1
Multivariate Data and Multivariate Analysis

1.1 Introduction

Multivariate data arise when researchers measure several variables on each "unit" in their sample. The majority of data sets collected by researchers in all disciplines are multivariate. Although in some cases it may make sense to isolate each variable and study it separately, in the main it does not. In most instances the variables are related in such a way that when analyzed in isolation they may often fail to reveal the full structure of the data. With the great majority of multivariate data sets, *all* the variables need to be examined simultaneously in order to uncover the patterns and key features in the data. Hence the need for the collection of multivariate analysis techniques with which this book is concerned.

Multivariate analysis includes methods that are largely descriptive and others that are primarily inferential. The aim of all the procedures, in a very general sense, is to display or extract any "signal" in the data in the presence of noise, and to discover what the data has to tell us.

1.2 Types of Data

Most multivariate data sets have a common form, and consist of a data matrix, the rows of which contain the units in the sample, and the columns of which refer to the variables measured on each unit. Symbolically a set of multivariate data can be represented by the matrix, $\mathbf{X}$, given by

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1q} \\ \vdots & & \ddots & \\ x_{n1} & x_{n2} & \cdots & x_{nq} \end{bmatrix}$$

where n is the number of units in the sample, q is the number of variables measured on each unit, and x_{ij} denotes the value of the jth variable for the ith unit.

The units in a multivariate data set will often be individual people, for example, patients in a medical investigation, or subjects in a market research study. But they

can also be skulls, pottery, countries, products, to name only four possibilities. In all cases the units are often referred to simply as "individuals," a term we shall generally adopt in this book.

A hypothetical example of a multivariate data matrix is given in Table 1.1. Here $n = 10$, $q = 7$, and, for example, $x_{33} = 135$. These data illustrate that the variables that make up a set of multivariate data will not necessarily all be of the same type. Four levels of measurement are often distinguished;

- *Nominal*—Unordered categorical variables. Examples include treatment allocation, the sex of the respondent, hair color, presence or absence of depression, and so on.
- *Ordinal*—Where there is an ordering but no implication of equal distance between the different points of the scale. Examples include social class and self-perception of health (each coded from I to V, say), and educational level (e.g., no schooling, primary, secondary, or tertiary education).
- *Interval*—Where there are equal differences between successive points on the scale, but the position of zero is arbitrary. The classic example is the measurement of temperature using the Celsius or Fahrenheit scales.
- *Ratio*—The highest level of measurement, where one can investigate the *relative magnitude* of scores as well as the differences between them. The position of zero is fixed. The classic example is the absolute measure of temperature (in Kelvin, for example) but other common examples include age (or any other time from a fixed event), weight and length.

The qualitative information in Table 1.1 could have been presented in terms of numerical codes (as often would be the case in a multivariate data set) such that sex = 1 for males and sex = 2 for females, for example, or health = 5 when very good and health = 1 for very poor, and so on. But it is vital that both the user and consumer of these data appreciate that the same numerical codes (1, say) will often convey completely different information.

In many statistical textbooks discussion of different types of measurements is often followed by recommendations as to which statistical techniques are suitable

Table 1.1 Hypothetical Set of Multivariate Data

Individual	Sex	Age (yr)	IQ	Depression	Health	Weight (lb)
1	Male	21	120	Yes	Very good	150
2	Male	43	NK	No	Very good	160
3	Male	22	135	No	Average	135
4	Male	86	150	No	Very poor	140
5	Male	60	92	Yes	Good	110
6	Female	16	130	Yes	Good	110
7	Female	NK	150	Yes	Very good	120
8	Female	43	NK	Yes	Average	120
9	Female	22	84	No	Average	105
10	Female	80	70	No	Good	100

NOTE: NK = not known.

for each type; for example, analyses of nominal data should be limited to summary statistics such as the number of cases, the mode, and so on. And in the analysis of ordinal data, means and standard deviations are not really suitable. But Velleman and Wilkinson (1993) make the important point that restricting the choice of statistical methods in this way may be a dangerous practice for data analysis; the measurement taxonomy described is often too strict to apply to real-world data. This is not the place for a detailed discussion of measurement, but we take a fairly pragmatic approach to such problems. For example, we will often not agonize over treating variables such as a measure of depression, anxiety, or intelligence as if they were interval-scaled, although strictly they fit into the ordinal category described above.

Table 1.1 also illustrates one of the problems often faced by statisticians undertaking statistical analysis in general, and multivariate analysis in particular, namely the presence of *missing values* in the data, that is, observations and measurements that should have been recorded, but, for one reason or another, were not. Often when faced with missing values, practitioners simply resort to analyzing only *complete cases*, since this is what most statistical software packages do automatically. In a multivariate analysis, they would, for example, omit any case with a missing value on any of the variables. When the incomplete cases comprise only a small fraction of all cases (say, 5 percent or less) then case deletion may be a perfectly reasonable solution to the missing data problem. But in multivariate data sets in particular, where missing values can occur on any of the variables, the incomplete cases may often be a substantial portion of the entire dataset. If so, omitting them may cause large amounts of information to be discarded, which would clearly be very inefficient.

But the main problem with complete-case analysis is that it can lead to a serious bias in both estimation and inference unless the missing data are *missing completely at random* (see Chapter 9 and Little and Rubin, 1987, for more details). In other words, complete-case analysis implicitly assumes that the discarded cases are like a random subsample. So at the very least complete-case analysis leads to a loss, and perhaps a substantial loss in power, but worse, analyses based on just complete cases might in some cases be misleading.

So what can be done? One answer is to consider some form of *imputation*, the practice of "filling in" missing data with plausible values. At one level this will solve the missing-data problem and enable the investigator to progress normally. But from a statistical viewpoint careful consideration needs to be given to the method used for imputation; otherwise it may cause more problems than it solves. For example, imputing an observed variable mean for a variable's missing values preserves the observed sample means, but distorts the covariance structure, biasing estimated variances and covariances toward zero. On the other hand imputing predicted values from regression models tends to inflate observed correlations, biasing them away from zero. And treating imputed data as if they were "real" in estimation and inference can lead to misleading standard errors and p-values, since they fail to reflect the uncertainty due to the missing data.

The most appropriate way to deal with missing values is a procedure suggested by Rubin (1987), known as *multiple imputation*. This is a Monte Carlo technique

in which the missing values are replaced by $m > 1$ simulated versions, where m is typically small (say 3–10). Each of the simulated complete datasets is analyzed by the method appropriate for the investigation at hand, and the results are later combined to produce estimates and confidence intervals that incorporate missing-data uncertainty. Details are given in Rubin (1987) and more concisely in Schafer (1999). An S-PLUS® library for multiple imputation is available; see Schimert et al. (2000). The greatest virtues of multiple imputation are its simplicity and its generality. The user may analyze the data by virtually any technique that would be appropriate if the data were complete. However, one should always bear in mind that the imputed values are not real measurements. We do not get something for nothing! And if there is a substantial proportion of the individuals with large amounts of missing data one should clearly question whether *any* form of statistical analysis is viable.

1.3 Summary Statistics for Multivariate Data

In order to summarize a multivariate data set we need to produce summaries for each of the variables separately and also to summarize the relationships between the variables. For the former we generally use *means* and *variances* (assuming that we are dealing with continuous variables), and for the latter we usually take pairs of variables at a time and look at their *covariances* or *correlations*. Population and sample versions of all of these quantities are now defined.

1.3.1 *Means*

For q variables, the population mean vector is usually represented as $\boldsymbol{\mu}' = [\mu_1, \mu_2, \ldots, \mu_q]$, where

$$\mu_i = E(x_i)$$

is the population mean (or *expected value* as denoted by the E operator in the above) of the ith variable. An *estimate of* $\boldsymbol{\mu}'$, based on n, q-dimensional observations, is $\bar{\mathbf{x}}' = [\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_q]$, where $\bar{x}_i$ is the sample mean of the variable x_i.

To illustrate the calculation of a mean vector we shall use the data shown in Table 1.2, which shows the heights (millimeters) and ages (years) of both partners in a sample of 10 married couples. We assume that the data are available as the data.frame `huswif` with variables labelled as shown in Table 1.2. The mean vector for these data can be found directly in R with the `mean` function and in S-PLUS by using the `apply` function combined with the `mean` function;

R: `mean(huswif)`

S-PLUS: `apply(huswif,2,mean)`

Table 1.2 Heights and Ages of Husband and Wife in 10 Married Couples

Husband age (Hage)	Husband height (Hheight)	Wife age (Wage)	Wife height (Wheight)	Husband age at first marriage (Hagefm)
49	1809	43	1590	25
25	1841	28	1560	19
40	1659	30	1620	38
52	1779	57	1540	26
58	1616	52	1420	30
32	1695	27	1660	23
43	1730	52	1610	33
47	1740	43	1580	26
31	1685	23	1610	26
26	1735	25	1590	23

The values that result are:

Hage	Hheight	Wage	Wheight	Hagefm
40.3	1728.9	38.0	1578.0	26.9

1.3.2 *Variances*

The vector of population variances can be represented by $\boldsymbol{\sigma}' = [\sigma_1^2, \sigma_2^2, \ldots, \sigma_q^2]$, where

$$\sigma_i^2 = E(x_i - \mu_i)^2.$$

An estimate of $\boldsymbol{\sigma}'$ based on n, q-dimensional observations is $\mathbf{s}' = [s_1^2, s_2^2, \ldots, s_q^2]$, where s_i^2 is the sample variance of x_i.

We can get the variances for the variables in the husbands and wives data set by using the `sd` function directly in R and again using the `apply` function combined with the `var` function in S-PLUS:

R:
```
sd(huswif)^2
```
S-PLUS:
```
apply(huswif,2,var)
```

to give

Hage	Hheight	Wage	Wheight	Hagefm
130.23	4706.99	164.67	4173.33	29.88

1.3.3 *Covariances*

The population covariance of two variables, x_i and x_j, is defined by

$$\text{Cov}(x_i, x_j) = E(x_i - \mu_i)(x_j - \mu_j).$$

If $i = j$, we note that the covariance of the variable with itself is simply its variance, and therefore there is no need to define variances and covariances independently in the multivariate case. The covariance of x_i and x_j is usually denoted by σ_{ij} (so the variance of the variable x_i is often denoted by σ_{ii} rather than σ_i^2).

With q variables, $x_1, x_2, \ldots, x_q$, there are q variances and $q(q-1)/2$ covariances. In general these quantities are arranged in a $q \times q$ symmetric matrix, $\boldsymbol{\Sigma}$, where

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1q} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2q} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{q1} & \sigma_{q2} & \cdots & \sigma_{qq} \end{pmatrix}.$$

Note that $\sigma_{ij} = \sigma_{ji}$. This matrix is generally known as the *variance–covariance matrix* or simply the *covariance matrix*. The matrix $\boldsymbol{\Sigma}$ is estimated by the matrix $\mathbf{S}$, given by

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$$

where $\mathbf{x}_i' = [\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iq}]$ is the vector of observations for the ith individual. The diagonal of $\mathbf{S}$ contains the variances of each variable.

The covariance matrix for the data in Table 1.2 is obtained using the `var` function in both R and S-PLUS,

```
var(huswif)
```

to give the following matrix of variances (on the main diagonal) and covariances (the off diagonal elements).

	Hage	Hheight	Wage	Wheight	Hagefm
Hage	130.23	−192.19	128.56	−436.00	28.03
Hheight	−192.19	4706.99	25.89	876.44	−229.34
Wage	128.56	25.89	164.67	−456.67	21.67
Wheight	−436.00	876.44	−456.67	4173.33	−8.00
Hagefm	28.03	−229.34	21.67	−8.00	29.88

1.3.4 *Correlations*

The covariance is often difficult to interpret because it depends on the units in which the two variables are measured; consequently, it is often standardized by dividing by the product of the standard deviations of the two variables to give a quantity called the *correlation coefficient*, ρ_{ij}, where

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}.$$

The correlation coefficient lies between -1 and $+1$ and gives a measure of the *linear* relationship of the variables x_i and x_j. It is positive if high values of x_i are

associated with high values of x_j and negative if high values of x_i are associated with low values of x_j. With q variables there are $q(q-1)/2$ distinct correlations which may be arranged in a $q \times q$ matrix whose diagonal elements are unity.

For sample data, the correlation matrix contains the usual estimates of the ρ's, namely Pearson's correlation coefficient, and is generally denoted by **R**. The matrix may be written in terms of the sample covariance matrix **S** as follows,

$$\mathbf{R} = \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2}$$

where $\mathbf{D}^{-1/2} = \text{diag}(1/s_i)$.

In most situations we will be dealing with covariance and correlation matrices of full rank, q, so that both matrices will be nonsingular (i.e., invertible).

The correlation matrix for the four variables in Table 1.2 is obtained by using the function `cor` in both R and S-PLUS,

```
cor(huswif)
```

to give

	Hage	Hheight	Wage	Wheight	Hagefm
Hage	1.00	−0.25	0.88	−0.59	0.45
Hheight	−0.25	1.00	0.03	0.20	−0.61
Wage	0.88	0.03	1.00	−0.55	0.31
Wheight	−0.59	0.20	−0.55	1.00	−0.02
Hagefm	0.45	−0.61	0.31	−0.02	1.00

1.3.5 *Distances*

The concept of distance between observations is of considerable importance for some multivariate techniques. The most common measure used in *Euclidean distance*, which for two rows, say row i and row j, of the multivariate data matrix, **X**, is defined as

$$d_{ij} = \left[\sum_{k=1}^{q}(x_{ik} - x_{jk})^2\right]^{1/2}.$$

We can use the `dist` function in both R and S-PLUS to calculate these distances for the data in Table 1.2,

```
dis<-dist(huswif)
```

This can be converted into the required distance matrix by using the function `dist2full` given in `help(dist)`:

```
dist2full<-function(dis) {
    n<-attr(dis,"Size")
    full<-matrix(0,n,n)
    full[lower.tri(full)]<-dis
```

```
    full+t(full)
}
dis.matrix<-dist2full(dis)
round(dis.matrix,digits=2)
```

The resulting distance matrix is

```
numeric matrix: 10 rows, 10 columns.
        [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]   [,8]   [,9]  [,10]
 [1,]   0.00  52.55 154.33  60.05 257.56 135.81  82.60  69.76 128.46  79.58
 [2,]  52.55   0.00 193.17  76.57 268.35 177.15 126.16 106.58 164.15 110.28
 [3,] 154.33 193.17   0.00 147.71 206.69  56.52  75.23  92.32  32.40  84.39
 [4,]  60.05  76.57 147.71   0.00 202.60 150.88  86.35  57.81 123.83  78.39
 [5,] 257.56 268.35 206.69 202.60   0.00 255.33 222.10 202.96 206.03 211.81
 [6,] 135.81 177.15  56.52 150.88 255.33   0.00  67.61  94.42  51.24  80.87
 [7,]  82.60 126.16  75.23  86.35 222.10  67.61   0.00  33.85  55.31  39.28
 [8,]  69.76 106.58  92.32  57.81 202.96  94.42  33.85   0.00  67.68  29.98
 [9,] 128.46 164.15  32.40 123.83 206.03  51.24  55.31  67.68   0.00  54.20
[10,]  79.58 110.28  84.39  78.39 211.81  80.87  39.28  29.98  54.20   0.00
```

But this calculation of the distances ignores the fact that the variables in the data set are on different scales, and changing the scales will change the elements of the distance matrix without preserving the rank order of pairwise distances. It makes more sense to calculate the distances *after* some form of standardization. Here we shall divide each variable by its standard deviation. The necessary R code is

```
#find standard deviations of variables

std<-sd(huswif)
#use sweep function to divide columns of data matrix
#by the appropriate standard deviation
huswif.std<-sweep(huswif,2,std,FUN=''/'')
dis<-dist(huswif.std)
dis.matrix<-dist2full(dis)
round(dis.matrix,digits=2)
```

(In S-PLUS `std` will have to be calculated using `apply` and `var`.)
The result is the matrix given below

```
numeric matrix: 10 rows, 10 columns.
       [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
 [1,]  0.00 2.73 3.51 1.44 4.10 2.80 2.08 1.05 2.88  2.71
 [2,]  2.73 0.00 4.66 3.64 5.60 2.80 3.97 3.00 2.80  1.79
 [3,]  3.51 4.66 0.00 3.87 4.19 2.96 2.22 2.83 2.43  3.26
 [4,]  1.44 3.64 3.87 0.00 3.17 3.71 2.02 1.45 3.67  3.57
 [5,]  4.10 5.60 4.19 3.17 0.00 5.07 3.67 3.37 4.57  4.89
 [6,]  2.80 2.80 2.96 3.71 5.07 0.00 2.99 2.36 1.01  1.35
 [7,]  2.08 3.97 2.22 2.02 3.67 2.99 0.00 1.58 2.88  3.18
 [8,]  1.05 3.00 2.83 1.45 3.37 2.36 1.58 0.00 2.29  2.38
 [9,]  2.88 2.80 2.43 3.67 4.57 1.01 2.88 2.29 0.00  1.07
[10,]  2.71 1.79 3.26 3.57 4.89 1.35 3.18 2.38 1.07  0.00
```

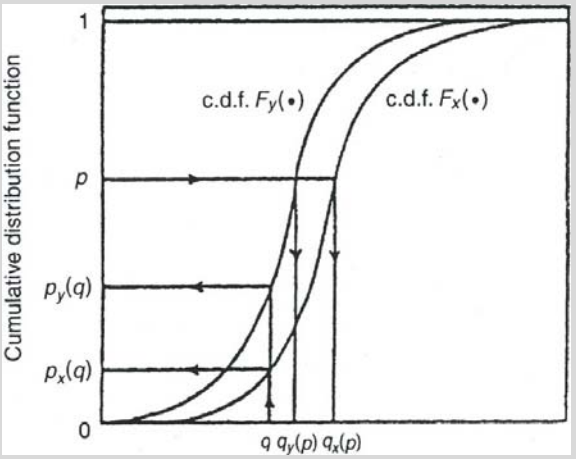
In essence, in the previous section `var` and `cor` have computed *similarities* between variables, and taking `1-cor(huswif)`, for example, would give a measure of distance between the variables. More will be said about similarities and distances in Chapter 5.

1.4 The Multivariate Normal Distribution

Just as the normal distribution dominates univariate techniques, the *multivariate normal distribution* often plays an important role in some multivariate procedures. The distribution is defined explicitly in, for example, Mardia et al. (1979) and is assumed by techniques such as *multivariate analysis of variance* (MANOVA); see Chapter 7. In practice some departure from this assumption is not generally regarded as particularly serious, but it may, on occasions, be worthwhile undertaking some test of the assumption. One relatively simple possibility is to use a *probability plotting* technique. Such plots are commonly applied in univariate analysis and involve ordering the observations and then plotting them against the appropriate values of an assumed cumulative distribution function. Details are given in Display 1.1

Display 1.1
Probability Plotting

- There are two basic types of plot for comparing two probability distributions, the *probability–probability plot* and the *quantile–quantile plot*. The diagram below may be used for describing each type.



- A plot of points whose coordinates are the cumulative probabilities $(p_x(q), p_y(q))$ for different values of q is a probability–probability plot, while a plot of the points whose coordinates are the quantiles $(q_x(p), q_y(p))$ for different values of p is a quantile–quantile plot.
- An example, a quantile–quantile plot for investigating the assumption that a set of data is from a normal distribution would involve plotting the ordered sample values $y_{(1)}, y_{(2)}, K, y_{(n)}$ against the quantiles of a standard normal distribution, $\Phi^{-1}[p(i)]$, where usually

$$p_i = \frac{i - \frac{1}{2}}{n} \qquad \text{and} \qquad \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}\, du.$$

- This is usually known as a *normal probability plot*.

For multivariate data such plots may be used to examine each variable separately, although marginal normality does not necessarily imply that the variables follow a multivariate normal distribution. Alternatively (or additionally), the multivariate observation might be converted to a single number in some way before plotting. For example, in the specific case of assessing a data set for multivariate normality, each q-dimensional observation $\mathbf{x}_i$, could be converted into a *generalized distance* (essentially *Mahalanobis distance*—see Everitt and Dunn, 2001), d_i^2 giving a measure of the distance of the particular observation from the mean vector of the complete sample, $\bar{\mathbf{x}}$; d_i^2 is calculated as

$$d_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}),$$

where $\mathbf{S}$ is the sample covariance matrix. This distance measure takes into account the different variances of the variables and the covariances of pairs of variables. If the observations do arise from a multivariate normal distribution, then these distances have, approximately, a *chi-squared distribution* with q degrees of freedom. So, plotting the ordered distances against the corresponding quantiles of the appropriate chi-square distribution should lead to a straight line through the origin.

First, let us consider some probability plots of a set of multivariate data constructed to have a multivariate normal distribution. We shall first use the R function `mvrnorm` (the MASS library will need to be loaded to make the function available) and the S-PLUS function `rmvnorm` to create 200 bivariate observation with correlation coefficient 0.5;

R:

```
#load MASS library
library(MASS)
#set seed for random number generation to get the same plots
set.seed(1203)
X<-mvrnorm(200,mu=c(0,0),Sigma=matrix(c(1,0.5,0.5,1.0),ncol=2))
```

S-PLUS:

```
set.seed(1203)
X<-rmvnorm(200,rho=0.5,d=2)
```

(The data generated by R and S-PLUS will not be the same. The results below are those obtained from the data generated by R.)

The probability plots for the individual variables are obtained using the following R and S-PLUS code:

```
#set up plotting area to take two side-by-side plots
par(mfrow=c(1,2))
qqnorm(X[,1],ylab="Ordered observations")
qqline(X[,1])
qqnorm(X[,2],ylab="Ordered observations")
qqline(X[,2])
```

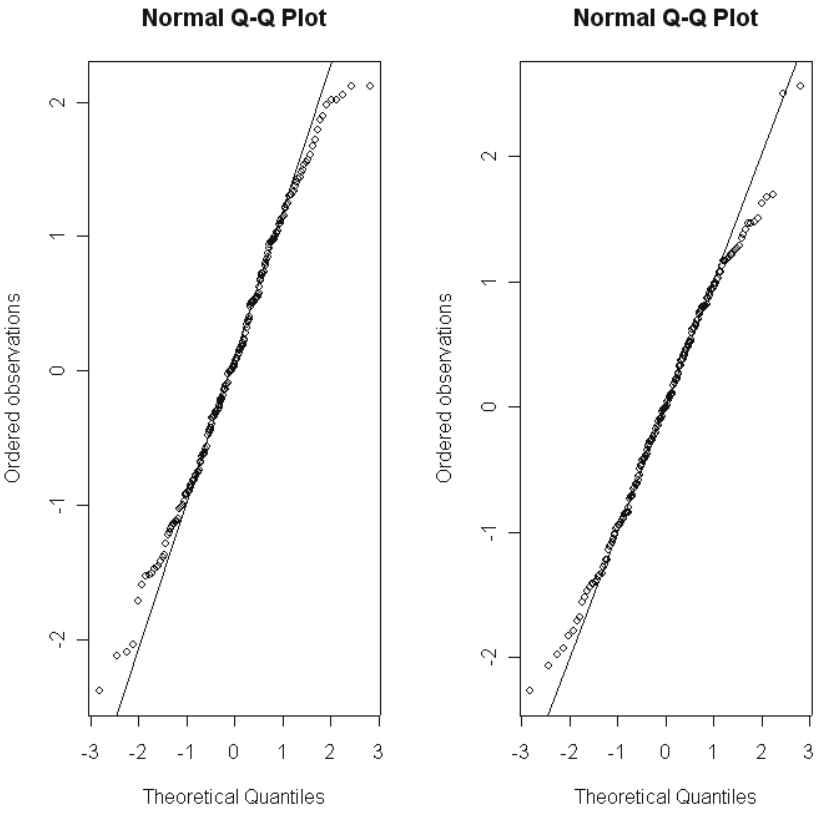


Figure 1.1 Probability plots for both variables in a generated set of bivariate data with $n = 200$ and a correlation of 0.5.

```
#qqnorm produces the required plot and qqline the line
#corresponding to a normal distribution
```

The resulting plots are shown in Figure 1.1. Neither probability plot gives any indication of a departure from linearity as we would expect.

The chi-square plot for both variables simultaneously can be found using the function `chisplot` given on the website mentioned in the preface. The required code is

```
par(mfrow=c(1,1))
chisplot(X)
```

Here the result appears in Figure 1.2. The plot is approximately linear, although some points do depart a little from the line.

If we now transform the previously generated data by simply taking the log of the absolute values of the generated data and then redo the previous plots, the results are shown in Figures 1.3 and 1.4. In each plot, there is a very clear departure from linearity, indicating the non-normality of the data.

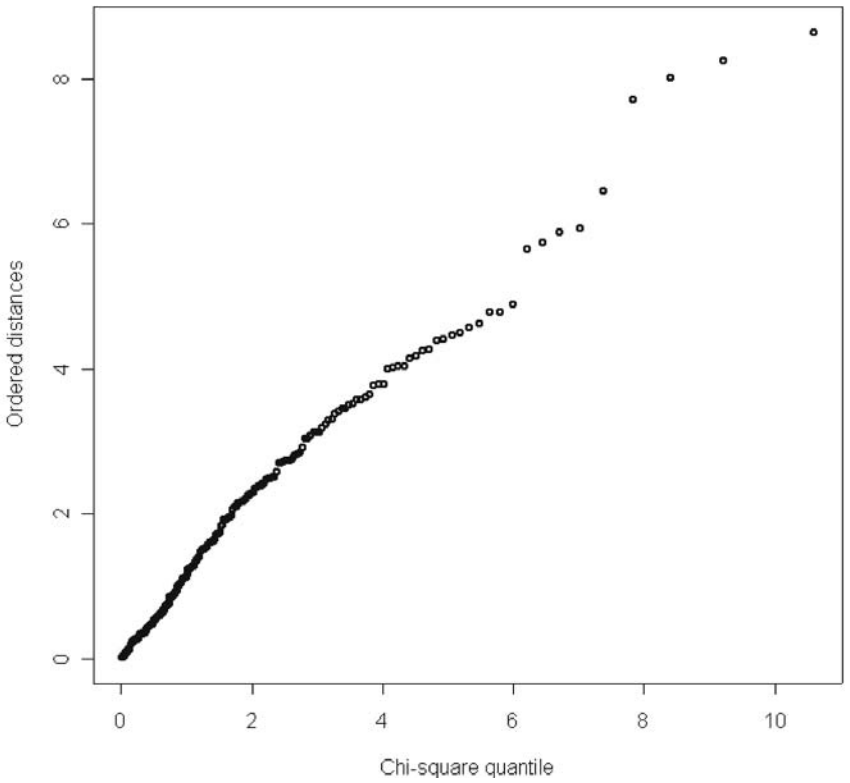


Figure 1.2 Chi-square probability plot of generated bivariate data.

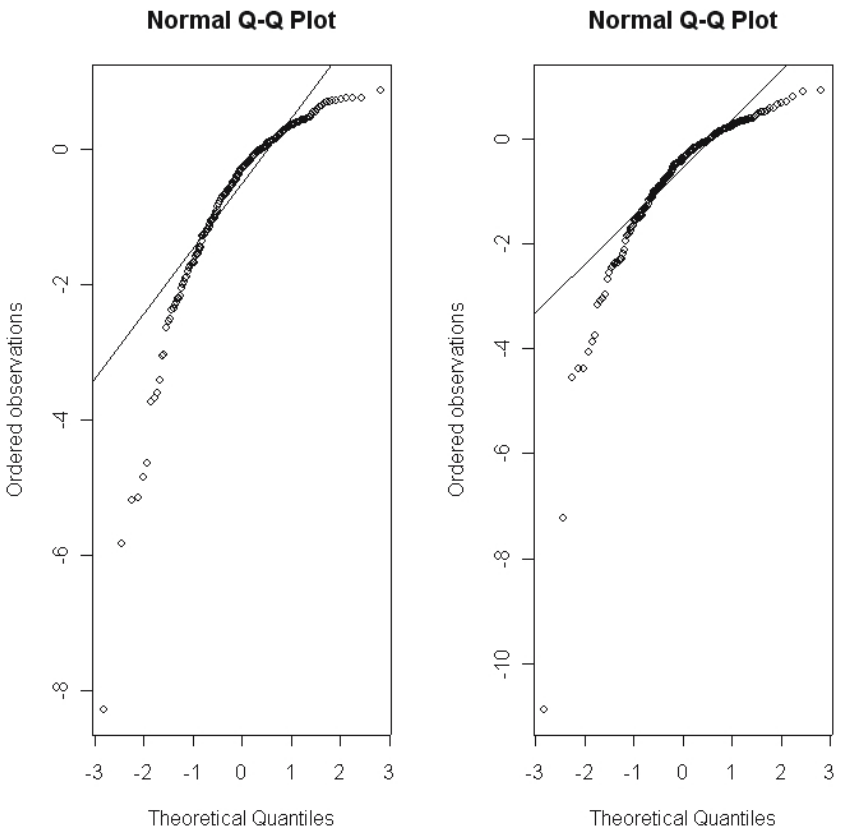


Figure 1.3 Probability plots of each variable in the transformed bivariate data.

1.5 The Aims of Multivariate Analysis

It is helpful to recognize that the analysis of data involves two separate stages. The first, particularly in new areas of research, involves *data exploration* in an attempt to recognize any nonrandom pattern or structure requiring explanation. At this stage, finding the question is often of more interest than seeking the subsequent answer. The aim of this part of the analysis being to generate possible interesting hypotheses for further study. (This activity is now often described as *data mining*.) Here, formal models designed to yield specific answers to rigidly defined questions are not required. Instead, methods are sought that allow possibly unanticipated patterns in the data to be detected, opening up a wide range of competing explanations. Such techniques are generally characterized by their emphasis on the importance of visual displays and graphical representations and by the lack of any associated stochastic model, so that questions of the statistical significance of the results are hardly ever of much importance.

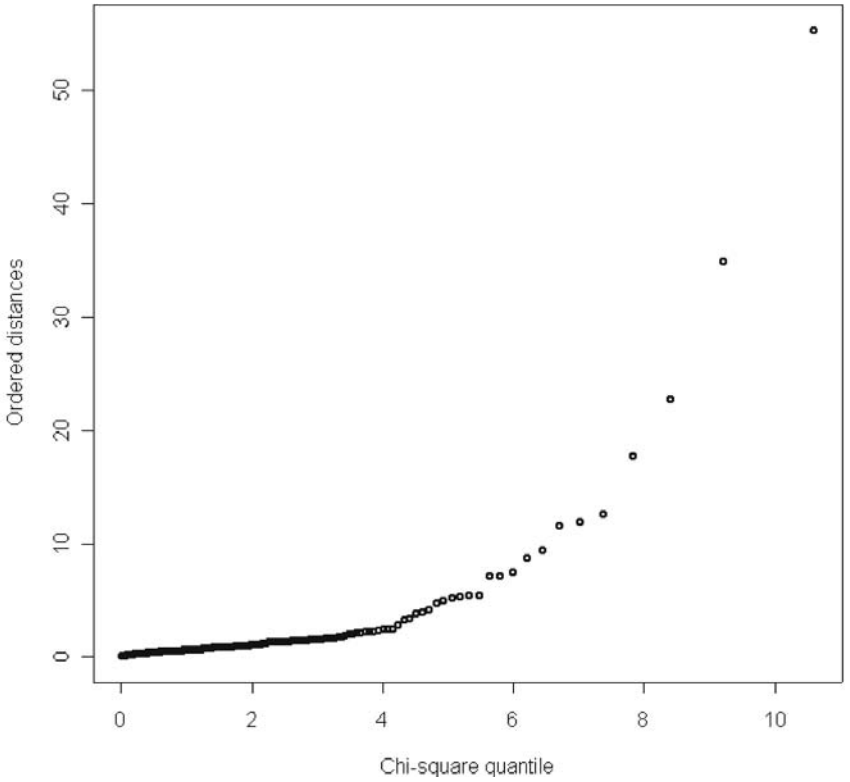


Figure 1.4 Chi-square plot of generated bivariate data after transformation.

A confirmatory analysis becomes possible after a research worker has some well-defined hypothesis in mind. It is here that some type of statistical significance test might be considered. Such tests are well known and, although their misuse has often brought them into some disrepute, they remain of considerable importance.

In this text, Chapters 2–6 describe techniques that are primarily exploratory, and Chapters 7–9 techniques that are largely confirmatory, but this division should not be regarded as much more than a convenient arrangement of the material to be presented, since any sensible investigator will realize the need for exploratory and confirmatory techniques, and many methods will often be useful in both roles. Perhaps attempts to rigidly divide data analysis into exploratory and confirmatory parts have been misplaced, and what is really important is that research workers should have a flexible and pragmatic approach to the analysis of their data, with sufficient expertise to enable them to choose the appropriate analytical tool and use it correctly. The choice of tool, of course, depends on the aims or purpose of the analysis.

Most of this text is written from the point of view that there are no rules or laws of scientific inference—that is, "anything goes" (Feyerabend, 1975). This implies

that we see both exploratory and confirmatory methods as two sides of the same coin. We see both methods as essentially tools for data exploration rather than as formal decision-making procedures. For this reason we do not stress the values of significance levels, but merely use them as criteria to guide a modelling process (using the term "modelling" as a method or methods of describing the structure of a data set). We believe that in scientific research it is the skillful interpretation of evidence and subsequent development of hunches that are important, rather than a rigid adherence to a formal set of decision rules associated with significance tests (or any other criteria, for that matter). One aspect of the scientific method, however, which we do not discuss in any detail, but which is the vital component in testing the theories that come out of our data analyses, is replication. It is clearly unsafe to search for a pattern in a given data set and to "confirm" the existence of such a pattern using the same data set. We need to validate our conclusions using further data. At this point our subsequent analysis might become truly confirmatory.

1.6 Summary

Most data collected in the social sciences and other disciplines are multivariate. To fully understand most such data sets the variables need to be analyzed simultaneously. The remainder of this book is concerned with methods that have been developed to make this possible, and to help discover any patterns or structure in the data that may have important implications in uncovering the data's message.