Appendix

An Aide Memoir for R and S-PLUS®

1. Elementary Commands

Elementary commands consist of either expressions or assignments. For example, typing the expression

```
 > 42 + 8
```

in the **Commands** window and pressing **Return** will produce the following output:

```
[1]   50
```

In the remainder of this chapter, we will show the command (preceded by the prompt >) and the output as they would appear in the **Commands** window together like this:

```
> 42 + 8
[1]   50
```

Instead of just evaluating an expression, we can assign the value to a scalar using the syntax `scalar <- expression`

```
> x <- 42 + 8
```

Longer commands can be split over several lines by pressing Return before the command is complete. To indicate waiting for completion of a command, a "+" occurs instead of the > prompt. For illustration, we break the line in the assignment above:

```
> x<-
+ 48+8
```

2. Vectors

A commonly used type of R and S-PLUS® object is a *vector*. Vectors may be created in several ways of which the most common is via the concentrate command, `c`, which combines all values given as arguments to the function into a vector. For example,

```
>x<-c(1, 2, 3,4)
>x
[1] 1 2 3 4
```

Here, the first command creates a vector and the second command, `x`, a short-form for `print(x)`, causes the contents of the vector to be printed. (Note that R and S-PLUS are case sensitive, and so, for example, x and X are different objects.) The number of elements of a vector can be determined using the `length()` function:

```
>length(x)
[1] 4
```

The `c` function can also be used to combine *strings* which are denoted by enclosing them in ".". For example,

```
>names <-c ("Brian", "Sophia", "Harry")
>names
[1] "Brian" "Sophia" "Harry"
```

The `c()` function also works with a mixture of numeric and string values, but in this case, all elements in the resulting vector will be converted to strings as in the following.

```
> mix <-c(names, 55, 33)
> mix
[1] "Brian" "Sophia"    "Harry" "55"    "33"
```

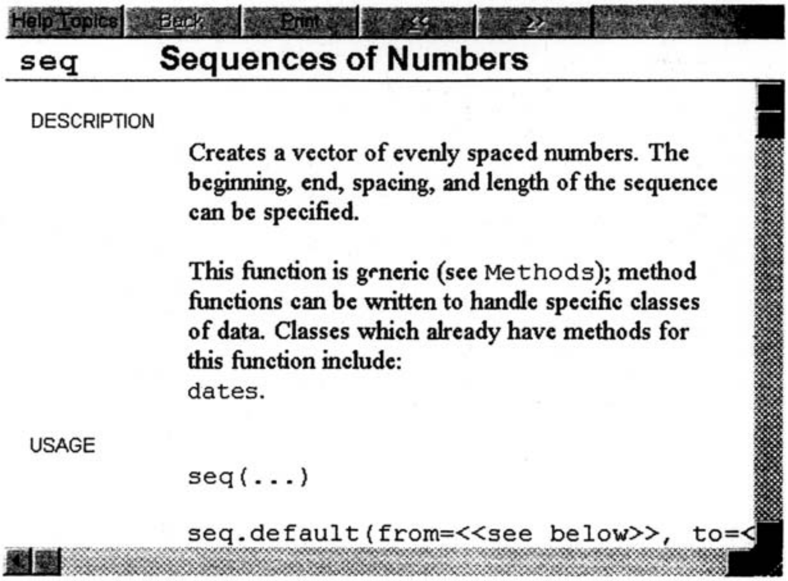
Vectors consisting of regular sequences of numbers can be created using the `seq()` function. The general syntax of this function is `seq(lower, upper, increment)`. Some examples are given below:

```
>seq (1, 5, 1)
[1] 1 2 3 4 5
>seq (2, 20, 2)
[1] 2 4 6 8 10 12 14 16 18 20
>x <-c(seq(1, 5, 1), seq (4, 20, 4))
>x
[1] 1 2 3 4 5 4 8 12 16 20
```

When the increment argument is one it can be left out of the command. The same applies to the lower value. More information about the `seq` function and all other R and S-PLUS functions can be found using the `help` facilities, e.g.,

```
>help(seq)
```

shows the following information:



seq **Sequences of Numbers**

DESCRIPTION

Creates a vector of evenly spaced numbers. The beginning, end, spacing, and length of the sequence can be specified.

This function is grneric (see `Methods`); method functions can be written to handle specific classes of data. Classes which already have methods for this function include:
`dates`.

USAGE

```
seq(...)

seq.default(from=<<see below>>, to=<
```

Sequences with increments of one can also be obtained using the syntax `first: last`, for example,

```
>1:5
[1] 1 2 3 4 5
```

A further useful function for creating vectors with regular patterns is the `rep` function, with general form `rep(pattern, number of times)`. For example,

```
>rep(10, 5)
[1] 10 10 10 10 10
>rep (1:3, 3)
[1] 1 2 3 1 2 3 1 2 3
> x <- rep(seq(5), 2)
> x
[1] 1 2 3 4 5 1 2 3 4 5
```

The second argument of `rep` can also be a vector of the same length as the first argument to indicate how often each element of the first argument is to be repeated as shown in the following;

```
> x <- rep(seq (3), c(1, 2, 3))
> x
[1] 1 2 2 3 3 3
```

Increasingly complex vectors can be built by repeated use of the `rep` function

```
> x <- rep (seq (3), rep (3, 3))
> x
[1] 1 1 1 2 2 2 3 3 3
```

We can access a particular element of a vector by giving the required position in square brackets- here are two examples

```
> x <- 1:5
> x[3]
[1] 3

>x[c(1, 4)]
[1] 1 4
```

A vector containing several elements of another vector can be obtained by giving a vector of required positions in square brackets:

```
> x[c(1, 3)]
[1] 1 3
> x[1:3]
[1] 1 2 3
```

We can carry out any of the arithmetic operations described in Table A.1 between two scalars, a vector and a scalar or two vectors. An arithmetic operation between two vectors returns a vector whose elements are the results of applying the operation to the corresponding elements of the original vectors. Some examples follow:

```
> x<- 1:3
> x+2
[1] 3 4 5

>x + x
[1] 2 4 6
> x* x
[1] 1 4 9
```

We can also apply mathematical functions such as the square root or logarithm, or the others listed in Table A.2, to vectors. The functions are simply applied to each element of the vector. For example,

```
> x <- 1:3
> sqrt (x*x)
[1] 1 2 3
```

Table A.1 Arithmetic Operators

Operator	Meaning	Expression	Result
$+$	Plus	$2+3$	5
$-$	Minus	$5-2$	3
$*$	Times	$5*2$	10
/	Divided by	10/2	5
$\wedge$	Power	$2 \wedge 3$	8

Table A.2 Common Functions

S-PLUS function	Meaning
`sqrt ()`	Square root
`log ()`	Natural logarithm
`log10 ()`	Logarithm base 10
`exp ()`	Exponential
`abs ()`	Absolute value
`round ()`	Round to nearest integer
`ceiling ()`	Round up
`floor ()`	Round down
`sin (), cos (), tan ()`	sine, cosine, tangent
`asin (), acos (), atan ()`	arc sine, arc cosine, arc tangent

3. Matrices

Matrix objects are frequently needed in R and S-PLUS and can be created by the use of the `matrix` function. The general syntax is

```
matrix (data, nrow, ncol, byrow = F)
```

The last argument specifies whether the matrix is to be filled row by row or column by column and takes on a logical value. The expression *byrow=F* indicates that `F` (false) is the default value. An example follows;

```
> x <-c(1, 2, 3)
> y <-c(4, 5, 6)
> xy <- matrix (c(x, y), nrow =2)
> xy
        [,1]    [,2]    [,3]
[1,]    1       3       5
[2,]    2       4       6
```

Here the number of columns is not specified and so is determined by simple division:

```
> xy <-matrix (c(x, y), nrow = 2, byrow =T)
xy

        [,1]    [,2]     [,3]
 [1,]   1       2        3
 [2,]   4       5        6
```

Here the matrix, is filled row-wise instead of by columns by setting the `byrow` argument to `T` for True. A square bracket with two numbers separated by a comma is used to refer to an element of a matrix. The first number specifies the row, and the second specifies the column.

```
>xy [1, 3]
[1] 3
```

The `[i,]` and `[,j]` nomenclature is used to refer to complete rows or columns of a matrix and can be used to extract particular rows or columns as shown in the following examples;

```
> xy [1,]
[1] 1 2 3
> xy [,2]
[1] 2 5

>xy [, c(1, 3)]

        [,1]     [,2]
[1,]    1        3
[2,]    4        6
```

As with vectors, arithmetic operations operate element by element when applied to matrices, for example;

```
> xy* xy

        [,1]     [,2]     [,3]
[1,]    1        4        9
[2,]    16       25       36
```

Matrix multiplication is performed using the %*% operation as here

```
> xy %*% t(xy)

              [,1]     [,2]
[1,]    14    32
[2,]    32    77
```

Here the matrix *xy* is multiplied by its transpose (obtained using the *t*() function). An attempt to apply matrix multiplication to *xy* by *xy* would, of course, result in an error message. It is usually extremely helpful to attach names to the rows and columns to a matrix. This can be done using the `dimension()` function. We shall illustrate this in Section 5 after we have covered `list` objects.

As with vectors, matrices can be formed from numeric and string objects, but in the resulting matrix, all elements will be strings as illustrated below:

```
> Mix <- matrix(c(names, 55, 32, 30), nrow = 2
+ byrow = T)
> Mix

        [,1]        [,2]         [,3]
[1,]    "Brian"     "Sophia"     "Harry"
[2,]    "55"        "32"         "30"
```

Higher dimensional matrices with up to eight dimensions can be defined using the `array()` function.

4. Logical Expressions

So far, we have mentioned values of type numeric or character (string). When a numeric value is missing, it is of type NA. (Complex numbers are also available.) Another type in R and S-PLUS is logical. There are two logical values, `T` (true) and

Table A.3 Logical Operators

Operator	Meaning
<	Less than
>	Greater than
<=	Less than or equal to
>=	Greater than or equal to
==	Equal to
!=	Not equal to
&	And
|	Or
!	not

F (false), and a number of logical operations that are extremely useful when making comparisons and choosing particular elements from vectors and matrices.

The symbols used for the logical operations are listed in Table A.3. We can use a logical expression to assign a logical value (T or F) to x:

```
> x <-3 = = 4
> x
[1] F
> x <-3 < 4
> x
[1] T
> x < - 3 = = 4 & 3 < 4
> x
[1] F
> x < - 3 = = 4 | 3 < 4
> x
[1] T
```

In addition to logical operators, there are also logical functions. Some examples are given below:

```
> is.numeric (3)
[1] T
> is.character (3)
[1] F
> is.character ("3")
[1] T
> 1/0
[1] Inf
>is.numeric (1/0)
[1] T
>is.infinite (1/0)
[1] T
```

Logical operators or functions operate on elements of vectors and matrices in the same say as arithmetic operators:

```
> is.na(c(1, 0, NA, 1))
[1] F F T F
```

```
> ! is.na (c(1, 0, NA, 1))
[1] T T F T
> x <- seq(20)
> x <10
[1] T T T T T T T T T F F F F F F F F F F F
```

A logical vector can be used to extract a subset of elements from another vector as follows:

```
> x[x <10]
[1] 1 2 3 4 5 6 7 8 9
```

Here, the elements of the vector less than 10 are selected as the values corresponding to *T* in the vector $x < 10$. We can also select element in *x* depending on the values in another vector *y*:

```
> x <-seq(50)
> y <- c(rep(0, 10), rep(1, 40))
> x[y = =0]
[1]  1 2 3 4 5 6 7 8  9 10
```

5. List Objects

`List` objects allow any other R or S-PLUS objects to be linked together. For example,

```
>x<-seq(10)
>y<- matrix(seq(10), nrow = 5
>xylist<-list (x,y)
>xylist
[[1]]:
[1] 1 2 3 4 5 6 7 8 9 10

[[2]]:

        [,1]     [,2]
[1,]    1        6
[2,]    2        7
[3,]    3        8
[4,]    4        9
[5,]    5        10
```

Note the elements of the list are referred to by a double square brackets notation; so we can print the first component of the list using

```
>xylist[[1]]
[1] 1 2 3 4 5 6 7 8 9 10
```

The components of the list can also be given names and later referred using the `list$name` notation,

```
>xylist <-list (X=x, Y=y)
>xylist$X
[1] 1 2 3 4 5 6 7 8 9 10
```

List objects can, of course, include other list objects

```
>newlist<-list(xy=xylist, z=rep(0,10))
>newlist$xy
$X
[1] 1 2 3 4 5 6 7 8 9 10

$Y:
              [,1]     [,2]
[1,]     1     6
[2,]     2     7
[3,]     3     8
[4,]     4     9
[5,]     5     10

>newlist$z
[1] 0 0 0 0 0 0 0 0 0 0
```

The rows and columns of a matrix can be named using the `dimnames()` function and a list object

```
>x<-matrix(seq(12), nrow=4)
> dimnames(x)<-list(c("R1","R","R3","R4"),
+c("C1", "C2", "C3"))
>x

      C1     C2     C3
R1    1      5      9
R2    2      6      10
R3    3      7      11
R4    4      8      12
```

The names can be created more efficiently by using the `paste()` function, which combines different strings and numbers into a single string

```
> dimnames(x)<-list(paste("row", seq (4)),
+paste ("col", seq(3)))
>x
        col 1     col 2     col 3
row1    1         5         9
row2    2         6         10
row3    3         7         11
row4    4         8         12
```

Having named the rows and columns, we can, if required, refer to elements of the matrix using these names,

```
>x["row 1", "col 3"]
[1] 9
```

6. Data Frames

Data sets in R and S-PLUS are usually stored as matrices, which we have already met, or as data frames, which we shall describe here.

Data frames can bind vectors of different types together (e.g., numeric and character), retaining the correct type of each vector. In other respects, a data frame is like a matrix so that each vector should have the same number of elements. The syntax for creating a data frame is `data.frame(vector1,vector 2, ...)`, and an example of how a small data frame can be created is as follows:

```
>height<-c(50, 70, 45, 80, 100)
>weight<-c(120, 140, 100, 200, 190)
>age<-c(20, 40, 41, 31, 33)
>names<-c("Bob", "Ted", "Alice", "Mary", "Sue")
sex<-c("Male", "Male", "Female", "Female")
>data<-data.frame(names, sex, height, weight, age)
>data
```

```
     names    sex       height    weight    age
1    Bob      Male      50        120       20
2    Ted      Male      70        140       40
3    Alice    Female    45        100       41
4    Mary     Female    80        200       31
5    Sue      Female    100       190       33
```

Particular parts of a data frame can be extracted in the same way as for matrices

```
>data[,c(1,2,5)]
```

```
     names    sex       age
1    Bob      Male      20
2    Ted      Male      40
3    Ali      Female    41
4    Mary     Female    31
5    Sue      Female    33
```

Column names can also be used

```
>data[,"age"]
[1] 20 40 41 31 33
```

Variables can also be accessed as in lists:

```
>data$age
[1] 20 40 41 31 33
```

It is, however, more convenient to “attach” a data frame and work with the column names directly, for example,

```
>attach (data)
>age
[1] 20 40 41 31 33
```

Note that the `attach()` command places the data frame in the 2nd position in the search path. If we assign a value to `age`, for example,

```
>age <-10
>age
[1] 10
```

This creates a new object in the first position of the search path that “masks” the `age` variable of the data frame. Variables can be removed from the first position in the search path using the `rm()` function:

```
>rm(age)
```

To change the value of age within the data frame, use the syntax

```
>data$age<-c(20, 30, 45, 32, 32)
```

References

Afifi, AA, Clark, VA and May, S. (2004) Computer-Aided Multivariate Analysis, (4th ed.). London: Chapman and Hall.

Alon, U, Barkai, N, Notterman, DA, Gish, K, Ybarra, S, Mack, D, and Levine, AJ (1999) Broad patterns of gene expressions revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Cell Biology, 99, 6745–6750.

Anderson, JA (1972) Separate sample logistic discrimination. Biometrika, 59, 19–35.

Banfield, JD and Raftery, AE (1993) Model-based Gaussian and non-Gaussian clustering. Biometrics, 49, 803–821.

Bartholomew, DJ (1987) Latent Variable Models and Factor Analysis. Oxford: Oxford University Press.

Bartlett, MS (1947) Multivariate analysis. Journal of the Royal Statistical Society B, 9, 176–197.

Becker, RA and Cleveland, WS (1994) S-PLUS Trellis Graphics User's Manual Version 3.3. Seattle: Mathsoft, Inc.

Benzécri, JP (1992) Correspondence Analysis Handbook. New York: Marcel Dekker.

Blackith, RE and Rayment, RA (1971) Multivariate Morphometrics. London: Academic Press.

Carpenter, J, Pocock, SJ, and Lamm, CJ (2002) Coping with missing data in clinical trials: A model-based approach applied to asthma trials. Statistics in Medicine, 21, 1043–1066.

Chambers, JM, Cleveland, WS, Kleiner, B, and Tukey, PA (1983) Graphical Methods for Data Analysis. Belmont, CA: Wadsworth.

Chatfield, C and Collins, AJ (1980) Introduction to Multivariate Analysis. London: Chapman and Hall.

Cleveland, WS (1979) Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association, 74, 829–836.

Cleveland, WS and McGill, ME (1987) Dynamic Graphics for Statistics. Belmont, CA: Wadsworth.

Crowder, MJ (1998) Nonlinear growth curve. In Encyclopedia of Biostatistics (eds. P Armitage and T Colton). Chichester: Wiley.

Dalgaard, P (2002) Introductory Statistics with R. New York: Springer.

Davis, CS (2002) Statistical Methods for the Analysis of Repeated Measurements. New York: Springer.

Dempster, AP, Laird, NM and Ruben, DB (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, B, 39, 1–38.

de Leeuw, J (1985) Book review. Psychometrika, 50, 371–375.

Diggle, PJ (1998) Dealing with missing values in longitudinal studies. In Statistical Analysis of Medical Data (eds. BS Everitt and G Dunn). London: Arnold.

Diggle, PJ, Heagerty, P, Liang, KY, and Zeger, SL (2002) Analysis of Longitudinal Data. Oxford: Oxford University Press.

Diggle, PJ and Kenward, MG (1994) Informative dropout in longitudinal analysis (with discussion) Applied Statistics, 43, 49–93.

Dunn, G, Everitt, BS, and Pickles, A (1993) Modelling Covariances and Latent Variables Using EQS. London: Chapman and Hall.

Everitt, BS (1984) An Introduction to Latent Variable Models. Boca Raton, Florida: CRC/Chapman and Hall.

Everitt, BS (1987) Statistics in Psychiatry. Statistical Science, 2, 107–134.

Everitt, BS (2002) Modern Medical Statistics: A Practical Guide. London: Arnold.

Everitt, BS and Bullmore, ET (1999) Mixture model mapping of brain activation in functional magnetic resonance images. Human Brain Mapping, 7, 1–14.

Everitt, BS and Dunn, G (2001) Applied Multivariate Data Analysis (2nd ed.). London: Arnold.

Everitt, BS, Gourlay, J, and Kendall, RE (1971) An attempt at validation of traditional psychiatric syndromes by cluster analysis. British Journal of Psychiatry, 119, 299–412.

Everitt, BS, Landau, S, and Leese, M (2001) Cluster Analysis (4th ed.). London: Arnold.

Everitt, BS and Rabe-Hesketh, S (1997) The Analysis of Proximity Data. London: Arnold.

Feyerabend, P (1975) Against Method. London: Verso.

Fisher, NI and Switzer, P (1985) Chi-plots for assessing dependence. Biometrika, 72, 253–265.

Fisher, NI and Switzer, P (2001) Graphical assessment of dependence: Is a picture worth 100 tests? The American Statistician, 55, 233–239.

Fisher, RA (1936) The use of multiple measurements on taxonomic problems. Annals of Eugenics, 7, 179–188.

Fraley, C and Raftery, AE (1998) How many clusters? Which cluster method? Answers via model-based cluster analysis. The Computer Journal, 41, 578–588.

Fraley, C and Raftery, AE (1999) MCLUS: Software for the model-based cluster analysis. Journal of Classification, 16, 297–306.

Fraley, C and Raftery, AE (2002) Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association, 97, 611–631.

Frets, GP (1921) Heredity of head form in man. Genetica, 3, 193–384.

Friedman, HP and Rubin, J (1967) On some invariant criteria for grouping data. Journal of the American Statistical Association, 62, 1159–1178.

Friedman, JH (1989) Regularized discriminant analysis. Journal of the American Statistical Association, 84, 165–175.

Goldberg, KM and Iglewicz, B (1992) Bivariate extensions of the boxplot. Technometrics, 34, 307–320.

Gordon, AD (1999) Classification (2nd ed.). Boca Raton, Florida: Chapman & Hall/CRC Press.

Gower, JC (1966) Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika, 53, 325–338.

Greenacre, M (1992) Correspondence analysis in medical research. Statistical Methods in Medical Research, 1, 97–117.

Hancock, BW, Aitken, M, Martin, JF, Dunsmore, IR, Ross, CM, Carr, I, and Emmanuel, IG (1979) Hodgkin's disease in Sheffield (1971–1976). Clinical Oncology, 5, 283–297.

Hand, DJ (1998) Discriminant analysis, linear. In Encyclopedia of Biostatistics (eds. P Armitage and T Colton). Chichester: Wiley.

Hawkins, DM, Muller, MW, and ten Krooden, JA (1982) Cluster analysis. In Topics in Applied Multivariate Analysis (ed. DM Hawkins). Cambridge: Cambridge University Press.

Heitjan, DF (1997). Ignorability, sufficiency and ancillarity. Journal of the Royal Statistical Society, Series B (Methodological) 59, 375, 381.

Henderson, HV and Velleman, PF (1981) Building multiple regression models interactively. Biometrics, 37, 391–411.

Hendrickson, AE and White, PO (1964) Promax: A quick method for rotation to oblique simple structure. British Journal of Mathematical Statistical Psychology, 17, 65–70.

Heywood, HB (1931) On finite sequences of real numbers. Proceedings of the Royal Society, Series A, 134, 486–501.

Hills, M (1977) Book review. Applied Statistics, 26, 339–340.

Hotelling, H (1933) Analysis of a complex of statistical variables into prinicpal components. Journal of Educational Psychology, 24, 417–441.

Huba, GJ, Wingard, JA, and Bentler, PM (1981) A comparison of two latent variable causal models for adolescent drug use. Journal of Personality and Social Psychology, 40, 180–193.

Hyvarinen, A, Karhunen, J, and Oja, E (2001) Independent Component Analysis. New York: Wiley.

Jennrich, RI and Sampson, PF (1966) Rotation for simple loadings. Psychometrika, 31.

Johnson, RA and Wichern, DW (2003) Applied Multivariate Statistical Analysis: Prentice-Hall.

Jolliffe, IT (1972) Discarding variables in a principal components analysis I: Artificial data. Applied Statistics, 21, 160–173.

Jolliffe, IT (1986) Principal Components Analysis. New York: Springer-Verlag.

Jolliffe, IT (1989) Roation of ill-defined principal components. Applied Statistics, 38, 139–148.

Jolliffe, IT (2002) Principal Component Analysis (2nd ed.). New York: Springer.

Jones, MC and Sibson, R (1987) What is projection pursuit? Journal of the Royal Statistical Society A, 150, 1–36.

Kaiser, HF (1958) The varimax criterion for analytic rotation in factor analysis, Psychometrika, 23, 187–200.

Keyfitz, N and Flieger, W (1971) Population: The Facts and Methods of Demography. San Francisco: W. H. Freeman.

Krause, A and Olsen, M (2002) The Basics of S-PLUS (3rd ed.). New York: Springer.

Krzanowski, WJ (1988) Principles of Multivariate Analysis. Oxford: Oxford University Press.

Krzanowski, WJ (2004) Canonical correlation. In Encyclopedic Companion to Medical Statistics (eds. BS Everitt and C. Palmer). London: Arnold.

Lackey, NR and Sullivan, J (2003) Making Sense of Factor Analysis: The Use of Factor Analysis for Instrument Development in Health Care Research. Sage Publications.

Lawley, DN and Maxwell, AE (1971) Factor Analysis as a Statistical Method (2nd ed.). London: Butterworths.

Little, RJA and Rubin, DB (1987) Statisticial Analysis with Missing Data. New York: Wiley.

Longford, N (1993) Inference about variation in clustered binary data. In Multilevel Conference. Los Angeles.

Mardia, KV, Kent, JT, and Bibby, JM (1979) Multivariate Analysis. London: Academic Press.

Marriott, FHC (1982) Optimization methods of cluster analysis. Biometrika, 69, 417–421.

Mayor, M, Frei, P-Y, and Roukema, B (2003) New Worlds in the Cosmos: The Discovery of Exoplanets. English language edition, Cambridge University Press 2003; originally published as Les Nouveaux Mondes du Cosmos, Editions du Seuil 2001.

McDonald, GC and Schwing, RC (1973) Instabilities of regression estimates relating air pollution to mortality. Technometrics, 15, 463–482.

Morant, GM (1923) A first study of the Tibetan skull. Biometrika, 14, 193–260.

Morrison, DF (1990) Multivariate Statistical Methods (3rd ed.). New York: McGraw-Hill.

Murray, GD and Findlay, JG (1988) Correcting for the bias caused by dropouts in hypertension trials. Statistics in Medicine, 7, 941–946.

Muthen, LK and Muthen, BO (1998) Mplus Users Guide.

O'Sullivan, JB and Mahon, CM (1966) Glucose tolerance test: variability in pregnant and non-pregnant women. American Journal of Clinical Nutrition, 19, 345–351.

Pearson, K (1901) On lines and planes of closest fit to systems of points in space. Philosophical Magazine, 2, 559–572.

Proudfoot, J, Goldberg, D, Mann, A, Everitt, BS, Marks, I, and Gray, J (2003) Computerised, interactive, multimedia cognitive behavioral therapy for anxiety and depression in general practice. Psychological Medicine, 33, 217–227.

Rawlings, JO, Pantula, SG, and Dickey, AD (1998) Applied Regression Analysis. New York: Springer.

Rencher, AC (1995) Methods of Multivariate Analysis. New York: Wiley.

Rohlf, FJ (1970) Adaptive hierarchical clustering schemes. Systematic Zoology, 19, 58–82.

Rousseeuw, PJ (1985) Multivariate estimation with high breakdown point. In Mathematical Statistics and Applications (eds. W Grossman, G Pfilug, I Vincze, and W Wertz). Dordrecht: Reidel.

Rousseeuw, PJ and van Zomeren, B (1990) Unmasking multivariate outliers and leverage points (with discussion). Journal of the American Statistical Association, 85, 633–651.

Rubin, DB (1976) Inference and missing data. Biometrika, 63, 581–592.

Rubin, DB (1987) Multiple Imputation for Nonresponse in Surveys. New York: Wiley.

Schafer, JL (1999) Multiple imputation: A primer. Statistical Methods in Medical Research, 8, 3–15.

Schimert, J, Schafer, JL, Hesterberg, T, Fraley C, and Clarkson, DB (2000) Analysing Data with Missing Values in S-PLUS. Seattle: Insightful Corporation.

Scott, AJ and Symons, MJ (1971) Clustering methods based on likelihood ratio criteria. Biometrics, 37, 387–398.

Sibson, R (1979) Studies in the robustness of multidimensional scaling. Perturbational analyis of classical scaling. Journal of the Royal Statistical Society B, 41, 217–229.

Silverman, BW (1986) Density Estimation for Statistics and Data Analysis. London: Chapman and Hall.

Spearman, C (1904) General intelligence objectively determined and measured. American Journal of Psychology, 15, 201–293.

Spicer, CC, Laurence, GJ, and Southall, DP (1987) Statistical analysis of heart rates and subsequent victims of sudden infant death syndrome. Statistics in Medicine, 6, 159–166.

Tabachnick, BG and Fidell, B (2000) Using Multivariate Statistics (4th ed.). Upper Saddle River, NJ: Allyn & Bacon.

Thurstone, LL (1931) Multiple factor analysis. Psychology Review, 39, 406–427.

Tubb, A, Parker, NJ, and Nickless, G (1980) The analysis of Romano-British pottery by atomic absorption spectroplotomy. Archaeometry, 22, 2, 153–171.

Tufte, ER (1983) The Visual Display of Quantitative Information. Cheshire, CT: Graphics Press.

Velleman, PF and Wilkinson, L (1993) Normal, ordinal, internal and ratio typologies are misleading. The American Statistician, 47, 65–72.

Verbyla, AP, Cullis, BR, Kenward, MG, and Welham, SJ (1999) The analysis of designed experiments and longitudinal data using smoothing splines (with discussion). Applied Statistics, 48, 269–312.

Wand, MP and Jones, CM (1995) Kernel Smoothing. London: Chapman and Hall.

Ward, JH (1963) Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 58, 236–244.

Wastell, DG and Gray, R (1987) The numerical approach to classification: a medical application to develop a typoloty of facial pairs. Statistics in Medicine, 6, 137–164.

Wermuth, N (1976) Exploratory analyses of multidimensional contingency tables. In Proceedings 9th International Biometrics Conference, pp. 279–295: Biometrics Society.

Young, G and Householder, AS (1938) Discussion of a set of points in terms of their mutual distances. Psychometrika, 3, 19–22.

Zerbe, GO (1979) Randomization analysis of the completely randomized design extended to growth and response curves. Journal of the American Statistical Association, 74, 215–221.

Index

agglomerative hierarchical clustering
  cluster analysis and, 115–123
  complete linkage clustering, 118
  data partitions in, 116–117
  dendrograms and, 117, 121
  evolutionary trees and, 117
  group average clustering and, 118
  intercluster dissimilarity, 118–123
  single linkage clustering, 118
airline distances, 98–99
air pollution, studies of, 17–22, 49–61, 159
average linkage methods, 134

Bartlett test, 170
Bayes theorem, 189
Bayesian Information Criterion (BIC), 130
Beating the Blues (BtB) program, 191
Beck Depression Inventory, 191
best linear unbiased predictions (BLUPs), 190
BIC. *See* Bayesian Information Criterion
bivariate methods, 29–32
  boxplots and, 25–29
  convex hull approach, 22–23
  correspondence analysis and, 104–105
  density estimates and, 29–32
  kernel function, 29
  See also specific topics
BLUPs. *See* best linear unbiased predictions
boxplots, 16, 25–29
bubbleplots, 32

canonical correlation analysis (CCA), 160–167
canonical variates, 149–155
CCA. *See* canonical correlation analysis
chi-square distributions, 10–12, 149
  chiplot function, 23–25, 149
  chi-squared test, 104
  distance in, 104–108
  plots of, 10–11, 23–25, 149
classification functions, 149–155
classification maximum likelihood methods, 115, 129–134
cluster analysis, 29, 115–136
  agglomerative hierarchical clustering, 115–123
  classification maximum likelihood and, 129–134
  k-means clustering, 123–128
  model-based clustering, 128–134
  three types of, 115
cognitive behavioral therapy, 191
common factors, 65–68
communality, of variables, 67–68
complete-case analysis, 3
complete linkage methods, 118, 134
compound symmetry structure, 175
conditioning plots, 37–39
confirmatory factor analysis, 88
contingency tables, 104
continuous variables, 23–24
convex hull method, 22–23
correlation coefficients, 6–7, 22
correspondence analysis, 104–112
  applications of, 109–112
  multidimensional scaling and, 91–114
  principal components analysis and, 104

covariances, 3, 5–6, 43–46, 69, 143.
See specific methods
cross-validation estimate, 153

data mining, 13
data types, 1–4
dendrograms, 117, 121
density estimation, 29. *See also* scatterplots
depression, study of, 164–167, 191
discriminant function analysis, 142–156
assessment of, 146
Fisher linear discriminant, 142–146, 151
linear discriminant functions, 142–146, 151
plug-in estimates, 146
dissimilarity, 91, 118–123
distances, measures of
chi-squared distance, 104
covariance matrix and, 69
defined, 7–9
Euclidean, 91, 94
generalized distance, 10
non-Euclidean, 98
dropouts
classification of, 196
in longitudinal data, 190–198
drug usage, studies of, 82–85

economic theory, 41
Egyptian skull data, 137, 151
eigenvalues, 43–44, 46, 95
EM algorithm, 130
equal mean vectors, hypothesis of, 147
error sum-of-squares criterion (ESS), 136
Euclidean distances, 7, 91, 94, 105, 108
evolutionary trees, 117
exoplanet analysis, 130
expected value, defined, 4

F-statistics, 148
factor analysis
common factors, 65–68
communalities and, 68
criticism of, 72, 88
degrees of freedom, 70
deriving factor scores, 76
examples of, 70–76
factor loadings, 66
factor rotation, 71–76
interpretation and, 73
iterative approach, 69
k-factor model, 67, 69
latent variables, 65
manifest variables, 65, 68
mathematics of, 66
maximum likelihood approach, 69
numbers of factors, 69
oblimin rotation, 75
oblique rotation, 74
orthogonal rotation, 74
principal components analysis, 68, 85–88
promax rotation, 75
quartimax rotation, 75
rotational indeterminacy, 76
types of rotation, 74
varimax rotation, 75
fence, elliptical, 25–26
finite mixture densities, 134
Fisher linear discriminant function, 142–146, 151

graphical procedures, 16–40
bivariate boxplots, 25–29
chiplots, 23–25
clusters and, 29–32
density estimates, 29
outliers. *See* outliers
scatterplots. *See* scatterplots
group average clustering, 118–119
grouped multivariate data, 137–156
Fisher linear discriminant, 142–146
Hotelling T^2 test, 137–142
MANOVA and, 147–149
t tests, 137, 139, 141
two-group analysis, 137–146, 149

head measurements, 162–164
Heywood case, 69
hierarchical classification, 115.
See agglomerative hierarchial clustering
hinge, elliptical, 25–26
histograms, 16
Hodgkin's disease, 111–112
horseshoe effect, 112
Hottelling T^2 test, 137–142

imputation, 3–4
independence, 1, 24–25, 174.
  See specific tests
independent components analysis, 61
intelligence, estimate of, 70–71
interpretation, 53, 55, 73
intraclass correlation, 175

jittering, 19–20
Jolliffe rule, 47

k-means clustering, 123–128
Kaiser rule, 47

labelling, of components, 53
Lagrange multipliers, 43
latent variables, 65, 197
Lawley-Hotelling trace, 148
leave-one-out method, 146, 153
life expectancies, 77
linear discriminant function, 142–146
linear mixed effects models, 174–190
linear regression, 19–21
local independence assumption, 174
locally weighted regression, 21
log-eigenvalue diagram, 47
logistic discrimination, 146
long form, 176
longitudinal data, 171, 190–198
low-dimensional representation, 97
lowest fit, 21

Mahalanobis distances, 102
manifest variables, 65
MANOVA technique, 147–149
map, 91. *See also* scatterplot
maximum likelihood approach, 69, 70, 77
MDS. *See* multidimensional scaling
means, defined, 4
medical studies, 134
minimum volume ellipsoid, 56
missing-data problem, 3
model-based clustering, 128–134
modelling process, 15
Monte Carlo method, 4
multicenter design, 198
multidimensional scaling (MDS), 93–104, 94
  classical, 96–104
  examples of, 96–104
  horseshoe effect, 112
  mathematical methods, 94
  principal components and, 113
multiple correlation coefficient, 159
multiple regression, 157–160
multivariate analysis, 9
  aims of, 13–15
  complete cases, 3
  example of, 2
  F-statistics, 148
  graphical procedures, 16–40
  grouped multivariate data, 137–156
  MANOVA technique, 147–149
  missing-data problem, 3
  multivariable methods, 157
  multivariate data, 1–15
  normal distribution, 9–13
  normality and, 9–13, 149
  summary statistics for, 4–9
  test statistics for, 148
  See also specific topics

noise, 1, 19
non-linear models, 158
non-normal distributions, 146
normal distributions, 9–13
normal probability plot, 10
normality, assumption of, 76
normalization constraint, 49

oblimin rotation, 75
ordinal data, 2
outliers, 29, 51
  correlation coefficient and, 22
  fence ellipse, 25
  hinge ellipse, 25
  minimum volume ellipsoid, 56
  robust estimators and, 25–26
  scatterplots and, 22

Pearson coefficient, 7
Pillai trace, 148
plug-in estimates, 146
point density, 29
polynomial fitting, 21
pottery, analysis of, 123–124

principal components analysis, 41–64, 97
  air pollution studies, 49–61
  algebraic basics of, 42–49
  application of, 41, 48–49
  basic aim of, 41
  of bivariate data, 48–49
  calculating, 47–48
  correspondence analysis, 104
  factor analysis, 68, 85–88
  introduction of, 61
  Lagrange multipliers, 43
  modern competitors to, 61
  multidimensional scaling, 113
  number of components, 46–47
  rescaling, 45–46
  rotation techniques, 76
probability plotting, 9–10
probability-probability plot, 9–10
projection pursuit method, 61
promax rotation, 75
proximity matrix, 91

quadratic discrimination function, 146
quadratic maximization, 75
quantile-quantile plot, 9–10
quartimax rotation, 75–77

random effects, 189
random intercept model, 174–175
random scatter, 24
ranking, 41
regularized discriminant analysis, 146
reification, 53
repeated measures data, 171–199
restricted maximum likelihood, 176
robust estimation, 22, 25–26
rotation techniques, 75
Roy greatest root, 148
R statistical software
  bivbox function, 27
  bivden function, 30
  bkde2D function, 32, 40
  chiplot function, 23–25, 149
  coplot function, 37
  cor function, 7
  cov.mve function, 60
  data frames in, 208–210
  discrim function, 151
  dist function, 7, 96
  elementary commands, 200
  graphical procedures, 16
  KernSmooth library, 32, 40
  lattice library, 36
  lda function, 144
  list objects in, 207
  lme function, 176
  logical expressions, 205–207
  lqs library, 60
  MASS library, 144
  matrices in, 203–205
  mvrnorm function, 10
  pairs function, 33
  predict function, 146
  random.effects function, 187
  rep function, 202
  resid function, 187
  strings in, 201
  vectors in, 201
  wireframe function, 40
  See also specific functions, topics

Sagan, Carl, 16
scatterplots, 91
  chi-plots and, 24
  continuous variables and, 23
  convex hull, 22–23
  density estimates and, 29–30
  extra variables, 32–33
  jittering and, 20
  lowest fit, 21
  maps and, 91–92
  matrix for, 17–29, 33–35
  outliers, 22
  polynomial fitting and, 21
  representing variables, 32–33
  xy scatterplots, 17–29
scientific method, 14–15
Scott-Simon approach, 128
scree diagram, 47, 69
similarity measure, 91
simple structure, 73
single linkage, 118, 134
skulls, study of, 99, 137, 146
smoking, studies of, 109–111
smooth functions, 21–22
specific variance, 66–67
S-PLUS statistical software
  apply function, 4, 5

bivbox function, 27
bivden function, 30
chiplot function, 25, 149
click-and-point features, 33
cloud function, 36
coplot function, 37
cor function, 7
cov.mve function, 60
data frames in, 208–210
discrim function in, 144
dist function, 7, 96
elementary commands, 200
graphical procedures, 16
group effect in, 185
lda function, 144, 151
list objects in, 207
lme function, 176
logical expressions, 205–207
MASS library, 144
matrices in, 203–205
mean function, 4
pairs function, 33
predict function, 146
random.effects function, 187
rep function, 202
resid function, 187
rmvnorm function, 10
strings in, 201
var function, 5
vectors in, 201
See also specific functions, topics
stem-and-leaf plots, 16
structural relationships, 44
Student t-test, 137, 139
sum of squares constraint, 43

t tests, 137, 139, 141. *See also* Hotelling T^2 test
three-dimensional plots, 35–36
trace, minimization of, 130
trellis graphics, 37–39
two-group analysis, 137–146, 148, 149

unique variance, 67

variance, analysis of, 5, 66–67, 137–156. *See specific methods, tests*
variance component model, 175
variance-covariance matrix, 6
varimax rotation, 75–77

Wilks determinantal ratio, 148