change. Thus, the dependence structure is preserved, while the marginals change.

Two standard nonparametric correlation measures, the Spearman's correlation coefficient and the Kendall's correlation coefficient, can be expressed solely in terms of the copula function. The commonly used Pearson's correlation coefficient, however, depends not only on the copula but also on the marginal distributions, so it is affected by changes of scale.

7.2.1 Statistical analysis based on copula

In practice, when data are available, we can choose marginal distributions based on univariate data, using graphical tools such as histograms and QQ-plots. For example, if data from X are roughly normally distributed and if data from Y are skewed lifetime data, then we may assume X follows a normal distribution and assume Y follows a Weibull distribution. Then, we can choose an appropriate copula, by comparing nonparametric and parametric quantiles using QQ-plots, to incorporate the multivariate dependence. The resulting copula model may be viewed as a joint model which incorporates the association between the marginal variables.

Once the marginal distributions and the copula are chosen, we can then use the maximum likelihood method to estimate the model parameters. The estimated association parameter in the copula function reflects the dependence between the marginal variables. Estimates of the parameters in the marginal distributions based on the copula model often have smaller standard errors than those based on separate fitting of the marginal distributions, since there is often an efficiency gain in using the joint model.

When several copulas are all reasonable choices for a given dataset, the most appropriate copula model may be obtained by comparing Akaike's Information Criteria (AIC) values for the models. The model with the smallest AIC value is generally preferred. Other similar model selection methods may also be used to choose a copula.

7.3 Measures of Dependence

Measuring the association or dependence between different variables is important in multivariate analysis, since the key consideration in multivariate analysis is to incorporate the association or dependence between variables. The most common measure of association between two continuous variables X and Y is the Pearson's correlation coefficient r, which is defined as

$$r = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}.$$

The correlation r is a good measure of association when the data or variables are approximately normally distributed or are symmetric. However, the correlation r has some disadvantages: i) it only measures linear association, not nonlinear associations, so $r = 0$ does not mean that the two variables are independent; ii) it is not invariant under variable transformations, e.g., the correlation between X and Y is different from the correlation between $\log(X)$ and $\log(Y)$ (assuming X and Y are positive); iii) it is inappropriate when the variances tend to infinite, e.g., heavy-tailed distributions which are common for financial data. Therefore, alternative measures of association or dependence are required.

Copula provides an alternative approach to measure association or dependence. It determines the dependence relationship by joining the marginal distributions together to form a joint distribution. The dependence relationship is determined by the copula function, so, with different choices of the copula function, we can measure different dependence structures. For example, the Gumbel copula can be written as

$$C(u,v) = \exp\left[-\left((-\log u)^{\theta} + (-\log v)^{\theta}\right)^{\frac{1}{\theta}}\right],$$

and the Clayton copula can be written as

$$C(u,v) = \left(u^{-\theta} + v^{-\theta} - 1\right)^{-\frac{1}{\theta}}.$$

These two copula functions provide two different measures of dependence. There are also many other choices of copula function. Therefore, copulas can be used to measure various dependence structures.

The copula has many advantages. First, a copula can be used when the distributions of data or variables are heavy tailed or skewed. For example, in finance, the credit risk and operational risk are usually modelled with heavy skewed distributions because of occasional and extreme losses, so in these cases the copula is more desirable for measuring dependence than the correlation r. Second, the copula is invariant under strictly increasing transformations of the variables. For example, the copula of $\log(X)$ and $\log(Y)$ is the same as the copula of X and Y. Third, copula can be used to measure dependence between different types of variables, such as a continuous variable and a discrete variable. Therefore, copulas allow us to model various types of dependence.

7.4 Applications in Actuary and Finance

Actuaries often try to understand stochastic outcomes of financial security systems and thus frequently face problems involving multivariate outcomes, since these outcomes are usually measured in several ways due to the complexity of financial security systems. Understanding relationships among different outcomes is a basic actuarial technique for explaining the behaviour of financial security systems to business and public policy decision-makers. Therefore, it is important to adequately model these multivariate outcomes in order to avoid biased results. Copula is a useful tool for relating different outcomes. For example, when two lives (e.g., husband and wife) are subject to failure, such as under a joint life insurance or annuity policy, we may be concerned with the joint distribution of lifetimes, not in isolation of one another. In the following, we briefly describe some of these applications, with a focus on actuarial science.

In actuarial studies, it is of interest to examine the joint mortality pattern of more than a single individual, such as the joint mortality pattern of a husband and wife or a family with children. For example, suppose that we are interested in the joint survival distribution of a husband and wife. The (marginal) survival distribution of a husband may be different from that of the wife, due to gender difference and other factors, but the two marginal distributions are likely to be associated because of shared common risk factors such as common environmental factors and life styles. Similarly, the survival distribution of a father may be different from that of a daughter, due to the gender and age differences, but the two distributions may also be associated because of genetic factors as well as environmental factors. In these cases, we may build a joint multivariate survival distribution for the husband and wife (or for the father and daughter) based on their marginal distributions and a copula function which incorporates the association between the marginal distributions.

7.4.1 Mixed effects or frailty models

Mixed effects models or random effects models are useful in practice since they allow us to incorporate the large variation between individuals or they allow us to model heterogeneity. The random effects in a mixed effects model may be viewed as individual effects. Moreover, dependence between different outcomes may be incorporated by the random effects in a mixed effects model. Specifically, the random effects in a mixed effects model may be interpreted as unobserved common risk factors, as described in the above paragraph, or interpreted as latent variables shared by a group of individuals. The dependence among the individuals within a group is thus induced through shared random effects. Survival models with random effects are also called *frailty models*. These frailty models have become quite popular for modelling clustered survival data, such as the ones described in the previous paragraph. Note that many frailty models can be written as copulas (Marshall and Olkin, 1988). In other words, there is a close relationship between frailty models and copulas.

7.4.2 Competing risk models

Competing risk problems arise in survival analysis, systems reliability, medical studies, and actuarial science. It is called *multiple decrement theory* in actuarial science. For example, a person may die because of one of several possible causes such as cancer, heart disease, smoking, and so on. A piece of wood may break because of one of several possible causes such as knots, shakes, grain, split, and so on. Let T_k denote the lifetime due to the k-th cause of failure, $k = 1, 2, \cdots, p$. Then, the observed quantity is

$$T = \min(T_1, T_2, \cdots, T_p).$$

The dependence between $T_1, T_2, \cdots, T_p$ may be modelled using copula or may be introduced by a shared frailty or random effect. When using a random effect model, we can then assume that $T_1, T_2, \cdots, T_p$ are conditionally independent given the random effects, which greatly simplifies the models. For example, a patient may die because either cancer or a heart disease. Let T_1 be the survival time due to cancer and let T_2 be the survival time due to a heart disease. The observed survival time in practice is $T = \min(T_1, T_2)$. Here T_1 and T_2 are likely to be correlated since they are survival times for the same patient, although due to different causes. The survival distributions of T_1 and T_2 may also be different. Thus, we can use copula to build a joint distribution of (T_1, T_2) or use a mixed effects model to incorporate the correlation between T_1 and T_2.

7.4.3 Loss and expense in insurance company

Consider insurance company's indemnity claims. Each claim typically consists of an indemnity payment (called the loss X_1) and an allocated loss adjustment expense (called ALAE or expense X_2) such as lawyers' fees and claim investigation expenses. It is of interest to understand the association and joint distribution of the losses and expenses. Note that such data are often censored due to a possible maximal claim amount. For the univariate distributions, we can choose a lognormal distribution for expenses and a longer tail distribution, such as Pareto distribution, for losses associated with the claim. Then, we can use copula to build a joint distribution of (X_1, X_2). For the choice of the copula function, we may consider the Gumbel copula and Frank copula. The resulting copula model may be viewed as a joint model for expense and loss, which incorporates the association between expense and loss. Note that ignoring the dependence between the loss and expense may lead to mispricing, so it is important to incorporate the dependence and focus on the multivariate model instead of the univariate models.

7.4.4 Financial risk management

Copulas have been widely used in quantitative finance such as risk/portfolio management. The concept of risk is based on the uncertainty about future outcomes. That is, risk indicates any uncertainty that might lead to losses. In the financial world, there are three types of individual risks: market risk, credit risk, and operational risk. Market risk suggests that the value of an investment may decrease due to adverse movements in market factors. Credit risk suggests that a company or individual may be unable to pay the contractual interest or principal on its debt obligations. Operational risk refers to the risk of loss resulting from failed internal systems or from external events. The market risk has portfolio value distributions that are often approximated by a normal distribution, but the credit and operational risks may be better approximated by skewed distributions due to extreme losses.

In risk or portfolio management, copulas are usually used to perform stress-tests and robustness checks, where panic copulas are glued with market estimates of the marginal distributions to analyze the effects of panic regimes on the portfolio profit and loss distribution. Thus, Copula models are very useful in financial risk assessment, especially the pricing of collateralized debt obligations (CDOs). It is claimed that one of the reasons behind the global financial crisis in 2008 is the application of the Gaussian copula to credit derivatives, which is an industry standard model for pricing CDOs but alternative copula models should also be considered. Copulas have also been applied to other asset classes as a flexible tool in analyzing multi-asset derivative products, and are used in pricing and risk management of options on

multi-assets in the presence of volatility smile/skew, in equity, foreign exchange and fixed income derivative business. Interested readers can find extensive information in this area.

7.4.5 Other applications

Copulas have also been used for the reliability analysis of highway bridges and various multivariate simulation studies in civil and mechanical engineering, as well as in medical research and climate and weather research. Copulas are useful in the following situations: i) we wish to build multivariate distributions or models when standard multivariate distributions such as the multivariate normal distributions may be inappropriate or when the marginal distributions are of different types; and ii) we wish to study correlation or association between different variables when the usual correlation measures such as the Pearson's correlation coefficient may be inappropriate.

7.5 Applications in Longitudinal and Survival Data*

In this section, we briefly describe applications of Copula for modelling longitudinal or clustered data and multivariate survival data. In a longitudinal study, variables are repeatedly measured over time. For example, information about a patient is collected repeatedly everytime when the patient visits the doctor, so the data for this patient are longitudinal data. The repeated measurements for an individual are usually correlated. When the measurement schedules are fixed and are common for all individuals, one approach is to assume a multivariate distribution for the repeated measurements, such as a multivariate normal distribution. However, when the number of repeated measurements is large, a multivariate normal distribution may contain too many parameters. Alternatively, the multivariate distribution can be built from univariate marginal distributions based on the copula method. The Copula approach may be particularly appropriate for non-normal longitudinal data, since for non-normal multivariate data there may be no suitable choice for the multivariate joint distribution but it may be easy to specify univariate marginal distributions. In this case, a copula model built from the marginal distributions may be considered, as illustrated below.

Let $y_{i1}, \cdots, y_{id}$ be d repeated measurements on individual i, $i = 1, 2, \cdots, n$. In modelling these repeated measurements, we may view $\{(y_{i1}, \cdots, y_{id}),\ i = 1, 2, \cdots, n\}$ as multivariate data on d variables $y_1, \cdots, y_d$, which may be discrete or continuous. Let $F_j(y_j)$ be the cdf of the marginal distribution of variable y_j. Then, we may consider the following copula model

$$F(y_1, \cdots, y_d) = \Phi_d(\Phi^{-1}(F_1(y_1)), \cdots, \Phi^{-1}(F_d(y_d)); R),$$

where $\Phi_d(\cdot)$ denote the cdf of the d-dimension multivariate normal distribution $N_d(\mathbf{0}, R)$ with correlation matrix R, and $\Phi(\cdot)$ is the cdf of the univariate standard normal distribution $N(0,1)$. This copula model allows a wide range of dependence between variables $y_1, \cdots, y_d$, since it inherits the dependence structure of the multivariate normal distribution.

The copula method can also be used to build multivariate survival models. To illustrate, we consider the bivariate case. Let T_1 and T_2 be the times to two events of interest respectively, and let $S_1(t_1)$ and $S_2(t_2)$ be the corresponding survival functions. Then, a survival function of the bivariate survival time (T_1, T_2) can be obtained from the copula method

$$S(t_1, t_2) = C(S_1(t_1), S_2(t_2); \theta),$$

where $C(\cdot)$ is a copula function and θ is the association parameter. For example, we may consider the Weibull distribution for the univariate survival time T_j. For convenience, we reparameterize the Weibull distribution and write its survival function as

$$S_j(t_j) = \exp\left[-\left(\frac{t_j}{\eta_j}\right)^{\gamma_j}\right], \qquad j = 1, 2.$$

We may choose the Gumbel copula and obtain the following bivariate survival function

$$S(t_1, t_2) = \exp\left[-\left\{\left(\frac{t_1}{\eta_1}\right)^{\gamma_1\theta} + \left(\frac{t_2}{\eta_2}\right)^{\gamma_2\theta}\right\}^{1/\theta}\right].$$

The above method may be extended to more general cases with higher dimensions.

7.6 Example in R

We give a simple illustration in R for building a multivariate (bivariate) distribution based on known marginal distributions using a copula. Suppose that the random variable X_1 follows a normal distribution with mean $\mu = 8$ and standard deviation $\sigma = 3$, i.e., $X_1 \sim N(8, 3^2)$, and the random variable X_2 follows the exponential distribution with parameter (rate) $\lambda = 3$. In this case, there is no existing multivariate distribution for the random vector (X_1, X_2) with the known marginal distributions. However, we can build a multivariate (bivariate) distribution for (X_1, X_2) using a copula function. As an illustration, we first consider the Gumbel copula with parameter $\theta = 4$. In the following R example, we build the bivariate distribution using the Gumbel copula, simulate a sample of size $n = 100$ from this bivariate distribution, estimate the parameters in this bivariate distribution based on the maximum likelihood method (assuming the parameters are unknown), and then we compare

the parameter estimates to their true values to see how close the estimates are to the true values. We use the R package "copula" for the illustration.

```
> install.packages("copula")  # download the R package "copula"
> library(copula)

# create the two dimensional Gumbel copula object with parameter 4
> copula.g <- gumbelCopula(param=4, dim=2)
# obtain the cdf of the bivariate distribution for (X1, X2)
> my.cdf1 <- mvdc(copula.g, margins=c("norm","exp"),
                  paramMargins=list(list(mean=8,sd=3),list(rate=3)))
# simulate a sample of 100 from the bivariate distribution with cdf my.
    cdf1
> x1 <- rmvdc(my.cdf1,100)   # x1 is a 100*2 data matrix
# Find MLEs of parameters based on "data" x1 and model "my.cdf1"
> fit1 <- fitMvdc(x1, my.cdf1, c(4, 9,3,4))  # c(4,9,3,4) are starting
    values
# print fitting results
> fit1
The Maximum Likelihood estimation is based on 100 observations.
Margin 1 :
        Estimate Std. Error
m1.mean    8.115      0.267
m1.standard deviation      3.025      0.126
Margin 2 :
        Estimate Std. Error
m2.rate    2.825      0.271
Copula:
      Estimate Std. Error
param    4.741      0.471
The maximized loglikelihood is -134.8161
Optimization converged
Number of loglikelihood evaluations:
function gradient
      55       19

# contour plot of the bivariate cdf
> contour(my.cdf1, dmvdc, xlim=c(0,15),ylim=c(0,1.0))
```

From the above results, we see that the parameter estimates are close to their true values: $\hat{\mu} = 8.115$ (with standard deviation 0.267, while the true value is $\mu = 8$), $\hat{\sigma} = 3.025$ (with standard deviation 0.126, while the true value is $\sigma = 3$), $\hat{\lambda} = 2.825$ (with standard deviation 0.271, while the true value is $\lambda = 3$), and $\hat{\theta} = 4.741$ (with standard deviation 0.471, while the true value is $\theta = 4$). All the true parameter values are within 95% confidence intervals of the corresponding parameters. Note that, if the algorithm does not converge, we can choose different starting values for the parameters. Note also that, in real data analysis, $x1$ can be replaced by a real dataset. Figure 7.1 (left) shows the contour of the bivariate cdf.

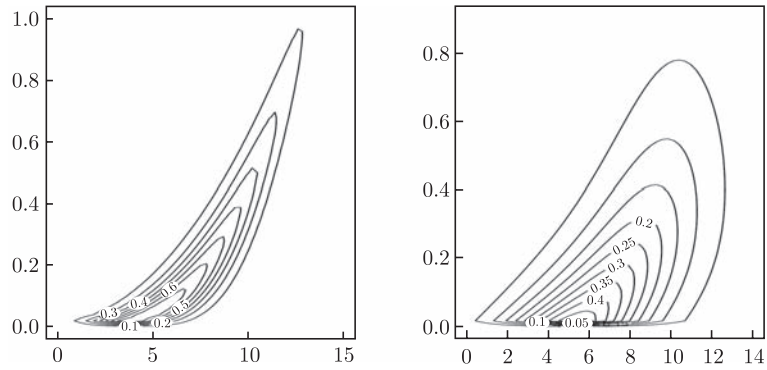


Figure 7.1 Contour plots of the bivariate cdf based on the Gumbel copula (left) and the Frank copula (right) respectively.

For comparison, we can also build the bivariate distribution using the Frank copula (with $\theta = 4$). The results are presented below.

```
# Frank copula (with theta=4)
> copula.f <- frankCopula(param=4, dim=2)
> my.cdf2 <- mvdc(copula.f, c("norm","exp"),
                  list(list(mean=8,sd=3), list(rate=3)))
# simulate a sample of 100 from my.cdf2
> x2 <- rmvdc(my.cdf2,100)
# Find MLEs of parameters based on "data" x2
> fit2<-fitMvdc(x2, my.cdf2, c(4, 9,3,4))
# print fitting results
> fit2
The Maximum Likelihood estimation is based on 100 observations.
Margin 1 :
        Estimate Std. Error
m1.mean    7.882      0.245
m1.standard deviation      2.576      0.182
Margin 2 :
        Estimate Std. Error
m2.rate    3.426      0.341
Copula:
      Estimate Std. Error
param    3.478      0.686
The maximized loglikelihood is -195.087
Optimization converged
Number of loglikelihood evaluations:
function gradient
      43       18

# contour plot of the bivariate cdf
> contour(my.cdf2,dmvdc,xlim=c(0,14),ylim=c(0,0.9))
```

We see that the Frank copula gives similar results. Figure 7.1 (right) shows the contour of the bivariate cdf. From Figure 7.1, we see that the density functions for the two bivariate distributions based on the Gumbel and Frank copula are quite different. In real data analysis, we can choose the copula function by comparing the corresponding AIC values.

Exercises 7

7.1. Find the bivariate copula based on the Clayton generator.

7.2. Find the bivariate copula based on the Gumbel generator.

7.3. Find the bivariate copula based on the Frank generator.

7.4. Spearman's correlation coefficient can be written as a function of the copula function as follows

$$\rho(X_1, X_2) = 12E((F_1(x_1) - 1/2)(F_2(x_2) - 1/2)) = 12 \iint (C(u, v) - uv)\mathrm{d}u\mathrm{d}v.$$

Find Spearman's ρ in terms of parameters from the Frank copula.

7.5. Consider a claim random variable X that, given a risk classification parameter γ, can be modelled as an exponential distribution, i.e.,

$$P(X \leqslant x|\gamma) = 1 - e^{-\gamma x}.$$

(1) Show the following well known result in credibility theory: if γ has a gamma distribution $gamma(\alpha, \lambda)$, then the marginal distribution of X is a Pareto distribution with cdf given by

$$F(x) = 1 - \left(1 + \frac{x}{\lambda}\right)^{-\alpha}.$$

(2) Suppose, conditional on the risk class γ, that X_1 and X_2 are independent and identically distributed. Assuming that they come from the same risk class γ induces a dependency. Write the bivariate distribution function as a copula.

Chapter 8

Linear and Nonlinear Regression Models

8.1 Introduction

In the analysis of multivariate data, there are two general approaches. One approach is to treat each variable *equally*, and the goal is to understand the correlation structure between the variables or to reduce the dimension of the data space. Examples include principal component analysis, factor analysis, and cluster analysis. Another approach is to treat one variable as a *response* and the other variables as *predictors*, and the goal is to understand the variation in the response that can be partially explained by predictors. Such models are called *regression models*. Regression analysis is an important component of multivariate analysis, since it allows researchers to focus on the effects of predictors on the response. Regression analysis also is widely applied in actuarial science and finance. It is a required educational component of the two main actuarial bodies in the U.S. and Canada, the Society of Actuaries and the Casualty Actuarial Society.

Regression models attempt to partially explain the variation in the response by the predictors. In other words, regression models attempt to find the approximate relationship between the response and predictors. For example, we may wish to find the approximate relationship between income (response) and education, age, experience, gender, etc (predictors). Or we may wish to find the approximate relationship between success (response) and age, education, gender, IQ score, attitude, etc (predictors). A regression model is a useful statistical tool to determine such an approximate relationship.

In practice, the true relationship between the response and predictors may be highly complicated and may not be known exactly, but an approximation to the true relationship is possible, based on the observed data. When the response variable is continuous, the simplest approximation is a linear approximation, which assumes that the response and the predictors have an approximate linear relationship. The resulting model is called a *linear regression model*. Then, we use observed data to estimate the linear relationship and hope that such a linear approximation will be satisfactory. As such, linear regression models are empirical models which only de-

scribe the observed data, without a true understanding of the underlying mechanism which generate the data. In many practice situations, linear models may provide reasonable approximations to the true relationship, even though the true relationships are unknown and complicated. When necessary, some variables may be transformed to make the linear approximations more reasonable. Therefore, linear regression models are the simplest but are also the most widely used regression models.

Nonlinear regression models, on the other hand, are typically based on the underlying mechanisms which generate the data, so derivations of nonlinear models require good understanding of the scientific problems. Thus, nonlinear models are usually closer to the true relationships than linear models, and predictions based on nonlinear models are more reliable than linear models. However, for many practical problems, it may be difficult to derive nonlinear models since the underlying data generation mechanisms may be highly complicated.

In this chapter, we briefly review both linear and nonlinear regression models, with a focus on linear models. References for linear regression models are extensive, including Draper and Smith (1998) and Weisberg (2005), among others. Interested readers can find more detailed discussions of linear models in these references.

8.2 Linear Regression Models

Suppose that data are available on $p+1$ variables $(y, x_1, x_2, \cdots, x_p)$, where variable y is chosen as a *response* based on scientific interest and the other variables are treated as *predictors* or *covariates*. If the data is a sample of size n, then the data may be denoted as $\{(y_i, x_{i1}, x_{i2}, \cdots, x_{ip}),\ i = 1, 2, \cdots, n\}$. We wish to build a regression model which describes the approximate relationship between the response and the predictors.

The simplest regression model is a linear regression model, where the response and predictors are assumed to have a linear relationship. A linear regression model can be written as follows

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, 2, \cdots, n, \tag{8.1}$$

or

$$E(y_i) = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta},$$

where y_i is the response for individual i, β_j's are unknown parameters, x_{ij} is the j-th predictor for individual i, $\mathbf{x}_i = (1, x_{i1}, \cdots, x_{ip})^{\mathrm{T}}$ is a collection of all predictors, $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_p)^{\mathrm{T}}$ is a vector of regression parameters, and ε_i's are random errors with mean zeros, $i = 1, 2, \cdots, n$. In linear model (8.1), the response and the predictors are assumed to have a linear relationship, with the unexplained variation accounted by the random error ε_i.

Linear regression models are widely used in practice because they are simple and are easy to interpret, even though they may not exactly represent the true relationship between the response and covariates. For example, regression parameter β_j may be interpreted as the effect of predictor x_j on the response y: one unit change in x_j is associated with β_j units change in y. Such a simple linear form also allows us to derive the distribution of y based on the assumed distribution of ε and study the properties of parameter estimates. Moreover, in practice, such a linear relationship assumption may be reasonable and useful, especially when some predictors are transformed.

The linear model (8.1) may be written in a more compact matrix form. Let

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Model (8.1) can then be written in a matrix form as follows

$$\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon. \tag{8.2}$$

Common assumptions for models (8.1) or (8.2) are

- the errors ε_i's are *independent*,
- the errors ε_i's have mean zero, i.e., $E(\boldsymbol{\epsilon}) = \mathbf{0}$, as well as a *constant variance* σ^2, i.e., $Var(\boldsymbol{\epsilon}) = \sigma^2 I_n$, where I_n is the $n \times n$ identity matrix,
- the errors ε_i's are normally distributed, i.e., $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I_n)$.

These assumptions are required for statistical inference. In data analysis, these assumptions must be checked to see if they are valid or not. Based on these assumptions, the (marginal) distribution of the response $\mathbf{y}$ is given by

$$\mathbf{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n). \tag{8.3}$$

Note that, in a standard regression model, the predictors $\mathbf{x}_i$ are assumed to be fixed (i.e., they are not random variables), while the response y is assumed to be a random variable.

Once a linear model is assumed for the variables, the next step is to estimate the unknown parameters and make statistical inference, based on the observed data. The *least squares method* for estimating parameters $\boldsymbol{\beta}$ is to minimize

$$Q(\boldsymbol{\beta}) = (\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

The resulting parameter estimates, called the *least square estimates*, are given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\mathbf{y}, \qquad \hat{\sigma}^2 = \frac{RSS}{n-p-1} = \frac{\mathbf{r}^{\mathrm{T}}\mathbf{r}}{n-p-1}, \tag{8.4}$$

where

$$\mathbf{r} = \mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = (y_1 - \hat{y}_1, \cdots, y_n - \hat{y}_n)^{\mathrm{T}}$$

is a vector of *residuals*, and

$$RSS = \mathbf{r}^{\mathrm{T}}\mathbf{r} = \sum_i (y_i - \hat{y}_i)^2$$

is called the *residual sum of squares*. Residuals represent the differences between the fitted values based on the assumed model and the observed values of the response, so they can be used to check if the assumed model fits the data well or not. They play an important role in model checking or model diagnostics. It can be shown that

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta},\ \sigma^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}). \tag{8.5}$$

This result can be used to construct confidence intervals and hypothesis testing for $\boldsymbol{\beta}$.

A more general method for parameter estimation in regression models is the maximum likelihood method. For linear model (8.1) or (8.2), however, the least square estimates are identical to the maximum likelihood estimates (MLEs). This may not be true for other regression models.

The *coefficient of determination* is defined as

$$R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (y_i - \bar{y})^2},$$

where $\bar{y} = \sum_i y_i/n$. It is the proportion of variation in the response that is explained by the regression or by the predictors, so it indicates the usefulness of the regression model.

8.3 Model Selection

In regression models (8.1) or (8.2), some predictors may not have significant effects on the response, so they should be removed from the models. This is called model selection or variable selection. In regression analysis, parsimonious or simple models are preferred over complex models. In other words, unimportant or non-significant predictors should be removed from the models in order to reduce the number of unknown parameters and increase the accuracies of the estimates for the remaining parameters in the models. In data analysis, we should avoid models with too many predictors or avoid large models.

In model selection or variable selection, we need to check the significance of each predictor in the model. This is equivalent to comparing a smaller model to a larger model with more predictors to see if the two models differ significantly. For example, we may compare model I with predictors x_1, x_2 to a larger model II with predictors x_1, x_2, x_3. Let Ω be a larger model and ω be a smaller model so that model ω is nested within model Ω. We can perform a hypothesis test to compare the two nested models. The null hypothesis is that the two models are not significantly different, while the alternative hypothesis is that the larger model is significantly better. A commonly used test statistic is

$$F = \frac{(RSS_\omega - RSS_\Omega)/(p-q)}{RSS_\Omega/(n-p)},$$

where p and q are the numbers of parameters in model Ω) and model ω respectively, and RSS_Ω and RSS_ω denote the residual sum of squares of model Ω) and model ω respectively. We reject the null hypothesis if $F > F_\alpha(p-q, n-p)$, i.e., the larger model is significantly better, so we should choose the larger model Ω. If we fail to reject the null hypothesis, i.e., the two models are not significantly different, we should choose the smaller model ω.

The above F-test can also be used to test the significance of a single predictor. To test a continuous predictor x_j, an alternative approach is to use the usual t-test, with the test statistic given by

$$t_j = \frac{\hat{\beta}_j}{s.e.(\hat{\beta}_j)},$$

where $s.e.(\hat{\beta}_j)$ is the standard error of the parameter estimate $\hat{\beta}_j$ associated with covariate x_j. The result is equivalent to the F-test.

A more general test for comparing nested models is the likelihood ratio test. The likelihood ratio test statistic has an asymptotic χ^2 distribution. It is a very general testing procedure and can be used to compare other nested regression models such as nonlinear and generalized linear models, while the above F-test is often used for linear regression models.

In model selection or variable selection, sometimes we need to compare several models which may or may not be nested. That is, the models to be compared may contain different sets of predictors or may have different functional forms, not necessarily that one model is nested within the other model. For example, we may compare a model with predictors x_1 and x_2 to a model with predictors x_3 and x_4. In this case, we may use some general and more commonly used model selection criteria, such as AIC and BIC (see Section 8.7). We will discuss these general criteria in Section 8.7.

Note that model selection should not completely rely on statistical criteria. We should also consider scientific issues for the problems under consideration. For example, if our main goal is to evaluate a treatment effect for a disease, we should keep the treatment variable in the model, whether it is significant or not. Moreover, the final model should also make scientific sense. Therefore, model selection is usually a compromise between statistical criteria and scientific consideration. If several models are similar based on statistical criteria, such as similar AIC values, we should choose the model which is scientifically reasonable or the model which makes sense in practice.

8.4 Model Diagnostics

Regression models are assumed, with certain assumptions, so they may not reflect the true data-generation mechanisms. Thus, in data analysis, the assumed models must be checked to see if the models fit the observed data well and if the model assumptions are satisfied. This is called model diagnostics. In model diagnostics, residuals play an important role since they measure how far away the fitted response values to the observed response values. A residual plot usually plots the residuals against the fitted response values. Other plots may also be used. These graphical methods are the basic tools for model diagnostics.

More specifically, for model diagnostics we should check the following features of the assumed model

- goodness of fit, i.e., whether the model fits the observed data well. This can be checked based on residual plots. If the model fits well, the residuals should all be close to zero without clear patterns.
- constant variance assumption. If the residual plot shows some clear patterns, such as increasing or decreasing patterns, the variance may not be constant. In this case, we may try to make a transformation on the response, such as a log-transformation, or other ways to improve the model.
- normality assumption. We can use a normal quantile-quantile (QQ) plot. If the normality assumption holds, the QQ plot should show a roughly straight line. When the normality does not hold, sometimes a transformation on the response may be a good idea.
- outliers. This can be seen from the residual plot: observations with unusually large or small residuals may be outliers.
- influential observations, i.e., observations which may have big impacts on parameter estimates but are not necessary outliers. They can be checked based on the Cook's distances: observations with unusually large Cook's distances may be influential.

When outliers or influential observations are identified, they should be removed and studied separately.

In summary, model diagnostics can generally be done informally based on graphical tools such as residual plots, QQ plots, and Cook's distance plots, but some more formal methods are also available and may be used in some cases. Note that there is no perfect model. Model diagnostics are important, but we should also consider the interpretation and simplicity of the models. In other words, we can accept a model that is reasonable, simple, and easy to interpret.

8.5 Data Analysis Examples with R

Example 1. The Current Population Survey (CPS) is used to supplement census information between census years. These data consist of a random sample of 534 persons from the CPS, with information on wages and other characteristics of the workers, including sex, number of years of education, years of work experience, occupational status, region of residence and union membership. We wish to determine whether wages are related to these characteristics. The data file contains 534 observations on 11 variables sampled from the CPS. Variable names are:

```
EDUCATION: Number of years of education.
SOUTH: Indicator variable for Southern Region
       (1=Person lives in South, 0=Person lives elsewhere).
SEX: Indicator variable for sex (1=Female, 0=Male).
EXPERIENCE: Number of years of work experience.
UNION: Indicator variable for union membership (1=Union member, 0=Not
    union member).
WAGE: Wage (dollars per hour).
AGE: Age (years).
RACE: Race (1=Other, 2=Hispanic, 3=White).
OCCUPATION: Occupational category (1=Management, 2=Sales, 3=Clerical,
            4=Service, 5=Professional, 6=Other).
SECTOR: Sector (0=Other, 1=Manufacturing, 2=Construction).
MARR: Marital Status (0=Unmarried,  1=Married)
```

In the following, we analyze the data in R using linear regression models. Note that factors (or categorical variables) such as "race" need to be made clear in R. Otherwise, the computer cannot recognize them and may treat them as numerical variables, which may lead to misleading results.

```
wage.dat <- read.table("wage.data", head=T)
attach(wage.dat)
# Categorical variables must be declared as follows
race <- factor(race, labels=c("other", "hispanic", "white"))
# categorical variable
occupation <- factor(occupation,  labels=c("management","sale",
```

```
"clerical", "service","professional","other"))   # categorical variable
sector <- factor(sector, labels=c("other","manufacturing","construction
    "))
# categorical variable

# Fit a full linear regression model
fit1 <- lm(wage ~ education+south+sex+experience+union+age+
                  race+occupation+sector+marr)
summary(fit1)
....
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)              2.2781     6.6976   0.340  0.73390
education                0.8128     1.0869   0.748  0.45491
south                   -0.5627     0.4198  -1.340  0.18070
sex                     -1.9425     0.4194  -4.631 4.60e-06 ***
experience               0.2448     1.0818   0.226  0.82103
union                    1.6017     0.5127   3.124  0.00188 **
age                     -0.1580     1.0809  -0.146  0.88382
racehispanic             0.2314     0.9915   0.233  0.81559
racewhite                0.8379     0.5745   1.458  0.14532
occupationsale          -4.0638     0.9159  -4.437 1.12e-05 ***
occupationclerical      -3.2682     0.7626  -4.286 2.17e-05 ***
occupationservice       -3.9754     0.8108  -4.903 1.26e-06 ***
occupationprofessional  -1.3336     0.7289  -1.829  0.06791 .
occupationother         -3.2905     0.8005  -4.111 4.59e-05 ***
sectormanufacturing      1.0409     0.5492   1.895  0.05863 .
sectorconstruction       0.4774     0.9661   0.494  0.62141
marr                     0.3005     0.4112   0.731  0.46523
Residual standard error: 4.282 on 517 degrees of freedom
Multiple R-squared: 0.3265,     Adjusted R-squared: 0.3056
F-statistic: 15.66 on 16 and 517 DF,  p-value: < 2.2e-16
```

The above full model, with 10 predictors, explains about 31% variation in wage. An F-test of the full model against the null model (i.e., the model without any predictors) gives a very small p-value, indicating the full model is useful. However, we see that many predictors in the full model are not significant, so some predictors may be removed from the model. In other words, we need to do model selection or variable selection. We consider a stepwise method for variable selection using R function *step*(). The stepwise method is a combination of the forward method, which adds one predictor at each step, and the backward method, which deletes one insignificant predictor at each step.

```
fit2 <- step(fit1, direction="both")
summary(fit2)
.......
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             1.97952    1.71053   1.157 0.247696
```

```
education                 0.67229    0.09904    6.788 3.10e-11 ***
south                    -0.68858    0.41504   -1.659 0.097701 .
sex                      -1.84527    0.41523   -4.444 1.08e-05 ***
experience                0.09370    0.01656    5.657 2.54e-08 ***
union                     1.51738    0.50836    2.985 0.002970 **
occupationsale           -3.97544    0.91420   -4.349 1.65e-05 ***
occupationclerical       -3.34712    0.76002   -4.404 1.29e-05 ***
occupationservice        -4.14818    0.80534   -5.151 3.68e-07 ***
occupationprofessional   -1.26791    0.72703   -1.744 0.081754 .
occupationother          -2.79902    0.75655   -3.700 0.000239 ***
Residual standard error: 4.284 on 523 degrees of freedom
Multiple R-squared: 0.3181,     Adjusted R-squared: 0.305
F-statistic: 24.39 on 10 and 523 DF,  p-value: < 2.2e-16
```

The above model (called Model 2) is smaller than the full model and contains only 6 predictors, but it can still explain about 31% variation in wage. In other words, the above smaller model is as good as the full model, so it is preferred since it contains less predictors and is thus simpler.

Note that, for categorical variables, we can test their significances using the R function *drop1*(), as shown below.

```
drop1(fit1, test="Chi")
Single term deletions
           Df Sum of Sq     RSS    AIC   Pr(Chi)
<none>                   9480.8 1570.1
education   1     10.26  9491.0 1568.7  0.447366
south       1     32.95  9513.7 1570.0  0.173481
sex         1    393.34  9874.1 1589.8 3.176e-06 ***
experience  1      0.94  9481.7 1568.2  0.818076
union       1    178.95  9659.7 1578.1  0.001578 **
age         1      0.39  9481.2 1568.1  0.881884
race        2     44.55  9525.3 1568.6  0.286020
occupation  5    641.71 10122.5 1595.1 1.523e-06 ***
sector      2     65.90  9546.7 1569.8  0.157309
marr        1      9.79  9490.6 1568.7  0.457771
```

It seems that only sex, union, and occupation are significant. So let's try to fit the linear model with these three covariates, and then we can use the R function *anova* to compare nested models using the F-test.

```
fit3 <- lm(wage ~ sex+union+occupation)
# Compare three nested models
anova(fit3, fit2, fit1)
Analysis of Variance Table
Model 1: wage ~ sex + union + occupation
Model 2: wage ~ education + south + sex + experience + union +
    occupation
Model 3: wage ~ education + south + sex + experience + union + age +
    race + occupation + sector + marr
```

```
  Res.Df      RSS      Df       Sum of Sq        F             Pr(>F)
1    526 10753.1
2    523  9599.4       3         1153.69 20.9708      7.795e-13 ***
3    517  9480.8       6          118.59  1.0778          0.3746
```

We see that Model 2 is almost as good as the full Model 3 since the p-value from the F-test is 0.374 (i.e., the two models are not significantly different), but Model 1 is significantly worse than Model 2 (very small p-value). Thus, we should choose the larger Model 2, since the additional covariates in Model 2 explain substantial extra variation than Model 1.

Finally, we should do model diagnostics for the final Model 2, using graphical tools as described earlier.

```
par(mfrow=c(2,1))
plot(fitted(fit2), resid(fit2), main="Residual Plot",
     xlab="fitted value", ylab="residuals")  # residual plots
abline(a=0,b=0)
qqnorm(resid(fit2))  # Normal QQ  plot
```

From these diagnostic plots (see Figure 8.1), we see that Model 2 does not fit the data well, since the residual plot shows some pattern (increasing trend), indicating possibly non-constant variance, and the QQ plot also shows possible non-normality since some points deviate from a straightline.

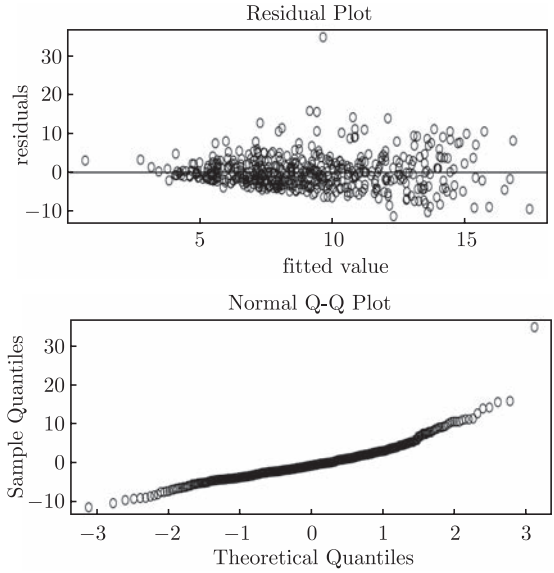


Figure 8.1 Residual plot and QQ plot for Model 2.

To improve model fitting, we can try to make a log transformation on the response and then re-fit the model with the new response. The results are shown below.

```
wage2 <- log(wage)  # a log-transformation of "wage" as the new
    response
fit4 <- lm(wage2 ~ education+south+sex+experience+union+age+race+
occupation+sector+marr)   # re-fit the model with new response
fit5<- step(fit4, direction="both")  # variable selection
summary(fit5)  # Model 5
........
Coefficients:
                          Estimate  Std.Error t value    Pr(>|t|)
(Intercept)               1.224289   0.172070   7.115    3.73e-12 ***
education                 0.068838   0.009912   6.945    1.14e-11 ***
south                    -0.102588   0.041668  -2.462    0.014139 *
sex                      -0.213602   0.041842  -5.105    4.65e-07 ***
experience                0.009494   0.001723   5.510    5.65e-08 ***
union                     0.202720   0.051009   3.974    8.06e-05 ***
occupationsale           -0.355381   0.091448  -3.886    0.000115 ***
occupationclerical       -0.209820   0.076149  -2.755    0.006068 **
occupationservice        -0.385680   0.080855  -4.770    2.40e-06 ***
occupationprofessional   -0.047694   0.072746  -0.656    0.512351
occupationother          -0.254277   0.079781  -3.187    0.001523 **
sectormanufacturing       0.111458   0.054845   2.032    0.042636 *
sectorconstruction        0.099777   0.096481   1.034    0.301541
marr                      0.065464   0.041036   1.595    0.111257
Residual standard error: 0.4283 on 520 degrees of freedom
Multiple R-squared: 0.3573,     Adjusted R-squared: 0.3412
F-statistic: 22.24 on 13 and 520 DF,  p-value: < 2.2e-16
```

The above model (called Model 5) explains about 36% variation in wage (in log-scale), so it seems better than model 2. Let's check to see if it indeed provides a better fit than model 2 for the observed data, based on model diagnostics.

```
par(mfrow=c(2,1))
plot(fitted(fit5), resid(fit5), main="Residual Plot",
     xlab="fitted value", ylab="residuals")
abline(a=0,b=0)
qqnorm(resid(fit5))
```

The diagnostic plots are shown in Figure 8.2. We see that Model 5 fits the observed data better than Model 2. Thus, we can choose Model 5 as our final model. Note that better models may exist, such as models with interaction terms or transformed predictors, but it is also reasonable to simply choose Model 5, due to its simplicity. In other words, Model 5 may not be the best model, but it is a reasonable one. From Model 5, we see that the significant predictors are education, south, sex, experience, union, occupation, sector, and marr. A smaller model may be also possible. For example, the predictor "marr" may be removed from the model.

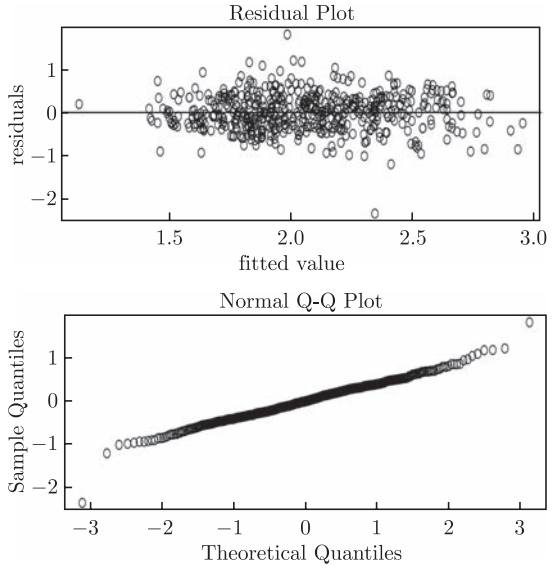


Figure 8.2 Residual plot and QQ plot for Model 5.

8.6 Nonlinear Regression Models

Linear regression models have been widely used because of their simplicity, which is an important advantage before modern computers become available. However, linear models usually only provide description of observed data, rather than trying to understand data, since they are usually chosen based on goodness-of-fit of the observed data. In other words, linear models usually provide little understanding of the data-generation mechanism. Nonlinear regression models, on the other hand, attempt to understand the mechanics of data generation, so they are often called mechanistic models or scientific models. There are some advantages of nonlinear models. First, nonlinear models may provide better predictions outside the range of observed data than that of linear models. Second, parameters in nonlinear models often have natural physical interpretations. Third, nonlinear models may require few parameters than the corresponding linear models that fit the data equally well. Note that, however, in many practical situations we do not know the data-generating mechanisms. In these cases, linear models would be good choices.

Unlike linear models, for nonlinear models there are typically no analytic or closed-form expressions for parameter estimates, so iterative algorithms are generally required obtain parameter estimates. Moreover, in fitting nonlinear models it is important to choose good *starting values* for the iterative algorithms since some likelihoods may have multiple modes.

Let y_i and $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})^{\mathrm{T}}$ be the response and predictors for individual i respectively, $i = 1, 2, \cdots, n$. A general nonlinear regression model can be written as

$$y_i = h(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_i, \qquad i = 1, 2, \cdots, n, \tag{8.6}$$

where h is a known nonlinear function, $\boldsymbol{\beta}$ is a vector of regression parameters, and ε_i is the random error. Assumptions for a standard nonlinear regression model are the same to those for a standard linear model, i.e., (i) the errors ε_i's are independent, (ii) the errors ε_i's have mean zero and constant variance σ^2, and (iii) the errors ε_i's are normally distributed.

Statistical inference for a nonlinear regression model can be based on the least squares method or the likelihood method. The ordinary least-squares estimator for parameter $\boldsymbol{\beta}$ is to minimize the sum of squares $\sum_{i=1}^{n}(y_i - g(\mathbf{x}_i, \boldsymbol{\beta}))^2$. This can be achieved by solving the following estimating equation

$$\sum_{i=1}^{n} \frac{\partial g(\mathbf{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} [y_i - g(\mathbf{x}_i, \boldsymbol{\beta})] = 0. \tag{8.7}$$

An iterative algorithm such as the Newton-Raphson method is often needed to solve the above equation.

Alternatively, under the normality assumption for the errors, i.e., ε_i i.i.d. $\sim N(0, \sigma^2)$, the MLE of $\boldsymbol{\beta}$ can be obtained by maximizing the likelihood function

$$L(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_i - g(\mathbf{x}_i, \boldsymbol{\beta}))^2}{2\sigma^2} \right].$$

So the MLE of $\boldsymbol{\beta}$ satisfies the following likelihood equation

$$\frac{\partial \log L(\boldsymbol{\beta}, \sigma^2 | \mathbf{y})}{\partial \boldsymbol{\beta}} = 0,$$

which is identical to the least-squares equation (8.7). Therefore, for nonlinear regression models the ordinary least-squares estimator of $\boldsymbol{\beta}$ is also the same as the MLE of $\boldsymbol{\beta}$, and estimation for a nonlinear regression model is analogous to that for a linear regression model.

For nonlinear regression models, analytic or closed-form expressions for parameter estimates are unavailable. However, statistical inference can still be carried out based on the standard asymptotic results of likelihood methods under the usual regularity conditions. That is, under some regularity conditions, MLEs of the model parameters are consistent, asymptotically normal, and asymptotically most efficient.

Confidence intervals and hypothesis testing can be based on the asymptotic normality of the MLEs. Therefore, with the availability of modern computers and software, statistical inference for nonlinear models does not offer much more difficulties than that for linear models.

In many cases, nonlinear models can be derived from a set of differential equations based on the understanding of the underlying data-generation mechanisms, as shown in the examples below. The developments of nonlinear models require close collaboration between statisticians and subject-area scientists, but such models may not be always available since the true data-generation mechanisms can be highly complex. Note that, in principle, any smooth nonlinear functions can be approximated by a high-order polynomial based on Taylor series expansions, if the functions are sufficiently smooth. However, high order polynomials are often unstable in replications of the data so they are generally not recommended.

Nonlinear models have been widely used in practice, such as HIV viral dynamics, pharmacokinetics, pharmacodynamics, molecular genetics, and growth or decay. More detailed discussions of nonlinear models can be found in Bates and Watts (1988), Seber and Wild (2003), and Wu (2009).

8.6.1 Example 1 (Growth curve models)

In the analysis of *growth curves*, nonlinear models are usually necessary. There are various growth curve models. Here we consider a simple monomolecular growth function. Let $y(t)$ be the size at time t (e.g., size of an animal), and let $\mu(t) = E(y(t))$. Suppose that the growth rate is proportional to the remaining size. Then $\mu(t)$ satisfies the following differential equation:

$$\frac{\mathrm{d}\mu(t)}{\mathrm{d}t} = \beta_1(\beta_0 - \mu(t)), \qquad \beta_1 > 0,$$

which can be solved analytically, with solution

$$\mu(t) = \beta_0 + \beta_2 e^{-\beta_1 t}.$$

Thus, given an observed sample, we can consider the following nonlinear regression model for estimating the parameters

$$y_{ij} = \beta_0 + \beta_2 e^{-\beta_1 t_{ij}} + e_{ij}, \qquad i = 1, \cdots, n,\ j = 1, \cdots, n_i, \tag{8.8}$$

where y_{ij} is the size for individual i at measurement time t_{ij} and e_{ij} is the corresponding measurement error. Note that, when $\beta_0 + \beta_2 = 0$ (i.e., when the initial size is 0), the above model is called the *von Bertalanffy growth curve*, which is often used in ecology to describe animal growth.

8.6.2 Example 2 (Pharmacokinetics)

Studies of *pharmacokinetics* are important in drug developments. Pharmacokinetics studies the course of absorption, distribution, metabolism, and elimination of some substance in the body over time, given drug dose, i.e., how the drug moves through the body. Suppose that a substance enters the body via ingestion. Let $y(t)$ be the concentration of the substance in the body at time t (usually measured in the blood), and let $\mu(t) = E(y(t))$. Let $\mu_0(t)$ be the amount at the absorption site (e.g., stomach). A commonly used one-compartment model is based on the following differential equations

$$\begin{aligned}\frac{\mathrm{d}\mu(t)}{\mathrm{d}t} &= \beta_1\mu_0(t) - \beta_2\mu(t),\\ \frac{\mathrm{d}\mu_0(t)}{\mathrm{d}t} &= -\beta_1\mu_0(t),\end{aligned}$$

where β_1 is the absorption rate and β_2 is the elimination rate. The above differential equations have an analytic solution given by

$$\mu(t) = \frac{\beta_1 x}{(\beta_1 - \beta_2)\beta_3}\left(e^{-\beta_2 t} - e^{-\beta_1 t}\right),$$

where x is the dose of the substance and β_3 is the volume of distribution. Therefore, given an observed sample, we can consider the following nonlinear regression model for estimating the parameters

$$\begin{gathered}y_{ij} = \frac{\beta_1 x_i}{(\beta_1 - \beta_2)\beta_3}(e^{-\beta_2 t_{ij}} - e^{-\beta_1 t_{ij}}) + e_{ij}, \qquad (8.9)\\ i = 1, \cdots, n, \quad j = 1, \cdots, n_i,\end{gathered}$$

where y_{ij} is the concentration for individual i at time t_{ij} and e_{ij} is the corresponding random error. This nonlinear model is widely used in pharmacokinetics.

8.7 More on Model Selection

In regression analysis, model selection or variable selection is an important step. In practice, there are often many potential predictors or covariates which may be included in a regression model. Too many predictors in a model may cause many potential problems such as multi-collinearity and poor parameter estimates. Generally, simple or parsimonious models are preferred, but simple models may not fit the data as well as complex models. A good model selection method should achieve a *balance* between simplicity and goodness of fit. Moreover, a model should make scientific sense. That is, a model should be meaningful in the application under consideration and should be easy to interpret.

There are many methods or criteria for model selection. Here we focus on the following commonly used methods: Akaike information criterion (AIC), Bayesian information criterion (BIC), the likelihood ratio test (LRT), and least absolute shrinkage and selection operator (LASSO). For illustration purpose, we focus on linear regression model selections, but the methods may also be used in other regression models such as nonlinear and generalized linear models.

Consider the following linear regression model with p predictors

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, 2, \cdots, n, \tag{8.10}$$

where y_i is the response for individual i, β_j's are unknown parameters, x_{ij} is the j-th covariate (predictor) for individual i, $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_p)^{\mathrm{T}}$ is a vector of regression parameters, and ε_i's are random errors. We assume that ε_i i.i.d. $\sim N(0, \sigma^2)$. Then, $y_i \sim N(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}, \sigma^2)$, and the likelihood function $L(\boldsymbol{\beta}|\mathbf{y})$ for the y_i's can be easily obtained. Let $L_n(p)$ be the maximized value of the likelihood function

$$L_n(p) = \max_{\boldsymbol{\beta}} L(\boldsymbol{\beta}|\mathbf{y}).$$

The value of *Akaike information criterion* (AIC) is given by

$$AIC = -2\log(L_n(p)) + 2p,$$

where the first term measures the goodness of fit and the second term is a penalty for the number of parameters in the model. Thus, AIC describes a tradeoff between accuracy and complexity of the model or between bias and variance. Given a set of candidate models, the model with the smallest AIC value is preferred. In other words, we can start with several plausible models, find the models' corresponding AIC values, and then choose the model with smallest AIC value.

Another commonly used and closely related criteria is called the *Bayesian information criterion* (BIC) or the *Schwarz criterion*, which is derived using Bayesian arguments. The value of BIC is given by

$$BIC = -2\log(L_n(p)) + p\log(n).$$

Given a set of candidate models, the model with the smallest BIC value is preferred. Note that BIC penalizes the number of parameters more strongly than does the AIC. It has been shown that AIC is asymptotically optimal in selecting the model with the least mean square error, while BIC is not asymptotically optimal.

Note that AIC and BIC do not provide significance tests of models. That is, the best model selected by AIC/BIC may be better than other models, but not necessarily significantly better. For *nested* models, the likelihood ratio test (LRT)

can be used to compare models with a given significance level. Suppose that we want to compare model (8.10) with a *smaller* model given by

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{iq} + \varepsilon_i \qquad i = 1, 2, \cdots, n, \tag{8.11}$$

where $q < p$. Let $L_n(p)$ and $L_n(q)$ be the maximized values of the likelihood functions of models (8.10) and (8.11) respectively. Then, the LRT statistic is given by

$$T = 2(\log(L_n(p)) - \log(L_n(q))),$$

which asymptotically follows a $\chi^2(p-q)$ distribution. The LRT allows us to compare two nested model to see if one is significantly better than the other.

There are other model selection methods, such as Mallows' C_p, false discovery rate, and cross-validation. In data analysis, a common used strategy is to use the *stepwise* method: it is a combination of a forward selection and a backward elimination procedure. In each step, the selection may be based AIC, BIC, LRT, and other model selection criteria.

Note that the foregoing model selection or variable selection methods select subsets of all predictors, so they are discrete processes, i.e., predictors are either retained or dropped from the models (in other words, regression coefficients are either zeros or non-zeros). Thus, these methods can be quite unstable. For example, small changes in the data can result in very different models being selected. Moreover, a model which fits the data well may not be good for prediction. An important criterion for evaluating a model is its prediction error. The *prediction error* (PE) for a data point $\mathbf{x}_0$ for linear model (8.10) is given by

$$\begin{aligned} \text{PE} &= E((\mathbf{y} - \hat{\mathbf{y}})^2 \mid \mathbf{x} = \mathbf{x}_0) \\ &= \sigma^2 + \text{bias}^2(\hat{\mathbf{y}}) + \text{Var}(\hat{\mathbf{x}}|\mathbf{x}_0), \end{aligned}$$

where $\hat{\mathbf{y}}$ is the fitted value. The above decomposition is known as the *bias-variance tradeoff*. As a model becomes more complex (i.e., more terms are added in the model), the coefficient estimates suffer from higher variance.

The idea of ridge regression is to regularize the coefficients (i.e., control how large the coefficients grow). That is, instead of minimizing the residuals sum of squares as for the least-square estimates, we minimize

$$\sum_{i=1}^{n} (y_i - \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta})^2 \qquad \text{such that} \qquad \sum_{j=1}^{p} \beta_j^2 \leqslant t,$$

where the predictors $\mathbf{x}_i$ are standardized and t is a tuning parameter. This is equiv-

alent to minimize the following *penalized residual sum of squares* ($PRSS$):

$$PRSS(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2 + \lambda\sum_{j=1}^{p}\beta_j^2.$$

The resulting ridge estimate is given by

$$\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}^{\mathrm{T}}\mathbf{y},$$

where $\boldsymbol{X}$ is the standardized design matrix and $\mathbf{y}$ is centered. For models with large p and small n, matrix $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ may not be invertible (or matrix $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ may be singular), while introducing parameter λ makes the problem non-singular. This is the original motivation for ridge regression. The tuning parameter may be selected using the cross-validation method.

Note that, in a linear regression model, when there are too many predictors, multi-collinearity may happen and parameter estimates may become unstable. A ridge regression circumvents this problem. It makes the parameter estimates somewhat biased, but the variances of the estimates are smaller than that of the least square estimates, and their mean square errors may also be smaller than that of the least square estimates. Thus, the idea behind ridge regression is about bias-variance tradeoff. Ridge regression is a continuous process that *shrinks* regression coefficients and are thus more stable than subset selection methods (such as AIC/BIC/LRT), but it usually does not set regression coefficients to zeros (i.e., it does not select predictors).

In recent years, the *least absolute shrinkage and selection operator* ($LASSO$) has become increasingly popular, especially for high-dimensional data (i.e., large p, small n). The LASSO combines shrinkage and selection methods for linear regression. Unlike ridge regression, small value of t in LASSO will set some coefficients exactly to 0, which performs variable selection. In other words, it shrinks some regression coefficients and sets others to 0, so it retains the good features of both subset selection and ridge regression.

Specifically, consider the linear model (8.10). Suppose that the predictors x_{ij}'s are standardized so that they have mean 0 and variance 1. The LASSO estimate $\hat{\boldsymbol{\beta}}$ is defined by

$$\hat{\boldsymbol{\beta}} = \arg\min\left\{\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p}\beta_j x_{ij}\right)^2\right\} \qquad \text{subject to} \quad \sum_j |\beta_j| \leqslant t,$$

where $t \geqslant 0$ is a tuning parameter. This is equivalent to minimize the loss function

$$PRSS(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2 + \lambda\sum_{j=0}^{p}|\beta_j|.$$

The tuning parameter t (or λ) controls the amount of shrinkage applied to the estimates. When the tuning parameter t is sufficiently large, the constraint has no effect and the LASSO solutions are just the usual least squares estimates. On the other hand, small values of t will cause shrinkage towards 0 (some regression coefficients may be exactly 0). That is, for small values of t, the solutions are shrunken versions of the least squares estimates. Therefore, choosing the tuning parameter is like choosing the number of predictors. Cross validation methods can be used to estimate the best value of t. Standard errors of the LASSO estimates can be obtained using bootstrap methods.

Unlike ridge regression, LASSO estimates have no analytic expressions. The LASSO solutions can be obtained using a quadratic programming method based on standard numerical analysis algorithms. But a better approach is the least angle regression method, which provides an efficient way to compute the LASSO solutions simultaneously for all values of t. The R package **lars** implements the LASSO. For problems with more predictors than observations, the least square method may not have solutions, but both ridge regression and the LASSO have solutions.

Finally, in model selection or variable selection, several models may have similar performances. In this case, the decision should be based on scientific considerations and simplicity or interpretations of the models.

Exercises 8

8.1. Show that the prediction error (PE) for a data point $\mathbf{x}_0$ based on a linear model can be decomposed as follows

$$\mathrm{PE}(\mathbf{x}_0) = E((\mathbf{y} - \hat{\mathbf{y}})^2 \mid \mathbf{x} = \mathbf{x}_0) = \sigma^2 + \mathrm{bias}^2(\hat{\mathbf{y}}) + \mathrm{Var}(\hat{\mathbf{y}}|\mathbf{x}_0).$$

8.2. For the linear model (8.11), find the MLEs of $\boldsymbol{\beta}$ and σ^2. Is the MLE of σ^2 unbiased? Why?

8.3. For the linear model (8.11), show that the MLE of $\boldsymbol{\beta}$ has the following sampling distribution

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta},\ \sigma^2(X^{\mathrm{T}}X)^{-1}).$$

8.4. For the linear model (8.11), derive the LRT test statistic for $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$.

8.5. The environmental Kuznets curve is a hypothesized relationship between environmental quality and economic development: various indicators of environmental degradation tend to get worse as modern economic growth occurs until average income reaches a certain point over the course of development. Table 8.1 consists of gross regional product (GDP), total volume of waste water discharge (WasWater), total volume of industrial waste gas emission (WasGas), WasSolid (volume of industrial solid wastes produced) of 31 regions in China in year 2010. Fit a quadratic linear regression model for each pollution indicator and GDP respectively. Perform model diagnostics. Do the models fit the data well? Use the analysis results to investigate if the relationships between pollutions and economic growth

is a inverted U-shaped curve. *Hint: The quadratic models here are in fact linear models since the models are linear in the parameters.*

Table 8.1 Gross reginal product and pollution indicators of 31 regions in China

multirow2*Region	GDP (100 million yuan)	WasWater (10000 tons)	WasGas (100 million cu.m)	WasSolid (10000 tons)
Beijing	14113.6	8198	4750	1269
Tianjin	9224.5	19680	7686	1862
Hebei	20394.3	114232	56324	31688
Shanxi	9200.9	49881	35190	18270
Inner Mongolia	11672.0	39536	27488	16996
Liaoning	18457.3	71521	26955	17273
Jilin	9128.6	38656	8240	4642
Heilongjiang	10368.6	38921	10111	5405
Shanghai	17166.0	36696	12969	2448
Jiangsu	41425.5	263760	31213	9064
Zhejiang	27722.3	217426	20434	4268
Anhui	12359.3	70971	17849	9158
Fujian	14790.4	124168	13507	7487
Jiangxi	9451.3	72526	9812	9407
Shandong	39169.9	208257	43837	16038
Henan	23092.4	150406	22709	10714
Hubei	16182.3	94593	13865	6813
Hunan	16038.0	95605	14673	5773
Guangdong	46013.1	187031	24092	5456
Guangxi	9569.9	165211	14520	6232
Hainan	2064.5	5782	1360	212
Chongqing	7925.6	45180	10943	2837
Sichuan	17185.5	93444	20107	11239
Guizhou	4602.2	14130	10192	8188
Yunnan	7224.2	30926	10978	9392
Tibet	507.5	736	16	11
Shaanxi	10123.5	45487	13510	6892
Gansu	4120.8	15352	6252	3745
Qinghai	1350.4	9031	3952	1783
Ningxia	1689.7	21977	16324	2465
Xinjiang	5437.5	25413	9310	3914

Source: China Year Book 2011.

8.6. For the job applicants dataset described in Chapter 2 (Example 1), we can choose "suitability (SUIT)" as a response variable and the other variables as predictors. Fit a linear regression model to the data, perform model selections, conduct model diagnostics, and report your conclusions. What variables are most predictive for suitability? Are your conclusions reasonable? Explain the results in a simple language understandable by non-statisticians.

Chapter 9

Generalized Linear Models

9.1 Introduction

Both linear and nonlinear regression models typically assume that the response variable is continuous and follows a normal distribution. Nonlinear regression models extend linear regression models by allowing nonlinear relationships between the response and predictors, but the response is still assumed to be normally distributed as in linear models. In practice, however, there are different types of response variables, and many of them are not continuous variables and are unlikely to follow normal distributions, even after variable transformations. For example, if the response is a binary variable taking only two possible values (say, male or female, pass or fail, success or failure, etc), then the response variable cannot follow a normal distribution, no matter what transformation is used. In this case, linear or nonlinear regression models cannot be used. In this section, we describe a class of regression models for which the response variables can be binary, count, continuous but skewed, and more. This class of regression models is called generalized linear models.

Generalized linear models (GLMs) extend linear models by allowing the response variable to follow distributions in the *exponential family*, which includes a wide range of commonly used distributions such as normal, binomial, and Poisson distributions. In other words, in a GLM, the response variable can be continuous, discrete, and count. The covariates or predictors still enter the model in a linear fashion, but the response and predictors are linked by a nonlinear link function. A main advantage of a GLM is that it can be used to build a regression model when the response is discrete such as "pass/fail" and "success/failure". Moreover, GLMs include linear regression models as a special case. Therefore, GLMs greatly extend the applicability and popularity of regression models.

In this chapter, we first describe GLMs in a general form. Then, we discuss the two most popular GLMs, the logistic regression models and the Poisson regression models, in greater details.

9.2 The Exponential Family

In a GLM, the response variable is assumed to follow a distribution from the exponential family. The exponential family consists of many parametric distributions which share some common characteristics, including many of the most well known distributions. It is the most popular class of parametric distributions, although it is a small subset of all parametric distributions. We briefly describe the exponential family as follows.

A distribution in the *exponential family* has the following general form for its probability density function (pdf)

$$f(y|\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}, \tag{9.1}$$

where $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are known functions, θ is called the canonical parameter representing the location, and ϕ is called the *dispersion parameter* representing the scale. It can be shown that the mean and variance of a distribution from the exponential family are given by

$$E(y) = \mu = \frac{\partial b(\theta)}{\partial \theta}, \qquad Var(y) = a(\phi)\frac{\partial^2 b(\theta)}{\partial \theta^2}.$$

The following distributions are in the exponential family: normal distribution, binomial distribution, Poisson distribution, gamma distribution, inverse Gaussian distribution, and some other distributions. We focus on the three most important ones: normal, binomial, and Poisson. For the normal distribution $N(\mu,\sigma^2)$ with mean μ, variance σ^2, and pdf

$$f(y|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right],$$

we have

$$\theta = \mu, \qquad \phi = \sigma^2,$$
$$a(\phi) = \phi, \qquad b(\theta) = \theta^2/2, \qquad c(y,\phi) = -(y^2/\phi + \log(2\pi\phi))/2.$$

For the binomial distribution with probability distribution

$$P(y=k) = \left(\begin{array}{c} n \\ k \end{array}\right)\mu^k(1-\mu)^{n-k},\ k = 0,1,\cdots,n,$$

where $0 < \mu < 1$, we have

$$\theta = \log(\mu/(1-\mu)), \qquad \phi = 1,$$
$$b(\theta) = -n\log(1-\mu), \qquad c(y,\phi) = \log\left(\begin{array}{c} n \\ y \end{array}\right),$$
$$E(y) = n\mu, \qquad Var(y) = n\mu(1-\mu).$$

Note that the binomial distribution reduces to the Bernoulli distribution when $n = 1$, i.e.,

$$P(y = k) = \mu^k(1 - \mu)^{1-k}, \qquad k = 0, 1,$$

with

$$E(y) = \mu = P(y = 1), \qquad Var(y) = \mu(1 - \mu).$$

For the Poisson distribution with probability distribution

$$P(Y = k) = (k!)^{-1}e^{-\mu}\mu^k, \qquad k = 0, 1, 2, \cdots,$$

where $\mu > 0$, we have

$$\theta = \log(\mu), \qquad \phi = 1,$$
$$a(\phi) = 1, \qquad b(\theta) = e^\theta, \qquad c(y, \phi) = -\log(y!),$$

with $E(y) = Var(y) = \mu$.

The above three distributions are the most well known distributions in the exponential family. They are also most commonly used in GLMs. They can be used to model continuous, discrete, and count response data. Thus, GLMs can cover a wide variety of practical situations where regression analysis is needed. Note that, in a regression model, the predictors or covariates are viewed as fixed (i.e., not viewed as random variables) and thus can have any types. In other words, the type of regression model is determined by the type of the response variable, not the predictors.

9.3 The General Form of a GLM

A regression model links the mean of the response to a set of predictors or covariates. The set of predictors or covariates are usually combined in a linear way, which is called the linear predictor. Specifically, let

$$\mu_i = E(y_i)$$

be the mean response. Let $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})^{\mathrm{T}}$ be a vector of predictors or covariates, $i = 1, \cdots, n$. We call the following linear combination

$$\eta_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip}\beta_p$$

the *linear predictor*, where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^{\mathrm{T}}$ is a $p\times 1$ vector of unknown parameters. In other words, the linear predictor η_i combines the predictors (covariates) in a linear form, which is the simplest and most common way to combine predictors. In a linear regression model, we link the mean response and the linear predictors as

$$E(y_i) = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}, \qquad i.e., \qquad \mu_i = \eta_i.$$

That is, for linear regression models we use the identity function $g(x) = x$ to link the mean response μ_i and the linear predictor η_i, i.e., $g(\mu_i) = \eta_i$.

The identity link function $g(x) = x$ for linear regressions, however, may not be appropriate for other types of response. For example, if y_i is a binary variable, then the mean response $\mu_i = P(y_i = 1)$, which is a number between 0 and 1, while the value of the linear predictor η_i can take any value from $-\infty$ to ∞, so we cannot use the identity link to link the mean response to the linear predictor. In other words, for binary responses we should use other link functions to link the mean response to the linear predictor in regression modelling. For example, we may consider the following link function for binary response

$$g(x) = \log\left(\frac{x}{1-x}\right).$$

Then, a regression model for the binary response y can be written as $g(\mu_i) = \eta_i$, i.e.,

$$\log\left(\frac{P(y_i = 1)}{1 - P(y_i = 1)}\right) = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}, \qquad i = 1, 2, \cdots, n,$$

or

$$P(y_i = 1) = \frac{\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})}, \qquad i = 1, 2, \cdots, n,$$

which is the well known logistic regression model – the most popular generalized linear model. As an example, suppose that $y_i = 1$ if a person gets a cancer and $y_i = 0$ otherwise. Let x_i be the smoking status. Then, the above logistic regression model can be used to study the relationship between the probability of getting a cancer if a person smokes.

A *generalized linear model* (GLM) can be written as

$$g(\mu_i) = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}, \qquad i = 1, 2, \cdots, n, \tag{9.2}$$

where $g(\cdot)$ is a monotone and differentiable function, called the *link function*. Thus, a GLM has two components:

- the response y_i follows a distribution in the exponential family;
- the link function $g(\cdot)$ describes how the mean response μ_i is related to the linear predictor η_i.

The choice of the link function $g(\cdot)$ depends on the type of the response y_i. If the response is a binary variable and is assumed to follow the Binomial distribution, the most common choice of the link function is the following link, called *logit link*,

$$g(x) = \log\left(\frac{x}{1-x}\right).$$

The logit link function offers nice interpretation: $P(y_i = 1)/(1 - P(y_i = 1))$ may be interpreted as "odds of success" (success is defined as "$y_i = 1$" here). Other link functions for binary response are also available, but their interpretations are less attractive. If the response is count and is assumed to follow the Poisson distribution, the most common choice of the link function is the log link:

$$g(x) = \log(x).$$

Both the logit link and the log link functions are called *canonical link functions*, since they can be naturally obtained from the assumed distributions written in the standard form of an exponential family (see, e.g., McCullagh and Nelder, 1989).

Since the link function in a GLM is often a nonlinear function, GLMs are special nonlinear regression models. The linear predictor is in a linear form, like in linear regression models. The nonlinear part is the link function. Note that, for nonlinear regression models, the predictors are not combined in a linear form but can be in any nonlinear forms in the models. So GLMs are essentially like linear models. In other words, GLMs are essentially "empirical models" rather than "scientific models". On the other hand, true nonlinear models are "scientific models", which are derived based on the data-generation mechanisms, and true nonlinear models usually do not contain linear predictors. Thus, GLMs are different from usual nonlinear models. Moreover, the response variable in a GLM is assumed to follow a distribution in the exponential family, while the response variable in a nonlinear model is assumed to follow the normal distribution. GLMs include linear regression models as a special case when the response distribution is normal and the link function is the identity function. In summary, GLMs are still restrictive in that they involve essentially linear models and only cover distributions from the exponential family.

9.4 Inference for GLM

A common approach for parameter estimation and inference for GLMs is to use the likelihood method. For a general GLM, the log-likelihood function is given by

$$l(\boldsymbol{\beta}, \phi) = \sum_{i=1}^{n} \left\{ \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\}.$$

Note that the regression parameters $\boldsymbol{\beta}$ is implicit in the loglikelihood function $l(\boldsymbol{\beta}, \phi)$ since

$$g(E(y_i)) = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}, \qquad E(y_i) = \partial b(\theta_i)/\partial \theta_i.$$

The likelihood equation for $\boldsymbol{\beta}$ is given by

$$\frac{\partial l(\boldsymbol{\beta}, \phi)}{\partial \boldsymbol{\beta}} = 0.$$

The resulting solution is a candidate for the MLE of $\boldsymbol{\beta}$. Since the loglikelihood $l(\boldsymbol{\beta}, \phi)$ is nonlinear in the parameters $\boldsymbol{\beta}$ and ϕ, MLEs are obtained using an iterative algorithm such as the Newton-Raphson method, which is equivalent to the iteratively reweighted least squares method described in McCullagh and Nelder (1989). Since an iterative algorithm is used, sometimes convergence of the algorithm can be an issue, such as non-convergence, especially when the observed data are poor or the model is too complex.

The MLEs of the model parameters in a GLM share the usual asymptotic properties of MLEs: they are consistent, efficient, and asymptotically normally distributed. However, for finite samples in practice, the performance of the MLEs may depend on the sample size. For example, the standard errors of parameter estimates obtained based on the asymptotic Fisher information matrix may not always be reliable for small samples. Similarly, the asymptotic normality and the likelihood ratio test (deviance test) for small samples may not always perform well. For small samples, a better approach to obtain standard errors may be the bootstrap method. In summary, we should be careful in interpreting computer outputs of the parameter estimates and their standard errors for GLMs since these results are often based on asymptotic theory which may not hold for small samples.

Statistical inference for GLMs is often based on the *deviance*, which can be defined as the difference between the log-likelihoods for the full model and for the fitted model. Consider a model A. The deviance for model A is defined as

$$D(\mathbf{y}) = -2\left[\log(f(\mathbf{y}|\hat{\theta}_A)) - \log(f(\mathbf{y}|\hat{\theta}_F))\right],$$

where $\hat{\theta}_A$ is the parameter estimate under model A and $\hat{\theta}_F$ is the parameter estimate under the full model. The *full model* (or the *saturated model*) is the most complex model where the data is explained exactly (i.e., it represents the data as being entirely systematic such as having n parameters for n data points). The *null model* is the smallest model where there is no relationship between the predictors and the response (i.e., it represents the data as being entirely random).

Note that the deviance is simply –2 times the log-likelihood ratio statistic of model A compared to the full model, so the deviance measures how close model A is to the full model (the perfect model). For linear regression models, the deviance is simply the residual sum of squares $RSS = \sum_i (y_i - \hat{y}_i)^2$, which usually measures the goodness-of-fit of the model or the discrepancy between observed data and fitted values. An alternative measure of discrepancy is the so-called *Pearson*'s χ^2 statistic defined as

$$\chi^2 = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{Var(\hat{\mu}_i)}.$$

9.5 Model Selection and Model Diagnostics

Similar to linear regression models, for GLMs we should also do model selection or variable selection in data analysis. The basic ideas and methods for model selection of GLMs are similar to those for linear regression models, such as the criteria of AIC, BIC, and LRT. Models with small values of AIC or BIC are preferred, but AIC or BIC do not test the significance of a model over an alternative. Hypothesis tests can be used to compare models with significance. Two types of hypothesis tests are often considered:

- *Goodness of fit test*: test if the current model fits the observed data well by comparing the current model with the full model;
- *Compare two nested models*: compare a smaller model with a larger model.

These two types of test are described in more details below.

For goodness of fit tests, under some regularity conditions, the scaled deviance $D(\mathbf{y})/\phi$ and the Pearson's χ^2 statistics are both asymptotically distributed as the $\chi^2(d)$ distribution, where d is the number of parameters in model A. Note that, for binary data, this χ^2 approximation is poor. When comparing two *nested* models, under the null hypothesis of no difference between the two models, the difference in the deviances of the two models asymptotically follows the $\chi^2(d)$ distribution, with degrees of freedom d being the difference of the number of parameters in the two models being compared, i.e.,

$$D_S - D_L \to \chi^2(d), \qquad \text{as } n \to \infty,$$

where D_S and D_L are the deviances for the small model and the large model respectively. This result can be used for model comparison and model selection. For example, if the p-value is large (say larger than 0.05), we prefer the smaller model since there is no significant difference between the two models. When the p-value is small (say, less than 0.05), we prefer the larger model since the larger model fits significantly better than the smaller model. Note that the above χ^2 approximations are more accurate when comparing nested models than for the goodness of fit statistic. However, in both cases, the results are only approximate, and model selection should be combined with scientific interpretation of the selected model.

For testing the significances of individual predictors in a model, we can also consider a Wald-type test. Suppose that we wish to test the significance of predictor x_j, i.e., testing the hypotheses $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$. The test statistic is given by

$$z = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \sim N(0,1), \qquad \text{asymptotically under } H_0.$$

So H_0 is rejected (i.e., predictor x_j is significant) if the p-value is small (say, less than 0.05). Since the χ^2 tests and the Wald-type tests are approximate, it is generally desirable to consider both types of test and compare the results to see if they agree or not, rather than relying on a single test. Generally, the deviance test is preferred.

9.5.1 Model Diagnostics

In model diagnostics for regression models, residuals play an important role. In linear regression models or Gaussian models, the residuals are defined as $r_i = y_i - \hat{\mu}_i,\ i = 1, 2, \cdots, n$, which are called *response residuals* in GLMs. For most GLMs, however, these residuals are not appropriate since the variance of the response is not constant so any patterns in the residual plots do not necessarily indicate lack of fit of the models. In other words, for some common distributions in the exponential family, such as the binomial distribution and the Poisson distribution, the variances depend on the mean parameters, while this is not the case for normal or Gaussian distributions in which the variance parameter σ is independent of the mean parameter μ. Thus, for most GLMs, the usual response residuals may not be appropriate. We need some modifications when defining the residuals for GLMs.

For GLM model diagnostics, we can use the following *Pearson residuals*

$$r_{ip} = \frac{y_i - \hat{\mu}_i}{\sqrt{Var(\hat{\mu}_i)}}, \qquad i = 1, 2, \cdots, n,$$

which are usual residuals scaled by the standard deviations. Note that $\sum_i r_{ip}^2$ is simply the familiar Pearson statistic. Another type of residuals, called the *deviance residuals* r_{iD}, are defined such that the deviance may be written as the sum of r_{iD}^2, i.e.,

$$\text{Deviance} = \sum_{i=1}^{n} r_{iD}^2 \equiv \sum_{i=1}^{n} d_i.$$

Thus, we can write

$$r_{iD} = sign(y_i - \hat{\mu}_i)\sqrt{d_i}, \qquad i = 1, 2, \cdots, n.$$

Residuals allow us to check the goodness-of-fit of the models. Influential observations, on the other hand, are observations which have big effects on the resulting estimates of model parameters. In other words, removing a few influential observations may lead to big changes in parameter estimates. Thus, in data analysis influential observations should be removed and studied separately. In GLMs, influential observations may be checked using the *Cook's distances*, which are defined in a similar way to that in linear regression models. They are essentially the changes

in model parameter estimates when the corresponding observations are removed in model fitting. Thus, observations with large values of Cook's distances may be influential.

When fitting a GLM to a dataset, model diagnostics often include the following: (i) check if there are any outliers or influential observations, and if so compare analysis without these observations to analysis with these observations; (ii) check whether the structure form of the model is reasonable, which includes choices of predictors and possible transformations of the predictors; (iii) check whether the stochastic part of the model is reasonable, such as the distributional assumptions or the nature of the variance. Common diagnostic methods include

- residual plots: we can plot the deviance residuals against the estimated linear predictor $\hat{\eta}_i$. This may help the choice or transformations of the predictors, but for a binary response the residual plot is not very useful due to the nature of the response data;
- transformation of the predictors: polynomial terms or interaction terms or other transformations may be considered;
- Cook's distance plot: check influential observations.

For data analysis using GLMs, an important component of model diagnostics is to check the so-called *over-dispersion problem* described below. This problem does not exist for linear regression models but is common for GLMs, due to possible relationship between mean and variance in GLMs. For Gaussian models such as linear or nonlinear regression models, there is no over-dispersion problem because the mean and variance are independent in normal distributions.

9.5.2 Over-Dispersion Problem

For some most common distributions in the exponential family, such as the binomial distribution and the Poisson distribution, the variance is determined by the mean. For example, if y_i follows a Bernoulli distribution, we have

$$Var(y_i) = E(y_i)(1 - E(y_i)),$$

and if y_i follows a Poisson distribution, we have

$$Var(y_i) = E(y_i).$$

This is very different from a linear regression model where the response is assumed to a normal distribution in which the variance is unrelated to the mean. That is, if y_i follows a normal distribution $N(\mu, \sigma^2)$, the mean μ and the variance σ^2 are independent and both can vary freely, which allows great flexibility in modelling real data. On the other hand, for Binomial and Poisson distributions, the strong

relationship between the mean and variance is very restrictive in practice since the variation in the data may *not* agree with the theoretical variance which is determined by the mean. In other words, in a GLM we assume a mean structure $g(E(y_i)) = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}$ and assume that y_i follows a distribution in the exponential distribution, but the observed variation in the data may be different from the theoretical variance obtained from the assumed distribution. For example, for the Poisson regression model, the theoretical variance is the same as the mean, which is $E(y_i) = g^{-1}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})$, but the variation in the data may be much larger or much smaller than the theoretical variance, so the assumed model is inappropriate.

If the variation in the data is larger (or smaller) than the theoretical variance determined by the assumed distribution, the problem is called an *over-dispersion* (or a *under-dispersion*) problem. When over-dispersion or under-dispersion problem arises in data analysis, the assumed distribution for the GLM does not hold, so this problem must be addressed for correct inference. Usually over-dispersion problems are more common than under-dispersion problems. Overdispersion problems can arise in longitudinal or clustered data if the correlation within clusters are not incorporated in the models. One way to address the over-dispersion problem is to specify the mean and variance functions separately in a GLM, without a distributional assumption. This approach is called the *quasi-likelihood* method, which is closely related to the generalized estimating equation (GEE) method. In data analysis using a statistical software, we may consider the "quasi-likelihood" option and compare the results with that from a standard GLM fit to see if the results agree. Alternatively, we can address the over-dispersion problem by introducing a dispersion parameter.

9.6 Logistic Regression Models

Logistic regression models are perhaps the most widely used models in the GLM family. A logistic regression model is used when the response y is a binary variable taking only two possible values (say, 0 or 1), which is very common in practice such as success/failure, cancer/health, death/alive, etc. In this case, it may be reasonable to assume that y follows a binomial or Bernoulli distribution.

Recall that the main idea of a regression model is to link the mean response to a set of predictors. When the response y is a binary variable, the mean response is $E(y) = P(y = 1)$, which is a value between 0 and 1, while the predictors may take any real values. Thus, a link function making the data ranges on both sides consistent is needed. Moreover, the link function should make interpretation easy. There are several choices for the link function. The most popular choice of the *logit link*

$$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right),$$

where $\mu = E(y) = P(y = 1)$. With the logit link, $g(\mu)$ can take any real value. The resulting GLM is the following *logistic regression model*

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip}\beta_p, \qquad i = 1, \cdots, n, \tag{9.3}$$

where $\mu_i = E(y_i) = P(y_i = 1)$, and y_i is assumed to follow a Binomial distribution with mean μ_i.

A main advantage of the logistic regression model is its attractive interpretation: $\mu_i/(1-\mu_i)$ can be interpreted as the *odds* of event "$y_i = 1$", so the parameter (say) β_j may be interpreted as the change of odds in log-scale when predictor x_j is changed by 1 unit. Other link functions for binary responses include probit link and complementary log-log link, but they do not have the attractive interpretation as the logit link.

A logistic regression model is used when the response is a binary variable, which is very different from a continuous response variable as in a linear regression model. For a binary variable, the two values of the response variable represent two categories rather than real values. Thus, residuals are not well defined for logistic regression models. In other words, residual plots are not very useful for checking the goodness-of-fit of logistic regression models.

Model selection or variable selection for logistic regression models can be based on the χ^2 test of deviances, which is similar to the likelihood ratio test. Note that, for binary data, the deviance does not assess goodness of fit and it is not approximately χ^2 distributed. Thus, bootstrap methods may be preferred. For comparing two nested models, however, the χ^2 test based on the difference of the two deviances is still reasonable for binary data. For testing the significance of a single continuous predictor, say testing $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$, the usual Wald-type z-test based on

$$z = \hat{\beta}_j / se(\hat{\beta}_j)$$

can still be used, although the deviance test is preferred. Other common model selection methods, such as AIC or BIC criteria, can also be used for model or variable selections. That is, models with small values of AIC or BIC are preferred.

Example 1. Consider the "wage data" described in the previous chapter again. Suppose that hourly wages below \$7.78 are considered to be "low" and hourly wages above \$7.78 are considered to be "high" (7.78 is the median wage). We wish to determine whether low/high wages are related to the variables in the dataset. This is an alternative way to study the relationship between wage and other variables, and

the results may be compared with the linear regression models in the previous chapter. Note that, by converting continuous data to binary data, some information may be lost, but binary data sometimes have more attractive and simpler interpretation such as high or low wages rather than the actual wage values.

```
wage.dat <- read.table("wage.data", head=T)
attach(wage.dat)
## For categorical variables, the function "factor" should be used to
    identify them
race <- factor(race, labels=c("other", "hispanic", "white"))   #
    categorical variable
occupation <- factor(occupation,  labels=c("management","sale",
    "clerical","service","professional","other"))
    # categorical variable
sector <- factor(sector, labels=c("other","manufacturing",
                "construction"))   # categorical variable
# We convert the original response "wage" into a binary variable
# based on whether "wage>7.78".
wage3 <- as.numeric(wage>7.78)
```

Note that categorical variables "race, occupation, sector" need to be declared in R using the **factor()** function so that they will not be treated as numerical variables by computer. When a continuous variable (such as "wage") is converted into a binary variable, some information will be lost, but it sometimes gains easier interpretation in practice (such as high or low wage).

```
# Fit a logistic regression model with all predictors
fit1.glm <- glm(wage3 ~ education+south+sex+experience+union+age+
                        race+occupation+sector+marr, family=binomial)
summary(fit1.glm)
.......
Coefficients:
                        Estimate  Std.Errorz    value      Pr(>|z|)
(Intercept)              16.8137   803.1175    0.021      0.983297
education                 3.7579   133.8528    0.028      0.977603
south                    -0.5260     0.2323   -2.265      0.023528 *
sex                      -0.8326     0.2350   -3.543      0.000395 ***
experience                3.5065   133.8528    0.026      0.979101
union                     1.2282     0.2911    4.219      2.45e-05 ***
age                      -3.4642   133.8528   -0.026      0.979352
racehispanic             -0.6049     0.5864   -1.032      0.302259
racewhite                 0.1849     0.3202    0.577      0.563776
occupationsale           -1.4599     0.5110   -2.857      0.004277 **
occupationclerical       -0.5582     0.4173   -1.338      0.181032
occupationservice        -1.3227     0.4532   -2.918      0.003519 **
occupationprofessional   -0.3700     0.4201   -0.881      0.378434
occupationother          -1.0871     0.4442   -2.447      0.014395 *
sectormanufacturing       0.5849     0.3028    1.931      0.053438 .
sectorconstruction        0.8012     0.5295    1.513      0.130261
```

```
marr                    0.3883     0.2251   1.725     0.084518 .
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 740.27  on 533  degrees of freedom
Residual deviance: 577.33  on 517  degrees of freedom
AIC: 611.33
```

Note that, for a categorical variable with k categories, there are $k-1$ estimates, with each estimate being the contract between that category and the baseline category (the R default is the first category). In this case, we can test the significance of whole categorical variable using R function drop1(), which compares nested models with one variable being dropped at each time.

```
drop1(fit1.glm, test="Chi")
Single term deletions
           Df Deviance         AIC          LRT       Pr(Chi)
<none>          577.33      611.33
education   1   580.57      612.57       3.2365      0.0720146 .
south       1   582.52      614.52       5.1828      0.0228115 *
sex         1   590.21      622.21      12.8727      0.0003334 ***
experience  1   579.19      611.19       1.8565      0.1730255
union       1   596.54      628.54      19.2022      1.176e-05 ***
age         1   579.00      611.00       1.6642      0.1970358
race        2   579.99      609.99       2.6526      0.2654604
occupation  5   592.81      616.81      15.4761      0.0085106 **
sector      2   582.39      612.39       5.0559      0.0798228 .
marr        1   580.32      612.32       2.9812      0.0842359 .
```

Thus, variables "south, sex, union, and occupation" seem to be highly predictive for "low/high wage" when they are tested *individually*. Next, let's do a variable selection based on AIC values using a stepwise method to see which variables will be selected simultaneously.

```
# Stepwise method for variable selection
fit2.glm <- step(fit1.glm)
 summary(fit2.glm)
......
Coefficients:
                         Estimate Std. Error z value     Pr(>|z|)
(Intercept)             -3.945607   0.984340   -4.008     6.11e-05 ***
education                0.301113   0.057976    5.194     2.06e-07 ***
south                   -0.570361   0.229312   -2.487     0.012873 *
sex                     -0.816617   0.233073   -3.504     0.000459 ***
experience               0.042208   0.009787    4.313     1.61e-05 ***
union                    1.190237   0.287789    4.136     3.54e-05 ***
occupationsale          -1.419285   0.506888   -2.800     0.005110 **
occupationclerical      -0.559147   0.413983   -1.351     0.176807
occupationservice       -1.321571   0.450408   -2.934     0.003344 **
occupationprofessional  -0.344214   0.416351   -0.827     0.408385
occupationother         -1.043318   0.440786   -2.367     0.017936 *
```

```
sectormanufacturing      0.581945   0.301277   1.932      0.053409 .
sectorconstruction       0.844982   0.529516   1.596      0.110542
marr                     0.389090   0.223787   1.739      0.082095 .
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 740.27  on 533  degrees of freedom
Residual deviance: 581.72  on 520  degrees of freedom
AIC: 609.72
```

We see that the stepwise method based on AIC values only removes variables "age" and "race" from the original model. Let's compare the two nested models using the deviance test to see if they are significantly different.

```
# Compare the two nested models using deviance test
anova(fit1.glm, fit2.glm)
Analysis of Deviance Table
Model  Resid.Df Resid.Dev Df  Deviance
1           517    577.33
2           520    581.72  3    4.3823
```

We see that the difference in deviances from the two models is 4.3823, and the two models differ by 3 parameters (3 degrees of freedom). Compared with a $\chi^2(3)$-distribution 5th-percentile critical value, we see that models 1 and 2 are not significantly different (at 5% level), and the p-value of the χ^2 test is 0.22. Thus, we should choose the smaller model, which is model 2.

Note that model selections or variable selections often involve comparing nested models. R function anova() can also be used to test each covariate sequentially, i.e., it can compare models by dropping/adding one covariate at a time using a χ^2 test. This approach is especially helpful for categorical covariates since the Wald-type z-tests may not be good choice for categorical variables.

```
anova(fit2.glm, test="Chi")
Analysis of Deviance Table
Terms added sequentially (first to last)
           Df  Deviance  Resid.Df Resid. Dev     P(>|Chi|)
NULL                         533     740.27
education   1   49.794        532     690.48     1.707e-12 ***
south       1   10.001        531     680.48      0.001564 **
sex         1   16.838        530     663.64     4.072e-05 ***
experience  1   38.594        529     625.05     5.218e-10 ***
union       1   18.576        528     606.47     1.632e-05 ***
occupation  5   16.282        523     590.19      0.006083 **
sector      2    5.442        521     584.75      0.065803 .
marr        1    3.029        520     581.72      0.081773 .
```

From the above results, we see that covariates education, south, sex, experience, union, and occupation seem significant when tested sequentially, while covariates sector and marr are not.

Influential observations can be checked using R function cooks.distance(), as shown in Figure 9.1.

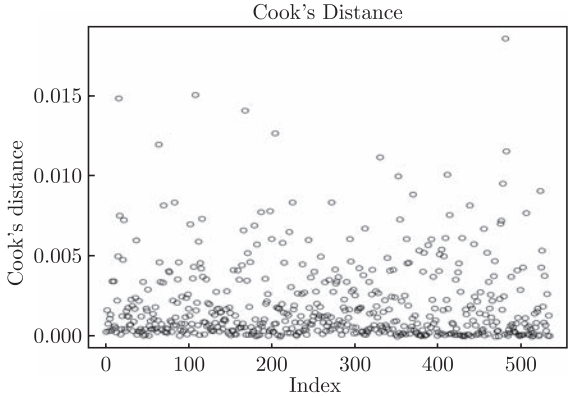


Figure 9.1 Cook's distance plot.

```
cooks.distance(fit2.glm)
           1            2            3            4            5
3.082100e-04 1.651538e-03 8.009031e-04 3.612468e-04 1.119035e-03
           6
1.380721e-03
.....

plot(cooks.distance(fit2.glm), ylab="Cook's distance", main="Cook's
    Distance")
```

There is one observation which may be influential. We performed an analysis with this observation removed, but we obtained similar results as that based on all observations. Thus, there seem no obvious influential observations for this dataset.

Recall that this dataset was also analyzed based on linear regression models in previous chapter, where a similar set of significant covariates was selected, but the interpretations of the regression coefficients are different for linear regressions and logistic regressions. These results indicate that the selected covariates may be related to wage, whether wage is viewed as a continuous variable or a discrete variable.

The final selected model should not just be based on statistical criteria such as AIC values or χ^2 tests. We should also take into account scientific considerations and interpretations of the results, especially when choosing several models with similar AIC values.

9.7 Poisson Regression Models

If the response y is a count, it may be reasonable to assume that y follows a Poisson distribution. Then, the Poisson GLM is a natural choice. For the Poisson GLM, the standard link function is the following log-link

$$g(\mu) = \log(\mu),$$

where $\mu = E(y)$. The resulting GLM is called a *Poisson GLM* and it can be written as follows

$$\log(\mu_i) = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip}\beta_p, \tag{9.4}$$

where $\mu_i = E(y_i)$, and y_i is assumed to follow a Poisson distribution with mean μ_i. The regression parameters in the Poisson GLM have attractive interpretation: parameter (say) β_j represents the change in the mean response (in log-scale) when covariate x_j is changed by 1 unit, which is similar to that for linear models but in log-scale.

When the response is a count, it may be reasonable to assume that the count follows a Poisson distribution and thus the Poisson GLM is a natural choice. However, in practice, the observed count data do not necessarily follow a Poisson distribution. In other words, the Poisson distributional assumption may not hold in practice. This can be seen when the variation in the observed data is much larger or smaller than the mean value, since the variance and the mean should be the same if the distribution is Poisson. That is, when over-dispersion or under-dispersion problems arise in a given situation, the count data do not follow a Poisson distribution.

We focus on the over-dispersion problem since it's more common. Over-dispersion arises when the observed variance in the response data is greater than the theoretical variance $Var(y_i) = \mu_i$. If an over-dispersion problem exists, the parameter estimates based on an assumed Poisson GLM will still be consistent, but the standard errors will be wrong. Thus, the overdispersion problem must be addressed in order for the statistical inference to be valid. If there is an over-dispersion problem, we can introduce a dispersion parameter ϕ such that $Var(y) = \phi E(y)$. This dispersion parameter can be estimated from the data. We will illustrate this in the example below. Where there is an over-dispersion problem, an alternative model of choice is the *negative binomial* GLM, rather than Poisson GLM.

To check the goodness of fit of a Poisson model, we can check the deviance against a χ^2 distribution. For comparing two nested models, we can also use a χ^2 distribution based on the difference of the deviances of the two models. For Poisson models, an alternative goodness of fit measure is the Pearson's χ^2 statistic

$$\chi^2 = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i},$$

which is like scaled (squared) differences between observed counts and estimated counts based on the assumed model, so a large value indicates a poor fit of the assumed model.

The above model diagnostic methods are illustrated in the example below.

Example 2. The dataset contains 1681 householders in Copenhagen who were surveyed on the type of rental accommodation they occupied, the degree of contact they had with other residents, their feeling of influence on apartment management and their level of satisfaction with their housing conditions. These correspond to the following variables: Type, Contact, Influence, and Satisfaction. Satisfaction (Sat) and Influence (Infl) are three-level categorical variables: Low, Medium, High. Contact (Cont) is a 2-level categorical variable: Low and High. Type has 4 categories: Tower, Apartment, Atrium, Terrace. We model Frequencies (counts) using a Poisson GLM and relate the counts to the variables of interest.

```
> library(MASS)
> attach(housing) # the dataset is available in the R MASS library.
# partial data is shown below
> housing
      Sat    Infl       Type Cont Freq
1     Low     Low      Tower  Low   21
2  Medium     Low      Tower  Low   21
3    High     Low      Tower  Low   28
4     Low  Medium      Tower  Low   34
5  Medium  Medium      Tower  Low   22
...........

# Fit a Poison GLM with 4 categorical covariates
> house.glm0 <- glm(Freq~Infl+Type+Cont+Sat, family=poisson)
> summary(house.glm0)
.....
Coefficients:
             Estimate Std. Error z value     Pr(>|z|)
(Intercept)   3.03545    0.06576  46.160      < 2e-16 ***
InflMedium    0.04978    0.05579   0.892      0.37226
InflHigh     -0.46206    0.06424  -7.193     6.34e-13 ***
TypeApartment 0.64841    0.06170  10.509      < 2e-16 ***
TypeAtrium   -0.51500    0.08176  -6.299     2.99e-10 ***
TypeTerrace  -0.36745    0.07817  -4.701     2.59e-06 ***
ContHigh      0.30575    0.04935   6.195     5.82e-10 ***
Sat.L         0.11592    0.04038   2.871      0.00409 **
Sat.Q         0.26292    0.04515   5.824     5.76e-09 ***
(Dispersion parameter for Poisson family taken to be 1)
    Null deviance: 833.66  on 71  degrees of freedom
Residual deviance: 295.35  on 63  degrees of freedom
AIC: 654.32
```

The above results show that all covariates are highly significant. Note that, for categorical covariates, the baseline category is not shown. The estimate for each category of a categorical variable is relative to the baseline category, which by software default is the first category (this default can be changed by users).

For a Poisson GLM, if the assumed Poisson distribution holds, the mean should be equal to the variance. Let's check to see if this is consistent with the observed data by plotting the estimated variance in the data against the estimated mean.

```
# plot estimated mean versus estimated variance (in log-scale)
> plot(log(fitted(house.glm0)), log((Freq-fitted(house.glm0))^2),
     xlab=expression(hat(mu)), ylab=expression((y-hat(mu))^2),
     xlim=c(0,8),ylim=c(-8,8))
> abline(0,1)
# We can estimate the dispersion parameter by
dp <- sum(residuals(house.glm0,type="pearson")^2)/house.glm0$df.res
> dp
[1] 4.85598
```

Figure 9.2 shows the plot. From this plot, we see that the estimated variance is much larger than the estimated mean (both in log-scale), so there is an over-dispersion. The estimated dispersion parameter is 4.85, which is much larger than 1, so it confirms the over-dispersion problem. Therefore, the assumed Poisson distribution for the Poisson GLM does not hold. This over-dispersion problem must be addressed for reliable inference.

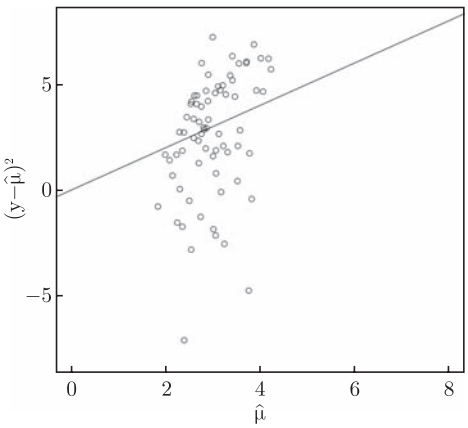


Figure 9.2 Estimated mean (in log-scale) versus estimated variance (in log-scale). Over-dispersion is obvious.

The over-dispersion problem will only affect the standard errors of the parameter estimates in the GLM, so for correct inference of the parameters we should adjust

the standard errors. That is, the parameter estimates for the GLM are still consistent even if there is overdispersion, but the standard errors may be incorrect. The following results give more reliable standard errors by incorporating the dispersion parameter.

```
# Introducing a dispersion parameter "dp" to correct standard errors
    and p-values
> summary(house.glm0, dispersion=dp)
........
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.03545    0.14491  20.947  < 2e-16 ***
InflMedium    0.04978    0.12294   0.405  0.68555
InflHigh     -0.46206    0.14156  -3.264  0.00110 **
TypeApartment 0.64841    0.13597   4.769 1.85e-06 ***
TypeAtrium   -0.51500    0.18016  -2.859  0.00426 **
TypeTerrace  -0.36745    0.17225  -2.133  0.03291 *
ContHigh      0.30575    0.10875   2.811  0.00493 **
Sat.L         0.11592    0.08898   1.303  0.19266
Sat.Q         0.26292    0.09949   2.643  0.00822 **
(Dispersion parameter for poisson family taken to be 4.85598)
    Null deviance: 833.66  on 71  degrees of freedom
Residual deviance: 295.35  on 63  degrees of freedom
AIC: 654.32
```

The above standard errors, and thus the corresponding z-values and p-values, are more reliable than the original ones since the dispersion problem is addressed by introducing a dispersion problem.

For categorical predictors, we can test the significance of each predictor by the F-test as follows, which compares nested models by dropping one predictor at a time.

```
> drop1(house.glm0, test="F")
Single term deletions
       Df Deviance      AIC F value     Pr(F)
<none>      295.35  654.32
Infl    2   373.87  728.84  8.3740 0.0005958 ***
Type    3   671.65 1024.62 26.7556 2.797e-11 ***
Cont    1   334.18  691.15  8.2831 0.0054578 **
Sat     2   340.01  694.98  4.7628 0.0118512 *
```

We can see that all the categorical predictors are significant at 5% level. So all the predictors may be included in the Poisson GLM.

Exercises 9

9.1. For a random variable Y following a distribution in the exponential family, prove that

$$E(Y) = \mu = \partial b(\theta)/\partial\theta, \qquad Var(Y) = a(\phi)\partial^2 b(\theta)/\partial\theta^2.$$

9.2. Show that, for linear regression models, the deviance is simply the familiar residual sum of squares $RSS = \sum_i (y_i - \hat{\mu}_i)^2$.

9.3. In Example 1, the Cook's distance plot shows that there is one observation which may be influential. Fit a GLM model without this observation and compare the results obtained in Example 1.

9.4. In biochemistry, the Michaelis-Menten model is the one of the simplest and best-known approaches to enzyme kinetics. The model takes the form of an equation describing the rate of enzymatic reactions by relating reaction rate y to x which is the concentration of a substrate:

$$y = \frac{\alpha x}{\beta + x},$$

where α and β are unknown parameters. What transformation will convert the above model to a linear model?

9.5. In data analysis, it is desirable to fit different models to the same dataset and compare the results. This is because each model has its own assumptions which may not hold, so we should not rely conclusions on a single model. Moreover, different models allow us to see the problem in different ways, which may give us additional insights. If different models lead to the same or similar conclusions, we are more confident about these conclusions than those based on a single model. In the following, you are asked to analyze a dataset using different models.

The dataset contains crime-related and demographic statistics for 47 US states in 1960. The data were collected from the FBI's Uniform Crime Report and other government agencies to determine how the variable "crime rate" depends on the other variables measured in the study. Variable names and definitions for the dataset are as follows (in the order from left to right in the dataset):

1. R: Crime rate: number of offenses reported to police per million population
2. Age: The number of males of age 14-24 per 1000 population
3. S: Indicator variable for Southern states (0 = No, 1 = Yes)
4. Ed: Mean number of years of schooling x 10 for persons of age 25 or older
5. Ex0: 1960 per capita expenditure on police by state and local government
6. Ex1: 1959 per capita expenditure on police by state and local government
7. LF: Labor force participation rate per 1000 civilian urban males age 14-24
8. M: The number of males per 1000 females
9. N: State population size in hundred thousands
10. NW: The number of non-whites per 1000 population
11. U1: Unemployment rate of urban males per 1000 of age 14-24
12. U2: Unemployment rate of urban males per 1000 of age 35-39
13. W: Median value of transferable goods and assets or family income (in tens of dollars)
14. X: The number of families per 1000 earning below 1/2 the median income

The main objective is to determine which variables are significantly predictive for the crime rate. Please use three different regression models to answer this question.

a) Fit a linear regression model to the crime rate data. What variables are most predictive for the crime rate?

b) A crime rate may be viewed as "high" if it is above 95 and "low" otherwise (a different cutoff value, such as 100, may be used if convergence is a problem). Fit a logistic regression model to the data. What variables are most predictive for a "high" crime rate?

c) Round off crime rate numbers to the nearest integers and then fit a Poisson GLM to the new crime rate data (this is roughly OK here since these crime rates may be viewed as counts). What variables are most predictive for the crime rate?

d) Compare the results from 1) – 3), and comment on what you find. What do you learn from the analysis? What is your final conclusion?

9.6. Likelihood method is the standard approach for parameter estimation and inference for GLMs. MLEs have nice asymptotic properties such as consistency, efficiency, and asymptotic normality. Standard errors of the MLEs are based on asymptotic formulae. In practice, however, the sample size may not be large enough, which implies that the asymptotic results may not hold exactly and the estimates may not be optimal. In this assignment, you are asked to evaluate the performances of MLEs for logistic regression models via a simulation study.

Consider the following simple logistic regression model

$$\text{logit}\left(\frac{P(y_i=1)}{1-P(y_i=1)}\right)=\beta_0+\beta_1 x_i, \qquad i=1,\cdots,n.$$

MLEs of the parameters β_0 and β_1 are available in most software, such as the "glm" function in R. To evaluate the performances of the MLEs when the sample size n is not large, we can conduct a simulation study as follows: Assume $n=20, \beta_0=1, \beta_1=2$, $x_i \sim N(0,1)$, and y_i follows a Binomial (Bernoulli) distribution. Generate 1000 datasets $\{(x_i,y_i), i=1,2,\cdots,n\}$ from the assumed distributions. For each generated dataset, obtain the MLEs of β_0 and β_1. Then compute the average biases and mean-square-errors (MSE) of the MLEs by comparing them with the true values $\beta_0=1, \beta_1=2$ over the 1000 replications (so, e.g., bias of $\hat{\beta}_1=\sum_{j=1}^{m}(\hat{\beta}_{1j}-\beta_1)/m$, MSE of $\hat{\beta}_1=\sum_{j=1}^{m}(\hat{\beta}_{1j}-\beta_1)^2/m$, where $\hat{\beta}_{1j}$ is the MLE of β_1 from the j-th iteration, $m=1000$). Repeat the above process again with $n=40$. Summarize your results in tables (or figures), and report your conclusions. Do the performance of MLE improves as n increases?

Chapter 10

Multivariate Regression and MANOVA Models

10.1 Introduction

In a standard linear regression model or an analysis of variance (ANOVA) model, there is only one continuous response variable, and that response variable is then linked to multiple predictors or covariates. In other words, in these models there is a single response but multiple predictors. An ANOVA model is a special linear regression model where all the predictors are discrete or categorical. These models may be called *univariate* regression models or *univariate* ANOVA models.

In data analysis, sometimes we need to consider more than one responses simultaneously, together with more than one predictors. For example, we may wish to study if income is related to education, age, and gender. In the meantime, we also wish to study if happiness is related to education, age, and gender. In this case, both income and happiness can be treated as response variables, and education, age, gender may be treated as predictors, so there are two response variables and three predictors. When there are more than one response variables, we can study each response variable separately and fit a univariate regression model to each response variable. However, the two response variables here, income and happiness, may be correlated. It would be desirable to incorporate this correlation in regression analysis, instead of ignoring this information. In other words, when several response variables are correlated, it is more efficient to incorporate the correlation between responses in regression analysis than separate regression analysis. That is, fitting a univariate regression model to each response separately may be inefficient and sometimes may even be biased. Thus, we need to consider regression models which contain multiple response variables and incorporate the correlation between these response variables. Univariate regression models for single responses can be extended to regression models with two or more responses (i.e., multivariate responses). Regression models with two or more response variables are called *multivariate regression models*. ANOVA models with two or more response variables are called *multivariate ANOVA* ($MANOVA$) *models*.

The extension from a univariate regression model to a multivariate regression model is conceptually not difficult. However, theoretical developments of these models can be quite tedious. In addition, some assumptions are required. For example, each univariate regression model in a multivariate regression model is assumed to contain the same set of predictors, and the response variables are assumed to follow a multivariate normal distribution. In this chapter, we briefly describe multivariate regression models and MANOVA models, with the focus on conceptually understanding of these models without much technical details. For more details, readers are referred to some classic multivariate textbooks such as Anderson (2003). In this chapter, we also give more details on MANOVA models since these models seem to be more often used in real data analysis. Finally, note that a multivariate regression model is different from a *multiple* regression model. A multiple regression model refers to a regression model with one response but more than one predictors, while a multivariate regression model refers to a regression model with several responses as well as possibly several predictors.

10.2 Multivariate Regression Models

A major advantage of a multivariate regression model over several univariate regression models is that the multivariate regression model incorporates the correlation between the responses. Thus, a multivariate regression model may provide more efficient inference than separate univariate regression models. The efficiency gain depends on the strength of the correlation. On the other hand, a multivariate regression model contains extra parameters, i.e., the parameters for the correlation structure, so statistical inference may be more complicated, while univariate regression models are easier to handle. Thus, it is not always better to use a multivariate regression model than a univariate regression model. The decision on which model to use may depend on the particular application, the strength of the correlation between responses, and the complexity of the models and data analysis.

Let $\mathbf{y}_i = (y_{i1}, y_{i2}, \cdots, y_{im})^{\mathrm{T}}$ be m continuous response variables on individual i, and let $\mathbf{z}_i = (z_{i1}, z_{i2}, \cdots, z_{ir})^{\mathrm{T}}$ be r predictor variables on individual i, $i = 1, 2, \cdots, n$. A *multivariate linear regression model* can be written as

$$\begin{aligned}
y_{i1} &= \beta_{01} + \beta_{11} z_{i1} + \cdots + \beta_{r1} z_{ir} + \varepsilon_{i1}, \\
y_{i2} &= \beta_{02} + \beta_{12} z_{i1} + \cdots + \beta_{r2} z_{ir} + \varepsilon_{i2}, \\
&\cdots\cdots \\
y_{im} &= \beta_{0m} + \beta_{1m} z_{i1} + \cdots + \beta_{rm} z_{ir} + \varepsilon_{im}, \\
&\qquad i = 1, 2, \cdots, n,
\end{aligned}$$

where the vector of random errors $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \cdots, \varepsilon_{im})^{\mathrm{T}}$ is assumed to follow a multivariate normal distribution $\boldsymbol{\varepsilon}_i \sim N_m(0, \Sigma)$. The covariance matrix Σ measures the variances and correlations between the m response variables, so it plays an important role. Note that the above multivariate regression model requires that the *same* set of predictor be used for all responses. The above model can be written in a compact matrix form as follows

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{B} + \boldsymbol{E},$$

where $\boldsymbol{Y} = (y_{ij})$ is a $n \times m$ matrix containing n observations on m response variables, $\boldsymbol{Z} = (z_{ij})$ is a $n \times (p+1)$ design matrix whose columns containing $p+1$ predictors (including the first column of 1's), $\boldsymbol{B} = (\beta_{ij})$ is a $(p+1) \times m$ matrix of regression parameters, and $\boldsymbol{E} = (\varepsilon_{ij})$ is a $n \times m$ matrix of random errors. This matrix form is convenient for theoretical developments.

Parameter estimation and inference for multivariate regression models can be based on the likelihood method or the least square method. The MLE or the least square estimate of $\boldsymbol{B}$ is given by

$$\widehat{\boldsymbol{B}} = (\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z})^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Y}.$$

Similar to univariate linear models, there is a decomposition of the total sum of squares (SS) into regression and residual sums of squares in multivariate regression models. Specifically, let SST be the $m \times m$ matrix of the total sum of squares, SSE be the matrix of residual (error) sum of squares, and SSR be the matrix of the regression sum of squares. Then, we have

$$\begin{aligned} SST &= \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y} - n\bar{\mathbf{y}}\bar{\mathbf{y}}^{\mathrm{T}} \\ &= \hat{\boldsymbol{E}}^{\mathrm{T}}\hat{\boldsymbol{E}} + (\hat{\boldsymbol{Y}}^{\mathrm{T}}\hat{\boldsymbol{Y}} - n\bar{\mathbf{y}}\bar{\mathbf{y}}^{\mathrm{T}}) \\ &= SSE + SSR, \end{aligned}$$

where $\bar{\mathbf{y}}$ is the $m \times 1$ vector of means for the response variables, $\hat{\boldsymbol{Y}} = \boldsymbol{Z}\hat{\boldsymbol{B}}$ is the matrix of fitted values, and $\hat{\boldsymbol{E}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$ is the matrix of residuals.

Statistical inference for multivariate regression models can be performed in a way similar to that for univariate regression models. Many hypothesis tests can be derived based on the matrices of sums of squares for nested models. Let SSH be the incremental SS matrix for a hypothesis. Many multivariate tests for the hypothesis can be based on the eigenvalues λ_j of the $m \times m$ matrix $SSH \times SSE^{-1}$. For example, the multivariate test statistic based on *Phillai-Bartlett Trace* is a function of

$$T_{PB} = \sum_{j=1}^{m} \frac{\lambda_j}{1 - \lambda_j},$$

the test statistic based on *Hotelling-Lawley Trace* is a function of

$$T_{HL} = \sum_{j=1}^{m} \lambda_j,$$

the test statistic based on *Wilks's Lambda* is a function of

$$\Lambda = \prod_{j=1}^{m} \frac{1}{1+\lambda_j}.$$

and the Roy's Maximum Root test statistic is λ_1 (the largest eigenvalue).

A hypothesis about regression coefficients can be written as the following general form

$$H_0 : \boldsymbol{LB} = \mathbf{0},$$

where $\boldsymbol{L}$ is a $q \times (p+1)$ matrix of constants, and $\mathbf{0}$ is a $q \times m$ matrix of 0's. Here the SSH matrix for the hypothesis is given by

$$SSH = \hat{\boldsymbol{B}}^{\mathrm{T}} \mathbf{L}^{\mathrm{T}} (\boldsymbol{L}(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z})^{-1}\boldsymbol{L}^{\mathrm{T}})^{-1} \boldsymbol{L}\hat{\boldsymbol{B}}.$$

The hypothesis can be tested based on the above test statistics. The null distributions of these test statistics can be approximated by F-distributions.

The multivariate linear regression model can also be used for the analysis of *repeated measurements* or longitudinal data, where a single response variable measured at different time points or at different circumstances for the same individuals may be viewed as multivariate response data. The unstructured covariance matrix Σ allows great flexibility in modelling the correlation among the repeated measurements. However, the repeated measurements need to be made at the same time points or same circumstances. Interests often lie in comparisons among the responses

$$H_0 : \boldsymbol{LBR} = \mathbf{0},$$

where $\boldsymbol{R}$ is a $m \times r$ matrix providing contrasts in the responses. For unbalanced repeated measurements or longitudinal data, which is more common in practice, more flexible models such as mixed effects models may be preferred.

In practice, univariate regression models are more commonly used than multivariate regression models, because many statistical methods for univariate regression models such as model selection and model diagnostics are well developed and are easy to use but these methods may not be as convenient or available for multivariate regression models. Moreover, some assumptions for multivariate regression models, such as the responses being multivariate normal, may be restrictive and difficult to check. However, multivariate regression models may still be valuable in some applications. In particular, they may be preferred when the responses are highly correlated and the sample size is reasonably large.

10.3 MANOVA Models

Analysis of variance (ANOVA) models may be viewed as special linear regression models in which all predictors are discrete or categorical and the response is a continuous variable. ANOVA models may also be viewed as an extension of the usual two-sample t-test to the comparison of more than two groups with continuous responses. Such models are often used in designed experiments where the conditions of the experiments are controlled and randomization is used to eliminate other possible confounding variables. The objective in such studies is usually to test the significance of any differences between groups. An ANOVA model with one response is often called a *univariate* ANOVA model.

A *multivariate analysis of variance* (MANOVA) model is a generalization of a univariate ANOVA model when there are two or more correlated response variables, similar to the generalization of a univariate linear regression to a multivariate linear regression. A MANOVA model incorporates the correlation between the response variables, as in multivariate regression models, so it may offer more efficient inference than univariate ANOVA models when the responses are correlated.

Parameter estimation and inference for univariate ANOVA models are often based on partition of variations. That is, we decompose the total variation in the response variable into the variation between groups (group effects) and the variation within groups (random errors), where the variations are measured by sums of squares (SS). Then we compare different sources of variations to check statistical significance of the differences between groups. If the between-group variation is much larger than the within-group variation (i.e., the error variation), we claim that there is significant difference between groups. In other words, the null hypothesis in an ANOVA is that there is no group differences, while the alternative hypothesis is that there is some group difference. The test statistic, which is a ratio of the between-group variation and the within-group variation, has an F-distribution under the null hypothesis and under the assumption that the response data are normally distributed with constant variance.

For a MANOVA model, parameter estimation and inference can also be based on partition of variations, similar to that for a univariate ANOVA model. However, for a MANOVA model, the mathematical expressions become more complicated. Moreover, compared to univariate ANOVA models, another complication for MANOVA models is that the distribution of the test statistics under the null hypothesis is hard to derive and it can only be approximated. Note that, in the case of comparing two groups, the MANOVA test reduces to the Hotelling's T^2 test. In other words, an MANOVA model may be viewed as an extension of the Hotelling's T^2 procedure to the case for comparing more than two groups.

Suppose that we wish to compare g groups, and we measure p response variables in each group. Let $\{\mathbf{y}_{l1}, \cdots, \mathbf{y}_{ln_l}\}$ be the n_l measurements on the p response variables in group l, where $\boldsymbol{y}_{lj} = (y_{lj1}, \cdots, y_{ljp})^{\mathrm{T}}$, $l = 1, 2, \cdots, g$. A MANOVA model can then be written as

$$\begin{aligned} &\mathbf{y}_{lj} = \boldsymbol{\mu}_l + \mathbf{e}_{lj}, \\ &\mathbf{e}_{lj} \sim N_p(0, \Sigma), \qquad j = 1, \cdots, n_l,\ l = 1, \cdots, g, \end{aligned}$$

where $\boldsymbol{\mu}_l = E(\mathbf{y}_{lj}) = (\mu_{1l}, \cdots, \mu_{pl})^{\mathrm{T}}$, $\mathbf{e}_{lj} = (e_{lj1}, \cdots, e_{ljp})^{\mathrm{T}}$, and $\boldsymbol{\mu}_l$ and Σ are unknown mean vector and covariance matrix respectively. We assume that the covariance matrix Σ is the same for each group. The hypothesis of interest is whether the g mean vectors are the same or not, i.e.,

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_g \qquad \text{versus} \qquad H_1 : \text{ not } H_0.$$

Let

$$\bar{\boldsymbol{y}}_l = \sum_{j=1}^{n_l} \boldsymbol{y}_{lj}/n_l, \qquad \bar{\boldsymbol{y}} = \sum_{l=1}^{g}\sum_{j=1}^{n_l} \boldsymbol{y}_{lj} / \sum_{l} n_l$$

be the response sample mean vector for group l and the total mean vector respectively. Let

$$B = \sum_{l=1}^{g} n_l(\bar{\mathbf{y}}_l - \bar{\mathbf{y}})(\bar{\mathbf{y}}_l - \bar{\mathbf{y}})^{\mathrm{T}}, \qquad W = \sum_{l=1}^{g}\sum_{j=1}^{n_l}(\mathbf{y}_{lj} - \bar{\mathbf{y}}_l)(\mathbf{y}_{lj} - \bar{\mathbf{y}}_l)^{\mathrm{T}},$$

be the *between-group sum of squares* (i.e., between group variation) and the *within-group sum of squares* (i.e., within group variation) respectively. A commonly used test statistic is given by

$$\Lambda = \frac{|W|}{|B + W|},$$

which is called the *Wilk's lambda*. We reject H_0 if Λ is too small, i.e., we reject H_0 if the between-group variation B is too large. The Wilk's lambda Λ reduces to an equivalent form of the usual F-statistic in univariate one-way ANOVA where $p = 1$. Computer software is needed to perform the test or to get the p-value, since the exact distribution of Λ under the null hypothesis is unavailable.

10.4 Examples in R

Example 1. To illustrate multivariate linear models in R, we again consider the quiz score dataset described in Chapter 1. We wish to check if there are any significant differences in the five quiz scores between gender and major, as well as any

interaction between gender and major. The five quiz scores are viewed as a multivariate response with $p = 5$ variables. Here the five quiz scores for each student are likely to be correlated since a good student may do well in all five quizes while a weak student may do poorly in all five quizes, so a MANOVA model may be useful as it incorporates the correlation between the five response variables. This dataset may also be viewed as repeated measurements data or longitudinal data, so there are different ways to analyze it. The R output is given below.

```
> quiz.dat <- read.table("class.dat2", head=T)
> quiz.dat2 <- quiz.dat[,c(4:8)]   # extract the five quiz scores
> attach(quiz.dat2)
# MANOVA is done via R function manova ()
> fit <- manova(as.matrix(quiz.dat2) ~ factor(gender)*factor(major))
> summary(fit)
                            Df  Pillai approx F num Df den Df Pr(>F)
factor(gender)               1 0.15606  1.55326      5     42 0.1943
factor(major)                2 0.19983  0.95463     10     86 0.4886
factor(gender):factor(major) 2 0.29162  1.46805     10     86 0.1654
Residuals                   46
# The last column contains the p-values.
```

We see that there are no significant differences in the five quiz scores between gender and major, and there is no significant interaction between gender and major, since all p-values are larger than 0.10. Therefore, the quiz scores are similar between male and female students and are also similar among all majors.

We can also fit a multivariate linear regression model to the above dataset. The R function **lm()** for univariate linear models can also be used for multivariate linear models – we can simply let the response to be multivariate (i.e., the response is a matrix with columns representing variables and rows representing observations), as shown below.

```
# multivariate linear model with 5 quizes as responses
> fit.m <- lm(cbind(quiz1,quiz2,quiz3,quiz4,quiz5)~gender*major, data=
    quiz.dat)
> summary(fit.m)
Response quiz1 :
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)         70.27       4.90   14.35   <2e-16 ***
genderM             -8.07       8.76   -0.92     0.36
majorMath          -10.27      10.58   -0.97     0.34
majorStat           -2.02       6.78   -0.30     0.77
genderM:majorMath   11.57      14.05    0.82     0.41
genderM:majorStat    2.90      10.91    0.27     0.79
......
Response quiz2 :
```

```
Coefficients:
                   Estimate  Std. Error t value Pr(>|t|)
(Intercept)           75.18        4.22   17.81   <2e-16 ***
genderM              -10.98        7.55   -1.45     0.15
majorMath             -9.85        9.12   -1.08     0.29
majorStat             -3.85        5.84   -0.66     0.51
genderM:majorMath      4.65       12.12    0.38     0.70
genderM:majorStat     10.26        9.40    1.09     0.28
......
Response quiz3 :
..........
```

The above results show the parameter estimates corresponding to each response variable. For example, for "quiz1", male students have an average of 8.07 lower scores than female students, while for "quiz2", male students have an average of 10.98 lower scores than female students. Similar results are also available for quiz3, quiz4, and quiz 5 (not shown here). The advantage of the multivariate linear regression model over the MANOVA model is that the multivariate linear regression model gives estimates of each regression coefficients, while the MANOVA model only indicates whether the differences between gender and major are statistically significant or not. From the above results, we also see that the scores between male and female students and scores between majors are not significantly different.

We can also use scores in quiz1, quiz2, and quiz3 to predict scores in quiz4 and quiz5, adjusting for gender and major. In this case, we can consider a multivariate linear regression model with (quiz4, quiz5) as a multivariate response and other variables as predictors.

```
# multivariate linear model with quiz4 and quiz5 as responses
> fit2.m <- lm(cbind(quiz4, quiz5)~quiz1+quiz2+quiz3+gender+major, data
    =quiz.dat)
> summary(fit2.m)
Response quiz4 :
Coefficients:
            Estimate  Std. Error t value Pr(>|t|)
(Intercept)  40.5951      7.6408    5.31  3.2e-06 ***
quiz1        -0.2413      0.1270   -1.90  0.06391 .
quiz2         0.3787      0.1464    2.59  0.01299 *
quiz3         0.3586      0.0935    3.83  0.00039 ***
genderM       2.6061      2.7968    0.93  0.35640
majorMath    -2.2586      3.9286   -0.57  0.56822
majorStat    -1.6850      3.0536   -0.55  0.58382
---
Response quiz5 :
Coefficients:
            Estimate  Std. Error t value Pr(>|t|)
(Intercept)  24.6832      7.7459    3.19  0.00262 **
quiz1         0.1882      0.1288    1.46  0.15073
```

```
quiz2          0.5851     0.1484     3.94  0.00028 ***
quiz3         -0.0247     0.0948    -0.26  0.79580
genderM        2.2406     2.8353     0.79  0.43353
majorMath      4.5683     3.9826     1.15  0.25742
majorStat     -0.0284     3.0956    -0.01  0.99272
---
```

We see that the first three quizes are significantly predictive for the last two quizes. In particular, quiz3 is highly predictive for quiz4 (p-value=0.00039), and quiz2 is highly predictive for quiz5 (p-value=0.00028). Again, there are no significant differences between gender and majors.

Finally, we illustrate the multivariate hypothesis testing. For example, we can test if there is a difference between male and female students in the quiz scores using multivariate tests. This is equivalent to test "genderM = 0" in the model since "genderF" (i.e., female) is chosen as the baseline category in the R function. The R package "car" is needed for multivariate tests.

```
> library(car)
> linearHypothesis(fit.m, "genderM")
...
Multivariate Tests:
                 Df test stat approx F number Df den Df Pr(>F)
Pillai            1      0.16      1.6       5     42    0.2
Wilks             1      0.84      1.6       5     42    0.2
Hotelling-Lawley  1      0.19      1.6       5     42    0.2
Roy               1      0.19      1.6       5     42    0.2
```

We see that there is no significant difference between male and female students (p-value=0.2). All multivariate tests (Pillai, Wilks, Hotelling-Lawley, Roy) give the same results, confirming the validity of the results. The results are also consistent with those in the multivariate linear regression models shown above.

We can also test if the averages of Math major students and Stat major students are significantly different from Comp major students. This is equivalent to test "0.5 majorMath+0.5 majorStat = 0" since the "majorComp" is chose as the baseline category in R.

```
> linearHypothesis(fit.m, "0.5*majorMath+0.5*majorStat")
...
Multivariate Tests:
                 Df test stat approx F number Df den Df Pr(>F)
Pillai            1      0.05     0.41       5     42    0.8
Wilks             1      0.95     0.41       5     42    0.8
Hotelling-Lawley  1      0.05     0.41       5     42    0.8
Roy               1      0.05     0.41       5     42    0.8
```

The above results show that there is no significant difference. Again, all multivariate tests give similar results, confirming the validity of the results.

A main advantage of multivariate linear models over univariate linear models is that multivariate linear models incorporate the correlations between response variables so that they should be more efficient. On the other hand, multivariate linear models contain more parameters than univariate linear models, so multivariate linear models may not be appropriate for small datasets. Moreover, the multivariate normality may not be easy to check. The differences are similar for MANOVA models and univariate ANOVA models.

An advantage of multivariate regression models over MANOVA models is that regression models allow for continuous covariates in the models to partially explain the variations in the response variables, while MANOVA models only allow discrete or categorical predictors. This is similar to the difference between univariate regression models and ANOVA models.

Example 2. We again consider the Chinese consumption dataset "consum2007". The variables East and West in the dataset are categorical variables, where a value of 1 indicates the region belonging to the corresponding area. If a region has a value of 0 for both East and West, it belongs to the central area. We wish to investigate if there exist a significant difference in consumption expenditures among different regions. From the plots in Chapter 4, we find that consumption expenditures on Food (food), Cloth (clothing), Resid (residence), TranC (transport and communications), Educ (education) seem larger than others. Moreover, these variables are likely to be correlated. So we use the MANOVA method to analyze the data to see if there is a significant difference on the five consumptions among the regions.

```
> consum.1<-read.table("consum2007.txt",head =T)
> consum.1[1:3,]
        Food Cloth Resid HousF Health TranC Educ Miscel East West
Beijing 4934  1513  1246   981   1294  2329 2384    650    1    0
Tianjin 4249  1024  1417   761   1164  1310 1640    464    1    0
Hebei   2790   976   917   547    834  1011  895    266    1    0
> consum.east<-subset(consum.1,East==1) [,c(1,2,3,6,7)]
> consum.west<-subset(consum.1,West==1) [,c(1,2,3,6,7)]
> consum.central<-subset(consum.1,(East!=1)&(West!=1)) [,c(1,2,3,6,7)]
# compute the mean vectors for three areas
> (consum.east.mean<-colMeans(consum.east))
 Food Cloth Resid TranC  Educ
 4233  1064  1162  1811  1630
> (consum.west.mean<-colMeans(consum.west))
 Food Cloth Resid TranC  Educ
 3184   979   776   942   979
> (consum.central.mean<-colMeans(consum.central))
 Food Cloth Resid TranC  Educ
```

```
 3008  1019   867   878  1059
> consum.east$Area<-1
> consum.central$Area<-2
> consum.west$Area<-3
> consum.s<-rbind(consum.east,consum.central,consum.west)
> Y<-as.matrix(consum.s[,1:5])  # 5 response variables
> manova.con<-manova(Y~factor(Area), consum.s)
> summary(manova.con)
             Df  Pillai  approx F num Df den Df Pr(>F)
factor(Area)  2  0.737      2.92     10     50 0.0059 **
Residuals    28
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
# The above result shows that there is a significant difference
# on the 5 responses among the 3 areas.
# We also display the anova results for each variable
> summary.aov(manova.con)
 Response Food :
             Df   Sum Sq Mean Sq F value      Pr(>F)
factor(Area)  2  9040519 4520259    10.5     0.00039 ***
Residuals    28 12043422  430122

 Response Cloth :
             Df  Sum Sq Mean Sq F value Pr(>F)
factor(Area)  2   41145   20572    0.43   0.65
Residuals    28 1326061   47359

 Response Resid :
             Df Sum Sq Mean Sq F value      Pr(>F)
factor(Area)  2 905662  452831    18.4     7.8e-06 ***
Residuals    28 687622   24558

 Response TranC :
             Df  Sum Sq Mean Sq F value  Pr(>F)
factor(Area)  2 5700361 2850180    12.8 0.00012 ***
Residuals    28 6254338  223369

 Response Educ :
             Df  Sum Sq Mean Sq F value Pr(>F)
factor(Area)  2 2749784 1374892    8.59 0.0012 **
Residuals    28 4479600  159986
```

From the above results, we see that there exist significant differences in the five consumption expenditures among the three areas. Furthermore, the ANOVA results show that, except Cloth, the differences in consumption expenditures on the other 4 categories are all significant for that area.

Exercises 10

10.1. Write the multivariate linear regression model in a matrix form, and derive the least square estimates of the parameters.

10.2. Derive the maximum likelihood estimates of the parameters in a multivariate linear regression model.

10.3. When $g = 2$, show that the MANOVA test is equivalent to the Hotelling's T^2 test.

10.4. For Example 1, perform univariate regression and ANOVA analyses on each quiz score separately. Do you get the same conclusions as that from multivariate regression and MANOVA analyses?

10.5. Use the dataset "consum2010" to check if there are any differences in consumption expenditures among the areas, using both multivariate regression models and MANOVA models.

10.6. Use the datasets "consum2000", "consum2007", and "consum2010" to check if there are any differences in consumption expenditures among the years. Are there any differences in the consumption structures among the years?

Chapter 11

Longitudinal Data, Panel Data, and Repeated Measurements

11.1 Introduction

Longitudinal studies are popular in practice. In a longitudinal study, individuals are followed over a period of time and data are collected for each individual at multiple time points. These data, which are collected repeatedly over time for each individual, are called *longitudinal data*. Longitudinal data are closely related to *repeated measurement data*, for which repeated or multiple measurements are obtained on each individual but these repeated measurements are not necessarily collected over time. For example, the repeated measurements can be collected over different locations of a city. In economics and sociology, longitudinal studies are often called *panel studies*, and longitudinal data are thus called *panel data*. Multivariate data, longitudinal data, and repeated measurement data are all examples of *correlated data*.

Examples of longitudinal data include measurements of students' quiz scores over time, measurements of individuals' incomes over time, and measurements of subjects' happiness levels over time. Figure 11.1 shows an example of longitudinal data, which displays the quiz scores of a class of students throughout a semester. In this example, each student has five quiz scores. These five quiz scores may be viewed as repeated measurements over time. The repeated measurements from each student are likely to be correlated. For example, a good student may have good marks for each quiz, while a weak student may have low marks for each quiz. The objective may be to compare performances between male and female students. This dataset may be analyzed using methods for multivariate data or methods for longitudinal data or repeated measurements. Different methods may provide different insights.

A key characteristic of longitudinal data and repeated measurement data is that the repeated measurements of a variable on each individual are likely to be *correlated*, since they are data collected over time from the same individuals. Ignoring this correlation in data analysis may lead to inefficient or biased results. Therefore, in the analysis of longitudinal data or repeated measurements, a major consideration is to

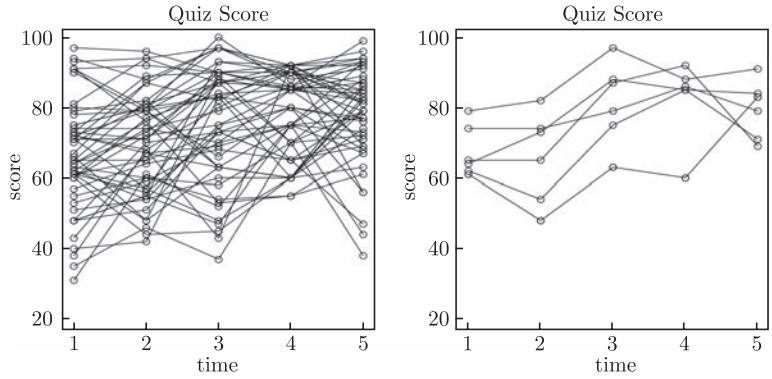


Figure 11.1 Quiz scores from a class of students throughout a semester. Left: all students. Right: six randomly selected students.

incorporate the correlation of the repeated measurements, as in multivariate analysis. Compared with usual multivariate analysis, a main advantage of longitudinal studies is that they allow us to study changes over time for variables of interests or allow us to study special structures in the mean parameters.

Longitudinal data are also related to *time series data*. In a time series, a *single* long series of measurements are observed over time, while in a longitudinal study *many* series of measurements are observed over time. Thus, a main advantage of longitudinal data over time series data is that, for longitudinal data, we can borrow information across different individuals in statistical inference. Such information is important for estimation of correlation structures over time. Longitudinal data may also be viewed as special *multivariate data*, if data on different time points are viewed as data on different variables. In fact, many methods for longitudinal data analysis combine elements from multivariate analysis methods and time series methods.

11.2 Methods for Longitudinal Data Analysis

Regression models for longitudinal data are very useful in which the systematic variation in the longitudinal response may be partially explained by predictors or covariates. That is, in regression models, covariates are introduced to partially explain the systematic between-individual variation in the response variable. Note that, in a regression model for longitudinal data, the response data are longitudinal measurements on a variable of interest, but covariates or predictors can either be longitudinal data (time-dependent covariates, such as time or other variables measured over time) or cross-sectional data (time-independent covariates, such as gender or other variables that do not change over time). A major consideration of statistical methods for longitudinal data analysis is to incorporate the within-individual

correlation.

Three approaches are commonly used in the analysis of longitudinal data. They incorporate the within-individual correlation in different ways:

- The first approach assumes that the repeated measurements within an individual are correlated because these measurements share the same unobserved characteristics of the individual. The unobserved characteristics of an individual can be represented by *random effects* of that individual. Such models are called *mixed effects models*.
- The second approach models the longitudinal mean process and the variance-covariance structure separately, based on a set of estimating equations. Such models are called *generalized estimating equation* (GEE) *models*.
- The third approach assumes that the repeated measurements within an individual are correlated because the longitudinal process may be viewed as a Markov process. Such models are called *transitional models*.

Each of the above three approaches has its advantages and limitations. In practice, the choice of the methods for data analysis is often based both on statistical considerations and on scientific considerations.

Mixed effects models can be obtained from the corresponding regression models for cross-sectional data by introducing random effects in the models. These models are particularly useful for longitudinal data with large between-individual variation since they allow model parameters to vary across individuals. They also allow for individual-specific inference. GEE models are based on estimating equations similar to likelihood equations but with no distributional assumptions for the data. A main advantage of GEE models is that they only require specifications of the mean and variance-covariance structures of the data, without distributional assumptions. GEE models may be particularly useful for non-normal data, such as binary data or count data, in which over-dispersion problems may arise so distributional assumptions may be inappropriate. Transitional models may be useful if certain Markov correlation structures are reasonable for the longitudinal processes. A transitional model has a similar form as a classical regression model for cross-sectional data, if previous response observations are viewed as covariates for the current response observation.

For the foregoing three modelling approaches for longitudinal data, mixed effects models may be preferred if the between-individual variation is large, GEE models may be preferred if distributional assumptions are questionable, and transitional models may be preferred if certain Markov structures are reasonable. In general, if distributional assumptions for the data are reasonable, models with distributional assumptions often produce more efficient estimates than models without distributional assumptions. In the analysis of longitudinal data, it is often desirable to consider different modelling approaches and then compare the results. If the results

are similar, conclusions based on these results may be reliable. Otherwise, further investigation of the models and methods may be needed. In this chapter, we briefly review mixed effects models and GEE models since they are most popular for the analysis of longitudinal data. Detailed discussion of these models can be found in Wu (2009).

11.3 Linear Mixed Effects Models

A linear mixed effects (LME) model can be obtained from a standard linear regression model for cross-sectional data by introducing random effects to the parameters that vary substantially across individuals (i.e., allowing these parameters to be individual-specific). To illustrate, let y_{ij} be the response value for individual i at time t_{ij}, $i = 1, 2, \cdots, n,\ j = 1, 2, \cdots, n_i$. Consider the following simple linear regression model for longitudinal data

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + e_{ij}, \qquad i = 1, 2, \cdots, n,\ j = 1, 2, \cdots, n_i, \tag{11.1}$$

where e_{ij} is a random error. If the data show that the intercept β_0 varies substantially for different individuals, we may introduce a random effect for the intercept β_0, i.e., we may allow the intercepts to be individual-specific. We thus obtain the following LME model

$$\begin{aligned} y_{ij} &= (\beta_0 + b_{0i}) + \beta_1 t_{ij} + e_{ij} \\ &= \beta_{0i} + \beta_1 t_{ij} + e_{ij}, \qquad i = 1, 2, \cdots, n,\ j = 1, 2, \cdots, n_i, \end{aligned} \tag{11.2}$$

where b_{0i} is a random effect for individual i, and $\beta_{0i} = \beta_0 + b_{0i}$ is the individual-specific intercept for individual i. The parameters β_0 and β_1 are called *fixed effects*, since they are fixed for all individuals. Thus, we can choose random effects informally based on the heterogeneous feature of the data. Formally, we can choose the random effects based on standard tests such as the likelihood ratio test or based on standard model selection methods such as AIC or BIC criteria.

General LME models can be described as follows. Let $\mathbf{y}_i = (y_{i1}, y_{i2}, \cdots, y_{in_i})^{\mathrm{T}}$ be the n_i repeated measurements of the response variable y on individual i, and let $\boldsymbol{e}_i = (e_{i1}, e_{i2}, \cdots, e_{in_i})^{\mathrm{T}}$ be the corresponding random errors of the repeated measurements, $i = 1, 2, \cdots, n$. A general form of LME models can be written as

$$\mathbf{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i, \qquad i = 1, 2, \cdots, n, \tag{11.3}$$

$$\boldsymbol{b}_i \sim N(0, \boldsymbol{D}), \qquad \boldsymbol{e}_i \sim N(0, \boldsymbol{R}_i), \tag{11.4}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)^{\mathrm{T}}$ is a $(p+1)\times 1$ vector of fixed effects, $\boldsymbol{b}_i = (b_{i0}, b_{i1}, \cdots, b_{iq})^{\mathrm{T}}$ is a $(q+1)\times 1$ vector of random effects, matrix $\boldsymbol{X}_i$ $(n_i \times (p+1))$ and matrix $\boldsymbol{Z}_i$ $(n_i \times (q+1))$ are known design matrices which often contain covariates (including

times), D is a $(q+1)\times(q+1)$ covariance matrix of the random effects, and $\boldsymbol{R}_i$ is a $n_i \times n_i$ covariance matrix of the within-individual random errors. The design matrices can be written as

$$\boldsymbol{X}_i = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n_i 1} & \cdots & x_{n_i p} \end{pmatrix}, \qquad \boldsymbol{Z}_i = \begin{pmatrix} 1 & z_{11} & \cdots & z_{1q} \\ 1 & z_{21} & \cdots & z_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & z_{n_i 1} & \cdots & z_{n_i q} \end{pmatrix},$$

and $\boldsymbol{Z}_i$ is often a submatrix of $\boldsymbol{X}_i$. The standard assumptions for models (11.3) and (11.4) are (i) the individuals are independent, (ii) the errors ε_i and the random effects $\boldsymbol{b}_i$ have mean zero, and (iii) the errors ε_i and the random effects $\boldsymbol{b}_i$ are independent and both are normally distributed.

For LME model (11.2) in the example, we have $p = 2$, $q = 1$, $\boldsymbol{b}_i = b_{0i}$, $\boldsymbol{R}_i = \boldsymbol{I}_i$ (the $n_i \times n_i$ identity matrix), and

$$\boldsymbol{X}_i = \begin{pmatrix} 1 & x_{i1} \\ 1 & x_{i2} \\ \vdots & \vdots \\ 1 & x_{in_i} \end{pmatrix}, \qquad \boldsymbol{Z}_i = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}. \tag{11.5}$$

We could also introduce a random effect in the slope β_1, if the data show such a need.

In LME model (11.3) and (11.4), the repeated measurements $\{y_{i1}, y_{i2}, \cdots, y_{in_i}\}$ of the response can be taken at different time points for different individuals, and the number of measurements n_i may vary across individuals. Thus, a LME model allows *unbalanced data* in the response. In other words, a LME model allows missing data in the response, assuming the missing data are missing at random. This is an advantage of a LME model over a multivariate normal model in which no missing data are allowed. In practice, we often assume that $R_i = \sigma^2 I_{n_i}$ in LME model (11.3) and (11.4), i.e., the within-individual measurements are assumed to be conditionally independent with constant variance given the random effects.

Statistical inference for LME models is typically based on the maximum likelihood method. MLE of the fixed parameters $\boldsymbol{\beta}$ can be obtained using an iterative algorithm such as the expectation-maximization (EM) algorithm. The random effects can also be estimated using the empirical Bayesian method.

Example 1. To study the growth of children, the distance from the pituitary gland to the pterygomaxillary fissure is measured every two years from 8 years of age until 14 years of age. A sample of 27 children – 16 males and 11 females was obtained by orthodontists from x-rays of the children's skulls. The dataset is denoted by "Orthodont". We use this example to illustrate statistical modelling procedures using LME models.

```
> library(nlme)  # get the NLME library for mixed effects models
> attach(Orthodont)  # get dataset (it's R an internal dataset)
# Here is part of the data
Orthodont
    distance age Subject    Sex
1       26.0   8     M01   Male
2       25.0  10     M01   Male
3       29.0  12     M01   Male
4       31.0  14     M01   Male
5       21.5   8     M02   Male
6       22.5  10     M02   Male
.............

# Now let's fit a LME model with random effects on both
# the intercept and the slope
# Add: (for outside datasets, do this first:
# >orthodont<- groupedData(distance~age
# Subject, data=orthodont))
> lme.fit1 <- lme(distance~age, data=Orthodont,
                  random = ~age  | Subject, method = "ML")
> summary(lme.fit1)
...
Fixed effects: distance ~ age
               Value Std.Error DF   t-value p-value
(Intercept) 16.761111 0.7678975 80 21.827278       0
age          0.660185 0.0705779 80  9.353997       0
...

# Add Sex and interaction
> lme.fit2 <- update(lme.fit1, fixed=distance~Sex*age)
> summary(lme.fit2)
...
Fixed effects: distance ~ Sex + age + Sex:age
                  Value      Std.Error      DF       t-value      p-value
(Intercept)   16.340625      0.9987521      79     16.361042       0.0000
SexFemale      1.032102      1.5647438      25      0.659598       0.5155
age            0.784375      0.0843294      79      9.301322       0.0000
SexFemale:age -0.304830      0.1321188      79     -2.307238       0.0237
......
```

We see that there seem significant difference between boys and girls since the interaction term is significant at 5% level. In other words, the growth patterns of male and female children over time may be different. Next, let's do an ANOVA test for comparing the two models to see whether the model with interaction fits the data significantly better.

```
> anova(lme.fit1, lme.fit2)
         Model df       AIC       BIC    logLik   Test  L.Ratio p-value
lme.fit1     1  6 451.2116 467.3044 -219.6058
lme.fit2     2  8 443.8060 465.2630 -213.9030 1 vs 2 11.40565  0.0033
```

We see that the model with interaction fits the data much better, so we should add covariate Sex and the interaction in the model. Finally, let's do model diagnostics.

```
# Check residual plot (you need not exactly follow the commands below)
> plot(lme.fit2, resid(.,type="p")~fitted(.) | Sex,id=0.05,adj=-0.3)
# residual plot looks OK, but there may be outliers
# Check normality for the within individual errors
> qqnorm(lme.fit2, ~resid(.) | Sex)
# Normality assumption OK for within subject errors
# Check normality assumption for b
> qqnorm(lme.fit2, ~ranef(.),id=0.1,cex=0.7)
# Normality assumption OK for random effects
```

The residual plots in Figure 11.2 shows that the LME model fits the data reasonably well since there seem not systematic patterns in the figures, but there may be a few outliers for the male plot. Figure 11.3 and Figure 11.4 show the Q-Q plots to check the normality assumptions for the random errors and random effects. We see that the normality assumptions seem reasonable, although there may be a few outliers, which corresponding to the ones in the residual plot.

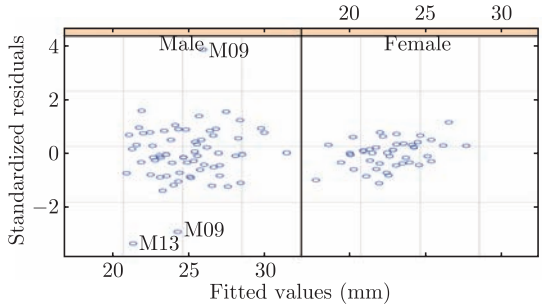


Figure 11.2 Residual plots.

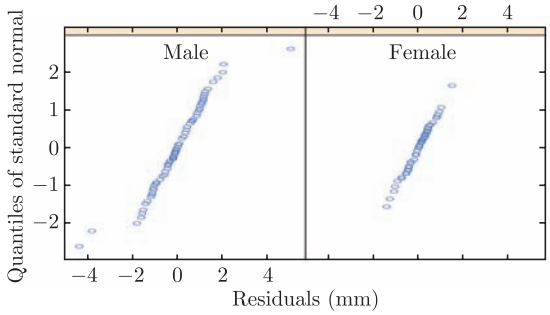


Figure 11.3 Q-Q plot for within-individual random errors.

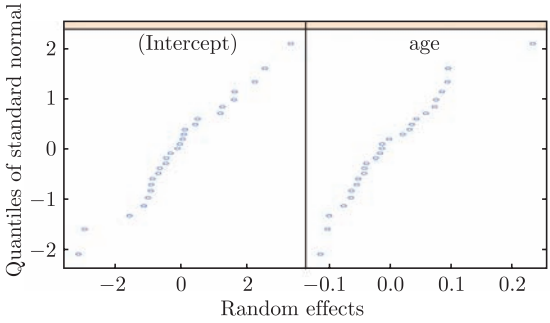


Figure 11.4 Q-Q plot for the random effects.

11.4 GEE Models

Mixed effects models assume distributions for the errors and random effects, but these assumptions sometimes do not hold in practice. GEE models, on the other hand, do not make distributional assumptions. They only make assumptions for the mean structure and variance-covariance structures, and then proceed with parameter estimation and inference. In GEE models, the mean structure and variance-covariance structure are specified separately, without any distributional assumptions for the data, and emphasis is placed on the correct specification of the mean structure. The specification of the variance-covariance structure only affects the *efficiency* of the estimates: the closer the variance-covariance structure to the true one, the more efficient the resulting GEE estimates. GEE models are particularly useful for non-normal data, such as binary data or count data.

For a GEE model, parameter estimators are obtained by solving a set of estimating equations. Specifically, let $\mathbf{y}_i = (y_{i1}, \cdots, y_{in_i})^{\mathrm{T}}$ be the longitudinal repeated measurements of the response variable on individual i, and let $\mathbf{x}_i$ be the corresponding covariates, $i = 1, 2, \cdots, n$. A (generalized) regression model can be written as

$$\boldsymbol{\mu}_i(\boldsymbol{\beta}) = E(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\beta}) = h(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}), \qquad i = 1, 2, \cdots, n, \tag{11.6}$$

where $\boldsymbol{\beta}$ is a vector of regression parameters, and $h(\cdot)$ is a known link function. For longitudinal data, we must incorporate the correlation among repeated measurements. To do this, we separately assume a variance-covariance structure for $\mathbf{y}_i$ as follows

$$Cov(\mathbf{y}_i) = \Sigma_i(\boldsymbol{\beta}, \boldsymbol{\alpha}), \tag{11.7}$$

where $\boldsymbol{\alpha}$ contains unknown parameters for the variance-covariance structure of $\mathbf{y}_i$.

The variance-covariance structure of the response vector $\mathbf{y}_i$ can be written in the following form

$$\Sigma_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = V_i^{1/2}(\boldsymbol{\beta}) \boldsymbol{R}_i(\boldsymbol{\alpha}) V_i^{1/2}(\boldsymbol{\beta}), \tag{11.8}$$

where $V_i(\boldsymbol{\beta}) = diag(\sigma_{i1}^2, \cdots, \sigma_{in_i}^2)$, with $\sigma_{ik}^2 = var(y_{ik}|\mathbf{x}_i, \boldsymbol{\beta})$, are the variances of the responses for individual i at different times, and the matrix $\boldsymbol{R}_i(\boldsymbol{\alpha})$ is called a *working correlation matrix*, which measures the correlations of the repeated response measurements on individual i. In data analysis, we may consider the following correlation structures

- the *independence* working correlation matrices:

$$\boldsymbol{R}_i(\boldsymbol{\alpha}) = \boldsymbol{I}_{n_i},$$

  where I_{n_i} is the $n_i \times n_i$ identity matrix;
- the *equicorrelation* (or exchangeable) working correlation matrices:

$$(\boldsymbol{R}_i(\boldsymbol{\alpha}))_{jk} = corr(y_{ij}, y_{ik}) = \alpha, \qquad \text{for } j \neq k,$$

  where $corr(y_{ij}, y_{ik})$ is the correlation between y_{ij} and y_{ik};
- the *stationary* working correlation matrices:

$$(\boldsymbol{R}_i(\boldsymbol{\alpha}))_{jk} = corr(y_{ij}, y_{ik}) = \alpha|t_{ij} - t_{ik}|, \qquad \text{for } j \neq k;$$

- the *unstructured* working correlation matrices

$$(\boldsymbol{R}_i(\boldsymbol{\alpha}))_{jk} = corr(y_{ij}, y_{ik}) = \alpha_{jk}, \qquad \text{for } j \neq k,$$

where α_{jk}'s are unknown parameters.

To estimate the unknown parameters in GEE models, we can construct a set of equations which are similar to the familiar likelihood equations but without distributional assumptions. Solving these equations lead to GEE estimates of the parameters. Specifically, let

$$\boldsymbol{\mu}_i(\boldsymbol{\beta}) = E(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\beta}), \qquad \Delta_i(\boldsymbol{\beta}) = \partial \boldsymbol{\mu}_i(\boldsymbol{\beta}) / \partial \boldsymbol{\beta}.$$

The *generalized estimating equation* (GEE) for estimating the mean parameters $\boldsymbol{\beta}$ is given by

$$S_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \left[\Delta_i(\boldsymbol{\beta}) \Sigma_i^{-1}(\boldsymbol{\beta}, \boldsymbol{\alpha}) (\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) \right] = 0. \tag{11.9}$$

GEE (11.9) can be solved by an iterative algorithm.

If the mean structure $\boldsymbol{\mu}_i(\boldsymbol{\beta})$ is correctly specified, it can be shown that the GEE estimator $\hat{\boldsymbol{\beta}}$ is consistent and asymptotically normal, even if the covariance matrix

$\Sigma_i(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is mis-specified. The choice of the covariance matrix $\Sigma_i(\boldsymbol{\beta}, \boldsymbol{\alpha})$ only affects the *efficiency* of the GEE estimator. This is a main advantage of GEE estimators. It suggests that for GEE models we can focus on correct specification of the mean structure. The idea of the GEE approach can be extended to a wide range of problems.

Example 2 (Children's height data). This dataset contains 26 children's heights measured at 9 different times. We use it to illustrate the GEE methods and mixed effects models. The variable "age" is standardized.

```
> oxboys.dat <- read.table("oxboys.dat",head=T)
> attach(oxboys.dat)
> library(gee)   # A R package for GEE
> library(nlme)  # A R package for mixed effects models

# Part of the data
> oxboys.dat[1:4,]
  id     age height occasion
1  1 -1.0000  140.5        1
2  1 -0.7479  143.4        2
3  1 -0.4630  144.8        3
........
# Fitting marginal GEE models with various assumptions
# of the working correlation structures so we can compare the results
fit.gee1 <- gee(height~age,id)   # independent correlation
fit.gee2 <- gee(height~age,id,corstr="AR-M") # AR-1 correlation
fit.gee3 <- gee(height~age,id,corstr="exchangeable")  # exchangeable
fit.gee4 <- gee(height~age,id,corstr="unstructured")  # unstructured
# Let's compare the estimates and SE's
> summary(fit.gee1)$coef
             Estimate Naive S.E.    Naive z Robust S.E.     Robust z
(Intercept) 149.371801  0.5285648 282.598864    1.554618     96.08266
age           6.521022  0.8169867   7.981797    0.329252     19.80557
> summary(fit.gee2)$coef
             Estimate Naive S.E.  Naive z Robust S.E. Robust z
(Intercept) 149.719096  1.5531285 96.39840   1.5847569 94.47449
age           6.547328  0.3177873 20.60286   0.3042478 21.51972
> summary(fit.gee3)$coef
             Estimate Naive S.E.  Naive z Robust S.E. Robust z
(Intercept) 149.371735  1.5615503 95.65605   1.5546081 96.08321
age           6.523916  0.1476602 44.18196   0.3295115 19.79875
> summary(fit.gee4)$coef
             Estimate Naive S.E.  Naive z Robust S.E. Robust z
(Intercept) 149.494178  1.5615644 95.73360   1.5533605 96.23920
age           6.052624  0.3726325 16.24288   0.3205602 18.88139
```

We see that the estimates based on different working correlation matrices are similar. However, the independent working correlation structure produce different SE estimates. This indicates the importance of incorporating the correlation in the

longitudinal data. The unstructured correlation is most general.

Next, let's try a mixed effects model for the same dataset and compare the results to the GEE results.

```
> oxboys.dat1 <- groupedData(height~age|id,data=oxboys.dat)
> fit.lme1 <- lme(fixed=height~age, random=~age, data=oxboys.dat1)
> summary(fit.lme1)
...
Fixed effects: height ~ age
                Value Std.Error  DF  t-value p-value
(Intercept) 149.37175 1.5854173 207 94.21605       0
age           6.52547 0.3363003 207 19.40370       0
......
```

We see that the LME model produces similar estimates as the GEE models. Note that the LME model requires normal distributional assumptions for the errors and random effects, while the GEE models do not require distributional assumptions. On the other hand, the LME model allows for individual-specific inference, while GEE models only provide population-average inference.

Exercises 11

11.1. Write down the likelihood function for the LME model (2). Suggest a method to find the MLEs of the model parameters.

11.2. Write down the likelihood function for the LME model (3) and (4).

11.3. For model (1), write down the GEE equation for estimating model parameters.

11.4. Derive the Generalized Estimating Equation, i.e. GEE (9).

11.5. For the Quiz dataset described in Chapter 1, perform the following analysis:

(a) Fit a LME model with a random intercept, and compare it with a LME model with both random intercept and random slope. Which model fits the data better?

(b) For the models in (a), is there a significant difference between male students and female students in quiz scores? Do students performance improve over time?

(c) Perform model diagnostics for the LME models in (a).

(d) Fig a GEE model with different correlation structures. Do you get similar results as the LME models?

11.6. Combine the three datasets "consum2000", "consum2007", and "consum2010" to form a longitudinal dataset (with repeated measurements at three time points).

(a) Fit a LME model to check if there is a time trend and if there are any differences in consumption expenditures among the three regions (East, West, and Central regions).

(b) Repeat the analysis in (a) using a GEE method and compare the results to (a).

Chapter 12
Methods for Missing Data

12.1 Missing Data Mechanisms

In practice, missing data are very common, since it is almost unlikely that all data are available for each individual on each variable in a multivariate or longitudinal dataset. For example, in a sample survey, some people may not report incomes, so some income data may be missing. In a class, some students may not hand in a few homework assignments, so the homework marks have some missing values. In the presence of missing data, statistical analysis becomes complicated since standard methods and software cannot be directly used. Ignoring the missing data in statistical analysis may lead to biased results or other complications. Thus, in real data analysis, we must find appropriate ways to handle missing data.

As an illustrating example, consider a subset of the "quiz dataset" described in Chapter 1. The *real* data may look like below (an "NA" represents a missing value):

```
ID gender major quiz1 quiz2 quiz3 quiz4 quiz5
1     M     Stat    90    79    90    90    93
2     M     Math    55    NA    58    NA    79
3     F     Stat    60    72    75    80    77
4     M     Math    66    48    NA    NA    NA
5     F     Stat    63    60    54    55    61
6     NA    Math    61    48    63    60    NA
7     M     NA      40    42    83    80    56
8     M     Math    50    44    11    NA    NA
9     M     Comp    75    80    93    90    85
10    F     Comp    57    64    NA    65    71
```

For the 10 students, 5 students missed at least one quiz. Students No. 4 and 8 may even have dropped out from class. Students may have missed quizes for various reasons, such as being sick or finding class too difficult. For whatever reasons, when analyzing the above dataset, a computer software by default will simply delete the students with NA's, which is called the complete-case method. So half of students (i.e., ID no. 2, 4, 6, 8, 10) will have to be removed, leading to a great loss of information and possibly biased results (the removed students may be the weak ones, so class averages or other results based on the remaining complete data are

not representative). In other words, if a student missed a quiz because he/she did not understand the course materials, analysis with this student removed will lead to biased results. Another simple method is to impute the missing scores by the class average, which is called the mean-imputation method, but this method is also undesirable if the missing scores were in fact quite low. Thus, it is important to use better statistical methods to handle these missing data.

As one can see from the foregoing example, for multivariate data, simply deleting observations with missing values will lead to substantial loss of information and may even lead to biased results. Imputing a missing value by a single (guessed) value, such as the mean, may not only lead to biased results but it also ignores missing data uncertainty, which will under-estimate standard errors of parameter estimates. That is, when a value is missing, there is some uncertainty about the possible true value. This uncertainty should be reflected in the estimation of standard errors. When choosing a good method for missing data, we should at least consider two questions: i) will the method lead to biased results or are the missing values random subset of all data? ii) how to incorporate the uncertainty about the missing data? The first question concerns about missing data mechanisms, while the second question is related to reliable estimation of standard errors.

To handle missing data appropriately, the first question is to check how the data are missing, i.e., the *missing data mechanism*. For example, for the quiz dataset given above, if a student missed a quiz because he was sick or forgot, statistical analysis without this student will still be valid, although the result may not be efficient due to some loss of information. However, if a student missed a quiz because he did not understand the course materials, data analysis without this student will be biased (e.g., the class average will be biased toward a high value). Thus, to choose an appropriate statistical method for missing data, we should first find out the possible missing data mechanism.

Although data may be missing for many reasons, from statistical point of view, there are only three possible missing data mechanisms for all missing data problems, depending on whether the missing values are related to other data or not. The three missing data mechanisms are

- *missing completely at random* ($MCAR$): missingness depends neither on the missing values nor on the observed values.
- *missing at random* (MAR): missingness does not depend on the missing values but may depend on observed values.
- *nonignorable missing*: missingness depends on the missing values and observed values.

That is, the missing data mechanism is based on whether the probability of a missing value may depend on the missing values or the observed value or not. This is

important since it will affect whether a statistical method may lead to biased results or not.

The MCAR is the strongest assumption. It basically says that the missing data are completely random, i.e., the individuals with missing values may be viewed as a random subsample of the original sample. In this case, the complete-case method which deletes all incomplete observations will still lead to valid statistical analysis, but it will be inefficient due to a reduced sample size. In many practical problems, the MCAR may be too strong and may not be a reasonable assumption. So we must be careful to make a MCAR assumption. The nonignorable missing mechanism has the greatest influence on analysis results, i.e., analysis ignoring the missing data mechanism will lead to severely biased results. In this case, the missing data contain valuable information, and this information must be incorporated in statistical analysis in order to obtain unbiased results. The MAR is a moderate assumption and it may be reasonable in many studies, or we can include more variables in data analysis in order to make MAR more reasonable, as we usually do in a multiple imputation method to be described later. When missing data are MCAR or MAR, if statistical analysis is based on likelihood methods, the missing data mechanism may be ignored, so MCAR or MAR are also called *ignorable missing*. Note that the complete-case method will still lead to biased results when the missing data are MAR.

To illustrate the missing data mechanisms, consider the quiz data given earlier. If a student missed a quiz because he forgot or was sick, the missing data may be MCAR, since the missingness does not depend on the missing quiz scores nor the observed (available) quiz scores. If the student missed a quiz because he did very poorly on the previous quiz (e.g., student No 2), the missing data may be MAR, since the missingness depends on the previous scores (observed value), but not on the missing scores. If the student missed a quiz because he did not understand the course materials to be tested in the quiz, the missing data may be nonignorable missing, since the missingness depends on the missing scores (i.e., if he did take the quiz, his score would likely to be low). As another example, in a sample survey, if a person did not report his income because he forgot, then the missing data is MCAR. If a person did not report his income because he is too old, then the missing data is MAR (the missingness depends on his age, but not income). If a person did not report his income because his income is too high or too low, then the missing data is nonignorable.

In practice, data analysts may not know the missing data mechanisms. Moreover, the nonignorable missing data mechanism cannot be tested based on the observed data. To formally address nonignorable missing data, we need to build a missing data model which represents the possible missing mechanism and then incorporate

the missing data model into formal statistical analysis, as illustrated in the next section. Note that, if the missing rate is low, say less than 10% missing data, it may be all right to handle the missing data in an informal way. However, if the missing rate is high, say at least 20% missing data, then statistical analysis must address the missing data problem more formally in order to obtain valid results. In the next section, we briefly discuss some general formal statistical methods for missing data.

12.2 Methods for Missing Data

Many statistical methods have been proposed for missing data problems. In this section, we briefly review the two most commonly used general methods for missing data. Note that, when a value is missing, its true value cannot be known for certain. However, since many variables are correlated, we can roughly guess the missing true value based on other observed data. For example, if a person's income is missing in a sample survey, we can roughly guess the person's income based on his education, age, job title, etc. Such a guess has some uncertainty, which can be incorporated through a statistical model. Formally, the two most popular formal statistical methods for missing data are

- the multiple imputation method.
- the expectation-maximization (EM) algorithm.

These two methods are quite general and can be used in a wide variety of missing data problems. We briefly describe the basic ideas of these two methods as follows. More details are provided in the next sections, and a comprehensive review may be found in Little and Rubin (2002) and Wu (2009).

Note that there are many simple or naive imputation methods for missing data which are commonly used in practice. For example, we may impute a missing value by the average of observed values or by a guessed value based on a regression model. These are called *single imputation methods*, i.e., imputing a missing value by *one* guessed value. A main problem with these single imputation methods is that they ignore the missing data uncertainty, i.e., the uncertainty associated with the missing values. In other words, a single guessed value, no matter how good it may be, is likely to be different from the true value if it were observed. Ignoring this missing data uncertainty will lead standard errors of parameter estimates to be under-estimated.

Multiple imputation methods, on the other hand, incorporate the missing data uncertainty by imputing *multiple* guessed values for *each* missing value. So multiple imputation methods are generally better than single imputation methods. Once each missing value is imputed by several guessed values, we obtain several "complete" datasets. Then we can use standard statistical methods and software to analyze these

"complete" datasets and combine the results to obtain an overall conclusion. The guessed values for each missing value are generated based on a possible relationship between different variables. For example, to generate an imputation for a missing value, we use a regression model which uses other observed data to predict the possible true value of the missing data.

For a multiple imputation method to work well, there are two key questions. The first question is how to generate good imputations (guessed values) for each missing value so that the "guessed values" are as good as we can get and they are statistically valid. The second question is how to combine multiple "complete data" analysis results to obtain an overall conclusion. When combining the results, a key consideration is to incorporate the missing data uncertainty, which can be roughly measured by the different imputations generated from a imputation model for each missing value. There are different methods to generate multiple imputations. Some methods may be better than other methods. One formal and valid approach is to use a Bayesian framework to generate "proper" imputations in the sense that it leads to valid statistical inference.

In multivariate analysis and other statistical analyses, the likelihood method is probably the most widely used approach. The maximum likelihood estimates (MLEs) of model parameters are obtained by maximizing the likelihood for the observed data. In the presence of missing data, however, it is often challenging to obtain the MLEs. The EM algorithm is a popular method which produces MLEs of model parameters in the presence of missing data. Since missing data are very common in practice and many statistical problems may also be formulated as "missing data" problems (such as mixed effects models and latent variable models), the EM algorithm has gained great popularity in statistical research. The EM algorithm is an iterative algorithm which iterates between an E-step and an M-step until converge. At convergence, the MLEs of model parameters can be obtained.

In some sense, the EM algorithm may be viewed as a multiple imputation method with infinite many imputations. However, a multiple imputation method does not require the use of the likelihood method. It is a quite general and intuitive method. Unlike multiple imputation methods, an EM algorithm is usually designed for a specific model and the programming for the algorithm may be non-trivial for many complicated problems. Thus, the EM algorithm is more commonly used in statistical research, but it may not be convenient in real data analysis since the EM algorithm may not be easily understood and intuitive. Moreover, general software for EM algorithm is limited.

An advantage of using likelihood method and EM algorithm for missing data analysis is that the missing data mechanism can be easily incorporated in the algo-

rithm. For example, suppose that $\{\mathbf{x}_i = (x_{i1}, \cdots, x_{ip}),\ i = 1, 2, \cdots, n\}$ is a sample on p variables with missing data. Let

$$r_{ij} = \begin{cases} 1 & \text{if } x_{ij} \text{ is observed,} \\ 0 & \text{if } x_{ij} \text{ is missing ,} \end{cases} \qquad i = 1, 2, \cdots, n;\ j = 1, 2, \cdots, p$$

be a missing data indicator, and let $\mathbf{r}_i = (r_{i1}, \cdots, r_{ip})^{\mathrm{T}}$. Then, a missing data model which approximates the missing data mechanism, denoted by $f(\mathbf{r}_i|\mathbf{x}_i, \boldsymbol{\phi})$, can be formulated as the following logistic regression model

$$\log \frac{P(r_{ij} = 1)}{1 - P(r_{ij} = 1)} = \phi_0 + \phi_1 x_{i1} + \cdots + \phi_p x_{ip}, \tag{12.1}$$

which describe how the missingness may be related to the missing values and observed values. Thus, if $\phi_j \neq 0$, the missing data are nonignorable since the missing probability depends on the missing values. If $\phi_j = 0$, the missing data are MAR since the missing probability does not depend on the missing values but may depend on other observed values. If $\phi_1 = \cdots = \phi_p = 0$, the missing data may be MCAR since the missing probability does not depend on any observed or missing data. Based on the observed missing data information and other observed data, the above logistic regression model can be fitted to the observed data and the parameters α_j's be estimated. Hypothesis testing can then be performed to test whether some parameters are zero or not. The above missing data model can be easily incorporated in likelihood inference and the EM algorithm can be used to estimate all model parameters, as shown in next section. Note that the missing data model is a secondary model, not the main model of interest, so we should avoid a too complicated missing data model.

When it is difficult to determine the missing data mechanism, a good strategy is to estimate the main model parameters based on different missing data mechanisms and then compare the resulting estimates. If the main model parameter estimates are similar under different missing data mechanisms, the results are not sensitive to missing data mechanism and thus may be reliable. Otherwise, further investigation is required to check how the results can be influenced by a missing data mechanism. This is called sensitivity analysis.

In addition to the above two methods for missing data, there are also other methods available for missing data. For example, if we use generalized estimation equation (GEE) methods for statistical inference, the weighted GEE methods may be used to handle missing data. There are also valid single imputation methods which incorporate missing data uncertainty by adjusting the variance formulas. For more details about various missing data methods, readers are referred to Little and Rubin (2002) or Wu (2009). From a practical point of view, the multiple imputation

method is probably most intuitive and easy to use, so it is a recommended method for missing data. It is described in more details in the next section.

12.3 Multiple Imputation Methods

The basic idea of a multiple imputation method is to impute each missing value by multiple predicted values based on a prediction model (or imputation model). Once missing data are multiply imputed, we obtain several "complete datasets". Then we analyze the complete datasets by standard methods and software, and combine the results to form an overall conclusion. There are two key considerations for a multiple imputation method: (i) how to generate good imputed values, and (ii) how to combine multiple results.

To generate good imputed values for each missing value, we should build a prediction or imputation model which uses other variables or observed data to help predicting the missing data. To illustrate the basic idea, consider the quiz example described at the beginning of this chapter. Suppose that we wish to generate imputations for the missing data in "quiz5". We can use other observed quiz scores to predict the missing data in "quiz5" based on the following prediction or imputation model

$$\text{quiz5} = \beta_0 + \beta_1 \text{ quiz1 } + \beta_2 \text{ quiz2 } + \beta_3 \text{ quiz3 } + \beta_4 \text{ quiz4} + e,$$

where e is the prediction error which reflects the uncertainty of the missing value. The parameters β_j's can be estimated from the data. To generate theoretically valid imputations, however, a Bayesian framework is usually used to estimate β_j's and generate imputations simultaneously, implemented by the Gibbs sampler (see description below). As another example, in a sample survey, if an income is missing, we may consider the following model to predict the missing income

$$\text{income } = \beta_0 + \beta_1 \text{ education } + \beta_2 \text{ age } + \beta_3 \text{ experience } + \beta_4 \text{ title } + e.$$

That is, we use education, age, working experience, and job title to help predicting the missing income. Since a missing value has some uncertainty, we can generate (say) 5 or 6 predicted values for each missing value. Such predicted values for the missing data should be closer to the (missing) true values than other simple guessed values such as averages.

More generally, a multiple imputation method consists of the following steps:

- For each missing value, we impute several possible values (say $m = 5$) based on a prediction or imputation model, then we obtain several (m) "complete datasets".
- Each of the m "complete datasets" is analyzed using the usual complete-data methods and software as if all data were observed, which leads to m analysis results.

- The m complete-data analysis results are combined to obtain an overall result and conclusion.

In the above procedure, the key is how to create good imputations and how to combine the results. We describe these below.

Consider a multiple imputation method for missing values in the $n \times p$ data matrix X, where each column of X contains n observations for the corresponding variable. Let X_{mis} be the missing parts of X and X_{obs} be the observed part. The prediction or imputation model is denoted by $f(X_{\text{mis}}|X_{\text{obs}})$. Proper imputations can be generated from the predictive distribution $f(X_{\text{mis}}|X_{\text{obs}})$ based on the following Bayesian framework

$$X_{\text{mis}} \sim f(X_{\text{mis}}|X_{\text{obs}}) = \int f(X_{\text{mis}}|X_{\text{obs}}, \beta) f(\beta|X_{\text{obs}}) d\beta$$

where $f(X_{\text{mis}}|X_{\text{obs}}, \beta)$ is the imputation (or prediction) model and $f(\beta|X_{\text{obs}})$ is a prior distribution for the parameters. Then, sampling from $f(X_{\text{mis}}|X_{\text{obs}})$ can be implemented by a Markov Chain Monte Carlo (MCMC) method such as the Gibbs sampler. See next section for a specific example. Such an approach is called *proper imputation*, and it will lead to theoretically valid results. Software is available to implement this approach. There are other simpler ways to generate imputations, but some of those methods may not be proper.

Suppose that m imputations are generated based on above procedure for each missing value, and the m "complete datasets" are analyzed using standard methods and software for complete data. Then, the next key question is how to combine the m results. Let θ contains parameters of main interest, e.g., the parameters in the main model which needs not be the same as the imputation model. Let $\hat{\theta}^{(i)}$ and $\text{Var}(\hat{\theta}^{(i)})$ be the parameter estimate and its variance based on the i-th imputed dataset, $i = 1, 2, \cdots, m$. The overall estimate of θ is given by

$$\hat{\theta} = \frac{1}{m} \sum_{i=1}^{m} \hat{\theta}^{(i)},$$

and its variance is given by

$$\begin{aligned} Var(\hat{\theta}) &= (1 + \frac{1}{m}) \frac{1}{m-1} \sum_{i=1}^{m} (\hat{\theta}^{(i)} - \hat{\theta})^2 + \frac{1}{m} \sum_{i=1}^{m} Var(\hat{\theta}^{(i)}) \\ &= \text{between imputation variation} + \text{within imputation variation} \end{aligned}$$

Note that the variance of the parameter estimate has two parts: between imputation variance and within imputation variance. The between imputation variation reflects the missing data uncertainty, while the within imputation variance is the variance

for complete data. Thus, a key difference between a multiple imputation method and a single imputation method (such as the mean imputation method) is that the multiple imputation method incorporates missing data uncertainty measured by the between imputation variation.

Note that the imputation model and the main model for data analysis need not to be the same. For the imputation model, we can include as many variables in X as possible to help predicting the missing data and to make the MAR assumption more reasonable since the imputation model allows the missing data to depend on other observed data. In practice, the number of imputation can be as few as 5 or 6, although the more the better. When the number of imputation is infinite, the procedure may be viewed as equivalent to the EM algorithm. When the missing data rate is not high, say less than 20%, the foregoing multiple imputation method works well even if the imputation model is mis-specified. This is because, in a multiple imputation method, the observed data remain unchanged, although the missing data may be imputed by different values. However, when the missing data rate is high, it is important to choose a good imputation model.

12.4 Multiple Imputation by Chained Equations

In the previous section, we describe a general approach for valid multiple imputations. In this section, we briefly describe a specific popular method for multiple imputations, called the *multivariate imputation by chained equations* ($MICE$). This method is widely used for multiple imputations and has been implemented in R packages **mice** and **mi**. Its basic idea is to impute incomplete multivariate data by fully conditional specifications, i.e., multiple imputation is done as a sequence of small steps in which one only needs to generate data from univariate distributions. This can be implemented via the Gibbs sampler. The method allows the variables in the dataset to be of any types, such as continuous and categorical, and it is computationally efficient.

We describe the basic idea of MICE as follows. Let $X = (X_1, \cdots, X_p)$ be p variables with distribution $P(X|\boldsymbol{\theta})$, and let $X_{j,\text{obs}}$ and $X_{j,\text{mis}}$ be the observed and missing parts of X_j respectively. The variables may be of different types such as continuous or discrete. Let $X_{-j} = (X_1, \cdots, X_{j-1}, X_{j+1}, \cdots, X_p)$ denote the collection of the $p-1$ variables in X without variable X_j. The chained equations method is based on a Bayesian approach, implemented by the Gibbs sampler via iteratively sampling from univariate conditional distributions $P(X_1|X_{-1}, \boldsymbol{\theta}_1), \cdots, P(X_p|X_{-p}, \boldsymbol{\theta}_p)$. Specifically, starting from a simple draw from observed marginal distributions, the t-th iteration of chained equations is based on the following Gibbs sampler which successively draws:

$$
\begin{aligned}
\boldsymbol{\theta}_1^{*(t)} &\sim P(\boldsymbol{\theta}_1 | X_{1,\text{obs}}, X_2^{(t-1)}, \cdots, X_p^{(t-1)}), \\
X_1^{*(t)} &\sim P(X_1 | X_{1,\text{obs}}, X_{-1}^{(t-1)}, \boldsymbol{\theta}_1^{*(t)}), \\
&\cdots\cdots \\
\boldsymbol{\theta}_p^{*(t)} &\sim P(\boldsymbol{\theta}_p | X_{p,\text{obs}}, X_1^{(t-1)}, \cdots, X_{p-1}^{(t-1)}), \\
X_p^{*(t)} &\sim P(X_p | X_{p,\text{obs}}, X_{-p}^{(t-1)}, \boldsymbol{\theta}_p^{*(t)}), \qquad t = 1, 2, 3, \cdots,
\end{aligned}
$$

where $X_j^{(t)} = (X_{j,\text{obs}}, X_j^{*(t)})$. Iterating the above procedure until convergence, at the last iteration we obtain one set of imputations $\{X_1^{*(t)}, \cdots, X_p^{*(t)}\}$ for the missing data. Note that, sampling from the above conditional distributions can be done using standard regression models. For example, if X_j is a continuous variable assuming a normal distribution, we can use linear regression model techniques to draw $X_j^{*(t)}$ from the linear regression model $P(X_j | X_{j,\text{obs}}, X_{-j}^{(t-1)}, \boldsymbol{\theta}_j^{*(t)})$. If X_j is a binary variable assuming a binomial distribution, we can use logistic regression model techniques to draw $X_j^{*(t)}$ from the logistic regression model $P(X_j | X_{j,\text{obs}}, X_{-j}^{(t-1)}, \boldsymbol{\theta}_j^{*(t)})$. Repeating the procedure m times, we create m multiple imputations.

The advantages of the above procedure are (i) at each draw (i.e., simulate a value for the missing value or parameter), we can use familiar univariate regression methods and software; (ii) the incompletely observed variable can be of any type; and (iii) convergence is usually fast. Once each missing data is imputed m times, we can use familiar complete-data methods to analyze the m "complete datasets" separately, and then we combine the m results using the formulas described in the previous section. Usually, $m = 5$ or 6 are good enough, although more is better.

12.5 The EM Algorithm

The likelihood method is a standard general approach for statistical inference, since the maximum likelihood estimates (MLEs) are consistent, efficient, and asymptotically normal, if the assumed model holds. Finding the MLE is thus a main task in statistical inference. In the presence of missing data, however, it may not be easy to find the MLE. The EM algorithm is designed to find MLEs when there are missing data. It has become a very popular method since many complicated statistical problems may be formulated as "missing data" problems, such as mixed effects models and latent variable models in which the unobservable random effects and latent variables may be viewed as "missing data".

The EM algorithm iterates between an E-step and an M-step as follows:

- E-step: we compute the conditional expectation of the "complete-data" log-likelihood given the observed data and the current parameter estimates, where the "complete-data" consist of both the observed data and the missing data.

- M-step: we maximize the conditional expectation in the E-step with respect to the unknown parameters to obtain updated estimates of the parameters.

Given suitable starting values, we iterate between the E-step and the M-step until the parameter estimates converge. At convergence, the final estimates of the parameters are the desired MLEs or candidates for the MLEs.

Specifically, suppose that $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^{\mathrm{T}}$ are data on p variables measured on individual i, with missing values, $i = 1, 2, \cdots, n$. We can write $\mathbf{x}_i = (\mathbf{x}_{\mathrm{obs},i}, \mathbf{x}_{\mathrm{mis},i})$, which consists of the observed part and missing part. Let the main model of interest be $f(\mathbf{x}_i|\theta)$, and let

$$r_{ij} = \begin{cases} 1 & \text{if } x_{ij} \text{ is observed,} \\ 0 & \text{if } x_{ij} \text{ is missing,} \end{cases} \quad i = 1, \cdots, n,\ j = 1, \cdots, p \tag{12.2}$$

be a missing data indicator, with $\mathbf{r}_i = (r_{i1}, \cdots, r_{ip})^{\mathrm{T}}$. To approximate a possible missing data mechanism, we assume a model for the missing data indicator $f(\mathbf{r}_i|\mathbf{x}_i, \boldsymbol{\phi})$, which describes how the probability of missing value may depend on the observed and unobserved values (see model (12.1) for an example). Then, the likelihood for the observed data $\{(\mathbf{r}_i, \mathbf{x}_{\mathrm{obs},i}),\ i = 1, 2, \cdots, n\}$ is given by

$$L_{\mathrm{obs}}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \int \left[f(\mathbf{x}_i|\boldsymbol{\theta}) f(\mathbf{r}_i|\mathbf{x}_i, \boldsymbol{\phi}) \right] \mathrm{d}\mathbf{x}_{\mathrm{mis},i}, \tag{12.3}$$

and the observed-data loglikelihood is $l_{\mathrm{obs}}(\boldsymbol{\theta}) = \log L_{\mathrm{obs}}(\boldsymbol{\theta})$. The observed-data likelihood $L_{\mathrm{obs}}(\boldsymbol{\theta})$ generally does not have a closed form expression, except for linear models, but we can use the EM algorithm to find the MLE of $\boldsymbol{\theta}$ as follows.

The "complete data" is

$$\{(\mathbf{r}_i, \mathbf{x}_i),\ i = 1, 2, \cdots, n\} = \{(\mathbf{r}_i, \mathbf{x}_{\mathrm{obs},i}, \mathbf{x}_{\mathrm{mis},i}),\ i = 1, 2, \cdots, n\},$$

and the "complete data" log-likelihood is

$$l_{\mathrm{com}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left[\log f(\mathbf{r}_i|\mathbf{x}_i, \boldsymbol{\phi}) \log f(\mathbf{x}_i|\theta) \right]. \tag{12.4}$$

Let $\boldsymbol{\theta}^{(0)}$ be the starting value of parameter estimates. At the k-th EM iteration ($k = 0, 1, 2, \cdots$), the E-step computes the conditional expectation of the complete-data loglikelihood given the observed data and the current parameter estimates, i.e.,

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) &= E\left(l_{\mathrm{com}}(\boldsymbol{\theta}) \big| \mathbf{r}_i, \mathbf{x}_{\mathrm{obs},i}, \boldsymbol{\theta}^{(k)} \right) \\ &= \sum_{i=1}^{n} \int \left[\log f(\mathbf{r}_i|\mathbf{x}_i, \boldsymbol{\phi}) + \log f(\mathbf{x}_i|\theta) \right] f(x_{\mathrm{mis},i}|\mathbf{r}_i, \mathbf{x}_{\mathrm{obs},i}, \boldsymbol{\theta}^{(k)})\ \mathrm{d}\mathbf{x}_{\mathrm{mis},i}. \end{aligned}$$

The M-step of the EM algorithm is to maximize $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ with respect to $\boldsymbol{\theta}$ to produce an updated estimate $\boldsymbol{\theta}^{(k+1)}$. The M-step can usually be accomplished by standard optimization procedures such as the Newton-Raphson method. Iterating between the E-step and the M-step, we can show that the likelihood is increasing (or non-decreasing) at each iteration, so eventually the EM algorithm will converge to a (possibly local) maximum. Thus, the MLE can be obtained at last iteration.

To implement an EM algorithm, the main challenge is often the E-step, since it may involve intractable and high-dimensional integrals. Often, Monte Carlo methods or numerical methods may be needed to approximate the integral in the E-step. The M-step is usually easier to implement since standard optimization procedures can be used. We can see that an EM algorithm is designed specifically for the main model of interest $f(\mathbf{x}_i|\theta)$, which makes general software difficult. But for a multiple imputation method, the imputation model can be different from the main model of interest, so general software is available and easy to use in practice. On the other hand, as we can see from the above description, the missing data mechanism $f(\mathbf{r}_i|\mathbf{x}_i, \boldsymbol{\phi})$ can be easily incorporated in an EM algorithm.

In the following, we illustrate the EM algorithms using a simple example, although in this example the EM algorithm is in fact not needed since closed-form MLEs can be obtained. Let $y_1, y_2, \cdots, y_n$ be an i.i.d. sample from the normal distribution $N(\mu, \sigma^2)$, and let $\boldsymbol{\theta} = (\mu, \sigma^2)$ be the unknown parameters. Suppose that $y_1, y_2, \cdots, y_r$ are observed, but $y_{r+1}, y_{r+2}, \cdots, y_n$ are missing, where $r < n$. Assume that the missing data are MAR or MCAR, so the missing data mechanism can be ignored in likelihood inference. Let $\mathbf{y}_{\text{obs}} = (y_1, y_2, \cdots, y_r)$ be the observed data and let $\mathbf{y}_{\text{mis}} = (y_{r+1}, y_{r+2}, \cdots, y_n)$ be the missing data. Then, the "complete data" are $\mathbf{y}_{\text{com}} = (\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}}) = (y_1, y_2, \cdots, y_n)$. The observed-data log-likelihood is given by

$$l_{\text{obs}}(\boldsymbol{\theta}) = -\frac{r}{2}\log(2\pi r\sigma^2) - \frac{1}{2}\sum_{i=1}^{r}\frac{(y_i-\mu)^2}{\sigma^2},$$

and the "complete-data" log-likelihood is given by

$$l_{\text{com}}(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi n\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\sigma^2}.$$

Let $\boldsymbol{\theta}^{(k)}$ be the parameter estimate from the $(k-1)$th EM iteration, $k = 1, 2, 3, \cdots$, and let $\boldsymbol{\theta}^{(0)}$ be the starting values. At k-th EM iteration, the E-step computes the conditional expectation of the "complete-data" log-likelihood given the current

parameter estimates $\boldsymbol{\theta}^{(k)}$ and the observed data $\mathbf{y}_{\text{obs}}$. That is, the E-step computes

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) &= E\left(l_{\text{com}}(\boldsymbol{\theta})|\boldsymbol{\theta}^{(k)}, \mathbf{y}_{\text{obs}}\right) \\ &= \left[-\frac{n}{2}\log(2\pi n\sigma^2) + \frac{n\mu^2}{2\sigma^2}\right] - \frac{1}{2\sigma^2}\left[E\left(\sum_{i=1}^{n} y_i^2 \big| \boldsymbol{\theta}^{(k)}, \mathbf{y}_{\text{obs}}\right)\right. \\ &\quad \left. -2\mu E\left(\sum_{i=1}^{n} y_i \big| \boldsymbol{\theta}^{(k)}, \mathbf{y}_{\text{obs}}\right)\right], \end{aligned}$$

where

$$\begin{aligned} E\left(\sum_{i=1}^{n} y_i | \boldsymbol{\theta}^{(k)}, \mathbf{y}_{\text{obs}}\right) &= \sum_{i=1}^{r} y_i + (n-r)\mu^{(k)}, \\ E\left(\sum_{i=1}^{n} y_i^2 | \boldsymbol{\theta}^{(k)}, \mathbf{y}_{\text{obs}}\right) &= \sum_{i=1}^{r} y_i^2 + (n-r)(\mu^{(k)2} + \sigma^{(k)2}). \end{aligned}$$

The M-step then updates the parameter estimates by maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ with respect to $\boldsymbol{\theta}$, which leads to

$$\begin{aligned} \mu^{(k+1)} &= \frac{1}{n} E\left(\sum_{i=1}^{n} y_i \big| \boldsymbol{\theta}^{(k)}, \mathbf{y}_{\text{obs}}\right) = \frac{1}{n}\left[\sum_{i=1}^{r} y_i + (n-r)\mu^{(k)}\right], \\ (\sigma^{(k+1)})^2 &= \frac{1}{n} E\left(\sum_{i=1}^{n} y_i^2 \big| \boldsymbol{\theta}^{(k)}, \mathbf{y}_{\text{obs}}\right) - (\mu^{(k+1)})^2 \\ &= \frac{1}{n}\left[\sum_{i=1}^{r} y_i^2 + (n-r)(\mu^{(k)2} + \sigma^{(k)2})\right] - (\mu^{(k+1)})^2. \end{aligned}$$

Iterating the E-step and the M-step until convergence, we obtain the following MLEs

$$\hat{\mu} = \frac{1}{r}\sum_{i=1}^{r} y_i, \qquad \hat{\sigma}^2 = \frac{1}{r}\sum_{i=1}^{r} y_i^2 - \hat{\mu}^2,$$

which are the MLEs in the presence of missing data.

12.6 Example in R

In this section, we illustrate the missing data methods using a simple example. To explain the main ideas, we simply use the (partial) "quiz dataset" shown at the beginning of this chapter and treat it as a real dataset with a sample size of 10. Our main purpose here is to demonstrate the use of R packages for missing data. There are several R packages for missing data, including "mice", "norm", and "mi". The R

package "mice" is used to generate multiple imputations based on chained equations (i.e., the MICE method described earlier). The R package "norm" is an earlier software for missing data, and it contains EM algorithm and multiple imputation for multivariate normal data (for multivariate categorical data, there is a R package "cat", and for panel or longitudinal data, there is a R package "pan" – all developed by the same author). The R package "mi" is similar to "mice", although it has some other features. An advantage of "mice" or "mi" over "norm" is that they allow different types of data, without the need to separately treat continuous and discrete data. For simplicity, we assume that the missing data are MCAR or MAR.

```
> install.packages("mice") # (install R package "mice" once)
> library(mice) # do this every time
> options(digits=2)

# "class.dat" is the raw data.
> class.dat2 <- class.dat[,c(4:8)] # extract the five quiz scores

# CC method (delete all incomplete observations)
> class.cc.dat <- class.dat2[!apply(is.na(class.dat2),1,any),]
> mean.cc <- apply(class.cc.dat,2,mean)  # sample means
> sd.cc <- sqrt(apply(class.cc.dat,2,var)/nrow(class.cc.dat)) # sample
    SD
> cor.cc <- cor(class.cc.dat)  # sample correlation matrix
# results based on CC method
> mean.cc
quiz1 quiz2 quiz3 quiz4 quiz5
   66    67    79    79    74
> sd.cc
quiz1 quiz2 quiz3 quiz4 quiz5
  8.3   7.1   7.0   6.4   7.0
> cor.cc
      quiz1 quiz2 quiz3 quiz4 quiz5
quiz1  1.00  0.89  0.33  0.38  0.90
quiz2  0.89  1.00  0.37  0.46  0.93
quiz3  0.33  0.37  1.00  0.98  0.62
quiz4  0.38  0.46  0.98  1.00  0.70
quiz5  0.90  0.93  0.62  0.70  1.00
```

The above results show the results based on the complete-case (CC) method, which simply deletes all incomplete observations (so half the data have to be deleted). This will clearly lead to loss of information, so the CC method will at least be inefficient. In the following, we use the MICE multiple imputation method to impute the missing data. For illustration, we impute each missing value by $m = 3$ "guessed values".

```
# we generate 3 imputations for each missing quiz score
> imp1 <- mice(class.dat2, m=3)
> imp.dat1 <- complete(imp1)  # the imputed dataset
```

```
> complete(imp1)  # first imputed dataset
   quiz1 quiz2 quiz3 quiz4 quiz5
1     90    79    90    90    93
2     55    64    58    60    79
3     60    72    75    80    77
4     66    48    63    55    79
......
> complete(imp1,2) # second imputed dataset
   quiz1 quiz2 quiz3 quiz4 quiz5
1     90    79    90    90    93
2     55    64    58    65    79
3     60    72    75    80    77
4     66    48    58    55    79
......
> complete(imp1,3)  # third imputed dataset
   quiz1 quiz2 quiz3 quiz4 quiz5
1     90    79    90    90    93
2     55    64    58    55    79
3     60    72    75    80    77
4     66    48    75    60    77
......
> imp1$imp$quiz5   # check for imputed values of quiz5
   1  2  3
4 79 79 77
6 61 79 56
8 84 85 56
```

We see that the multiple imputation method only imputes the missing data while keeps the observed data unchanged. For example, for scores of quiz5, there are missing values for three students (No. 4, 6, 8). For student No. 6, the three imputed values for the (one) missing quiz score are 61, 79, and 56. Note that these imputed values may change if the method is run again, which reflects the missing data uncertainty, but the range of the imputed values is within certain limit (determined by the observed data). Sometimes some imputed values may be the same (e.g., the first two imputed values for student No. 4).

Next, we do some analyses and modelling based on imputed data.

```
> apply(imp.dat1,2,mean)  # sample mean
quiz1 quiz2 quiz3 quiz4 quiz5
   62    60    66    69    75
> sqrt(apply(imp.dat1,2,var)/10) # sample SD
quiz1 quiz2 quiz3 quiz4 quiz5
  4.3   4.5   7.5   4.6   3.8
# sample correlation matrix based on each imputed dataset
> cor1 <- with(imp1, cor(cbind(quiz1,quiz2,quiz3,quiz4,quiz5)))
[[1]]
      quiz1 quiz2   quiz3 quiz4   quiz5
quiz1  1.00  0.73  0.4428  0.44  0.6267
quiz2  0.73  1.00  0.5834  0.65  0.5608
```

```
quiz3  0.44  0.58  1.0000  0.78 -0.0083
quiz4  0.44  0.65  0.7804  1.00  0.3200
quiz5  0.63  0.56 -0.0083  0.32  1.0000

[[2]]
      quiz1 quiz2  quiz3 quiz4  quiz5
quiz1  1.00  0.73 0.4659  0.41 0.6714
quiz2  0.73  1.00 0.5720  0.65 0.4562
quiz3  0.47  0.57 1.0000  0.77 0.0077
quiz4  0.41  0.65 0.7690  1.00 0.2890
quiz5  0.67  0.46 0.0077  0.29 1.0000
......

# Fit a linear regression, with quiz5 as response.
> fit1 <- with(imp1, lm(quiz5~quiz1+quiz2+quiz3+quiz4))
 ## summary of imputation 1 :
       ...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   26.851     17.720    1.52    0.190
quiz1          0.476      0.326    1.46    0.204
quiz2          0.268      0.370    0.72    0.501
quiz3         -0.440      0.208   -2.12    0.088
quiz4          0.459      0.362    1.27    0.261
     ...
 ## summary of imputation 2 :
       ...
Coefficients:
            Estimate Std.Error t value Pr(>|t|)
(Intercept)   25.860     17.231    1.50    0.194
quiz1          0.739      0.295    2.50    0.054
quiz2         -0.124      0.334   -0.37    0.725
quiz3         -0.389      0.181   -2.15    0.085
quiz4          0.534      0.340    1.57    0.177

............
```

Compared with the CC method, the analysis results based on the multiple imputation method can be quite different. For example, the average score for "quiz4" based on the CC method is 79, while it is 69 based on the multiple imputation method (one possible explanation is that the students missed quiz4 are weak ones). In particular, the standard deviations based on the multiple imputation method are much smaller than those based on the CC method, since the CC method has a much smaller sample size due to deletion of data. The estimated correlations are also different. For example, the estimated correlation between quiz1 and quiz5 is 0.90 based on the CC method, but this value is 0.63 based on the first imputed dataset. Regression analysis can also be performed, as shown above. Since we have imputed 3 values for each missing value, we have 3 results for the correlation and regression

analysis. We can then combine these results to obtain an overall (one) results.

In the above analysis, we only consider the continuous data (quiz scores). The "mice" package can also be used to impute discrete data. For example, we can impute the missing "gender" and "major" in the original data. In this case, we can specify the types of variables using the "method" option. For variable "ID", we don't need to impute any values, so we use option " ". For variable "gender", since it is a binary variable, we use a logistic regression model to generate the imputations (i.e., using the quiz scores to predict missing "gender") – this can be done using option "logreg". For variable "major", it is a three-category discrete variable, so we use option "polyreg". The quiz scores are continuous variables, so we use option "norm".

```
> imp2 <- mice(class.dat, m=3,
               method=c("","logreg","polyreg","norm","norm","norm","
                   norm","norm"))
> imp2$imp$gender
  1 2 3
6 F F F
> imp2$imp$major
     1    2    3
7 Comp Comp Comp
```

We see that the missing "gender" is predicted as "female" based on the quiz scores, and the missing "major" is predicted as "comp" (computer), based on all three imputations.

In the following, we show the results based on the EM algorithm using R package "norm". The EM algorithm may be viewed as a multiple imputation method with infinite many imputations.

```
> install.packages("norm")
> library(norm)

> s <- prelim.norm(as.matrix(class.dat2)) # preliminary manipulation
> thetahat <- em.norm(s) # parameter estimates based on EM
> em0 <- getparam.norm(s,thetahat, corr=T) # get the parameter
    estimates
> mean.em <- em0$mu  # sample mean
> sd.em <- em0$sdv/sqrt(nrow(class.dat))  # sample SD
> cor.em <- em0$r  # sample correlations
# print the EM results
> mean.em
[1] 62 61 64 67 68
> sd.em
[1] 4.1 4.5 7.4 6.2 4.6
> cor.em
     [,1] [,2] [,3] [,4] [,5]
```

```
[1,] 1.00 0.66 0.45 0.42 0.70
[2,] 0.66 1.00 0.51 0.65 0.94
[3,] 0.45 0.51 1.00 0.96 0.69
[4,] 0.42 0.65 0.96 1.00 0.82
[5,] 0.70 0.94 0.69 0.82 1.00
```

We see that the EM algorithm gives results different from the CC method and the multiple imputation method. The EM results are closer to the multiple imputation results than to the CC results. Note that, when the number of imputation increases, the multiple imputation results are increasingly close to the EM results.

An advantage of the multiple imputation method is that, once the missing data are imputed, the "complete datasets" can be analyzed using any models or methods, which may be different from the models used to generate the imputations. However, the EM algorithm is specific to the model, i.e., different models and methods require different implementations of EM algorithms.

Exercises 12

12.1. Find an example where the missing data are MCAR, MAR, and nonignorable respectively.

12.2. Derive the EM algorithm for two-dimensional multivariate normal distribution with missing data.

12.3. For Example 1, try to use different imputation models by removing some variables in the imputation model and then compare the results. What do you find?

12.4. Show that, for likelihood inference, the missing data mechanism may be ignored when the missing data are MCAR or MAR.

Chapter 13

Robust Multivariate Analysis

13.1 The Need for Robust Methods

An *outlier* is an observation which appears to be inconsistent with the rest of the data, i.e., an outlier is an unusual observation in a dataset. An outlier can have a great impact on statistical analysis and may completely change analysis results. In other words, statistical analysis with and without an outlier can be quite different. When we make decisions based on statistical analysis, we should not let a few outliers in the dataset to completely change the conclusions since these outliers do not represent the population.

As a simple example, suppose that a dataset consists of five values 2,3,5,7, 19. Here, the value of 19 is clearly an outlier. The means (variances) with and without the outlier are 7.2 (47.2) and 4.25 (4.92) respectively, which are quite different. This leads to very different confidence intervals and test statistics about the population mean: the 95% confidence intervals are (–6.26, 20.66) and (–0.09, 8.59) respectively, and the p-values are 0.15 and 0.03 respectively. Thus, the results with and without the outlier are quite different, and the result with the outlier can be quite misleading. This simple example illustrates how an outlier can completely change analysis results. Note that the *median* for this dataset is 5 with or without outlier, so the median is robust to outliers. However, many standard statistics and methods are very sensitive to outliers. For example, the mean, variance (standard deviation), correlation, and the maximum likelihood estimates are all very sensitive to outliers, i.e., statistical analysis results based on these statistics or methods can be completely changed by a few outliers in the data. Robust statistics, such as the median, are not sensitive to outliers, so they can produce better and more reliable results.

For univariate data, outliers are relatively easy to detect using graphical tools. For example, a histogram, a boxplot, or a QQ plot of a univariate continuous data may reveal outliers. However, for multivariate data, outliers can be hard to detect, since it may be difficult to use graphical tools to display multivariate data. Moreover, a multivariate outlier need not be an outlier in any of the coordinates when considered separately. Figure 13.1 shows an example where a multivariate outlier is

not an outlier on either the x-coordinate nor the y-coordinate. This figure displays a bivariate dataset with incomes for 25 years old college graduates and incomes for 45 years old college graduates. The sample correlation coefficients r with the outlier and without the outlier are -0.51 and -0.80 respectively, which is a big difference due to one outlier in the data. However, this outlier may not be easy to detect since it's not an outlier when just looking at the incomes for the 25 years old or the incomes for the 45 years old separately. Finding multivariate outliers using some tests, such as the χ^2 test mentioned in Chapter 1, may not work well when there are more than one outliers or when the sample size is small. Therefore, for multivariate data, it is particularly important to use robust statistical methods which automatically detect or downweight outliers in the data, if any.

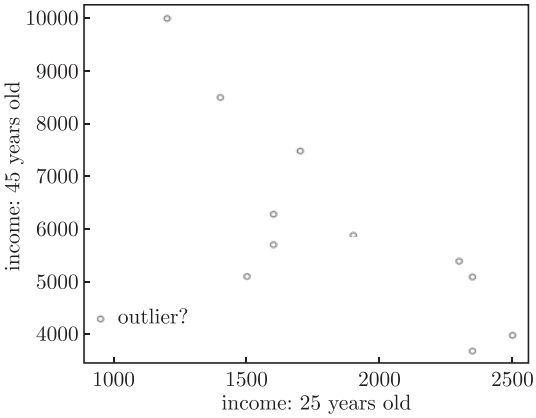


Figure 13.1 Outlier for two-dimensional data. The correlations with and without outlier are –0.51 and –0.80 respectively.

The consequence of outliers in multivariate data is more complex than in the univariate case. A multivariate outlier can distort estimates of means, variances, and correlations, and thus distort estimate of the covariance matrices. As we have seen, the means, variances, correlations, and the covariance matrices play key roles in multivariate analysis, but their estimates can be greatly influenced by a few outliers in a dataset (i.e., their estimates with and without a few outliers can be very different). Therefore, in multivariate analysis, it is very important to obtain reliable estimates of the means, variances, correlations, and the covariance matrices by addressing possible outliers in a dataset. On the other hand, for multivariate dataset with dimension (or number of variables) higher than 2, a graphical display of the data would be difficult, unlike the univariate case where an outlier can be detected graphically. So outliers in multivariate data are hard to detect, but they may lead to very misleading results.

Since a few outliers may lead to very misleading results in multivariate analysis but it may be hard to detect these outliers, robust statistical methods which address outliers automatically are quite valuable. There has been extensive research on robust statistical methods for multivariate data. See Maronna, Martin, and Yohai (2006) for a comprehensive review of these methods. The basic ideas of these robust methods are to either accommodate outliers via assuming appropriate heavy-tail distributions or automatically downweight potential outliers in parameter estimation. In other words, these robust methods are designed so that we do not need to detect outliers before data analysis. Instead, we only need to use these robust methods for data analysis which automatically incorporate possible outliers if they exist. Moreover, if no outliers exist in a dataset, these robust methods are equivalent to standard statistical methods. Therefore, in data analysis, we can simply use these robust methods and compare the analysis results with those obtained using standard statistical methods to see if the results differ substantially or not. If the results are very different, there are possibly outliers in the dataset and we should use the robust methods for data analysis. Otherwise, if the results are similar, there may be no outliers in the dataset, and we can simply use standard statistical methods for data analysis.

In the next section, we briefly review the basic ideas of the two most common robust methods. These basic ideas are very general and thus apply to both univariate data and multivariate data.

13.2 General Robust Methods

There are two general approaches for robust statistical analysis. One approach is to assume heavy-tail distributions, such as t-distributions instead of normal distributions, for the data. Because the assumed distributions have heavy-tails, they allow very small or very large values in a dataset to occur more often than lighter tail distributions. Another approach, which is more frequently used, is to automatically down-weight outliers in a dataset if they exist. That is, outliers receive less weights in parameter estimation so that their influences on the estimates are reduced, leading to more reliable parameter estimates. Specifically, the two most commonly used robust approaches are

- assuming robust distributions for the data, such as assuming t-distributions instead of normal distributions for the data since t-distributions have heavier tails than normal distributions;
- using robust estimation methods, such as the M-estimators which downweight outliers in parameter estimation.

In this section, we briefly describe the basic ideas of these two methods.

13.2.1 The *t*-distribution

Normal distributions are the most common assumptions for many statistical models and methods. However, a normal distribution has thin-tails which may not accommodate outliers in the data well. For example, if data are assumed to follow a normal distribution, then 99% of the data should fall within 3 standard deviations of the mean. That is, if $X \sim N(\mu, \sigma^2)$, then $P(\mu - 3\sigma < X < \mu + 3\sigma) > 0.99$. This assumption may be violated if there are outliers in the data which are outside 3 standard deviations of the mean. A t-distribution with small degrees of freedom, however, has heavier tails than a corresponding normal distribution with the same mean, so a t-distribution can accommodate some outliers. A t-distribution approaches to a normal distribution as its degrees of freedom increase, so one can choose the degree of freedom in a t-distribution to either increase or decrease its robustness to outliers. A t-distribution is similar to a corresponding normal distribution (with the same mean) in shape, i.e., they are both bell-shaped and symmetric. Therefore, we can use a t-distribution to replace the corresponding normal distribution for robust inference.

The probability density function of the *t-distribution* with k degrees of freedom, denoted by $t(k)$, can be written as

$$f(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\,\Gamma\left(\frac{k}{2}\right)}\left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}}, \qquad -\infty < x < \infty,$$

where $\Gamma(x)$ is the Gamma function $\Gamma(x) = \displaystyle\int_0^\infty u^{x-1}e^{-u}\mathrm{d}u$. Its mean is 0 and its variance is $k/(k-2)$ for $k > 2$. Since a t-distribution is symmetric, its mean and its median are the same. Figure 13.2 shows the density functions of standard t-distributions, with degrees of freedom being 1 and 4 respectively, and the standard normal distribution $N(0, 1)$. We see that the three distributions are similar in shape but the t-distributions have heavier tails than the normal distribution.

The multivariate t-distribution is a multivariate generalization of the univariate t-distribution. Let $\boldsymbol{X} = (x_1, x_2, \cdots, x_p)^{\mathrm{T}}$ be a p-dimension random vector. The probability density function of $\boldsymbol{X}$ following a *multivariate t-distribution with k degrees of freedom and parameters* $\boldsymbol{\mu}$ *and* Σ, denoted by $t_p(\boldsymbol{\mu}, \Sigma, k)$, is given by

$$f(\mathbf{x}) = \frac{\Gamma\left(\frac{k+p}{2}\right)}{\Gamma(k/2)k^{p/2}\pi^{p/2}|\Sigma|^{1/2}\left[1 + \frac{1}{k}(\boldsymbol{X} - \boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{X} - \boldsymbol{\mu})\right]^{(k+p)/2}}, \qquad \boldsymbol{x} \in \mathbf{R}^p,$$

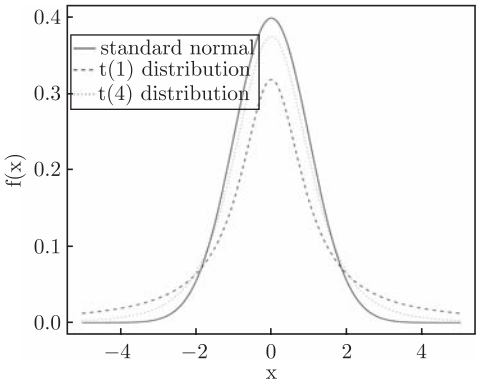


Figure 13.2 Probability density functions for the standard normal distribution and the t distributions with degrees of freedom of 1 and 4 respectively.

where $\boldsymbol{\mu}$ is a $p \times 1$ vector and Σ is a $p \times p$ matrix. The mean vector and the variance-covariance matrix of $\boldsymbol{X}$ are given by $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and $Cov(\mathbf{X}) = \left(\frac{k}{k-2}\right)\Sigma$ respectively. Note that the variance-covariance matrix of $\boldsymbol{X}$ exists only for $k > 2$. For robust inference, we can replace the multivariate normal distributions assumed in a model by the corresponding multivariate t-distributions.

Although the t-distribution may be more desirable than the normal distribution to accommodate outliers, it has several limitations. First, the t-distributional assumption may not hold for some data; second, assuming t-distributions makes models and methods more complicated than those assuming normal distributions; Third, assuming t-distribution may not accommodate outliers well for some problems. Thus, alternative robust methods are required. In the following, we introduce the M-estimators, which is a more popular robust method and is widely used in robust inference.

13.2.2 The M-estimator

The M-estimator and its extensions are probably the most popular formal robust methods in statistical literature. The basic idea of an M-estimator is to bound or down weight the influence of potential outliers in parameter estimation. Note that, in parameter estimation, we often need to solve a set of equations, such as the likelihood equations or the least square equations. The M-estimation method introduces a function which bounds or downweight outliers in a set of estimating equations for parameter estimation. Solving the bounded estimation equations leads to robust parameter estimates which are less sensitive to outliers than unbounded estimating equations.

Specifically, consider a model with distribution $f(y, \boldsymbol{\theta})$. The MLEs of parameters $\boldsymbol{\theta}$ based on data $\{y_1, y_2, \cdots, y_n\}$ can be obtained by minimizing the quantity

$$\sum_{i=1}^{n} \rho(y_i, \boldsymbol{\theta}),$$

where $\rho(y_i, \boldsymbol{\theta}) = -\log f(y_i, \boldsymbol{\theta})$, $\psi(y_i, \boldsymbol{\theta}) = -\partial\rho(y_i, \boldsymbol{\theta})/\partial\boldsymbol{\theta}$. This is equivalent to solving the following estimating equation

$$\sum_{i=1}^{n} \psi(y_i, \boldsymbol{\theta}) = 0. \tag{13.1}$$

For example, if $y_1, y_2, \cdots, y_n$ i.i.d. $\sim N(\mu, 1)$, then $\psi(y_i, \mu) = y_i - \mu$ and $\rho(y_i, \mu) = (y_i - \mu)^2/2$. This motivates the following M-estimators.

An *M-estimator* (maximum likelihood type estimator) are generalizations of MLEs in which the functions $\rho(y_i, \boldsymbol{\theta})$ or $\psi(y_i, \boldsymbol{\theta})$ in (13.1) are not necessarily related to a probability density function. Instead, function $\psi(y_i, \boldsymbol{\theta})$ can be chosen to downweight outliers in such a way that $\psi(x)$ is close to $|x|$ when $|x|$ is small and $\psi(x)$ remains small when $|x|$ is large. The most popular choice of $\psi(y_i, \boldsymbol{\theta})$ in (13.1) for an M-estimator is the following well-known *Huber's function*

$$\psi(x) = \begin{cases} x & \text{if } |x| \leqslant c, \\ c & \text{if } x > c, \\ -c & \text{if } x < -c \end{cases} \tag{13.2}$$

where the constant $c > 0$ is a turning point which controls the robustness of the M-estimator. The corresponding Huber's ρ-function is given by

$$\rho(x) = \begin{cases} x^2/2, & \text{for } |x| \leqslant c, \\ c|x| - x^2/2, & \text{for } |x| > c. \end{cases} \tag{13.3}$$

Thus, the Huber's function replaces the quadratic function in the normal likelihood by a function with a bounded derivative.

When the turning point c goes to ∞, the M-estimator approaches to the sample mean (the MLE of the population mean). When the turning point c goes to 0, the M-estimator approaches to the median. The larger the value of c, the closer the M-estimator to MLE. In other words, the smaller the value of c, the more robust the M-estimator. In practice, we can choose the turning point c which achieves a balance between robustness and efficiency, such as $c = 2$.

We illustrate the basic idea using a simple example. Let $\{y_1, y_2, \cdots, y_n\}$ be a sample from normal distribution $N(\mu, 1)$. The log-likelihood function for estimating parameter μ is given by

$$l(\mu) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^2.$$

Maximizing $l(\mu)$ is equivalent to minimizing $\sum_{i=1}^{n} \rho(y_i, \mu)$, where $\rho(y_i, \mu) = (y_i - \mu)^2/2$. MLE of μ can be obtained by solving the following likelihood equation

$$\sum_{i=1}^{n} \psi(y_i, \mu) = \sum_{i=1}^{n} (y_i - \mu) = 0,$$

where $\psi(y_i, \mu) = y_i - \mu$. The solution (MLE) of the above estimating equation has a unbounded influence function, i.e., the influence of an outlier on the MLE is unbounded. In other words, an outlier can have a large influence on the resulting MLE $\hat{\mu}$. For a robust estimator of μ, we can replace the above $\psi(y_i, \mu)$ by the Huber's function $\psi_c(y_i, \mu)$ given in (13.2). Then, we solve the following robust estimating equation

$$\sum_{i=1}^{n} \psi_c(y_i, \mu) = 0,$$

which bounds or downweights influences of large or small values of y_i's. The solution of the above robust estimating equation, $\tilde{\mu}_c$, is a robust estimator of μ. With an appropriate choice of the turning point c, the robust estimator $\tilde{\mu}_c$ will be less sensitive to outliers.

It can be shown that M-estimators are asymptotically normally distributed. M-estimators are popular in robust inference due to their generality and efficiency. There are many other robust methods such as R-estimators and L-estimators, but M-estimators may be the most popular. For more detailed discussion of robust methods, see Maronna, Martin, and Yohai (2006).

13.3 Robust Estimates of the Mean and Standard Deviation

Means and variances or standard deviations are most useful summary statistics for continuous data, but they are very sensitive to outliers. Robust versions of these statistics are available. We first consider the univariate case, and then we extend the estimates to the multivariate case. As noted earlier, a robust alternative to the sample mean is the sample *median*. Let $\boldsymbol{X} = \{x_1, x_2, \cdots, x_n\}$ be an i.i.d. sample. The *median* of the data $\boldsymbol{X}$, denoted by $Med(\boldsymbol{X})$, is the value such that half of the values in $\mathbf{X}$ are smaller than $Med(\boldsymbol{X})$ and half are larger than $Med(\boldsymbol{X})$. That is, the median is the 50th percentile. When there are even number of observations in $\boldsymbol{X}$ or there are ties, the median is the average of the middle points. Consider a continuous population with mean μ and variance σ^2. A robust estimate of the population mean μ is the median

$$\tilde{\mu} = Med(\boldsymbol{X}),$$

which is not sensitive to outliers. That is, the median estimate of the population mean is not influenced by outliers in the data. An M-estimator of μ may be obtained by solving the following equation

$$\sum_{i=1}^{n} \psi\left(\frac{x_i - \mu}{\tilde{\sigma}}\right) = 0,$$

where $\psi()$ can be chosen as the Huber's function and $\tilde{\sigma}$ is a robust estimate of the standard deviation as described below. The median $\tilde{\mu}$ may be viewed as an M-estimator which minimizes $\sum_{i=1}^{n} \rho(||x_i - \mu||)$ over all μ, where ρ is chosen as the Huber's function with turning point $c \to 0$.

A popular robust estimate of the standard deviation is the *median absolute deviation about the median (MAD)*, which is defined as

$$MAD(\boldsymbol{X}) = Med\{|\boldsymbol{X} - Med(\boldsymbol{X})|\},$$

i.e., $MAD(\boldsymbol{X})$ is the median of the values in $\{|\boldsymbol{X} - Med(\boldsymbol{X})|\} = \{|x_1 - \tilde{\mu}|, |x_2 - \tilde{\mu}|, \cdots, |x_n - \tilde{\mu}|\}$. Thus, a robust estimate of the population standard deviation σ is

$$\tilde{\sigma} = MAD(\boldsymbol{x}),$$

which is not sensitive to outliers.

There are also other robust estimates of the means and standard deviations. For example, an alternative robust estimate of the mean is the *trimmed mean*, which is the sample mean of the smaller dataset after discarding a small proportion of the largest and smallest values in the original dataset. An alternative robust estimate of the standard deviation is the *inter-quartile range* (IQR), which is the difference between the first quartile (25th percentile) and the third quartile (75th percentile), i.e., the length of the box in a boxplot.

In order to compare different robust estimators and methods, a measure of the robustness of an estimator or method is required. One such measure of robustness is the *breakdown point*, which can be roughly defined as the smallest fraction of data in a sample that can render the estimator useless. A breakdown point of zero means that the presence of even a single outlier can completely distort the estimate. For example, the sample mean $\bar{x}$ has a breakdown point of zero since a single outlier in the data can completely distort its value (e.g., the sample mean can become arbitrary large if one observation in the sample becomes arbitrary large while other observations remain the same). The highest breakdown point is 50%. The sample median has a breakdown point of 50% since the value of the median remains unchanged even if nearly 50% of the data in the sample are outliers (e.g., the

median remains the same even if nearly half the data become arbitrary large while the rest data remain unchanged). The breakdown points of the sample standard deviation, the MAD, and the IQR are 0, 50%, and 25% respectively, so the MAD is the most robust estimate while the sample standard deviation is the least robust estimate of the population standard deviation.

For multivariate data, we can use the above robust estimates for each component of the vector of p random variables. For example, let $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n\}$ be an i.i.d. sample from a multivariate population with mean vector $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_p)$ and covariance matrix $\Sigma = (\sigma_{ij})_{p\times p}$, where each $\mathbf{x}_k = (x_{1k}, x_{2k}, \cdots, x_{pk})^{\mathrm{T}}$, $k = 1, 2, \cdots, n$. Then, a robust estimate of the mean vector is

$$\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \cdots, \tilde{\mu}_p),$$

where each $\tilde{\mu}_j$ is the sample median of $\{x_{j1}, x_{j2}, \cdots, x_{jn}\}$, $j = 1, 2, \cdots, p$. Or, a robust estimate of the mean vector can be obtained as an M-estimator by minimizing

$$\sum_{i=1}^{n} \rho(||\boldsymbol{X}_i - \boldsymbol{\mu}||)$$

over all $\boldsymbol{\mu}$. Similarly, a robust estimate of the standard deviation $\sigma_{jj}^{1/2}$ is the median of the values in

$$\{|x_{j1} - \tilde{\mu}_j|, \cdots, |x_{jn} - \tilde{\mu}_j|\}, \quad j = 1, 2, \cdots, p.$$

A robust estimate of the covariance matrix Σ is more complex and is described next.

13.4 Robust Estimates of the Covariance Matrix

The Covariance matrix or the correlation matrix play a key role in multivariate analysis, so it is important to obtain reliable estimates of the covariance or correlation matrices which are not influenced by outliers in the data. Robust estimates of the covariance or correlation matrices are more complicated than those for the means and variances. A simple method is to use the following identity

$$Cov(x, y) = \frac{1}{4}(SD(x + y)^2 - SD(x - y)^2),$$

where $SD(x + y)$ is the standard deviation of $x + y$. A robust estimate of the covariance is then obtained by replacing $SD(x + y)$ by its robust version such as $MAD(x + y)$, i.e.,

$$\widetilde{cov}(x, y) = \frac{1}{4}(\tilde{\sigma}_{x+y}^2 - \tilde{\sigma}_{x-y}^2),$$

where $\tilde{\sigma}_{x+y} = MAD(x+y)$ and $\tilde{\sigma}_{x-y} = MAD(x-y)$. M-estimators are also available. Robust covariances and correlations based on M-estimators may be obtained as follows

$$\widetilde{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}\left[\psi\left(\frac{x-\tilde{\mu}_x}{\tilde{\sigma}_x}\right)\psi\left(\frac{x-\tilde{\mu}_y}{\tilde{\sigma}_y}\right)\right]\tilde{\sigma}_x\tilde{\sigma}_y,$$

$$\tilde{r}_{xy} = \widetilde{corr}(x,y) = \frac{\widetilde{cov}(x,y)}{\tilde{\sigma}_x\tilde{\sigma}_y}.$$

where $\psi()$ can be chosen as the Huber's function. Thus, robust estimates of the covariance matrix Σ and correlation matrix R are

$$\tilde{\Sigma} = (\tilde{\sigma}_{ij})_{p\times p}, \qquad \tilde{\boldsymbol{R}} = (\tilde{r}_{ij})_{p\times p}$$

respectively, where $\tilde{\sigma}_{ij} = \widetilde{cov}(x_i, x_j)$ and $\tilde{r}_{ij} = \widetilde{corr}(x_i, x_j)$, $i, j = 1, 2, \cdots, p$.

The above robust covariance matrices are simple and easy to use, but they may not be positive definite. There are better robust estimates of the covariance matrix, although they are more complicated. Note that, for multivariate data, it is desirable to estimate the mean vector and covariance matrix simultaneously. Here we describe M-estimates for the mean vector and covariance matrix. For either a multivariate normal distribution or a multivariate t-distribution, their density function can be written as

$$f(\boldsymbol{X}, \boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{|\Sigma|}}h(d(\boldsymbol{X}, \boldsymbol{\mu}, \Sigma)), \qquad \boldsymbol{X} \in \boldsymbol{R}^p,$$

where

$$d(\boldsymbol{X}, \boldsymbol{\mu}, \Sigma) = (\boldsymbol{X}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{X}-\boldsymbol{\mu})$$

measures the "distance" of observation $\boldsymbol{X}$ to the center $\boldsymbol{\mu}$, so a large distance may be defined as a potential outlier. For a p-variable multivariate normal distribution $N_p(\boldsymbol{\mu}, \Sigma)$, we have

$$h(d) = (2\pi)^{-p/2}\exp(-d/2).$$

For a multivariate t-distribution with k degrees of freedom $t_p(\boldsymbol{\mu}, \Sigma, k)$, we have

$$h(d) = \frac{c}{(d+k)^{(p+k)/2}},$$

and c is a constant. Let $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n\}$ be an i.i.d. sample, and let $d_i = d(\boldsymbol{X}_i, \boldsymbol{\mu}, \Sigma), i = 1, 2, \cdots, n$. The MLEs of $\boldsymbol{\mu}$ and Σ are solutions of

$$\sum_{i=1}^{n} W(d_i)(\boldsymbol{X}_i - \boldsymbol{\mu}) = 0,$$

$$\frac{1}{n}\sum_{i=1}^{n} W(d_i)(\boldsymbol{X}_i - \boldsymbol{\mu})(\boldsymbol{X}_i - \boldsymbol{\mu})^{\mathrm{T}} = \Sigma,$$

respectively, where $W(d) = -2\,\partial \log h(d)/\partial d$. We can write the solutions as weighted averages

$$\tilde{\boldsymbol{\mu}} = \sum_{i=1}^{n} \left[\frac{1}{\sum_{i=1}^{n} W(d_i)} \right] \mathbf{X}_i,$$

and $\tilde{\Sigma}$, which does not have a closed-form expression, may also be interpreted as a weighted covariance matrix. For the multivariate normal distribution $N_p(\boldsymbol{\mu}, \Sigma)$, $W(d_i) \equiv 1$ (i.e., each observation receives the same weight), and the resulting estimates are the familiar sample mean and sample covariance matrix. For the multivariate t-distribution $t_p(\boldsymbol{\mu}, \Sigma, k)$, which leads to more robust estimates, we have

$$W(d_i) = \frac{p + k}{d_i + k}.$$

For this weight function $W(d_i)$, it decreases as d_i increases, so it downweights observations $\mathbf{x}_i$ with large values of d_i (potential outliers). As the degree of freedom $k \to \infty$, we have $t_p(\boldsymbol{\mu}, \Sigma, k) \to N_p(\boldsymbol{\mu}, \Sigma)$ and $W(d) \to 1$. Therefore, robust estimates of the mean vector and the covariance matrix can be obtained as MLEs under the t-distribution assumption.

A perhaps more popular robust estimate of the covariance matrix is the so-called minimum volume ellipsoid estimator. To explain the basic idea of this estimator, note that robust estimates of $\boldsymbol{\mu}$ and Σ may be obtained by minimizing the distance $d(\boldsymbol{x}, \boldsymbol{\mu}, \Sigma)$, subject to the restriction $|\Sigma| = 1$ to avoid trivial estimates. Such an estimate of the covariance matrix is called the *minimum volume ellipsoid* (MVE) *estimator*. It is widely used in statistical literature. The name stems from the fact that, among all ellipsoids $\{\boldsymbol{x} : d(\boldsymbol{x}, \boldsymbol{\mu}, \Sigma) \leqslant 1\}$ containing at least half the data points, the MVE estimate has the minimum volume, i.e., the minimum $|\Sigma|$.

13.5 Robust PCA and Regressions

Robust versions of many multivariate analysis models and methods are available. In this section, we briefly discuss robust PCA and robust regression models.

Principal component analysis (PCA) is often the first step of multivariate data analysis. Standard PCA involves computing the eigenvectors and eigenvalues of the sample covariance or correlation matrices. However, as noted earlier, the sample covariance or correlation matrices are sensitive to outliers. In other words, in the presence of outliers, the sample covariance or correlation matrices may be unreliable, as shown in the example presented at the beginning of this chapter. Thus, PCA results based on the usual sample covariance or correlation matrices may be misleading if outliers may be present in the data, and a robust version of PCA is

desirable. A simple and intuitive approach is to perform PCA on the robust estimate of the covariance or correlation matrix. A robust estimate of the covariance or correlation matrix can be obtained using the methods described in the previous section. Then, we can conduct a PCA on the robust estimate. More complicated robust PCA methods are also available. However, since a PCA is usually a preliminary screening method in data analysis, a simple but reasonable robust PCA may be sufficient in initial data analysis.

For regression models, it is well known that the least square estimates or the maximum likelihood estimates are sensitive to outliers. In data analysis, residual plots and Cook's distance plots can be used to identify outliers, but sometimes it may be difficult to identify such outliers so robust methods are desirable. Consider the following linear regression model

$$y_i = \mathbf{x}_i^{\mathrm{T}}\beta + e_i, \qquad e_i \ \text{ i.i.d.} \ \sim N(0, \sigma^2), \qquad i = 1, 2, \cdots, n.$$

We focus on the case where outliers may be present in the responses y_i, but not in the predictors $\mathbf{x}_i$. The least square estimates or the maximum likelihood estimates of β are obtained by minimizing $\sum_{i=1}^{n}(y_i - \mathbf{x}_i^{\mathrm{T}}\beta)^2$, which is equivalent to solving the following estimating equation

$$\sum_{i=1}^{n}(y_i - \mathbf{x}_i^{\mathrm{T}}\beta)\mathbf{x}_i = 0, \qquad \text{or} \qquad \sum_{i=1}^{n}\psi\left(\frac{r_i}{\sigma}\right)\mathbf{x}_i = 0,$$

where $\psi(x) = x$ and $r_i = y_i - \mathbf{x}_i^{\mathrm{T}}\beta$. An outlier in the response y_i will lead to a large value of r_i. Here the function $\psi(x) = x$ is unbounded (i.e., a large value of x leads to a large value of $\psi(x)$). Thus the resulting parameter estimates will be influenced by an outlier. For robust analysis, we may choose a ψ function to bound large values of r_i (i.e., downweight large values of r_i) so that large values of r_i have less influence on the resulting parameter estimates. Specifically, the *M-estimator* of the parameter β can be obtained by solving the following estimating equation

$$\sum_{i=1}^{n}\psi\left(\frac{r_i}{\sigma}\right)\mathbf{x}_i = 0, \tag{13.4}$$

where the function $\psi(x)$ can be chosen to be the Huber's function which downweights large values of x. The solution to equation (13.4) is a robust estimate of β, and the robustness of this estimate can be controlled by choosing appropriate turning points in the Huber's function. Usually, the parameter σ is substituted by its robust estimate, but simultaneous estimates for β and σ are available by another estimating equation for σ.

Other robust approaches to regression analysis are also available. For example, the *least median of square estimators* are obtained by minimizing the median of $\{(y_i - \mathbf{x}_i\beta)^2, i = 1, 2, \cdots, n\}$. We may also consider robust regression analysis when there are outliers in the predictors $\mathbf{x}_i$.

13.6 Examples in R

Example 1. To illustrate robust methods for data analysis using R, we again consider the simple income dataset described at the beginning of this chapter and shown in Figure 13.1. From Figure 13.1, we see that multivariate outliers may be difficult to detect visually but they can have big impact on analysis results (e.g., the correlations with and without a single outlier can be as different as -0.51 and -0.80). Thus, robust methods which automatically incorporate outliers are useful in data analysis and may give more reliable results. We use R package "robustbase" for illustration (note that some methods in this package are not described in this chapter but are described in the help files of the R package).

```
> install.packages("robustbase")
> library(robustbase)
> options(digits=2)

# X contains the dataset. First, get summary statistics ignoring
    outliers.
# sample mean vector
> x_mean <- apply(X,2,mean)
> x_mean
yr25 yr45
1779 5958
# covariance matrix
> x_cov <- cov(X)
> x_cov
        yr25    yr45
yr25  252027 -483220
yr45 -483220 3529924
# Next, get robust summary statistics
> x_cov_robust <- covMcd(X)
> x_cov_robust
...
Robust Estimate of Location:
yr25  yr45
1890  6210
Robust Estimate of Covariance:
          yr25      yr45
yr25    404261  -1546879
yr45  -1546879   7447721

# robust correlation
```

```
> r_robust <- -1546879/sqrt(404261*7447721)
> r_robust
[1] -0.891
```

We see that the robust methods produce different means, variances, covariances, and correlations than the naive methods. The estimates based on robust methods should be more reliable since they downweight potential outliers.

In multivariate analysis, the Mahalanobs distance plays an important role, such as in cluster analysis. The Mahalanobs distance is sensitive to outliers since it involves means, variances, and covariances which are sensitive to outliers. So robust methods are needed to compute reliable Mahalanobs distances.

```
# Mahalanobs distances to the center
> d_m <- sqrt(mahalanobis(X, x_mean, x_cov))
> d_m
 [1] 0.52 2.60 0.88 1.35 2.15 0.26 0.36 1.07 1.35 1.15 1.48 1.03
# Robust Mahalanobis distances to the center
> d_r <- sqrt(mahalanobis(X,center=x_robust$center,cov=x_robust$cov))
> d_r
 [1] 1.39 4.69 0.54 0.84 1.43 0.22 0.94 0.89 0.94 0.89 0.96 2.19
```

We see that the robust Mahalanobis distances to the center (mean) are quite different from the non-robust Mahalanobis distances. For example, the first observation has a non-robust distance to the center of 0.52, but the corresponding robust distance is 1.39. These differences may affect many multivariate analysis results such as cluster analysis results.

Next, we illustrate robust principal component analysis (PCA). The covariance or correlation matrix play an important role in PCA. Since covariances or correlations are sensitive to outliers, PCA results are also influenced by outliers.

```
# PCA based on original covariance matrix
> x_pc <- princomp(X, cor=T)
> summary(x_pc)
Importance of components:
                       Comp.1 Comp.2
Standard deviation       1.23   0.70
Proportion of Variance   0.76   0.24
Cumulative Proportion    0.76   1.00
> x_cov_robust <- covMcd(X,cor=T)
# PCA based on robust covariance matrix
> x_rpc <- princomp(X, covmat=x_cov_robust, cor=T)
> summary(x_rpc)
Importance of components:
                       Comp.1 Comp.2
Standard deviation       1.38  0.329
Proportion of Variance   0.95  0.054
Cumulative Proportion    0.95  1.000
```

We see that the first PC explains 95% variation based on robust analysis, while the first PC explains only 76% variation based on non-robust analysis. The PC scores are also different (not shown here), since the robust estimates of the mean vector and covariance matrix are different from the corresponding non-robust estimates.

Finally, we illustrate robust regression analysis. The least square method and the maximum likelihood method in regression estimates are both sensitive to outliers, so robust regression analysis produces more reliable results when data contain outliers.

```
# Fit a linear regression using least square method
> fit1 <- lm(yr45~yr25, data=X)
> summary(fit1)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9369.59    1873.11    5.00  0.00054 ***
yr25           -1.92       1.02   -1.89  0.08858
....
# Fit a linear regression using robust method
> fit1_robust <- lmrob(yr45~yr25, data=X)
> summary(fit1_robust)
....
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10173.91    6446.17    1.58     0.15
yr25           -2.32       3.04   -0.76     0.46
```

We see that the robust method produces quite different estimates from the least square method. In particular, the standard errors of the parameter estimates are quite different, leading to different p-values. The robust method indicates that both the intercept and slope are not statistically significant, while this is not the case based on least square method. The results based on robust analysis should be more reliable, since the results based on non-robust analysis may be influenced by the outlier in the data.

Example 2. We re-analyze the Chinese consumption data (2007) based on robust methods and then compare the results with non-robust analysis. Such a comparison allows us to check wether previous non-robust analysis results are reliable or not.

```
# The data is contained in Y0. We consider the first 8 variables.
Y <- Y0[,1:8]
# Summary statistics based on non-robust analysis
> y_mean <- apply(Y,2,mean)  # sample means
> y_mean
  Food  Cloth  Resid  HousF Health  TranC   Educ Miscel
  3511   1020    937    566    675   1234   1231    351
> y_cov <- cov(Y)  # sample covariance matrix
> y_cov
```

```
         Food Cloth  Resid  HousF Health  TranC   Educ Miscel
Food   702798 46019 138015 101279  67802 475788 340941  73078
Cloth   46019 45574  19911  16151  25845  48214  56806  16190
Resid  138015 19911  53109  29740  33431 114267  92012  19985
HousF  101279 16151  29740  27769  20416  82291  73002  14523
Health  67802 25845  33431  20416  44096  61860  64825  15847
TranC  475788 48214 114267  82291  61860 398490 272579  57301
Educ   340941 56806  92012  73002  64825 272579 240979  50001
Miscel  73078 16190  19985  14523  15847  57301  50001  14576

# Robust summary statistics
> y_cov_robust <- covMcd(Y)
> y_cov_robust
Robust Estimate of Location:
  Food   Cloth   Resid   HousF  Health   TranC    Educ  Miscel
  3096    1012     843     537     617     960    1093     310
Robust Estimate of Covariance:
          Food   Cloth  Resid  HousF  Health  TranC    Educ  Miscel
Food    316998  -23911  18385  46746  -17548  66422  109948   -3679
Cloth   -23911   51669  17838   9506   17734  25620   10582    7656
Resid    18385   17838  27554  11146   17677  29575   25908    1502
HousF    46746    9506  11146  16221    5074  22639   23599    1410
Health  -17548   17734  17677   5074   30189  17562    9951    3631
TranC    66422   25620  29575  22639   17562  59038   46234    4618
Educ    109948   10582  25908  23599    9951  46234   80969    8598
Miscel   -3679    7656   1502   1410    3631   4618    8598    5775
```

We again see that the robust analysis results are quite different from non-robust analysis results. For example, the covariance between Food and Cloth based on non-robust method is 46019, but this number becomes -23911 (the corresponding correlation changes from 0.26 to -0.18), so the correlation changes direction! Similar results also occur for correlations between Food and Health and Miscel.

Next, let's check the PCA analysis results.

```
# Non-robust PCA results
> y_pc <- princomp(Y, cor=T)
> summary(y_pc)
Importance of components:
                       Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
Standard deviation       2.39   1.01  0.710  0.522  0.431   0.40  0.295
Proportion of Variance   0.71   0.13  0.063  0.034  0.023   0.02  0.011
Cumulative Proportion    0.71   0.84  0.904  0.938  0.962   0.98  0.993
                       Comp.8
Standard deviation     0.2416
Proportion of Variance 0.0073
Cumulative Proportion  1.0000
# Robust PCA results
> y_cov_robust <- covMcd(Y,cor=T)
> y_rpc <- princomp(Y, covmat=y_cov_robust, cor=T)
> summary(y_rpc)
```

```
Importance of components:
                          Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
Standard deviation          1.95   1.36   0.99  0.765  0.574  0.448  0.416
Proportion of Variance      0.48   0.23   0.12  0.073  0.041  0.025  0.022
Cumulative Proportion       0.48   0.71   0.83  0.902  0.943  0.968  0.990
                          Comp.8
Standard deviation          0.29
Proportion of Variance      0.01
Cumulative Proportion       1.00
```

Again, we see different results based on robust and non-robust analyses. For example, the first PC explains about 71% variation based on non-robust method, but the first PC only explains 48% variation based on robust method. Thus, the results based on non-robust method may be misleading due to potential outliers.

Finally, we perform regression analysis, choosing Health as the response and other variables as predictors.

```
# Non-robust regression analysis results
> fit2 <- lm(Health~., data=Y)
...
Coefficients:
             Estimate  Std. Error t value Pr(>|t|)
(Intercept) -12.03971  265.96923    -0.05    0.964
Food         -0.05926    0.07944    -0.75    0.463
Cloth         0.20967    0.18281     1.15    0.263
Resid         0.57152    0.21133     2.70    0.013 *
HousF        -0.00691    0.36571    -0.02    0.985
TranC        -0.09752    0.11853    -0.82    0.419
Educ          0.13984    0.19479     0.72    0.480
Miscel        0.27837    0.46727     0.60    0.557
# robust regression analysis results
> fit2_robust <- lmrob(Health~., data=Y)
...
Coefficients:
            Estimate  Std. Error t value Pr(>|t|)
(Intercept)  41.4882   410.1368    0.10   0.9203
Food         -0.0498     0.0781   -0.64   0.5296
Cloth         0.1412     0.1450    0.97   0.3403
Resid         0.7490     0.2632    2.85   0.0092 **
HousF        -0.1953     0.7341   -0.27   0.7926
TranC        -0.1080     0.1235   -0.87   0.3908
Educ          0.1338     0.1311    1.02   0.3181
Miscel        0.1022     0.6847    0.15   0.8826
```

We see that robust estimates of the regression coefficients are different from the non-robust estimates. For example, the intercept estimates are very different and have different signs. For other estimates, although their values are different based on different methods, the signs and significances remain the same, so the general conclusions remain the same. This robust analysis confirms that the non-robust

regression analysis conclusions may be reliable (although the estimates are somewhat different).

From the above examples, we see that it is important to perform robust analysis to check if non-robust analysis results are influenced by potential outliers or not. This is particularly important for multivariate analysis since outliers are difficult to detect in multivariate data. If the non-robust analysis results agree with robust analysis results, we are confident about conclusions from these results. Otherwise, results from robust analysis should be preferred.

Exercises 13

13.1. Show that the t-distribution $t(k)$ converges to the standard normal distribution $N(0,1)$ as $k \to \infty$.

13.2. If $x \sim N(\mu, \sigma^2)$, compute MAD(x) and IQR(x).

13.3. Show that the sample standard deviation has a breakdown point of zero.

13.4. Based on the definition of the ϕ function as in (13.1), find the ϕ function for the t-distribution $t(k)$. Plot this function to see how it downweights outliers and compare it to the Huber's function.

13.5. For the dataset "consum2010", compute the robust estimates of the mean vector and covariance matrix. Perform a robust PCA, and compare the results with a non-robust PCA. Do you find any differences?

Chapter 14
Selected Topics

14.1 Likelihood Methods

Likelihood methods are standard approaches in statistical inference. Likelihood methods are widely used because they are very general and have attractive asymptotic properties. The likelihood principle says that likelihoods contain all of the information in the data about unknown parameters in the assumed models. In the following, we provide a brief overview of likelihood methods.

For a likelihood method, the maximum likelihood estimates (MLEs) of parameters in a model can be obtained by maximizing the likelihood function. The MLEs are asymptotically consistent, asymptotically most efficient, and asymptotically normally distributed, under some regularity conditions. In other words, when the sample size is large, the MLE is approximately optimal and normally distributed. The asymptotic normality of MLEs can be used for (approximate) inference. For example, we may use the asymptotic normal distributions of MLEs to construct approximate confidence intervals and to perform hypothesis testing. The drawbacks of likelihood methods are that MLEs are sensitive to outliers and require distributional assumptions.

Let $y_1, y_2, \cdots, y_n$ be an i.i.d. sample from a distribution with a probability density function (for continuous variables) or a probability mass function (for discrete variables) $f(y; \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)^{\mathrm{T}}$ are unknown parameters. Let $\mathbf{y} = (y_1, y_2, \cdots, y_n)^{\mathrm{T}}$. The *likelihood function* for the observed data $\mathbf{y}$ is defined as

$$L(\boldsymbol{\theta}) = L(\boldsymbol{\theta}|\mathbf{y}) = \prod_{i=1}^{n} f(y_i; \boldsymbol{\theta}).$$

The *maximum likelihood estimate (MLE)* of $\boldsymbol{\theta}$, denoted by $\hat{\boldsymbol{\theta}}$, is the value of $\boldsymbol{\theta}$ which maximizes the likelihood $L(\boldsymbol{\theta})$, i.e., the MLE is the value of the parameter which makes the observed data most likely to occur. The corresponding *log-likelihood* is given by

$$l(\boldsymbol{\theta}) \equiv \log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(y_i; \boldsymbol{\theta}).$$

Since the log-likelihood $l(\boldsymbol{\theta})$ is a monotone function of the likelihood $L(\boldsymbol{\theta})$, maximization of the likelihood $L(\boldsymbol{\theta})$ is equivalent to maximization of the log-likelihood $l(\boldsymbol{\theta})$, but the log-likelihood is easier to handle since a summation is mathematically more manageable than a product. Thus, likelihood inference is often based on the log-likelihood.

The MLE $\hat{\boldsymbol{\theta}}$ satisfies the following estimating equation (likelihood equation)

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \frac{\partial \log f(y_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0.$$

The vector

$$\mathbf{s}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\partial \log f(y_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \left(\sum_{i=1}^{n} \frac{\partial \log f(y_i;\boldsymbol{\theta})}{\partial \theta_1}, \cdots, \sum_{i=1}^{n} \frac{\partial \log f(y_i;\boldsymbol{\theta})}{\partial \theta_p} \right)^{\mathrm{T}}$$

is called the *Fisher efficient score* or the *score*. It can be shown that $E(s(\boldsymbol{\theta})) = 0$. The *Fisher's information function (matrix)* is defined by

$$\boldsymbol{I}(\boldsymbol{\theta}) = -E\left(\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} \right) = (I(\boldsymbol{\theta})_{jk})_{p\times p}, \quad \text{with} \quad \boldsymbol{I}(\boldsymbol{\theta})_{jk} = -E\left(\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k} \right).$$

The information matrix $I(\boldsymbol{\theta})$ quantifies the expected amount of information in the data about the unknown parameters $\boldsymbol{\theta}$. The matrix

$$\boldsymbol{H}(\boldsymbol{\theta}) = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} = \frac{\partial \mathbf{s}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}},$$

is called the *Hessian matrix*.

Under some regularity conditions, the MLE is consistent, asymptotically efficient, and asymptotically normally distributed. These regularity conditions can be stated as follows:

R1. The parameter space Θ of $\boldsymbol{\theta}$ is an open subset of the whole space R^p.

R2. The set $A = \{y : f(y;\boldsymbol{\theta}) > 0\}$ does not depend on $\boldsymbol{\theta}$.

R3. The function $f(y;\boldsymbol{\theta})$ is three times continuously differentiable with respect to $\boldsymbol{\theta}$ for all y.

R4. The following equations hold

$$E(\partial l(y;\boldsymbol{\theta})/\partial \boldsymbol{\theta}) = 0, \qquad Cov(\partial l(y;\boldsymbol{\theta})/\partial \boldsymbol{\theta}) = I(\boldsymbol{\theta}), \qquad \text{for all } \boldsymbol{\theta}.$$

R5. The expectations of all the derivatives of $f(y;\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ exist and are finite.

The above regularity conditions are satisfied for a wide variety of models. Under the regularity conditions R1 – R5, the MLE $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ has the following large-sample properties:

- The MLE $\hat{\boldsymbol{\theta}}$ is *consistent*, i.e., $\hat{\boldsymbol{\theta}}$ converges in probability to $\boldsymbol{\theta}$

$$\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}, \qquad \text{as } n \to \infty;$$

- The MLE $\hat{\boldsymbol{\theta}}$ is *asymptotically efficient*, i.e., the asymptotic variance of $\hat{\boldsymbol{\theta}}$ attains the *Cramer-Rao lower bound*

$$Var(\hat{\boldsymbol{\theta}}) \geqslant I^{-1}(\boldsymbol{\theta}) \quad \text{asymptotically.}$$

- The MLE $\hat{\boldsymbol{\theta}}$ is *asymptotically normal*, i.e., $\hat{\boldsymbol{\theta}}$ converges in distribution to $\boldsymbol{\theta}$

$$\hat{\boldsymbol{\theta}} \xrightarrow{d} N(\boldsymbol{\theta}, I^{-1}(\boldsymbol{\theta})), \quad \text{as } n \to \infty.$$

Thus, the MLE is asymptotically *optimal*. Due to the above attractive asymptotic properties of MLEs, likelihood methods are widely used in statistical inference.

Based on the asymptotic normality of the MLE $\hat{\boldsymbol{\theta}}$, in practice when the sample size is large, an approximate level $1-\alpha$ confident interval for θ_j, the j-th component of $\boldsymbol{\theta}$, is given by

$$\hat{\theta}_j \pm z_{\alpha/2} \cdot s.e.(\hat{\theta}_j),$$

where $z_{\alpha/2}$ is the $1-\alpha/2$ percentile of the standard normal distribution $N(0,1)$ and $s.e.(\hat{\theta}_j) = I^{-1/2}(\hat{\boldsymbol{\theta}})_{jj}$ is the approximate standard error of the MLE $\hat{\theta}_j$. For hypothesis testing, the following three likelihood-based large-sample tests are widely used: the Wald test, the likelihood ratio test (LRT), and the efficient score test. These three tests are briefly described as follows.

Consider testing the hypotheses

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \qquad \text{versus} \qquad H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0.$$

The following three tests are based on asymptotic results and widely used in practice:

- *Wald-type test.* The Wald-type test statistic for testing H_0 versus H_1 is given by

$$T_W = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^{\mathrm{T}} \hat{\Sigma}^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0),$$

  where $\hat{\Sigma} = I(\hat{\boldsymbol{\theta}})^{-1}$ is an estimate of the covariance matrix of $\hat{\boldsymbol{\theta}}$. The test statistic $T_W \sim \chi^2_p$ asymptotically under H_0, where p is the dimension of parameter $\boldsymbol{\theta}$. To test an individual component of $\boldsymbol{\theta}$, say $H_{0j} : \theta_j = \theta_{j0}$ versus $H_1 : \theta_j \neq \theta_{j0}$, we may consider individual Wald-type test statistic

$$T_W^{(j)} = \frac{(\hat{\theta}_j - \theta_0)^2}{\widehat{var}(\hat{\theta}_j)}$$

  where $\widehat{var}(\hat{\theta}_j) = (I(\hat{\boldsymbol{\theta}})^{-1})_{jj}$. The test statistic $T_W^{(j)} \sim \chi^2_1$ asymptotically under H_{0j}.

- *Likelihood ratio test* (LRT). Let $\hat{\boldsymbol{\theta}}$ be the MLE of $\boldsymbol{\theta}$, and let $L(\boldsymbol{\theta}_0)$ and $L(\hat{\boldsymbol{\theta}})$ be the likelihood functions evaluated at $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}$ respectively. The LRT test statistic for testing H_0 versus H_1 is given by

$$T_L = -2\log\left(\frac{L(\boldsymbol{\theta}_0)}{L(\hat{\boldsymbol{\theta}})}\right) = 2\log L(\hat{\boldsymbol{\theta}}) - 2\log L(\boldsymbol{\theta}_0).$$

  The test statistic $T_L \sim \chi^2_p$ asymptotically under H_0.

- *Score test.* The score test statistic for testing H_0 versus H_1 is given by

$$T_S = \mathbf{s}(\boldsymbol{\theta}_0)^{\mathrm{T}} I(\boldsymbol{\theta}_0)^{-1} \mathbf{s}(\boldsymbol{\theta}_0),$$

  where $\mathbf{s}(\boldsymbol{\theta}_0)$ is the score function at $\boldsymbol{\theta}_0$. The test statistic $T_S \sim \chi^2_p$ asymptotically under H_0.

The above three tests are asymptotically equivalent, but they may differ with finite samples. The LRT is equivalent to the deviance test which is widely used in GLMs. The Wald test requires the least computational effort. The score test does not require computing the MLE since the test statistic is evaluated under the null hypothesis.

14.2 Bootstrap Methods

Bootstrap methods are very popular in statistical inference when i) the sample size is small, ii) the distributional assumption does not hold, and iii) analytic formula does not exist or the problem is highly complicated. For example, it is usually difficult to compute the variances of a sample median or a sample percentile or a sample correlation coefficient. In these cases, it is straightforward to use the bootstrap method to compute estimates of standard errors and confidence intervals of these estimators. Bootstrap methods are easy to implement, though may be computationally intensive, and can be applied to a wide variety of problems. Therefore, bootstrap has become a very popular statistical tool in modern statistics.

The basic idea of a bootstrap method is to repeatedly sample from the observed dataset *with replacement* (with the same sample size as the observed dataset) to estimate the variation in the observed data. For example, suppose that $(x_1, x_2, \cdots, x_n)$ is an observed dataset, and suppose that one wishes to estimate the variance of the sample median. A simple bootstrap method proceeds as follows. We can sample from this observed dataset with replacement. The resulting sample, denoted by $(x_1^*, x_2^*, \cdots, x_n^*)$, is called a *bootstrap sample*. Then, we compute the sample median of this bootstrap sample. Repeating this process B times (say $B = 100$), we obtain B median estimates from the B bootstrap samples. We then compute the sample

variance of these B median estimates and obtain a bootstrap estimate of the variance of the sample median from the original dataset.

As another example, we know that the MLE of a parameter is asymptotically normally distributed. In practice, this asymptotic distribution is often used to construct approximate confidence intervals and hypothesis testing. Since the sample size is finite in practice, we may want to know how close the distribution of the MLE is to normality, so that we can judge how reliable the approximate confidence intervals and testing results are. We can use a bootstrap method to check this. Suppose that we fit a model using the likelihood method, and we wish to check if the resulting MLEs of the parameters are approximately normal. A simple bootstrap method can be performed as follows:

- sample from the original dataset with replacement and obtain a bootstrap sample;
- fit the model to the bootstrap sample using the likelihood method and obtain MLEs of the parameters;
- Repeating the procedure B times, we obtain B sets of parameter estimates (MLEs).

The sampling distribution of the B estimates of a parameter is an approximation to the "true" sampling distribution of the MLE of this parameter based on the original dataset. We can then, for example, obtain an approximate confidence interval from the bootstrap samples by taking the α and $1-\alpha$ (say, $\alpha = 0.05$) quantiles of the B estimates. A bootstrap estimate of the standard error of the parameter estimate is the sample standard error of the B estimates.

14.3 MCMC Methods and the Gibbs Sampler

In modern statistics, Monte Carlo methods are widely used since analytic solutions to many complex problems are unavailable. For a Monte Carlo method, we often need to generate large numbers of samples from highly complicated and multi-dimensional distributions. *Markov chain Monte Carlo* (MCMC) methods are great tools for such tasks. MCMC methods are algorithms for generating samples from intractable distributions. The key idea of MCMC methods is to construct Markov chains that have the desired distributions as their stationary distributions. After a large number of steps, called a burn-in period, the Markov chain will converge to its stationary distribution, and thus the last state of the chain can be used as a sample from the desired distribution. MCMC methods have revolutionized Bayesian inference since they have made highly complicated Bayesian computations feasible. These MCMC methods are also very useful tools in likelihood inference since many likelihood computations encount similar problems as in Bayesian inference. The most useful MCMC method is probably the *Gibbs sampler*, which is briefly described below.

Gibbs sampling or the *Gibbs sampler* is an example of MCMC methods. The Gibbs sampler is typically used to obtain random samples from a multi-dimensional probability distribution, which is either intractable or is not known explicitly. The desired samples can be obtained by sequentially sampling from lower-dimensional conditional distributions which are easier to sample from. These samples then comprise a Markov chain, whose stationary distribution is the target distribution. The Gibbs sampler is widely used because it is often easier to sample from the *lower-dimensional* conditional distributions than the original distribution.

Suppose that we wish to generate samples from the probability distribution $f(\mathbf{u}|\boldsymbol{\theta})$, where $\mathbf{u} = (\mathbf{u}_1^{\mathrm{T}}, \mathbf{u}_2^{\mathrm{T}}, \cdots, \mathbf{u}_q^{\mathrm{T}})^{\mathrm{T}}$ is a random vector, with each component $\mathbf{u}_j$ being possibly also a random vector. Suppose also that $f(\mathbf{u}|\boldsymbol{\theta})$ is highly intractable or even not known explicitly, so it is difficult to generate samples from $f(\mathbf{u}|\boldsymbol{\theta})$ directly. Note that the components $\mathbf{u}_j$'s are typically *unobserved* quantities which may have different dimensions or different types. For example, let $\mathbf{y} \sim N(\boldsymbol{\mu}, \Sigma)$. Suppose that we want to simulate from the posterior distribution $f(\mu, \Sigma|\mathbf{y})$ in Bayesian inference. In this case, we can choose $q = 2, \quad \mathbf{u}_1 = \boldsymbol{\mu}, \quad \mathbf{u}_2 = \Sigma$.

In the following, we describe the Gibbs sampler method to generate samples from $f(\mathbf{u}|\boldsymbol{\theta})$, assuming $\boldsymbol{\theta}$ is known for simplicity. Let

$$\mathbf{u}_{-j} = (\mathbf{u}_1^{\mathrm{T}}, \cdots, \mathbf{u}_{j-1}^{\mathrm{T}}, \mathbf{u}_{j+1}^{\mathrm{T}}, \cdots, \mathbf{u}_q^{\mathrm{T}})^{\mathrm{T}}, \qquad j = 1, 2, \cdots, q,$$

be the sub-vector of $\mathbf{u}$ without component $\mathbf{u}_j$. It is often easier to generate samples from the lower-dimensional conditional distributions $f(\mathbf{u}_j|\mathbf{u}_{-j}, \boldsymbol{\theta}), j = 1, 2, \cdots, q$, which are called *full conditionals*. The Gibbs sampler proceeds as follows: beginning with starting values $(\mathbf{u}_1^{(0)}, \cdots, \mathbf{u}_q^{(0)})$, at step k,

- sample $\mathbf{u}_1^{(k)}$ from $f(\mathbf{u}_1|\mathbf{u}_2^{(k-1)}, \mathbf{u}_3^{(k-1)}, \cdots, \mathbf{u}_q^{(k-1)}, \boldsymbol{\theta})$;
- sample $\mathbf{u}_2^{(k)}$ from $f(\mathbf{u}_2|\mathbf{u}_1^{(k)}, \mathbf{u}_3^{(k-1)}, \cdots, \mathbf{u}_q^{(k-1)}, \boldsymbol{\theta})$;
- $\cdots$;
- sample $\mathbf{u}_q^{(k)}$ from $f(\mathbf{u}_q|\mathbf{u}_1^{(k)}, \cdots, \mathbf{u}_{q-1}^{(k)}, \boldsymbol{\theta}), \quad k = 1, 2, \cdots$.

The sequence $\{(\mathbf{u}_1^{(k)}, \cdots, \mathbf{u}_q^{(k)}),\ k = 1, 2, 3, \cdots\}$ then comprises a Markov chain with stationary distribution $f(\mathbf{u}|\boldsymbol{\theta})$. Therefore, when k is large enough (after a burn-in period), we can view $\mathbf{u}^{(k)} = (\mathbf{u}_1^{(k)}, \cdots, \mathbf{u}_q^{(k)})^{\mathrm{T}}$ as a sample generated from the target distribution $f(\mathbf{u}|\boldsymbol{\theta})$. Repeating the process m times, we obtain a sample of size m from the intractable distribution $f(\mathbf{u}|\boldsymbol{\theta})$. When m is large, we can approximate the mean and variance of the distribution $f(\mathbf{u}|\boldsymbol{\theta})$ by the sample mean and sample variance respectively, or we can approximate the density curve $f(\mathbf{u}|\boldsymbol{\theta})$ by the empirical density function based on the simulated samples.

Note that the Gibbs sampler or other MCMC methods can only approximate the target distributions. The accuracy of the approximation improves as the number of steps (burn-in period) increases. It may not be easy to determine the burn-in period for the Markov chain to converge to the stationary distribution within acceptable random errors. Determining the convergence criteria is an important issue. *WinBUGS* is a statistical software that is widely used to do Gibbs sampling. It is based on the BUGS project (Bayesian inference Using Gibbs Sampling).

Example. Suppose that we wish to generate samples from the bivariate normal distribution $\mathbf{u} = (u_1, u_2)^{\mathrm{T}} \sim N(\boldsymbol{\mu}, \Sigma)$ where $\boldsymbol{\mu} = (\mu_1, \mu_2)^{\mathrm{T}}$ and Σ is a 2×2 covariance matrix with diagonal elements being 1 and off-diagonal elements being ρ ($|\rho| < 1$). Assume that $\boldsymbol{\mu}$ and Σ are known. Here it is not difficult to sample from $N(\boldsymbol{\mu}, \Sigma)$, but we consider a Gibbs sampler for illustration purpose. Consider the Gibbs sampler to sample from the target distribution $N(\boldsymbol{\mu}, \Sigma)$. The full conditionals are

$$\begin{aligned} u_1|u_2, \boldsymbol{\mu}, \rho &\sim N(\mu_1 + \rho(u_2 - \mu_2),\ 1 - \rho^2), \\ u_2|u_1, \boldsymbol{\mu}, \rho &\sim N(\mu_2 + \rho(u_1 - \mu_1),\ 1 - \rho^2). \end{aligned}$$

The Gibbs sampler proceeds as follows: beginning with starting value $(u_1^{(0)}, u_2^{(0)})$, at k-th step,

- generate $u_1^{(k)}$ from $N(\mu_1 + \rho(u_2^{(k-1)} - \mu_2),\ 1 - \rho^2)$;
- generate $u_2^{(k)}$ from $N(\mu_1 + \rho(u_1^{(k)} - \mu_2),\ 1 - \rho^2)$, $\qquad k = 1, 2, \cdots$.

Then, the sequence $\{(u_1^{(k)}, u_2^{(k)}), k = 0, 1, 2, \cdots\}$ forms a Markov chain with stationary distribution $N(\boldsymbol{\mu}, \Sigma)$. Thus, when k is large (say 200), we may consider $(u_1^{(k)}, u_2^{(k)})$ as a sample from $N(\boldsymbol{\mu}, \Sigma)$.

14.4 Survival Analysis

Survival data, or time-to-event data, arise frequently in practice, such as time to death, time to cancer recurrence, and time to dropout. Statistical methods for the analysis of survival data have received great attention in the past few decades. Survival data typically have the following characteristics:

- survival data are often *censored*, i.e., the exact event times may not be observed for some subjects in the study, so the event times for these subjects are censored;
- survival data are often *skewed*, i.e., survival data are usually not symmetrically distributed (they are often skewed to the right);

Thus, survival data are different from longitudinal data and cross-sectional data, and special statistical techniques are required to analyze survival data.

A key characteristic of survival data is the *censoring* in the data. A typical situation leading to censoring is that the event times are not observed because subjects are

lost to follow-up or drop out or the study is terminated. These types of censoring are called *right censoring*, since the unobserved event times are larger than the censoring times. Another common type of censoring is *interval-censoring*, where the exact event times are not observed but the event times are known to fall in certain time intervals, such as the exact times of dropouts or disease infection. Statistical methods for survival analysis must address censoring.

When describing and summarizing survival data, the standard summary statistics such as means and standard deviations are not ideal since survival data are often skewed. Instead, survival data may be better summarized by percentiles, which include the median. In particular, *survival functions*, which are the probabilities (or percentages) of surviving or event-free at different times, are widely used in survival analysis. A closely related function, called *hazard function*, gives the risk of event at different times, may be more frequently used than the survival function, due to its attractive interpretation. In regression analysis of survival data, we are interested in the relationship between event risks and covariates (risk factors).

There are various approaches for analyzing survival data. From statistical methodology point of view, survival analysis methods may be broadly classified into three categories:

- *Nonparametric methods.* Nonparametric methods for survival data make *no* distributional assumptions about the data. For example, these methods can be used to estimate and compare survival functions. The most well-known nonparametric methods for survival data are perhaps the *Kaplan-Meier estimate* of a survival function and the *log-rank test* for comparing different survival functions.
- *Semiparametric methods.* Semiparametric methods are often used for survival regression models. These methods have both parametric components and nonparametric components. The most well-known semiparametric regression model for survival data is perhaps the *Cox proportional hazards model.*
- *Parametric methods.* Parametric methods assume parametric distributions for survival data. The most well-known parametric distribution for survival data is perhaps the *Weibull distribution.*

Each of these methods has its own advantages and limitations. Parametric methods may be more efficient if the assumed parametric distributions or parametric functions hold, nonparametric methods may be more robust if distributional assumptions are questionable, and semiparametric methods are compromises between the parametric and nonparametric methods.

A survival function is often estimated by the Kaplan-Meier estimate, which is a nonparametric estimate (it makes no distributional assumptions for the survival

data). The Kaplan-Meier estimate reduces to the usual empirical estimate of a survival function (or equivalently a cumulative distribution function) when there is no censoring in the data. When comparing survival experiences of two groups of individuals, say a treatment group and a control group, the log-rank test is often used, which is also a nonparametric method and makes no distributional assumptions for the survival data. If we wish to compare two groups and adjust for covariates such as age or gender, we may consider a survival regression model such as a Cox proportional hazards model. A Cox proportional hazards models also makes no distributional assumption for the survival data, and it is equivalent to the log-rank test when it is not adjusted for covariates.

Cox proportional hazards models are probably the most widely used regression models for survival data. A possible reason is that the regression parameters in a Cox proportional hazards model have attractive interpretations as relative risks, in addition to its desirable distribution-free property. Unlike usual regression models in which covariates are linked to the mean responses, in Cox proportional hazards models covariates are linked to the hazard functions. In a Cox proportional hazards model, covariates enter the model parametrically, while the hazard functions are left unspecified (nonparametric). A limitation of Cox proportional hazards models is that they *assume* the ratio of the hazard at any time and the baseline hazard is *constant*. This assumption may not hold in some situations, and in these cases alternative survival regression models such as accelerated failure time models may be considered instead.

Accelerated failure time (*AFT*) models are another type of popular regression models for survival data. From practical interpretation point of view, AFT models offer more desirable interpretations than Cox models. An AFT model has an appealing interpretation as "speed up" or "slow down" of disease progression. Moreover, it does not require the proportional hazards assumption as in a Cox proportional hazards model. An AFT model can also be written in a simple form as a log-linear regression model. Therefore, AFT models are good alternatives to Cox proportional hazards models.

A *Weibull distribution* is a parametric distribution which is widely used in survival analysis. In fact, Weibull distributions play a central role in parametric survival models, similar to the role of normal distributions play in classical statistical models. While the use of normal distributions can be justified by the Central Limit Theorem, the use of Weibull distributions can be justified by the extreme-value theory. Weibull distributions are often assumed for survival data in *parametric* survival models, including parametric survival regression models. If the distributional assumptions hold, parametric survival models provide more efficient inference than nonparametric or semiparametric survival models. Other parametric distributions

may also be assumed for survival data, such as log-normal distributions.

Survival data may also arise in clusters. For example, in a multi-center study, survival data from the same center may be correlated since these data may be more similar than data from different centers. In this case, each center may serve as a cluster. In analyzing clustered survival data, the within-cluster correlation should be addressed, as in longitudinal data. There may also be multiple (repeated) survival data from the same individuals, such as recurrent cancer. In this case, the multiple event times from the same individuals may be correlated, and each individual serves as a cluster, similar to longitudinal data. In the analysis of clustered survival data, we can also introduce random effects in the models to incorporate cluster-effect and within-cluster correlations, similar to mixed effects models for longitudinal data. Survival models with random effects are sometimes called *frailty models*.

14.5 Data Science, Big Data, and Data Mining

The modern world is generating massive amount of data or "big data" which have the potential to transform the ways business, government, and science are carried out. From governments to mobile networks, data is being collected at an unprecedented speed and scale. In this section, we briefly describe a few emerging new disciplines which are attracting great attentions and providing excellent opportunities for statisticians or data scientists.

14.5.1 Data Science

Data science is an emerging discipline which uses automated methods to analyze massive amount of data and extract useful knowledge from the data. It combines aspects of statistics, computer science, and mathematics, so it requires a versatile skill-set. In modern world, there is an explosion of new data generated from different sources such as smart devices, web, mobiles, and social media. These data consist of structured and unstructured data that large enterprises produce. Data scientists interpret rich data sources, manage large amounts of data, merge data sources together, create data visualizations, analyze the data using statistical methods, and extract useful information from these "big data". Data scientists rely on statistics, machine learning, and text retrieval to analyze data and interpret results.

Data scientists are an integral part of *competitive intelligence*, which is a newly emerging field that includes activities such as data mining and analysis to help businesses to gain a competitive edge. Data science technologies impact how we access data and conduct research across various domains, including social science and biological sciences. For example, data science plays an important role in security and fraud monitoring. From huge streams of data, we can use exploratory data

analysis, data visualization, and machine learning to gain insights from data to uncover unknown risks. That is, through extract useful information from big data, we can discover insider threats and prevent fraud. This area is called security data science.

Data scientists use statistical methods to summarize data, use portions of data (samples) to make inferences about the larger context, and visualize data by presenting it in tables and graphs. Data scientists play active roles in the design and implementation of four related areas: data architecture, data acquisition, data analysis, and data archiving. For example, much of real world data may be non-numeric and unstructured (e.g., data are not arranged in neat rows and columns), such as a web page full of photographs and short messages among friends. To manage and analyze such data, data scientists require a range of skills. These skills include, but not limited to,

- Communication skills – a data scientist must learn the needs and preferences of users, translate the technical terms of computing and statistics to simple and understandable language, and communicate the results of data analyses to general users.
- Data storage and presentation skills – a data scientist needs to have a clear understanding about how massive data can be stored and linked, and how to present the data effectively using graphs.
- Critical thinking skills – data scientists must understand how data will be used to make decisions and affect people's lives, important ethical issues such as privacy, and be able to communicate the limitations of data to try to prevent misuse of data or analysis results.

Therefore, data science is much more than simply analyzing data.

Data science is born in the first decade of the 21st century. It is an advanced discipline requiring proficiency in parallel processing, map-reducing computing, machine learning, and advanced statistics. In business and government today, data science should be practiced as a team, since a data scientist is most likely to be an expert in only one or two areas rather than an expert in all of the areas.

14.5.2 Big Data

Data science arises out of the "big data" world. Big data refers to huge and complex datasets that are difficult to store, manage, analyze, and visualize using traditional software and tools. That is, big data is a term used to describe the exponential growth and availability of structured and unstructured data. For example, VISA processes more than 172 million card transactions each day, United Parcel Service (UPS) receives 39.5 million tracking requests from customers on average

each day, and more than 5 billion people are calling, texting, tweeting and browsing websites on mobile phones. These all generate "big data". It is important to learn, analyze, and extract useful information from these big data. Big data analysis played an important role in the United States president Barack Obama's successful 2012 re-election campaign. In 2012, the Obama administration announced the Big Data Research and Development Initiative (BDRDI), which explores how big data could be used to address important problems faced by the government.

Big data are important to business since the data contain valuable information. Organizations can analyze data to find answers that enable cost and time reductions, new product development, and smarter business decision making. For example, by analyzing big data, we can

- determine root causes of failures and issues, which may potentially save billions of dollars.
- identify customers who matter the most.
- use data mining to detect fraudulent behavior.

Big data requires exceptional technologies to efficiently process large quantities of data, such as machine learning, signal processing, time series analysis, data-mining, and visualization. Real or near-real time information delivery is one of the defing characteristic of big data analysis.

14.5.3 Data Mining

Data mining is the process of discovering previously unknown information from large databases and using it to support strategic business decisions. The statistical techniques of data mining are familiar. They include linear regression, logistic regression, multivariate analysis, cluster analysis, and principal components analysis. Traditional approaches to statistical inference may fail with large databases, however, because with thousands or millions of cases and hundreds or thousands of variables there will be a high level of redundancy among the variables, there will be spurious relationships, and even the weakest relationships will be highly significant by any statistical test. The objective is to build a model with significant predictive power. It is not enough just to find which relationships are statistically significant.

Consider a campaign offering a product or service for sale, directed at a given customer base. Typically, about 1% of the customer base will be "responders," customers who will purchase the product or service if it is offered to them. A mailing to 100,000 randomly-chosen customers will therefore generate about 1000 sales. Data mining techniques enable customer relationship marketing, by identifying which customers are most likely to respond to the campaign. If the response can be raised from 1% to, say, 1.5% of the customers contacted, then 1000 sales could by achieved with only 66,666 mailings, reducing the cost of mailing by one-third.

In the above example, the response variable is y, which indicates whether or not a consumer responded to a direct mail campaign for a specific product: $y = 1$ if response, and $y = 0$ if non-response. So we have a binary response, and we may consider a logistic regression model to see what variables can predict the values of y. There may be (say) 200 predictors $x_1, x_2, \cdots, x_{200}$: gender (male/female), income, number of products ever purchased, number of months since first activity, etc. We may perform the following statistical analyses: i) exam the correlations between the variables in order to reduce the number of variables, and then build a logistic regression model, and ii) perform a principal component analysis, select a reduced set of variables from the PCA factors and build a logistic regression model from the factors. Other statistical methods can also be used to analyze the data and compare the results.

References

Agresti, A. 2012. *Categorical Data Analysis*, 3rd edition, New York; Wiley.

Anderson, T.W. 2003. *An Introduction to Multivariate Statistical Analysis*, 3rd edition, New York; Wiley.

Bates, D.M. and Watts, D.G. 2007. *Nonlinear Regression Analysis and Its Applications.* New York: John Wiley.

Draper, N.R. and Smith, H. 1998. *Applied Regression Analysis*, 3rd edition. New York: Wiley-Interscience.

Johnson, R.A. and Wichern, D. W. 2007. *Applied Multivariate Statistical Analysis*, 6th edition. Pearson.

Little, R.J.A. and Rubin, D.B. 2002. *Statistical Analysis with Missing Data*, 2nd edition. New York: Wiley.

Maronna, R., Martin, D., and Yohai, V., 2006. *Robust Statistics–Theory and Methods.* New York: Wiley.

Marshall, A.W. and Ingram, O. 1988. Families of multivariate distributions. *Journal of the American Statistical Association* 83, 834-841.

McCullagh, P. and Nelder, J. A. 1989. *Generalized Linear Models*, 2nd edition. New York: Chapman & Hall.

Seber, G.A.F. and Wild, C.J. 2003. *Nonlinear Regression.* New York: John Wiley.

Silvapulle, M.J. and Sen, P.K. 2004. *Constrained Statistical Inference: Inequality, Order, and Shape Restrictions.* New York: Wiley.

Weisberg, S. 2005. *Applied Linear Regression*, 3rd edition. Hoboken, NJ: Wiley-Interscience.

Wu, L. 2009. *Mixed Effects Models For Complex Data*, Chapman and Hall/CRC.