Chapter

10 *Interpretation of ecological structures*

10.0 Ecological structures

The previous chapters explained how to use the techniques of clustering and ordination to investigate relationships among objects or descriptors. What do these analyses contribute to the understanding of ecological phenomena? Ecological applications in Chapters 8 and 9 have shown how clustering and ordination can synthesize the variability of the data and present it in a format that is easily amenable to interpretation. It often happens, however, that researchers who are using these relatively sophisticated methods do not go beyond the description of the structures of multidimensional data matrices, in terms of clusters or gradients. The descriptive phase must be followed by interpretation, which is conducted using either the descriptors that were used to evidence the structure, or other ecological descriptors that have not yet been involved in the analysis.

Structure

From the previous chapters, it should be clear that the *structure* of a data matrix is the organization of the objects, or descriptors, along gradients in a continuum, or in the form of subsets (clusters). This organization characterizes the data matrix, and it is derived from it. The first phase of multidimensional analysis (i.e. clustering or/and ordination) thus consists in characterizing the data matrix in terms of a simplified structure. In a second phase, ecologists may use this structure to interpret the phenomenon that underlies the data matrix. To do so, analyses are conducted to quantify the relationships between the structure of the data matrix and potentially explanatory descriptors. The methods that are most often used for interpreting ecological structures are described in the present chapter and in Chapter 11.

During the interpretation phase, one must assume that the analysis of the structure has been conducted with care, using measures of association that were appropriate to the objects and/or descriptors of the data matrix (Chapter 7) as well as analytical methods that corresponded to the objectives of the study. Ordination (Chapter 9) is used when gradients are sought, and clustering (Chapter 8) when one is looking for a partition of the objects or descriptors into subsets. When the gradient is a function of a single or a pair of ordered descriptors, the ordination may be plotted in the original space of the descriptors. When the gradient results from the combined action of several descriptors, the ordination must be carried out in a reduced space using the methods discussed in Chapter 9. It may also happen that an ordination is used as a basis for visual clustering. Section 10.1 discusses the combined use of clustering and ordination to optimize the partition of objects or descriptors.

Explanation

Forecasting

Prediction

The interpretation of structures, in ecology, has three main objectives: (1) *explanation* (often called *discrimination*) of the structure of one or several descriptors, using the descriptors at the origin of the structure or, alternatively, a set of other descriptors that may potentially explain the structure; (2) *forecasting* of one or several descriptors (which are the response, or dependent, variables: Box 1.1), using a number of other descriptors (called the explanatory, or independent, variables); (3) *prediction* of one or several descriptors, using descriptors that can be manipulated experimentally or naturally exhibit environmental variation. The terms *forecasting* and *prediction*, which are not equivalent (Subsection 10.2.2), are often confused in the ecological and statistical literatures. Each of the above objectives covers a large number of numerical methods, which correspond to various levels of precision of the descriptors involved in the analysis.

Section 10.2 reviews the methods available for interpretation. The next sections are devoted to some of the methods introduced in Section 10.2. Regression and other scatterplot smoothing methods are discussed in Section 10.3. Section 10.4 deals with path analysis, which is used to assess causal relationships among quantitative descriptors. Section 10.5 discusses some methods developed to test the relationship between association or data matrices.

10.1 Clustering and ordination

Section 8.2 showed that single linkage clustering accurately accounted for the relationships between highly similar objects. However, due to its tendency to chaining, single linkage agglomeration is not very suitable for investigation of ecological questions. Because ecological data generally form a continuum in A-space (Fig. 7.2), it is often informative to use single linkage clustering in conjunction with an ordination of the objects. In the full multidimensional ordination space of principal component analysis (Section 9.1), Euclidean distances among the main clusters of objects are the same as in the original A-space. Other ordination methods (Sections 9.2 to 9.4) may be more appropriate in other cases. However, when only the first two or three dimensions are considered, ordinations in reduced space may misrepresent the structure by projecting together clusters of objects that are distinct in higher dimensions. Clustering methods allow one to separate clusters whose projections in reduced space may sometimes obscure the relationships between them.

Several authors (e.g. Gower & Ross, 1969; Rohlf, 1970; Schnell, 1970; Jackson & Crovello, 1971; Legendre, 1976) have independently proposed to take advantage of the characteristics of clustering and ordination by combining the results of the two types of analyses on the same diagram. The same similarity or distance matrix (Tables 7.4 to 7.6) is often used for the ordination and cluster analyses. Any clustering method may be used, as long as it is appropriate to the data. If linkage clustering is chosen, it is easy to draw the links between objects onto the ordination diagram, up to a given level of similarity. One may also identify the various similarity levels by using different colours or streaks (for example: solid line for $1.0 \geq S > 0.8$, dashed for $0.8 \geq S > 0.6$, dotted for $0.6 \geq S > 0.4$, etc., or any other convenient combination of codes or levels). If a divisive method or centroid clustering was used, a polygon or envelope may be drawn, on the ordination diagram, around the members of each cluster. This is consistent with the opinion of Sneath & Sokal (1973), who suggested to always simultaneously carry out clustering and ordination on a set of objects. Field *et al*. (1982) expressed the same opinion about marine ecological data. It is therefore recommended, as a routine procedure in ecology, to represent clustering results onto ordination diagrams.

The same approach can be applied to cluster analyses of descriptors. Clustering may be conducted on a dependence matrix among descriptors — especially species (Subsection 8.9.2) — in the same way as for an association matrix among objects. An ordination of species (e.g. Figs. 8.19 and 8.20) or other descriptors can be obtained using one of the ordination methods described in Chapter 9, depending on the measure of dependence among descriptors that is appropriate for the data under study. With quantitative physical or chemical descriptors of the environment, the method of choice is principal component analysis of the correlation matrix (Section 9.1); descriptors are represented by arrows in the ordination diagram. In some cases, before clustering, negative correlations among descriptors can be made positive because they are indicative of resemblance on an inverted scale.

When superimposed onto an ordination, single linkage clustering becomes a most interesting procedure for ecological interpretation. Single linkage clustering is the best complement to an ordination due to its contraction of the clustering space (Table 8.9, Fig. 8.24). Drawing single linkage results onto an ordination diagram provides both the correct positions for the main clusters of objects (from the ordination) and the fine relationships between closely similar objects (from the clustering). It is advisable to only draw the chain of primary connections (Section 8.2) on the ordination diagram because it reflects the changes in the composition of clusters. Otherwise, the groups of highly similar objects may become lost in the multitude of links drawn on the ordination. Ecological application 10.1 provides an example of this procedure.

Jackson & Crovello (1971) suggested to indicate the directions of the links on the ordination diagram (Fig. 10.1). This information may be useful when delineating clusters. In such diagrams, each link of the primary chain is drawn with an arrow. On a link from $\mathbf{x}_1$ to $\mathbf{x}_2$, an arrow pointing towards $\mathbf{x}_2$ indicates that object $\mathbf{x}_1$ has $\mathbf{x}_2$ as its closest neighbour in multidimensional A-space (i.e. in the association matrix among

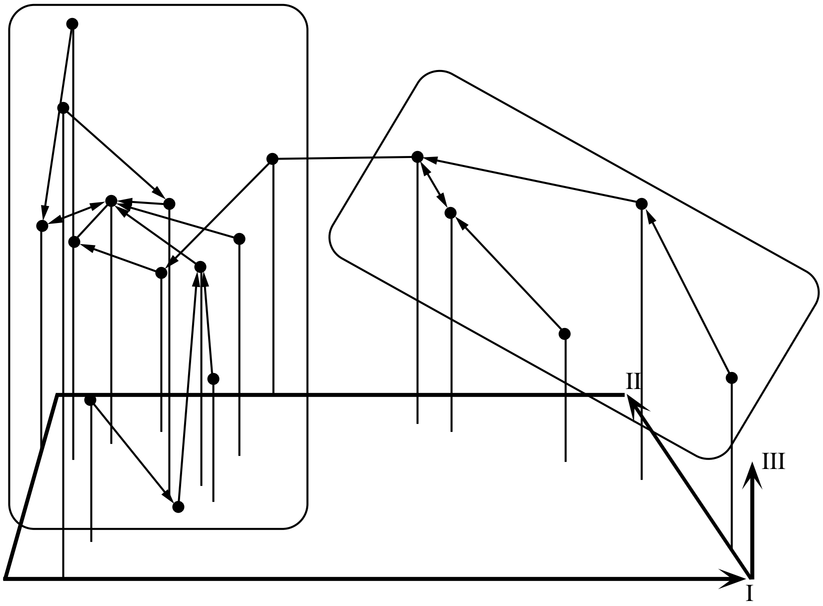


Figure 10.1 Three-dimensional ordination of objects (dots), structured by the primary connections of a single linkage clustering. The arrows (excluding those of the principal axes I to III) specify the directions of the relationships between nearest neighbours; see text. Modified from Jackson & Crovello (1971). Cutting the link without arrows determines two clusters (boxed points).

objects). When $\mathbf{x}_2$ also has $\mathbf{x}_1$ as its closest neighbour, the arrow goes both ways. When $\mathbf{x}_2$ has $\mathbf{x}_3$ as its closest neighbour, the arrow from $\mathbf{x}_2$ points towards $\mathbf{x}_3$. New links formed between objects that are already members of clusters do not receive arrows. These links may be removed to separate the clusters.

Ecological application 10.1

Single linkage clustering was illustrated by Ecological application 8.2 taken from a study of a group of ponds, based upon zooplankton. The same example (Legendre & Chodorowski, 1977) is used again here. Twenty ponds were sampled on islands of the St. Lawrence River, east and south of Montréal (Québec). Similarity coefficient S_{20} (eq. 7.27) was computed with $k = 2$. The matrix of similarities among ponds was used to compute both single linkage clustering and an ordination in reduced space by principal coordinate analysis. In Fig. 10.2, the chain of primary connections is superimposed onto the ordination, in order to evidence the clustering structure. The ponds are divided between a cluster of periodic ponds, which are dry during part of the year (encircled), and a cluster of permanent ponds. Ponds with identification numbers beginning with the same digit (which indicates the region) tend to be close to one another and to cluster first with one another. The second digit refers to the island on which a pond was located.

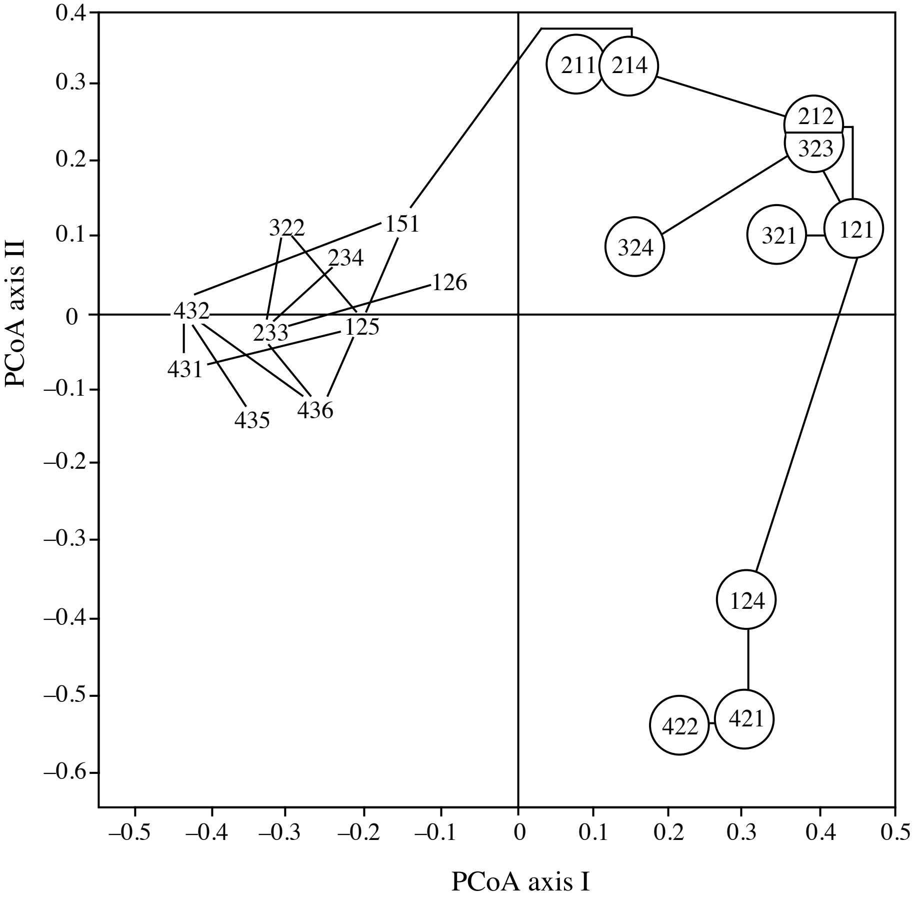


Figure 10.2 Comparison of 20 ponds on the basis of their zooplankton fauna. Ordination in a space of principal coordinates (principal axes I and II), and superimposition of the chain of primary connections obtained by single linkage clustering. The encircled ponds are periodic; the others are permanent. Adapted from Legendre & Chodorowski (l977).

When no clear clustering structure is present in the data but groups are still needed, for management purpose for instance, arbitrary groups may be delineated by drawing a regular grid on the reduced-space ordination diagram. This grid may be orthogonal (i.e. square or rectangular) or polar (division into triangles from the point of origin of the graph coordinates). Another method is to divide the objects according to the quadrants of the ordination in reduced space (in 2^d groups for a *d*-dimensional space); the result is the hierarchic classification scheme of Lefkovitch (1976) described in Subsection 8.7.3.

Figure 10.3 summarizes the steps involved in producing a cluster analysis and an ordination from a resemblance matrix. Description of the data structure is clearer when the clustering results are drawn onto the ordination. In order to assess to what extent the clustering and the ordination correspond to the resemblance matrix from which they originate, these representations may be compared to the original resemblance matrix using matrix correlation or related methods (Subsection 8.12.2).

10.2 The mathematics of ecological interpretation

The present section summarizes the numerical methods available for the interpretation of ecological structures. The most widely used of these techniques (regression, path analysis, matrix comparison, the fourth-corner method, and canonical analysis) are discussed in Sections 10.3 to 10.6 and in Chapter 11. A few other methods are briefly described in the present section.

The numerical methods presented in this section are grouped into three subsections, which correspond to the three main objectives of ecological interpretation, set in Section 10.0: explanation, forecasting, and prediction. For each of these objectives, there is a summary table (Tables 10.1 to 10.3) intended to facilitate the choice of methods best suited to the researchers' ecological objectives and the nature of their data.

Ecological interpretation, and especially the *explanation* and *forecasting* of the structure of several descriptors (i.e. multivariate data), may be conducted following two approaches, which are the indirect and direct comparison schemes (Fig. 10.4).

Indirect comparison

Indirect comparison proceeds in two steps. The structure (ordination axes, or clusters) is first identified from a set of descriptors (response data) of prime interest in the study. In a second step, the structure is interpreted using either (a) the descriptors that were analysed in the first step to identify the structure, or (b) another set of descriptors that may help explain the structure. In his chapter on ordination analysis, ter Braak (1987c) referred to this form of analysis as *indirect gradient analysis* because he was mostly concerned with the study of environmental gradients.

Direct comparison

In *direct comparison*, one simultaneously analyses the response and explanatory data matrices in order to identify how they are related. Direct comparison is done by the asymmetric methods of canonical analysis (Sections 11.1 and 11.2), which allow one to bring out the ordination structure of a response data set that is explained by another data set; ter Braak (1987c) refers to this approach as *direct gradient analysis*.

Other forms of direct comparison analysis are available. One can compare similarity or distance matrices, derived from the original data matrices, using the techniques of matrix comparison (Section 10.5); this type of comparison should, however, be restricted to test hypotheses that concern similarities or distances, not raw data (Subsection 10.5.1).

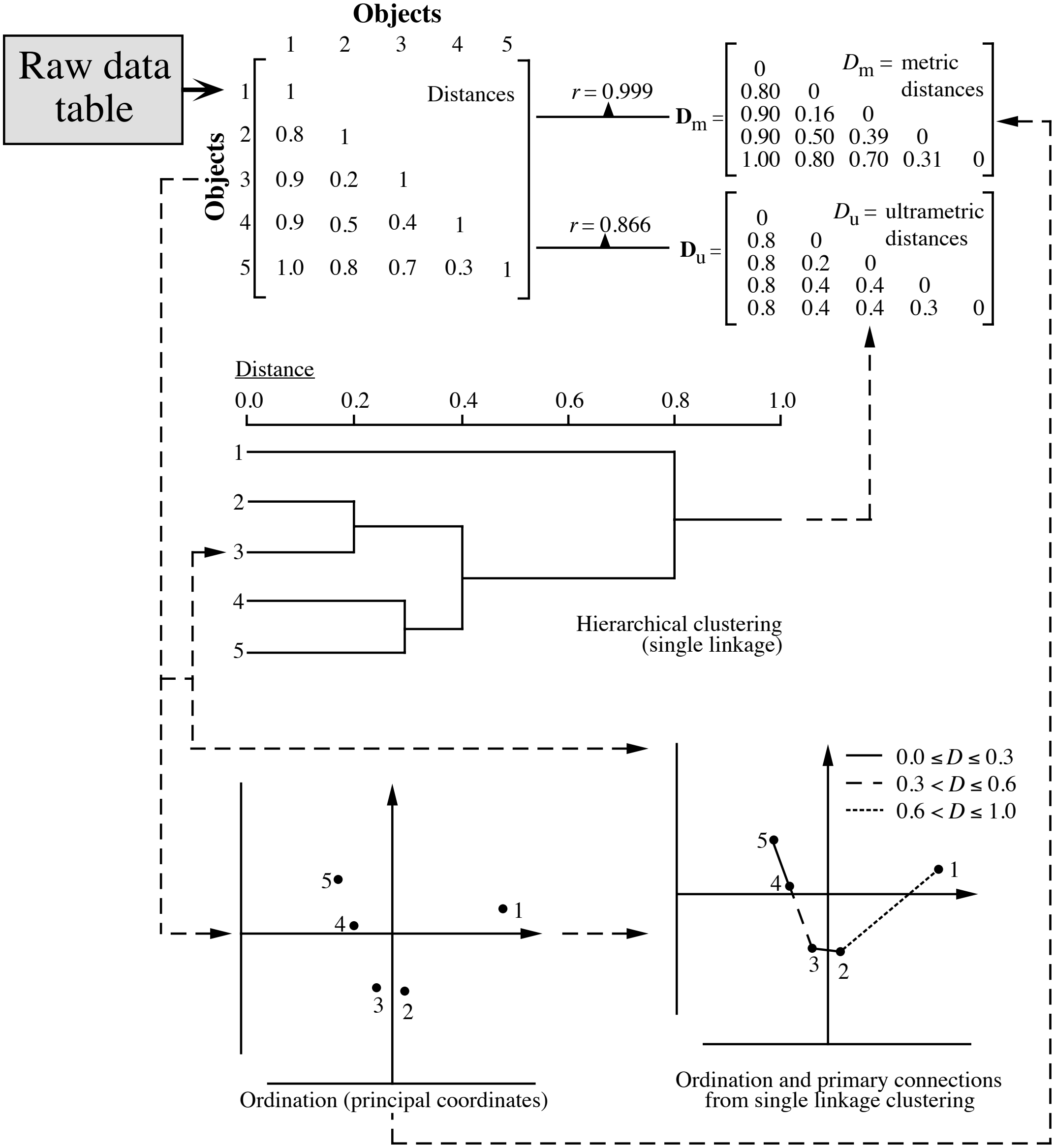


Figure 10.3 Identification of the structure of five objects, using clustering and ordination. Bottom right: the chain of primary connections is superimposed on a 2-dimensional ordination, as in Figs. 10.1 and 10.2. Top: the reduced-space ordination and the clustering results are compared to the resemblance matrix from which they originate. Upper right (top): a matrix of metric distances $\mathbf{D}_m$ is computed from the reduced-space ordination, and compared to the original distances using matrix correlation; $r = 0.999$ is a rather high score. Upper right (below): a cophenetic $\mathbf{D}_u$ matrix (Section 8.3) is computed from the dendrogram, and compared to the original distances using matrix correlation ($r = 0.866$).

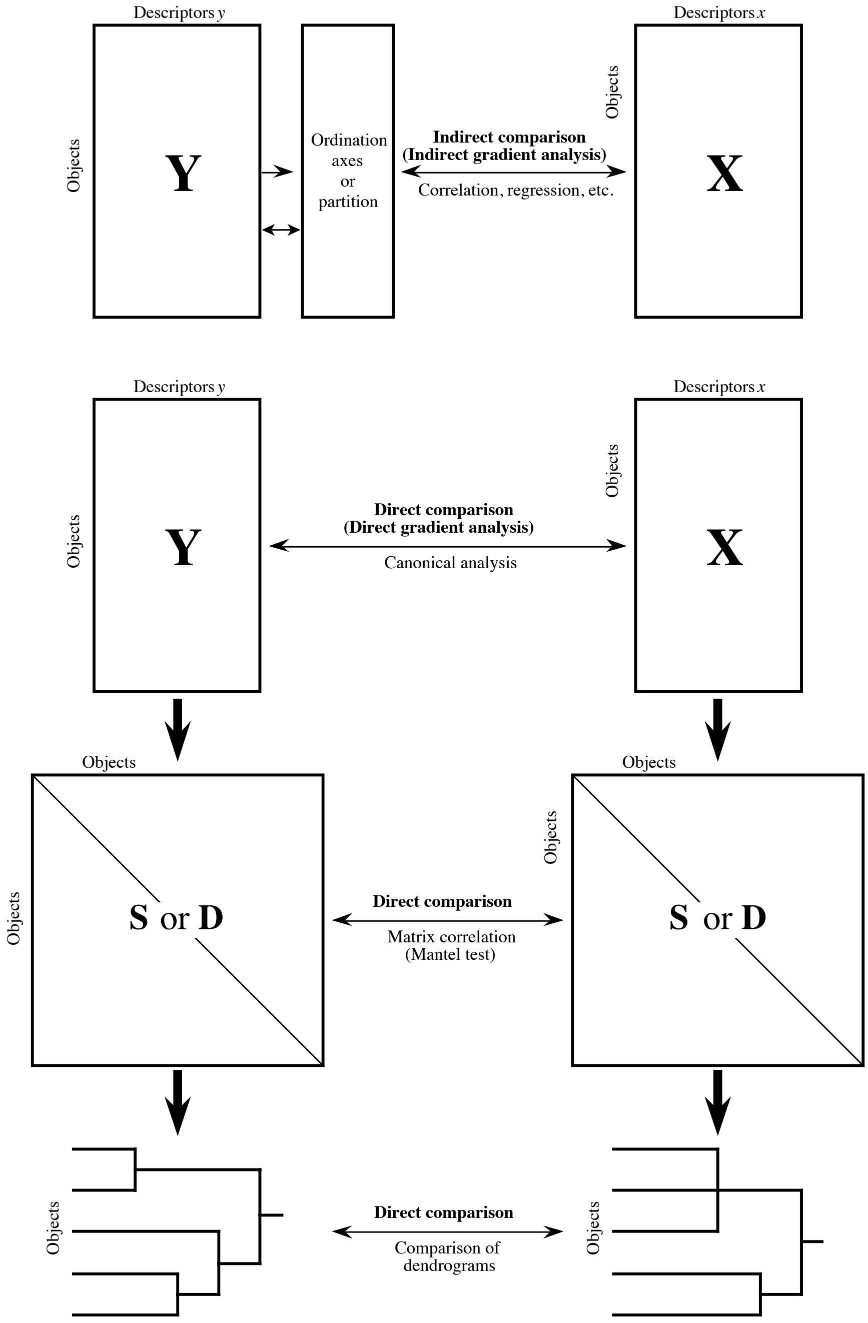


Figure 10.4 Indirect and direct comparison approaches for analysing and interpreting the structure of ecological data. Single thin arrow: inference of structure. Double arrow: interpretation strategy.

Consensus index

Permutation test

One can also compare dendrograms derived from resemblance matrices, using *consensus indices*; this approach should be restricted to test hypotheses that concern dendrograms. Two main approaches have been developed to test the significance of consensus statistics: (1) a probability distribution derived for a given consensus statistic may be used, or (2) a specific test may be carried out to assess the significance of the consensus statistic, in which the reference distribution is found by permuting the two dendrograms under study in an appropriate way (Lapointe & Legendre, 1995). Readers are referred to the papers of Day (1983, 1986), Shao & Rohlf (1983), Shao & Sokal (1986), Lapointe & Legendre (1990, 1991, 1992a, 1992b, 1995), and Steel & Penny (1993), where these methods are described. Lapointe & Legendre (1994) used the three forms of direct comparison analysis (i.e. comparison of raw data, distance matrices, and dendrograms; Fig. 10.4) on five data sets describing the same objects. In that study, all methods essentially led to similar conclusions, with minor differences.

The interpretation of a structure using the descriptors from which it originates makes it possible to identify which of these descriptors mainly account for the structuring of the objects. In some ordination methods (i.e. principal component and correspondence analysis), the eigenvectors readily identify the important descriptors. Other types of ordination, or the clustering techniques, do not directly provide this information, which must be found *a posteriori* using methods of indirect comparison. This type of interpretation does not allow one to perform formal tests of significance. The reason is that the structure under study is derived from the very same descriptors that are now used to interpret it; it is thus not independent of them.

Interpretation of a structure using external information (data matrix **X** in Fig. 10.4) is central to numerical ecology. This approach is used, for example, to diagnose abiotic conditions (response data matrix **Y**) from the available biological descriptors (explanatory data matrix **X**) or, alternatively, to forecast the responses of species assemblages (matrix **Y**) using available environmental descriptors (matrix **X**). In the same way, it is possible to compare two groups of biological descriptors or two matrices of environmental data. Until the mid-1980's, the indirect comparison scheme was favoured because of methodological problems with canonical correlation analysis, which was then the only method available in computer packages to analyse two sets of descriptors. When new methods and computer programs (including R functions) were made available, direct comparison became widely used in the ecological literature.

In the indirect comparison approach, the first set of descriptors is reduced to a single or a few one-dimensional variables, i.e. a partition resulting from clustering, or one or several ordination axes, the latter being generally interpreted one at a time. It follows that the methods of interpretation for univariate descriptors (e.g. correlation, regression) can also be used for indirect comparisons. This is the approach used in Tables 10.1 and 10.2.

1 — Explaining ecological structures

Table 10.1 summarizes the methods available to *explain* the patterns found in one or several ecological descriptors. *Explaining* is taken here in the sense of looking for correlations and using them to formulate hypotheses. The purpose is data exploration, not hypothesis testing. The first dichotomy of the table separates methods for univariate descriptors (used also in the indirect comparison approach) from those for multivariate data.

Methods used for explaining the structure of *univariate descriptors* belong to three major groups: (1) measures of dependence, (2) discriminant analysis, (3) and methods for qualitative descriptors. Methods used for explaining the structure of *multivariate descriptors* belong to two major types: (4) asymmetric canonical analysis using a response and an explanatory matrix, and (5) symmetric canonical analysis comparing two interchangeable data matrices. (6) Supplementary data associated with the sites and the species of a community composition data matrix can be related in fourth-corner analysis. The following paragraphs briefly review these groups of methods.

1. Various coefficients have been described in Chapters 4 and 5 to measure the *dependence* between two descriptors exhibiting linear or monotonic relationships (i.e. the parametric and nonparametric *correlation coefficients*). When there are more than two descriptors, one may use the *coefficients of partial correlation* (Section 4.5) or the *coefficient of concordance* (Section 5.4). The *coefficient of multiple determination* (R^2), computed in *multiple linear regression*, may be used to assess the dependence of a quantitative response descriptor on an explanatory matrix containing quantitative or mixed-level descriptors. *Dummy variable regression* is a special case of multiple regression where the explanatory matrix contains qualitative descriptors recoded into dummy variables, as explained in Subsection 1.5.7. These different types of regression are briefly discussed in Subsection 10.2.2, in relation with Table 10.2, and in more detail in Section 10.3.

2. Explaining the structure of a qualitative descriptor is often called *discrimination*, when the aim of the analysis is to identify explanatory descriptors that would allow one to discriminate among the various states of the qualitative descriptor. *Linear discriminant analysis* may be used when (1) the explanatory (or *discriminant*) descriptors are quantitative, (2) the distributions of the within-group residuals are not too far from normal, and (3) the within-group dispersion matrices are reasonably homogeneous. Linear discriminant analysis (LDA) is described in Section 11.3. Its use with species data is discussed in Section 11.6, where alternative strategies are proposed.

3. When both the descriptor to be explained and the explanatory descriptors are qualitative, one may use *multidimensional contingency table analysis*. It is then imperative to follow the rules, given in Section 6.3, concerning the models to use when a distinction is made between the explained and explanatory descriptors. When the response variable is binary, *logistic regression* is a better choice than multidimensional

Table 10.1 Numerical methods to *explain* the structure of descriptors, using either the descriptors from which the structure originates, or other, potentially explanatory descriptors. In parentheses, identification of the section where a method is discussed. Tests of significance cannot be performed when the structure of a descriptor is explained by the descriptors at the origin of that structure.

1) Explanation of the structure of a *single* descriptor, or *indirect comparison* .... **see 2**

  2) Structure of a quantitative or a semiquantitative descriptor .... **see 3**

    3) Explanatory descriptors are quantitative or semiquantitative............... **see 4**

      4) To *measure* the dependence between descriptors..................... **see 5**

        5) Pairs of descriptors: *Pearson r*, for quantitative descriptors exhibiting linear relationships (4.2); *Kendall* τ or *Spearman r*, for quantitative or semiquantitative descriptors exhibiting monotonic relationships (5.3)

        5) A single quantitative descriptor as a function of several others: *coefficient of determination* R^2 of *multiple regression* (4.5, 10.3.3)

        5) Several descriptors exhibiting monotonic relationships: *coefficient of concordance W* (5.4)

      4) To *interpret* the structure of a single descriptor: *partial Pearson r*, for quantitative descriptors exhibiting linear relationships (4.5); *partial Kendall* τ, for descriptors exhibiting monotonic relationships (5.3)

    3) Explanatory descriptors are qualitative: R^2 of *dummy variable regression* (10.3)

    3) Estimation of the dependence between descriptors of the sites and descriptors of the species (any precision level): *the fourth-corner method* (10.6)

  2) Structure of a qualitative descriptor (*or* of a classification) .... **see 6**

    6) Explanatory descriptors are quantitative: *linear discriminant analysis* (LDA, 11.3)

    6) Explanatory descriptors are qualitative: *multidimensional contingency table analysis* (6.3); *discrete discriminant analysis* (10.2)

    6) Explanatory descriptors are of mixed precision: *logistic regression* (in most cases, the explained descriptor is binary; 10.3)

1) Explanation of the structure of a *multivariate* data matrix .... **see 7**

  7) *Direct comparison*.. **see 8**

    8) Asymmetric analysis of a response matrix by an explanatory matrix: *redundancy analysis* (RDA, 11.1); *canonical correspondence analysis* (CCA, 11.2); multivariate regression tree analysis (MRT, 8.11). Basic statistic in RDA: *canonical* R^2

    8) Symmetric comparison of two data matrices: *canonical correlation analysis* (CCorA, 11.4), *co-inertia analysis* (CoIA, 11.5.1), *Procrustes analysis* (Proc, 11.5.2). Statistics: *RV* (11.5.1), Trace**W** and $m_{12}{}^2$ (10.5.4, 11.5.2)

    8) Compare classifications computed from two data matrices: *contingency table analysis* (6.2), *modified Rand index* (8.12)

  7) *Indirect comparison* .... **see 10**

    10) Ordination in reduced space: each axis is treated in the same way as a single quantitative descriptor .... **see 2**

    10) Clustering: each partition is treated as a qualitative descriptor .... **see 2**

contingency table analysis. An additional advantage is that logistic regression allows one to use explanatory variables presenting a mixture of precision levels. For qualitative variables, the equivalent of discriminant analysis is called *discrete discriminant analysis*. Goldstein & Dillon (1978) describe this form of analysis.

4. The standard approach for comparing two sets of descriptors is *canonical analysis* (Chapter 11). In ecology, the asymmetric forms of canonical analysis, where the two data matrices do not play the same role, are the most widely used. Asymmetric analyses involve a response matrix **Y** and an explanatory matrix **X**. The methods are called *redundancy analysis* (RDA, Section 11.1) and *canonical correspondence analysis* (CCA, Section 11.2). The difference between these two methods is the same as between principal component and correspondence analyses (Table 9.1). An alternative method of asymmetric analysis is multivariate regression tree analysis (MRT, Section 8.11), which looks for cutting points in the explanatory descriptors **X** that create compact groups in the response data **Y**.

5. It is also possible to compare two matrices that play the same role and can be interchanged in the analysis. These symmetric analyses are carried out by *canonical correlation analysis* (CCorA, 11.4), *co-inertia analysis* (CoIA, 11.5.1), and *Procrustes analysis* (Proc, 11.5.2).

6. Consider a (site × species) matrix containing community composition data (presence-absence or abundance), for which supplementary variables are known for the sites (e.g. habitat characteristics, spatial data) and for the species (e.g. biological or behavioural traits). The *fourth-corner method*, described in Section 10.6, offers a way of estimating the dependence between the supplementary variables of the rows and those of the columns and testing the resulting correlation-like statistics for significance.

2 — Forecasting ecological structures

Forecasting model

Predictive model

A distinction has to be made between *forecasting* and *prediction* in ecology. Forecasting models extend, into the future or to different situations, structural relationships among descriptors that have been quantified for a given data set. A set of relationships among variables, which simply describe the changes in one or several descriptors in response to changes in others as computed from a "training set", make up a *forecasting* model. In contrast, when the relationships are considered causal and to describe a mechanistic process, the model is *predictive*. A condition to successful forecasting is that the values of all important variables that have not been observed (or controlled, in the case of an experiment) be about the same in the new situation as they were during the survey or experiment. In addition, forecasting does not allow extrapolation beyond the observed range of the explanatory variables. *Forecasting models* (also called *correlative models*) are frequently used in ecology, where they are sometimes misleadingly called "predictive models". Forecasting models are useful provided that the above conditions are fulfilled. In contrast, predictive models describe known or assumed causal relationships. They allow one to estimate the effects, on

Table 10.2 Numerical methods to *forecast* one or several descriptors (response or dependent variables) using other descriptors (explanatory or independent variables). In parentheses, identification of the section where a method is discussed.

1) Forecasting the structure of a *single* descriptor, or *indirect comparison* .... **see 2**
    2) The response variable is quantitative .... **see 3**
        3) The explanatory variables are quantitative............................ **see 4**
            4) Null or low correlations among explanatory variables: *multiple linear regression* (10.3); *nonlinear regression* (10.3)
            4) High correlations among explanatory variables (collinearity): *ridge regression* (10.3); *regression on principal components* (10.3)
        3) The explanatory variables are qualitative: *dummy variable regression* (10.3)
    2) The response variable is qualitative (*or* a classification) .... **see 5**
        5) Response: two or more groups; explanatory variables are quantitative (but qualitative variables may be recoded into dummy variables): *identification functions in discriminant analysis* (11.3)
        5) Response: binary (presence-absence); explanatory variables are quantitative (but qualitative variables may be recoded into dummy var.): *logistic regression* (10.3)
    2) The response and explanatory variables are quantitative, but they display a nonlinear relationship: *nonlinear regression* (10.3)
1) Forecasting the structure of a *multivariate* data matrix.......................... **see 6**
    6) *Direct comparison*... **see 7**
        7) Linear modelling: *redundancy analysis* (RDA, 11.1); *canonical correspondence analysis* (CCA, 11.2)
        7) Find a tree-like decision model: *multivariate regression tree analysis* (MRT, 8.11)
    6) *Indirect comparison* .... **see 8**
        8) Ordination in reduced space: each axis is treated in the same way as a single quantitative descriptor .... **see 2**
        8) Clustering: each partition is treated as a qualitative descriptor .... **see 2**

some variables, of changes in other variables; they will be briefly discussed at the beginning of the next subsection.

Methods in Table 10.2 are used to *forecast* descriptors. As in Table 10.1, the first dichotomy in the table distinguishes the methods that allow one to forecast values of a single descriptor (*response* or *dependent* variable) from those that may be used to simultaneously forecast several descriptors. Forecasting methods belong to four major groups: (1) regression models, (2) identification functions, (3) asymmetric canonical analysis methods, and (4) multivariate regression trees.

1. Methods belonging to *regression* models are numerous. Several regression methods include measures of dependence that have already been mentioned in the discussion of Table 10.1: *multiple linear regression* (the explanatory variables are quantitative or mixed), *dummy variable regression* (a special case of multiple regression where the explanatory matrix contains qualitative descriptors (e.g. ANOVA factors) recoded into dummy variables, as explained in Subsection 1.5.7), and *logistic regression* (the explanatory variables may be of mixed levels of precision; the response variable is qualitative). Section 10.3 provides a detailed description of several regression methods.

2. *Identification functions* are part of linear discriminant analysis (Section 11.3), which was briefly described in the previous subsection. These functions allow the assignment of any object to one of the states of a qualitative descriptor, using the values taken by several quantitative variables (i.e. the explanatory or discriminant variables). As mentioned in the previous subsection, the distributions of the discriminant variables must not be too far from normality, and their within-group dispersion matrices must be reasonably homogeneous (i.e. about the same in all groups).

3. Canonical analysis, and especially *redundancy analysis* and *canonical correspondence analysis*, which were briefly discussed in the previous subsection (and in more detail in Sections 11.1 and 11.2), allow one to model a data matrix from the descriptors of a second data matrix; these two data matrices form the "training set". Using the resulting model, it is possible to forecast the position of any new observation among those of the "training set", for example along environmental gradients. The new observation may represent some condition that may occur in the future, or at a different but comparable location.

4. An alternative forecasting method of analysis is multivariate regression tree analysis (MRT, Section 8.11). This method produces a decision tree in which the response data **Y** are divided into groups, whereas the bifurcations of the tree correspond to splits in the explanatory variables **X** that can be used for forecasting the positions of new observations.

3 — Ecological prediction

As explained in the Preface, predictive modelling does not belong to numerical ecology *sensu stricto*. However, some methods of numerical ecology may be used to analyse causal relationships among a small number of descriptors, thus linking numerical ecology to predictive modelling. Contrary to the *forecasting* or *correlative models* (previous subsection), *predictive models* allow one to foresee how some variables of interest would be affected by changes in other variables. Prediction is possible when the model is based on causal relationships among descriptors (i.e. not only correlative evidence). Causal relationships are stated as hypotheses (theory) for modelling; they may also be validated through experiments in the laboratory or in the field. In *manipulative experiments*, one observes the responses of some descriptors to

Predictive model

Experiment

Table 10.3 Numerical methods for analysing causal relationships among ecological descriptors, with the purpose of *predicting* one or several descriptors using other descriptors. In parentheses, identification of the section where the methods are discussed. In addition, forecasting methods (Table 10.2) may be used for prediction when there are reasons to believe that the relationships between the explanatory and response variables are of causal nature.

1) The causal relationships among descriptors are given by hypothesis . . . . **see 2**

    2) Quantitative descriptors; linear causal relationships: *causal modelling using correlations* (4.5); *path analysis* (10.4)

    2) Qualitative descriptors: *logit* and *log-linear models* (6.3)

1) Hidden variables (latent variables, factors) are assumed to cause the observed structure of the descriptors: *confirmatory factor analysis* (not discussed in this book)

user-determined changes in other descriptors, by reference to a *control*. Besides manipulative experiments, which involve two or more treatments, Hurlbert (1984) recognizes *mensurative experiments*, which involve measurements made at one or more points in space or time and allow one to test hypotheses about patterns in space (Chapters 13 and 14) and/or time (Chapter 12). The numerical methods in Table 10.3 allow one to explore a network of causal hypotheses, using the observed relationships among descriptors. The design of experiments and analysis of experimental results are discussed by Mead (1988) who offers a statistically-oriented presentation, and by Underwood (1997) in a book emphasizing ecological experiments.

One may hypothesize that there exist causal relationships among the observed descriptors or, alternatively, that the observed descriptors are caused by underlying hidden variables. Depending on the hypothesis, the methods for analysing causal relationships are not the same (Table 10.3). Methods appropriate to the first case belong to the family of *path analysis* (Section 10.4). The second case leads to *confirmatory factor analysis*, which is not discussed in this book; see e.g. Brown (2006) or Harrington (2009) on this subject. The present chapter only discusses the former. In addition to these methods, techniques of forecasting (Table 10.2) may be used for predictive purposes when there are reasons to believe that the relationships between explanatory and response variables are of causal nature.

Fundamentals of *path analysis* are presented in Section 10.4. Path analysis is an extension of multiple linear regression and is thus limited to quantitative or binary descriptors (including qualitative descriptors recoded as dummy variables: Subsection 1.5.7). In summary, path analysis is used to decompose and interpret the relationships among a small number of descriptors, assuming that (a) there is a *(weak) causal order* among descriptors, and (b) the relationships among descriptors are *causally closed*. *Causal order* means, for example, that y_2 possibly (but not necessarily) affects y_3 but that, under no circumstance, y_3 would affect y_2 through the same process. Double causal "arrows" are allowed in a model only if different mechanisms may be hypothesized for the reciprocal relationships. Using this assumption, it is possible to set a causal order between y_2 and y_3. The assumption of *causal closure* implies independence of the residual causalities, which are the unknown factors responsible for the residual variance (i.e. the variance not accounted for by the observed descriptors). Path analysis is restricted to a small number of descriptors. This is not due to computational problems, but to the fact that the interpretation becomes complex when the number of descriptors in a model becomes large. When the analysis involves three descriptors only, the simple method of *causal modelling using correlations* may be used (Subsection 4.5.4).

For qualitative descriptors, Fienberg (1980; his Chapter 7) explains how to use *logit* or *log-linear models* (Section 6.3) to determine the signs of causal relationships among such descriptors, by reference to diagrams similar to the path diagrams of Section 10.4.

10.3 Regression

Random variable

The purpose of regression analysis is to describe the relationship between a *dependent* (or *response*) *random* variable* (y) and a set of *independent* (or *explanatory*) *variables*, in order to forecast or predict the values of y for given values of the independent variables $x_1, x_2, \ldots, x_m$. Box 1.1 gives the terminology used to refer to the dependent and independent variables of a regression model in an empirical or causal framework. The explanatory variables may be either random*, or controlled (and, consequently, known *a priori*). On the contrary, the response variable must of necessity be a random variable. That the explanatory variables be random or controlled will be important when choosing the appropriate computation method (model I or II).

Model

A *mathematical model* is simply a mathematical formulation (algebraic, in the case of regression models) of a relationship or a set of relationships among variables, whose parameters have to be estimated or tested against a hypothesis; in other words, it is a simplified mathematical description of a real-life system. Regression, with its many variants, is the first type of modelling method presented in this chapter for the analysis of ecological structures. It is also used as a platform to help introduce the principles of structure analysis. The same principles will apply to more mathematically advanced forms, collectively referred to as canonical analysis, which are discussed in Chapter 11.

[*] A random variable is a variable whose values are assumed to result from some random process (Section 1.0); these values are not known before observations are made. A random variable is *not* a variable consisting of numbers drawn at random; such variables, usually generated with the help of a pseudo-random number generator, are used by statisticians to assess the properties of statistical methods under some distribution hypotheses.

Regression modelling may be used for description, inference, or forecasting/prediction:

1. Description aims at finding the best functional relationship among variables in the model, and estimating its parameters, based on available data. In mathematics, a function $y = f(x)$ is a rule of correspondence, often written as an equation, that associates with each value of x one and only one value of y. A well-known functional relationship in physics is Einstein's equation $E = mc^2$, which describes the amount of energy E associated with given amounts of mass m; the scalar value c^2 is the parameter of the model, where c is the speed of light in vacuum.

2. Inference means generalizing the results of a set of observations to the whole target population, as represented by a sample drawn from that population. Inference may consist in estimating the confidence intervals within which the true values of the statistical population parameters are likely to be found, or testing *a priori* hypotheses about the values of model parameters in the statistical population. (1) The ecological hypotheses may simply concern the *existence* of a relationship, e.g. the slope or the intercept are different from 0. The test consists in finding the *two-tailed* probability of observing the slope (b_1) or intercept (b_0) values that have been estimated from the sample data, given the null hypothesis (H_0) stating that the slope (β_1) or intercept (β_0) parameters are zero in the statistical population. These tests are described in manuals of elementary statistics. (2) In other instances, the ecological hypothesis concerns the sign that the relationship should have. One then tests the *one-tailed* null statistical hypotheses (H_0) that the intercept or slope parameters in the statistical population are zero, against alternative hypotheses (H_1) that they have the signs (positive or negative) stated in the ecological hypotheses. For example, one might want to test Bergmann's law (1847), that the body mass of homeotherms, within species or groups of closely related species, *increases* with latitude. (3) There are also cases where the ecological hypothesis states specific values for the parameters. Consider for instance the isometric relationship specifying that mass should increase as the cube of the length in animals, or in log form: $\log(mass) = b_0 + 3\log(length)$. Length-to-mass relationships found in nature are most often allometric, especially when considering a multi-species group of organisms. Reviewing the literature, Peters (1983) reported allometric slope values from 1.9 (algae) to 3.64 (salamanders).

3. Forecasting (or prediction) consists in calculating values of the response variable using a regression equation. Forecasting (or prediction) is sometimes described as *the* purpose of ecology. In any case, ecologists agree that empirical or hypothesis-based regression equations are helpful tools for management. This objective is achieved by using the equation that minimizes the residual mean square error, or maximizes the coefficient of determination (r^2 in simple regression; R^2 in multiple regression).

A study may focus on one or two of the above objectives, but not necessarily all three. Satisfying two or all three objectives may call upon different methods for computing the regressions. In any case, these objectives differ from that of correlation

Correlation or regression analysis? Box 10.1

Regression analysis is a type of modelling. Its purpose is either to find the best functional model relating a response variable to one or several explanatory variables, in order to test hypotheses about the model parameters, or to forecast or predict values of the response variable.

The purpose of correlation analysis is quite different. It aims at establishing whether there is *interdependence*, in the sense of the coefficients of dependence of Chapter 7, between two random variables, without assuming any functional or explanatory-response or causal link between them.

In model I simple linear regression, where the explanatory variable of the model is controlled, the distinction is easy to make; in that case, a correlation hypothesis (i.e. interdependence) is meaningless. Confusion comes from the fact that the coefficient of determination, r^2, which is essential to estimate the forecasting value of a regression equation and is automatically reported by most regression programs, happens to be the square of the coefficient of linear correlation.

When the two variables are random (i.e. not controlled), the distinction is more tenuous and depends on the intent of the investigator. If the purpose is modelling (as broadly defined in the first paragraph of this Box), model II regression is the appropriate type of analysis; otherwise, correlation should be used to measure the interdependence between such variables. In Sections 4.5 and 10.4, the same confusion is rampant, since correlation coefficients are used as an *algebraic tool* for choosing among causal models or for estimating path coefficients.

analysis, which is to support the existence of a relationship between two random variables, without reference to any functional or causal link between them (Box 10.1).

This section does not attempt to present regression analysis in a comprehensive way. Interested readers are referred to general texts of (bio)statistics such as Sokal & Rohlf (1995), specialized texts on regression analysis (e.g. Draper & Smith, 1981; Neter *et al*., 1996), or textbooks such as those of Ratkowski (1983) or Ross (1990) for nonlinear estimation. The purpose here is to survey the main principles of regression analysis and, in the light of these principles, explain the differences among the regression models most commonly used by ecologists: simple linear (model I and model II), multiple linear, polynomial, partial, nonlinear, and logistic. Some smoothing methods will also be described. Several other types of regression will be

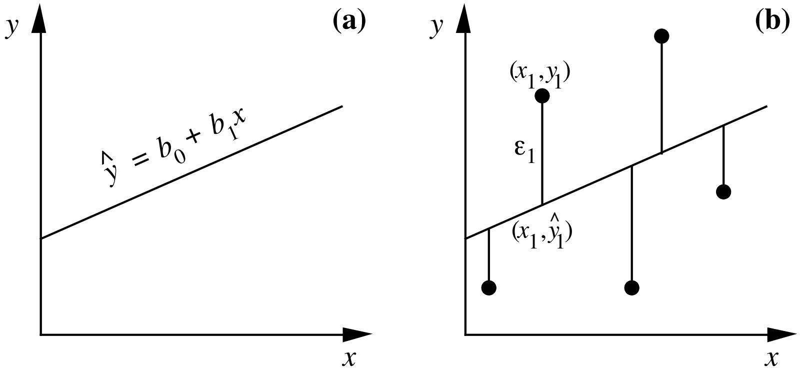


Figure 10.5 (a) Linear regression line, of equation $\hat{y} = b_0 + b_1 x$, fitted to the scatter of points shown in b. (b) Graphical representation of regression residuals ε_i (vertical lines); ε_1 is the residual for point 1 with coordinates (x_1, y_1).

mentioned, such as ridge regression, multivariate linear regression, and monotone or nonparametric regression.

Regression

Incidentally, the term *regression* has a curious origin. It was coined by the anthropologist Francis Galton (1889, pp. 95-99), a cousin of Charles Darwin, who was studying the relationship between the heights of parents and children. Galton observed "that the Stature of the adult offspring … [is] … more *mediocre* than the stature of their Parents", or in other words, closer to the population mean; so, Galton said, they *regressed* (meaning *going back*) towards the population mean. He called the slope of this relationship "the ratio of 'Filial Regression' ". For this historical reason, the slope parameter is now known as the regression coefficient.

1 — Simple linear regression: model I

Linear regression is used to compute the parameters of a first-degree equation relating variables y and x. The expression *simple linear regression* applies to cases where there is a single explanatory variable x. The equation (or model) for simple linear regression has the form:

$$\hat{y} = b_0 + b_1 x \qquad \textbf{(10.1)}$$

This corresponds to the equation of a straight line (hence the name *linear*) that crosses the scatter of points in some optimal way and allows the computation of an estimated value $\hat{y}$ (along the ordinate scale of the scatter diagram) for any value of x (abscissa; Fig. 10.5a). Parameter b_0 is the estimate of the intercept of the regression line with the *ordinate*; it is also called the y-intercept. Parameter b_1 is the slope of the regression line; it is also called the *regression coefficient*. In Subsection 10.3.4 on polynomial

Intercept
Slope

regression, a distinction will be made between linearity in parameters and linearity in response to the explanatory variables.

The intercept b_0 has the same physical dimensions as y, whereas the regression coefficient b_1 has the physical dimensions of $[y]/[x]$ (Section 3.1) so that b_1x has the same physical dimensions as y. As a consequence, the regression equation (eq. 10.1) is dimensionally homogeneous (Section 3.2).

When using this type of regression, one must be aware of the fact that a *linear model* is imposed on the data. In other words, one assumes that the relationship between variables may be adequately described by a straight line and that the vertical dispersion of observed values above and below the line is the result of a random process. The difference between the observed and estimated values along y, noted $\varepsilon_i = (y_i - \hat{y}_i)$ for every observation i, may be either positive or negative since the observed data points lie above and below the regression line. ε_i is called the *residual* value of observation y_i after fitting the regression line (Fig. 10.5b). Including ε_i in the equation allows one to describe exactly the ordinate value y_i of each point (x_i, y_i) in the data set; y_i is equal to the value $\hat{y}_i$ predicted by the regression equation plus the residual ε_i:

$$y_i = \hat{y}_i + \varepsilon_i = b_0 + b_1x_i + \varepsilon_i \qquad \textbf{(10.2)}$$

This equation is the *linear model* of the relationship. $\hat{y}_i$ is the predicted, or *fitted* value corresponding to each observation i. The model assumes that the only deviations from the linear functional relationship $y = b_0 + b_1x$ are vertical differences ("errors") ε_i on values y_i of the response variable, and that there is no "error" associated with the estimation of x. "Error" is the traditional term used by statisticians for deviations of all kind due to random processes, and not only measurement error. In practice, when it is known by hypothesis — or found by studying a scatter diagram — that the relationship between two variables is not linear, one may either try to linearise it (Section 1.5), or else use polynomial or nonlinear regression methods to model the relationship (Subsections 10.3.4 and 10.3.6, below).

Model I

Besides the supposition that the variables under study are linearly related, *model I regression* makes the following additional assumptions about the data:

1. The explanatory variable x is controlled, or it is measured without error. (The concepts of random and controlled variables have been briefly explained above.)

2. For any given value x_i of x, the values y in the statistical population are independently and normally distributed. This does not mean that the response variable y must be normally distributed, but instead that the "errors" ε_i are normally distributed about a mean of zero. One also assumes that the ε_i's have the same variance for all values of x in the range of the observed data (homoscedasticity: Box 1.3).

So, model I regression is appropriate to analyse results of controlled experiments, and also the many cases of field data where a response random variable y is to be related to sampling variables under the control of the researcher (e.g. location in time and space, volume of water filtered). The next subsection will show how to use model II regression to analyse situations where these assumptions are not met.

Least squares

OLS

In simple linear regression, one is looking for the straight line with equation $\hat{y} = b_0 + b_1 x$ that minimizes the sum of squares of the vertical residuals, ε_i, between the observed values and the regression line. This is the *principle of least squares*, first proposed by the mathematician Adrien Marie Le Gendre from France, in 1805, and later by Karl Friedrich Gauss from Germany, in 1809; these two mathematicians were interested in estimation problems of astronomy. This sum of squared residuals, $\Sigma (y_i - \hat{y}_i)^2$, offers the advantage of providing a unique solution, which would not be the case if one chose to minimize another function — for example $\Sigma |y_i - \hat{y}_i|$. It can also be shown that the straight line that meets the *ordinary least-squares* (OLS) criterion passes through the centroid, or centre of mass $(\bar{x}, \bar{y})$ of the scatter of points, whose coordinates are the means $\bar{x}$ and $\bar{y}$. The formulae for parameters b_0 and b_1 of the line meeting the least-squares criterion are found using partial derivatives. The solution is:

$$b_1 = s_{xy} / s_x^2 \quad \text{and} \quad b_0 = \bar{y} - b_1 \bar{x} \qquad \textbf{(10.3)}$$

where s_{xy} and s_x^2 are estimates of covariance and variance, respectively (Section 4.1). These formulae, written in full, are found in textbooks of introductory statistics. Least-squares estimates of b_0 and b_1 can also be computed directly from the **x** and **y** data vectors using eq. 2.19. Least-squares estimation provides the line of best fit for parameter estimation and forecasting when the explanatory variable is controlled.

Regressing y on x does not lead to the same least-squares equation as regressing x on y. Figure 10.6a illustrates this for two random variables, which would represent a case for model II regression discussed in the next subsection. Even when x is a random variable, the variables will continue to be called x and y (instead of y_1 and y_2) to keep the notation simple. Although the covariance s_{xy} is the same for the calculation of the regression coefficient of y on x ($b_{1\,(y \cdot x)}$) and that of x on y ($c_{1\,(x \cdot y)}$), the denominator of the slope equation (eq. 10.3) is s_x^2 when regressing y on x, whereas it is s_y^2 when regressing x on y. Furthermore, the means $\bar{x}$ and $\bar{y}$ play inverted roles when estimating the two intercepts, $b_{0\,(y \cdot x)}$ and $c_{0\,(x \cdot y)}$. This emphasizes the importance of clearly defining the explanatory and response variables when performing regression.

The two least-squares regression lines come together only when all observation points fall on the same line (correlation = 1). According to eq. 4.7, $r_{xy} = s_{xy} / s_x s_y$. So, when $r = 1$, $s_{xy} = s_x s_y$ and, since $b_{1\,(y \cdot x)} = s_{xy} / s_x^2$ (eq. 10.3), then $b_{1\,(y \cdot x)} = s_x s_y / s_x^2 = s_y / s_x$. Similarly, the slope $c_{1\,(x \cdot y)}$, which describes the same line in the transposed graph, is $s_x / s_y = 1 / b_{1\,(y \cdot x)}$. In the more general case where r is not equal to 1, $c_{1\,(x \cdot y)} = r_{xy}^2 / b_{1\,(y \cdot x)}$. When the two regression lines are drawn on the same graph, assuming that the variables have been standardized prior to the

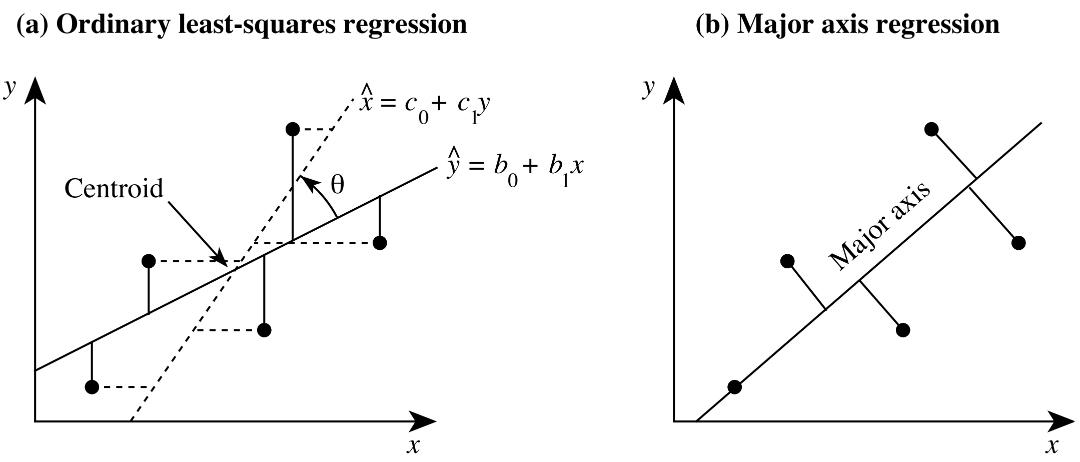


Figure 10.6 (a) Two least-squares regression equations are possible in the case of two random variables (called x and y here, for simplicity). When regressing y on x, the sum of *vertical* squared deviations is minimized (full lines); when regressing x on y, the sum of *horizontal* squared deviations is minimized (dashed lines). Angle θ between the two regression lines is computed using eq. 10.5. (b) In major axis regression, the sum of the squared Euclidean distances to the regression line is minimized.

computations, there is a direct relationship between the Pearson correlation coefficient r_{xy} and angle θ between the two regression lines:

$$\theta = 90^\circ - 2\tan^{-1} r, \quad \text{or} \quad r = \tan\left(\frac{90^\circ - \theta}{2}\right) \qquad \textbf{(10.4)}$$

If $r = 0$, the scatter of points is circular and angle $\theta = 90^\circ$, so that the two regression lines are at a right angle; if $r = 1$, the angle is 0°. Computing angle θ for non-standardized variables, as in Fig. 10.6a, is a bit more complicated:

$$\theta = 90^\circ - \text{sign}(r) \times [\tan^{-1}(r\, s_x/s_y) + \tan^{-1}(r\, s_y/s_x)] \qquad \textbf{(10.5)}$$

where sign(r) is the sign of the correlation coefficient.

Coefficient of determination

The *coefficient of determination* r^2 measures how much of the variance of each variable is explained by the other. This coefficient has the same value for the two regression lines. The amount of explained variance for y is the variance of the fitted values $\hat{y}_i$, calculated as:

$$s_{\hat{y}}^2 = \Sigma(\hat{y}_i - \bar{y})^2 / (n-1) \qquad \textbf{(10.6)}$$

whereas the total amount of variation in variable y is

$$s_y^2 = \Sigma (y_i - \bar{y})^2 / (n-1)$$

It can be shown that the coefficient of determination, which is the ratio of these two values (the two denominators (n – 1) cancel out), is equal to the square of the Pearson correlation coefficient r. It is thus designated by r^2:

$$r^2 = s_{\hat{y}}^2 / s_y^2 \qquad \textbf{(10.7)}$$

With two random variables, the regression of y on x makes as much sense as the regression of x on y. In that case, the coefficient of determination may be computed as the product of the two regression coefficients: $r^2 = b_{1\,(y \cdot x)} c_{1\,(x \cdot y)}$. The coefficient of correlation is then the geometric mean of the coefficients of linear regression of each variable on the other, to which the sign of one of the regression coefficients is imposed: $r = \text{sign}\,(b_{1\,(y \cdot x)}) \times (b_{1\,(y \cdot x)} c_{1\,(x \cdot y)})^{1/2}$; function sign() is described after eq. 10.5. It may also be computed as the square of r in eq. 4.7:

$$r^2 = \frac{(s_{xy})^2}{s_x^2\, s_y^2} \qquad \textbf{(10.8)}$$

Coefficient of non-determination

A value $r^2 = 0.81$, for instance, means that 81% of the variation in y is explained by x, and vice versa. In Section 10.4, the quantity (1 – r^2) will be called the *coefficient of nondetermination*; it measures the proportion of the variance of a response variable that is not explained by the explanatory variable(s) of the model.

When x is a controlled variable, one must be careful not to interpret the coefficient of determination in terms of interdependence, as one would for a coefficient of correlation, in spite of their algebraic closeness and the fact that one coefficient can, indeed, be calculated directly from the other (Box 10.1).

2 — Simple linear regression: model II

Model II

When both the response and explanatory variables of the model are random (i.e. *not* controlled by the researcher), there are errors associated with the measurements of both x and y. Such situations call for methods that are referred to as *model II regression*. As a parallel to model II ANOVA, which is concerned with the analysis of the effect of a random factor on a random variable (Sokal & Rohlf, 1995, Section 8.7), model II regression is concerned with the analysis of two random variables. In model II regression, different computational procedures are required for description and inference, as opposed to forecasting; these three objectives of regression analysis were described at the beginning of Section 10.3.

1. Model II regression can be used for description and inference, that is, to estimate the slope of a process (parametric estimation) corresponding to the linear relationship

between the measured variables, and compute confidence intervals around the slope or test its significance. Examples:

• In aquatic ecology, *in vivo* fluorescence is routinely used to estimate the amount of chlorophyll *a* in phytoplankton. These variables, which are both random and measured with error, must be related by model II regression to establish their functional relationship (slope). The slope can also be tested for significance. If the objective is to forecast chlorophyll *a* from fluorescence values, see point 2 below.

• In freshwater sediment, one may be interested in comparing the rate of microbial anaerobic methane production to total particulate carbon in two environments (e.g. two lakes) where several sites have been studied. Since total particulate carbon and methane production have been measured with error in the field, rates are given by the slopes of model II regression equations computed on the data from the two lakes separately; the confidence intervals of these slopes may serve to compare the two environments.

Model II regression can be used with the more simple purpose of drawing a line in a graph of two random variables. This can be done for the above examples.

2. Model II regression can also be used for forecasting, that is, for computing fitted values about one variable from the values of the other. The method to be used in that case is ordinary least squares (OLS). The reason is simple: OLS is the method that produces fitted values with the smallest error, defined as $\Sigma(y_i - \hat{y}_i)^2$ (Subsection 10.3.1). Hence, OLS is also one of the methods that can be used in model II situations, when the purpose is forecasting. Example:

• In microbial ecology, the concentrations of two substances produced by bacterial metabolism have been measured. One is of economical interest, but difficult to measure with accuracy, whereas the other is easy to measure. Determining their relationship by regression may allow ecologists to use the second substance as a proxy for the first. An OLS regression model can be used to estimate the concentrations of the first substance from the concentrations of the second.

3. Another application of model II regression concerns deterministic models, which are often used to describe ecological processes. In order to test how good a model is at describing reality, one can run the model with observed values of the control variables and compare the values predicted by the model to the observed values of the response variable. Since both sets of variables (control, response) are random, the values predicted by the model are just as random as the values of the response variables, so that they should be related and compared using model II regression. The hypothesis is one-tailed in this case; indeed, a model accurately reflects the field process only if its predictions are *positively* correlated with the field observations. Theory and examples are provided by Mesplé *et al.* (1996).

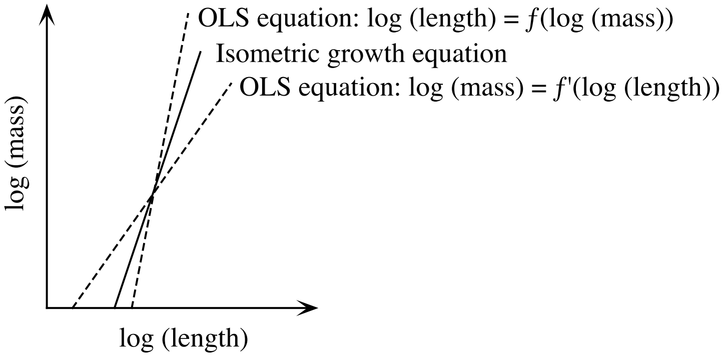

OLS equation: log (length) = ƒ(log (mass))
Isometric growth equation
OLS equation: log (mass) = ƒ'(log (length))
log (mass)
log (length)

Figure 10.7 Isometric growth is depicted by the functional relationship $\log(mass) = b_0 + 3\log(length)$. The ordinary least-squares (OLS) regression line of $\log(mass)$ on $\log(length)$ would suggest allometric growth of one type, while the OLS regression line of $\log(length)$ on $\log(mass)$ would suggest allometric growth of the opposite type.

In the descriptive examples above, one was interested in estimating the parameters of the equation that describes the functional relationship between pairs of random variables in order to quantify underlying physiological or ecological processes. When both variables are random, as in these examples, model II regression should be used for parameter estimation since the slope found by ordinary least squares (OLS) is too small in absolute value, due to the presence of measurement error in the explanatory variable. OLS regression should only be used when x is fixed by experiment (model I regression) or is a random variable measured with little error compared to y (see recommendation 1 at the end of this subsection). OLS regression (model I or II) should also be used when the objective of the study is forecasting (see recommendation 6).

To better understand the above assertion, let us consider the relationship between length and mass of adult animals of a given species. Let us further assume that the relationship is isometric ($mass = c \times length^3$) for the species under study; this equation would correspond to the case where all individuals, short or long, have the same shape (fatness). The same functional equation, in log form, is $\log(mass) = b_0 + 3\log(length)$, where b_0 is the log of parameter c. Since individual measurements are each subject to a large number of small genetic and environmental influences, presumably additive in their effects and uncorrelated among individuals, it is expected that both length and mass include random deviations from the functional equation; measurement errors must be added to this inherent variability. In such a system, the slope of the OLS regression line of $\log(mass)$ on $\log(length)$ would be smaller than 3 (Fig. 10.7; Ecological application 10.3a), which would lead one to conclude that the species displays allometric growth, with longer individuals thinner than short ones. On the contrary, the slope of the regression line of $\log(length)$ on $\log(mass)$, computed in the transposed space, would produce a slope smaller than 1/3; its inverse, drawn in Fig. 10.7, is larger than 3; this slope would lead to the opposite conclusion, i.e. that shorter individuals are thinner than long ones. This apparent paradox is simply due to the fact that OLS regression is inappropriate to describe the functional relationship between these variables.

Several methods have been proposed to estimate model II regression parameters, and a controversy has raged in the literature about which method was the best. The following methods are the most popular — although, surprisingly, the major statistical packages, except R, are still ignoring them (except method 4, OLS). For methods 1 to 3 described below, slope estimates can easily be calculated with a pocket calculator, from values of the means, variances, and covariance, computed with standard statistical software.

Methods 1, 2, and 4 are special cases of the *structural relationship*, which assumes that there is error ε_i on y and δ_i on x, ε_i and δ_i being independent of each other. As stated above, "error" means deviation of any kind due to a random process, not only measurement error. The maximum likelihood (ML) estimate of the slope for such data is (Madansky, 1959; Kendall & Stuart, 1966):

ML slope formula

$$b_{\mathrm{ML}} = \frac{s_y^2 - \lambda s_x^2 + \sqrt{(s_y^2 - \lambda s_x^2)^2 + 4\lambda s_{xy}^2}}{2 s_{xy}} \qquad \textbf{(10.9)}$$

where s_y^2 and s_x^2 are the estimated variances of y and x, respectively, s_{xy} is their covariance, and λ is the ratio $\sigma_\varepsilon^2 / \sigma_\delta^2$ of the variances of the two error terms.

When λ is large or s_{xy} is very small, another equation form may provide greater computational accuracy than eq. 10.9. It is derived from the property that the slope of the regression line of y on x is the inverse of the slope of the regression of x on y in the case of symmetric regression lines. After the proper substitutions, eq. 10.9 becomes:

$$b_{\mathrm{ML}} = \frac{2 s_{xy}}{s_x^2 - (s_y^2/\lambda) + \sqrt{[s_x^2 - (s_y^2/\lambda)]^2 + (4 s_{xy}^2/\lambda)}} \qquad \textbf{(10.10)}$$

The model II regression methods are derived from eq. 10.9 or eq. 10.10.

Major axis

1. *Major axis regression (MA)*. — In this method, the estimated regression line is the first principal component of the scatter of points (see principal component analysis, Section 9.1). The quantity that is minimized is the sum, over all points, of the squared *Euclidean distances* between the points and the regression line (Fig. 10.6b), instead of *vertical distances* as in OLS (Fig. 10.6a). In this method, one assumes that the two error variances (σ_ε^2 on y and σ_δ^2 on x) are equal, so that their ratio $\lambda = 1$. This assumption is strictly met, for example, when both variables have been measured using the same instrument and all of the error is measurement error (McArdle, 1988). The slope of the major axis is estimated by the following formula (Pearson, 1901; Jolicoeur, 1973; Sokal & Rohlf, 1995), which is a special case of eq. 10.9 for $\lambda = 1$:

$$b_{\mathrm{MA}} = \frac{s_y^2 - s_x^2 + \sqrt{(s_y^2 - s_x^2)^2 + 4\,(s_{xy})^2}}{2 s_{xy}} \qquad \textbf{(10.11)}$$

The positive square root is used in the numerator. A second equation is obtained by using the negative square root; it estimates the slope of the minor axis, which is the second principal component, of the bivariate scatter of points. When the covariance is near 0, b_{MA} is estimated using eq. 10.10 (with $\lambda = 1$) instead of eq. 10.11, in order to avoid numerical indetermination.

The slope of the major axis may also be calculated using estimates of the slope of the OLS regression line, b_{OLS} , and of the correlation coefficient, r_{xy}:

$$b_{\text{MA}} = \frac{d \pm \sqrt{d^2+4}}{2} \quad \text{where} \quad d = \frac{(b_{\text{OLS}})^2 - r_{xy}^2}{r_{xy}^2 \times b_{\text{OLS}}}$$

The positive root of the radical is used when the correlation coefficient is positive, and conversely when it is negative.

Just as with principal component analysis, this method is useful in situations where both variables are expressed in the same physical units or are dimensionless (naturally, or after standardization or ranging). Many natural ecological variables are not in the same physical units. Major axis regression has been criticized because, in that case, the slope estimated by major axis regression is not invariant under an arbitrary change of scale such as expansion (Section 1.5) and, after a change of scale, b_{MA} cannot be directly calculated using the change-of-scale factor. In these conditions, the actual *value* of the slope may be meaningless (Teissier, 1948; Kermack and Haldane, 1950; Ricker, 1973; McArdle, 1988) or difficult to interpret. By comparison, the slopes of the OLS, SMA, and RMA (described below) regression lines are not invariant either to change-of-scale transformations, but the slopes of the transformed data can easily be calculated using the change-of-scale factor. For example, after regressing a mass variable in g onto a length variable in cm, if the OLS slope is b_1 (in g/cm), then after rescaling the explanatory variable from cm to m, the OLS slope becomes $b'_1 = b_1 \times 100$.

Permutation test

Significance of b_{MA} estimates can be tested by permutation (Section 1.2); the values of one or the other variable (i.e. x or y) are permuted a large number of times and slope estimates are computed using eq. 10.11. The test should be carried out on the lesser of the two slopes in absolute value: b_1 of y on x, or $b'_1 = 1/b_1$ of x on y. If the objective is simply to assess the relationship between the two variables under study, the correlation coefficient should be tested for significance instead of the slope of a model II regression line.

When the variances s_y^2 and s_x^2 are equal, the slope estimated by eq. 10.11 is ±1, the sign being that of the covariance, whatever the value of s_{xy}. As in the case of SMA (below), permutations produce slope estimates of +1 or –1 in equal numbers, with a resulting probability near 0.5 whatever the value of the correlation. This result is meaningless. The practical consequence is that, if the slope estimate b_{MA} is to be tested by permutations, variables should not be standardized (eq. 1.12).

C.I. of MA slope

Alternatively, one may compute the confidence interval of the slope at a predetermined confidence level and check whether the value 0 (or, for that matter, any other value of interest) lies inside or outside the confidence interval. Computation of the confidence interval involves several steps; the formulae are given in Jolicoeur & Mosimann (1968), Jolicoeur (1990), and Sokal & Rohlf (1995, pp. 589-591), among others. When both n and the ratio of the eigenvalues of the bivariate distribution (see principal component analysis, Section 9.1) are small, limits of the confidence interval cannot be computed because it covers all 360° of the plane. Such a confidence interval always includes slope 0, as well as any other value. For example, when $n = 10$, the ratio of the eigenvalues must be larger than 2.21 for the 95% confidence interval to be real; for $n = 20$, the ratio must be larger than 1.63; and so on.

It frequently happens in ecology that a scatter plot displays a bivariate lognormal distribution; the univariate frequency distributions of such variables are positively skewed, with longer tails in the direction of the higher values. Such distributions may be normalized by applying a log transformation (Subsection 1.5.6; Fig. 1.11). This transformation also solves the problem of dimensionally heterogeneous variables and makes the estimate of the major axis slope invariant over expansion (multiplication or division by a constant: Section 1.5) — but not over translation. One should verify, of course, that the log-transformed data conform to a bivariate normal distribution before proceeding with major axis regression.

This property can easily be demonstrated as follows. Consider a model II functional equation describing the linear relationship between two log-transformed variables x and y:

$$\log(y) = b_0 + b_1 \log(x)$$

If x and y are divided by constants c_1 and c_2 respectively (expansion), one obtains new variables $x' = x/c_1$ and $y' = y/c_2$, so that $x = c_1 x'$ and $y = c_2 y'$. The functional equation becomes:

$$\log(c_2 y') = b_0 + b_1 \log(c_1 x')$$

$$\log(y') + \log(c_2) = b_0 + b_1 \log(c_1) + b_1 \log(x')$$

$$\log(y') = [b_0 + b_1 \log(c_1) - \log(c_2)] + b_1 \log(x')$$

which may be rewritten as

$$\log(y') = b_0' + b_1 \log(x')$$

where $b_0' = [b_0 + b_1 \log(c_1) - \log(c_2)]$ is the new intercept, while the slope of $\log(x')$ is still b_1. So, under log transformation, the slope b_1 is invariant for any values of expansion coefficients c_1 and c_2; it differs, of course, from the major axis regression coefficient (slope) of the untransformed variables.

Dividing x and y by their respective standard deviations, s_x and s_y, is an expansion which makes the two variables dimensionless. It thus follows that the major axis slope of the original log-transformed data is the same as that of the log of the standardized (dimensionless) data. This also applies to other standardization methods such as division by the maximum value or the range (eqs. 1.10 and 1.11).

Readers who prefer numerical examples can easily check the above derivation by computing a principal component analysis on a small data set containing two log-transformed variables only, with or without expansion (multiplication or division by a constant prior to the log transformation). The angles between the original variables and the first principal component are easily computed as the $\cos^{-1}$ of the values in the first normalized eigenvector (Subsection 9.1.3); the slopes of the major axis regression coefficients of $y = f(x)$ and $x = f(y)$, which are the tangents (tan) of these angles, remain the same over such a transformation.

Standard major axis

2. *Standard major axis (SMA).* — Regression using variables that are not dimensionally homogeneous produces results that vary with the scales of the variables. If the physical dimensions are arbitrary (e.g. length measurements that may indifferently be recorded in mm, cm, m, or km), the slope estimate is also arbitrary. In ordinary least-squares regression (OLS), the slope and confidence interval values change proportionally to the measurement units. For example, multiplying all y values by 10 produces a slope estimate ten times larger, whereas multiplying all x values by 10 produces a slope estimate 10 times smaller. This is not the case with MA; the major axis slope does not scale proportionally to the units of measurement. For that reason, it may be desirable to make the variables dimensionally homogeneous prior to model II regression.

Standard major axis regression is MA regression performed on standardized variables, which are thus dimensionally homogeneous. It is computed as follows:

- Standardize variables x and y using eq. 1.12.

- Compute MA regression on the standardized variables. The slope estimate is always +1 or –1; the sign is that of the covariance s_{xy} or correlation coefficient r_{xy}.

- Back-transform the slope estimate to the original units by multiplying it by (s_y/s_x).

As a consequence, the slope of the *standard major axis* (SMA), or *reduced major axis*, is computed as the ratio (Teissier, 1948):

$$b_{\text{SMA}} = \sqrt{s_y^2/s_x^2} = \text{sign}\,(r) \times (s_y/s_x) \qquad \textbf{(10.12)}$$

where sign(r) is the sign of the correlation coefficient. This formula is obtained from eq. 10.9 by assuming that the error variances σ_ε^2 and σ_δ^2 of y and x, respectively, are identically proportional to their respective variances σ_y^2 and σ_x^2; in other words, $\sigma_\varepsilon^2/\sigma_y^2 = \sigma_\delta^2/\sigma_x^2$. This assumption is unlikely to be strictly true with real data, except in cases where both variables are counts (e.g. numbers of organisms), raw or log-transformed (McArdle, 1988). Replacing variances σ_y^2 and σ_x^2 by their unbiased estimates s_y^2 and s_x^2 gives the following value to λ in eq. 10.9:

$$\lambda = \sigma_\varepsilon^2 / \sigma_\delta^2 = \sigma_y^2 / \sigma_x^2 = s_y^2 / s_x^2$$

Equation 10.9 then simplifies to eq. 10.12. Since the square root $\sqrt{s_y^2 / s_x^2}$ is either positive or negative, the slope estimate receives the sign of the Pearson correlation coefficient, which is the same as that of the covariance s_{xy} in the denominator of eq. 10.9 or that of the OLS slope estimate. The b_{SMA} estimate is also the geometric mean of the OLS regression coefficient of y on x and the *reciprocal* of the regression coefficient of x on y; this is why the method is also called *geometric mean regression*, besides a variety of other names.

From equations 4.7 (Pearson r), 10.3 (b_{OLS}) and 10.12 (b_{SMA}), one can show that

$$b_{\mathrm{SMA}} = |b_{\mathrm{OLS}}| / r_{xy} \quad \text{when} \quad r_{xy} \neq 0 \qquad \mathbf{(10.13)}$$

So, in addition to eq. 10.12, one can easily compute b_{SMA} from eq. 10.13 using values of b_{OLS} and r_{xy} provided by an OLS regression program. This equation also shows that, when the variables are highly correlated ($r \rightarrow 1$), $b_{\mathrm{SMA}} \rightarrow b_{\mathrm{OLS}}$. When they are not, b_{SMA} is always larger than b_{OLS} for positive values of r, and smaller for negative values of r; in other words, b_{OLS} is always closer to 0 than b_{SMA}.

When $r_{xy} = 0$, the b_{SMA} estimate obtained from eq. 10.12, which is the ratio of the standard deviations, is meaningless. It does not fall to zero when the correlation is zero, except in the trivial case where s_y is zero (Jolicoeur, 1975, 1990). Since the b_{SMA} estimate is independent of the presence of a significant covariance between x and y (eq. 10.12), users should always compute a Pearson correlation coefficient and test it for significance prior to computing the slope of a standard major axis regression line. If r is not significantly different from zero, b_{SMA} should not be computed.

The slope of the standard major axis cannot be tested for significance by a regular permutation test. There are two reasons for this.

Permutation test

- Consider permutation testing. The b_{SMA} slope estimate is $\pm s_y / s_x$ but, for all permuted data, s_y / s_x is a constant. Giving the signs of the permuted covariances to the permuted slope estimates inevitably produces a probability near 0.5 of obtaining, by permutation, a value as extreme as or more extreme than the estimate b_{SMA}.

- The confidence interval of the slope b_{SMA}, described below, is inappropriate to test the null hypothesis $\beta = 0$ because the ratio s_y / s_x cannot be zero unless s_y is equal to zero. This is a trivial case, unsuitable for regression analysis (Sokal & Rohlf, 1995).

McArdle (1988) suggests that the solution to this problem is to test the correlation coefficient r_{xy} for significance instead of testing b_{SMA}. Warton *et al.* (2006, Appendix F) describe a permutation test of the SMA slope based on residuals.

C.I. of SMA slope

When needed, an approximate confidence interval $[b_1, b_2]$ can be computed for b_{SMA} as follows (Jolicoeur & Mosimann, 1968):

$$b_1 = b_{\text{SMA}}[\sqrt{(B+1)} - \sqrt{B}]$$

$$b_2 = b_{\text{SMA}}[\sqrt{(B+1)} + \sqrt{B}]$$

where $$B = t^2(1-r^2)/(n-2)$$

and t is a two-tailed Student's $t_{\alpha/2}$ value for significance level α and $(n - 2)$ degrees of freedom.

Ranged major axis

3. *Ranged major axis regression (RMA)*. — An alternative transformation to make the variables dimensionally homogeneous is *ranging* (eqs. 1.10 and 1.11). This transformation does not make the variances equal and thus does not lead to the problems encountered with SMA regression. It leads to RMA, which proceeds as follows:

- Transform the y and x variables into y' and x', respectively, using eq. 1.11. For relative-scale variables (Subsection 1.4.1), which have zero as their natural minimum, the ranging transformation is carried out using eq. 1.10.

- Compute MA regression between the ranged variables y' and x'. Test by permutation if a test is required.

- Back-transform the estimated slope and confidence interval limits to the original units by multiplying them by the ratio of the ranges, $(y_{max} - y_{min})/(x_{max} - x_{min})$.

The RMA slope estimator has several desirable properties when variables x and y are not expressed in the same units. The slope estimator scales proportionally to the units of x and y. The estimator is not insensitive to the covariance, as is the case for SMA. Finally, it is possible to test the hypothesis that an RMA slope estimate is equal to a stated value, in particular 0 or 1. As in MA, this may be done either by permutations, or by comparing the confidence interval of the slope to the hypothetical value of interest. Thus, whenever MA regression cannot be used because of incommensurable units, RMA regression can be used. There is no reason, however, to use RMA when the variables are expressed in the same units.

Prior to RMA, one should check for the presence of outliers, using a scatter diagram of the objects. Outliers cause important changes to the estimates of the ranges of the variables. Outliers that are not aligned with the bulk of the objects may thus have an undesirable influence on the slope estimate. RMA should not be used in the presence of such outliers.

OLS method

4. *Ordinary least squares (OLS)*. — The OLS method is derived from eq. 10.10 by assuming that there is no error on x, so that the error variance on x, σ_δ^2, is zero and thus $\lambda = \sigma_\varepsilon^2 / \sigma_\delta^2 = \infty$. After simplification, the OLS slope is equal to (eq. 10.3)

$$b_{\text{OLS}} = s_{xy} / s_x^2$$

The remainder of the subsection is devoted to the description of general properties and the comparison of model II regression methods.

C.I. of intercept

With all methods of model II regression, an estimate of the intercept, b_0, can be computed from b_1 and the centroid of the scatter of points $(\bar{x}, \bar{y})$, using eq. 10.3. The same equation can be used to calculate *approximate estimates* of the confidence limits of the intercept. Warton *et al*. (2006) describe more precise estimates of these confidence limits.

The first three methods (MA, SMA, RMA) have the property that the slope of the regression $y = f(x)$ is the reciprocal of the slope of $x = f(y)$. This property of symmetry is desirable here since there is no functional distinction between x and y in a model II situation. OLS regression does not have that property (Fig. 10.6a).

Recom-mendations

Users of model II regression techniques are never certain that the assumptions of the various methods are met by the variables in the data sets (i.e. MA: $\sigma_\varepsilon^2 = \sigma_\delta^2$ so that $\lambda = \sigma_\varepsilon^2 / \sigma_\delta^2 = 1$; SMA: $\lambda = \sigma_y^2 / \sigma_x^2$; OLS: $\sigma_\delta^2 = 0$ so that $\lambda = \sigma_\varepsilon^2 / \sigma_\delta^2 = \infty$). For that reason, McArdle (1988) carried out an extensive simulation study to investigate the influence of the error variances, σ_ε^2 for y and σ_δ^2 for x, on the efficiency (i.e. precision of the estimation) of the MA, SMA and OLS methods, measuring how variable the estimated slopes were under various conditions. Likewise, Jolicoeur (1990) used simulations to investigate the effects of small sample sizes and low correlations on the slope estimates obtained by MA and SMA. D. J. Currie, P. Legendre and A. Vaudor (unpublished study) also used numerical simulations to investigate the relationship between slope estimate formulas. They compared MA to OLS and MA to SMA in the *correlation situation*, defined as that where researchers are interested in describing the slope of the bivariate relationship displayed by two correlated random variables, i.e. variables that are not controlled or error-free. The results of all these simulations lead to the following recommendations for the estimation of parameters of functional linear relationships between variables that are random (i.e. not controlled) and measured with error (Table 10.4). They were first presented in a guide (Legendre, 2008b) distributed with the R package LMODEL2.

1. If the magnitude of the random variation (i.e. the error variance*) on the response variable y is much larger (i.e. more than three times) than that on the explanatory variable x, use OLS as the model II regression method. Otherwise, proceed as follows.

* Contrary to the sample variance, the error variance on x or y cannot be estimated from the data. It can only be estimated from knowledge of the way the variables were measured.

Table 10.4 Recommendations for the application of the model II regression methods. The numbers refer to the corresponding recommendation paragraphs (recom.) in the text.

- The error on y is much larger than the error on x: use OLS (recom. 1)
- The data distribution is close to bivariate normal (recom. 2)
  - The variables are in the same physical units or dimensionless, the error variance is about the same for x and y: use MA (recom. 3)
  - The variables are not dimensionally homogeneous. The error variance along each axis is proportional to the variance of the corresponding variable (recom. 4)
    - There are no outliers in the scatter diagram: RMA can be used (recom. 4.1)
    - The Pearson correlation coefficient r is significant: SMA can be used (recom. 4.2)
- The data distribution is clearly not bivariate normal (recom. 2)
  - The relationship between x and y is linear: use OLS (recom. 5)
- The objective is to compute forecasted (i.e. fitted) values $\hat{y}$: use OLS (recom. 6)
- The objective is to compare observations to model predictions: use MA (recom. 7)

2. Check whether the data are approximately bivariate normal, either by examining a scatter diagram or by performing a formal test of significance. If they are not, attempt transformations to make the distribution bivariate normal. For data that are or can be made to be reasonably bivariate normal, consider recommendations 3 and 4. If not, see recommendation 5.

3. For bivariate normal data, if the two variables are expressed in the same physical units (untransformed variables that were originally measured in the same units) or are dimensionless (e.g. log-transformed variables), and if it can reasonably be assumed that the error variances of the variables are approximately equal, use major axis (MA) regression.

When no information is available on the ratio of the error variances and there is no reason to believe that it may differ from 1, MA may be used provided that the results are interpreted with caution. MA produces unbiased slope estimates and accurate confidence intervals (Jolicoeur, 1990).

MA can be used with dimensionally heterogeneous variables (1) when the purpose of the analysis is to compare slopes computed from these variables measured in an identical way in different systems (e.g. at two or more sampling sites). It may also be useful (2) when the objective of the study is to test the hypothesis that the slope of the major axis of the empirical data does not differ from a value given by theory.

4. For bivariate normal data, if MA cannot be used because the variables are not expressed in the same physical units or the error variances on the two axes differ, two methods are available to estimate the parameters of the functional linear relationship if it can be assumed that the error variance on each axis is proportional to the variance of the corresponding variable, i.e. (error variance of *y* / sample variance of *y*) ≈ (error variance of *x* / sample variance of *x*). This condition is often met with counts (e.g. number of plants or animals) or log-transformed data (McArdle, 1988). The two following methods can be used if their specific conditions are met by the data.

4.1. Ranged major axis regression (RMA) can be used if there are no outliers in the scatter of points. Prior to RMA, one should check for the presence of outliers, using a scatter diagram of the objects.

4.2. Standard major axis regression (SMA) can be used if the coefficient of linear correlation (Pearson *r*) is significant. SMA regression should not be computed when this condition is not met.

The SMA slope cannot be tested by a standard permutation test, but the correlation coefficient *r* can. See also Warton *et al*. (2006, Appendix F) for a permutation test of the SMA slope based on residuals. Confidence intervals should be used with caution: simulations have shown that, as the slope departs from ±1, the SMA slope estimate is increasingly biased and the confidence interval includes the true value less and less often. Even when the slope is near ±1, the confidence interval is too narrow if *n* is very small or if the correlation is weak.

5. If the distribution is not bivariate normal and the data cannot be transformed to satisfy that condition (e.g. if the distribution possesses two or several modes), one should wonder whether the slope of a regression line is really an adequate model to describe the functional relationship between the two variables. Since the distribution is not bivariate normal, there seems little reason to apply models such as MA, SMA or RMA, which primarily describe the first principal component of a bivariate normal distribution. So, (1) if the relationship is linear, OLS is recommended to estimate the parameters of the regression line. The significance of the slope should be tested by permutation, however, because the distributional assumptions of the parametric test are not satisfied. (2) If a straight line is not an appropriate model, polynomial or nonlinear regression should be considered.

6. When the purpose of the study is not to estimate the parameters of a functional relationship, but simply to forecast or predict values of *y* for given *x*'s, use OLS in all cases. OLS is the only method that minimizes the squared residuals in *y*. The OLS regression line itself is meaningless. Do not use the OLS standard error and confidence bands unless *x* is known to be free of error (Sokal and Rohlf, 1995: 545, Table 14.3).

7. Observations may be compared to the predictions of a statistical or deterministic model (e.g simulation model) in order to assess the quality of the model. If the model contains random variables measured with error, use MA for the comparison when the observations and model predictions are in the same units.

If the model fits the data well, the MA slope is expected to be 1 and the intercept 0. A slope that significantly differs from 1 indicates a *difference* between observed and simulated values that is proportional to the observed values. For relative-scale variables, a MA intercept that significantly differs from 0 suggests the existence of a systematic difference between observations and simulations (Mesplé *et al.*, 1996).

8. With all methods, the confidence intervals are large when n is small; they become smaller as n goes up to about 60, after which they change much more slowly. Model II regression should ideally be applied to data sets containing 60 observations or more.

Numerical examples illustrating the cases found in Table 10.4 are described in Legendre (2008b). The data and R script are found in the help file of the ***lmodel2()*** function (Section 10.7). Other interesting examples are found in Warton *et al.* (2006).

Ecological application 10.3a

Laws & Archie (1981) re-analysed data published in two previous papers that had quantified the relationships between the log of respiration rates and the log of biomass for zooplankton under various temperature conditions. The authors of the original papers had computed OLS slopes and confidence intervals (model I regression) of the biomass-respiration relationships for each temperature condition. They had come to the conclusions (1) that the *surface law*, which states that the slope of the log-log relationship should fall between 0.66 and 1.00, was not verified by the data, and (2) that the slope significantly varied as a function of temperature. Based on the same data, Laws & Archie recomputed the slopes using the standard major axis method. They found that all slopes were larger than estimated by OLS (same phenomenon as in Fig. 10.7) and that none of them was significantly outside the 0.66 to 1.00 interval predicted by the surface law. Furthermore, comparing the slopes of the different temperature data sets at $\alpha = 0.02$, they found that they did not differ significantly from one another.

3 — Multiple linear regression

When there are several explanatory variables $x_1, x_2, \ldots, x_m$, it is possible to compute a regression equation where the response variable y is a linear function of all explanatory variables x_j. The multiple linear regression model is a direct extension of simple linear regression:

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_m x_{im} + \varepsilon_i \qquad \textbf{(10.14)}$$

for object i. Equation 10.14 leads to the well-known formula for the fitted values:

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_m x_{im} \qquad \textbf{(10.15)}$$

Using ordinary least squares (OLS), the vector of regression parameters $\mathbf{b} = [b_j]$ is easily computed from matrix eq. 2.19: $\mathbf{b} = [\mathbf{X'X}]^{-1} [\mathbf{X'y}]$. If an intercept ($b_0$) must be estimated, a column of 1's is added to matrix $\mathbf{X}$ of the explanatory variables. QR decomposition (Section 10.7) is an alternative, computer-efficient method for the computation of regression coefficients in univariate or multivariate regression.

Equation 10.15 provides a model I estimation, which is valid when the x_j variables have been measured without error. This is the only method presently available in commercial statistical packages and, for this reason, it is the multiple regression model most widely used by ecologists. McArdle (1988) proposed a multiple regression method, the *standard minor axis*, to be used when the explanatory variables of the model are random (i.e. with measurement error or natural variability). McArdle's standard minor axis is the multivariate equivalent of the standard major axis (SMA) method described in the previous subsection.

Standard minor axis

Orthogonal distance regression

Another approach is *orthogonal distance regression* (ODR), computed through generalized least squares. The method minimizes the sum of the squares of the orthogonal distances between each data point and the curve described by the model equation; this is the multivariate equivalent of the major axis regression (MA) method described in the previous subsection. ODR is used extensively in econometrics. Boggs & Rogers (1990) give entry points to the numerous papers that have been published on the subject in the computer science and econometric literature and they propose an extension of the method to nonlinear regression modelling. They also give references to ODRPACK[*], a public-domain collection of FORTRAN subprograms for *weighted orthogonal distance regression*, which allows estimation of the parameters that minimize the sum of squared weighted orthogonal distances from a set of observations to the curve or surface determined by the parameters.

When the same multiple regression model is to be computed for several response variables $y_1, \ldots, y_i, \ldots, y_p$, regression coefficients can be estimated by ordinary least squares for all response variables simultaneously, using a single matrix expression:

$$\hat{\mathbf{B}} = [\mathbf{X'X}]^{-1}[\mathbf{X'Y}]$$

Multivariate linear regression

The procedure is called *multivariate linear regression* (Finn, 1974). In this expression, which is the multivariate equivalent of eq. 2.19, $\mathbf{X}$ is the matrix of explanatory variables, $\mathbf{Y}$ is the matrix of the p response variables, and $\hat{\mathbf{B}}$ is the matrix of regression coefficients. The coefficients found using this equation are the same as those obtained from multiple regressions computed in separate runs for each response variable. The multivariate matrix of fitted values is obtained by the following matrix expression, which will serve as the basis for redundancy analysis (eq. 11.3) in Section 11.1:

$$\hat{\mathbf{Y}} = \mathbf{X}\,[\mathbf{X'X}]^{-1}\,\mathbf{X'Y} \qquad \textbf{(10.16)}$$

Two types of regression coefficients can be computed in regression analysis.

- *Ordinary regression coefficients*, represented by symbols b, are computed on the original variables. The physical dimension of coefficient b_j associated with explanatory variable x_j is (dimension of y / dimension of x_j). These regression

[*] ODRPACK is available from the following Web site: <http://www.netlib.org/odrpack/>.

coefficients are useful when the regression equation is to be used to compute estimated values of y for objects that have not been involved in the estimation of the regression parameters, and for which y and x values are available. This is the case, for instance, when a regression model is validated using a new set of observations: estimates $\hat{y}$ are computed from the regression equation to be validated, using the observed values of the explanatory variables x_j, and they are compared to the corresponding observed y's, to assess how efficient the regression model is at calculating y for new data.

- In contrast, *standard regression coefficients*, often represented by symbols b', are computed on standardized variables **X** and **y**. Standard regression coefficients are dimensionless. These regression coefficients are useful as a means of assessing the relative importance of each explanatory variables x_j included in the regression model: the variables with the highest standard regression coefficients (in absolute values) are those that contribute the most to the estimated $\hat{y}$ values. The relationship between coefficients b and b' obtained by ordinary least-squares estimation is: $b'_{yx_j} = b_{yx_j} s_{x_j} / s_y$, where b_{yx_j} is the partial regression coefficient for explanatory variable x_j.

It is interesting and important to note that, for the objects that were used to estimate the regression parameters, the fitted values $\hat{y}$ computed from the ordinary regression coefficients (hence from the original variables) are identical to the fitted values computed by regressing **y** on the standardized variables **X**.

Partial regression coefficient

Both the ordinary and standard regression coefficients in multiple regression are *partial regression coefficients*. The term *partial* means that each regression coefficient is a measure, standardized or not, of the rate of change that variable y would have per unit of variable x_j, if all the other explanatory variables in the study were held constant. The concept of partial regression is further developed in Subsection 10.3.5. Partial regression coefficients can be tested by permutation using methods similar to those described in Subsection 11.1.8 for canonical redundancy analysis (RDA).

Collinearity

When the explanatory variables x_j of the model are uncorrelated, multiple regression is a straightforward extension of simple linear regression. In experimental work, controlled variables may satisfy this condition if the experiment has been planned with care and the design is balanced. With observational data, however, the explanatory variables used in multiple regression models are most often collinear (i.e. correlated to one another), and it will be seen that strong collinearity may affect the ability to correctly estimate the regression parameters. How to deal with this problem will depend on the purpose of the analysis. If one is primarily interested in forecasting, the objective is to maximize the coefficient of multiple determination (called R^2 in multiple regression); collinearity of the explanatory variables is not a concern. For description or inference, however, the primary interest is to correctly estimate the parameters of the model; the effect of multicollinearity on the estimates of the model parameters must then be minimized.

Identify collinear variables

Prior to regression, different methods can be used to identify fully or highly collinear variables.

- One can check if the group of explanatory variables is of full rank. This can be done by singular value decomposition (SVD) of the data matrix (Section 2.11, Application 1): the matrix is not of full rank if one or more of the singular values are 0. Alternatively, one can compute the determinant of the covariance matrix of a group of variables: the determinant is 0 if the group includes variables that are linearly dependent on other variables in the group (Section 2.6, property 5; Section 2.7).

- If the rank of the matrix is smaller than its order, check subgroups of explanatory variables. Place the variables in an order that seems suitable; for example, put the most ecologically informative or easy-to-measure variables first. Compute SVD of the matrix containing the first two variables, then the first three, and so on. SVD produces a singular value of zero when a variable that is fully collinear with the previous ones is included in the group. When identified, remove the fully collinear variable from the set of explanatory variables and resume the exploration of the remaining variables.

VIF

- For variables that are not fully collinear, compute the extent to which each variable is collinear with the other variables in the group. This is done by computing *variance inflation factors* (*VIF*; Neter *et al.,* 1996, their Sections 9.5 and 10.2). Each variable *j* is regressed, in turn, on all the other variables in the group and the coefficient of determination (R_j^2, eq. 10.20) of that regression model is noted. The *VIF* for variable *j* is computed as follows:

$$VIF_j = \frac{1}{1 - R_j^2} \qquad \textbf{(10.17)}$$

All *VIF* coefficients can actually be found in a single operation by computing the inverse of the correlation matrix, $\mathbf{R}^{-1}$, among the variables in the group under study; the diagonal elements of that inverse matrix are the *VIF* coefficients. VIF_j is 1 for a variable *j* that has correlations of 0 with all the other variables in the group, and is larger than 1 when the correlations between *j* and some or all the other variables differ from 0 (positive or negative correlation values). Variables that have high *VIF* coefficients can be scrutinized and considered as candidates for elimination from the group of explanatory variables. Different cut-off values have been proposed to identify highly collinear variables: *VIF* > 5, or > 10 (Neter *et al.,* 1996), or > 20 (ter Braak & Smilauer, 2002).

The effect of collinearity on the estimates of regression parameters may be described as follows. Let us assume that one is regressing *y* on two explanatory variables x_1 and x_2. If x_1 is uncorrelated to x_2, the variables form a well-defined Cartesian plane. If *y* is represented as an axis orthogonal to that plane, a multiple linear regression equation corresponds to a plane in the three-dimensional space; this plane represents the variation of *y* as a linear function of x_1 and x_2. If x_1 is strongly correlated (i.e. collinear) to x_2, the axes of the base plane form an acute angle instead of being at right angle. In the limit situation where $r\,(x_1, x_2) = 1$, they become a single axis. With such correlated explanatory variables, the angles determined by the slope coefficients (b_1 and b_2), which set the position of the regression plane in the x_1–x_2–y space, are more likely to be unstable; their values may change depending on the random component $\varepsilon_{\rm i}$ in $y_{\rm i}$. In other words, two samples drawn from the same statistical population may be modelled by regression equations with very different parameters — even to the point that the signs of the regression coefficients may change.

Simulation is the easiest way to illustrate the effect of collinearity on the estimation of regression parameters. Vectors $\mathbf{x_1}$ and $\mathbf{x_2}$ were generated, each containing 100 random normal deviates N(0,1), and assembled into an explanatory matrix **X**. Because the data were generated at random, vectors $\mathbf{x_1}$ and $\mathbf{x_2}$ should be uncorrelated. Actually, the correlation between them was –0.002. The *control data set* was completed by computing a response variable $\mathbf{y_1}$ as the sum of $\mathbf{x_1}$ and $\mathbf{x_2}$, to which a random component was added in the form of an error term ε composed of random normal deviates N(0,2):

$$y_{1i} = x_{i1} + x_{i2} + \varepsilon_i$$

For the *test data set*, two correlated explanatory variables $\mathbf{w_1}$ and $\mathbf{w_2}$ were created by multiplying matrix **X** by the square root of a correlation matrix stating that the correlation between $\mathbf{x_1}$ and $\mathbf{x_2}$ should be 0.8:

$$\mathbf{W} = [\mathbf{w_1}, \mathbf{w_2}] = \mathbf{XR}^{0.5} \quad \text{where} \quad \mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}] \text{ and } \mathbf{R}^{0.5} = \begin{bmatrix} 1.0 & 0.8 \\ 0.8 & 1.0 \end{bmatrix}^{0.5} = \begin{bmatrix} \sqrt{0.8} & \sqrt{0.2} \\ \sqrt{0.2} & \sqrt{0.8} \end{bmatrix}$$

$\mathbf{R}^{0.5}$ is computed using eq. 2.29; Cholesky factorization (Section 2.12) of **R** may be used instead of square root decomposition. Since $\mathbf{x_1}$ and $\mathbf{x_2}$ are N(0,1) random deviates, they have expected values of 0 and are orthogonal for large n. The covariance matrix of **W** can be developed as follows, which shows that its expected value is equal to the imposed correlation matrix **R**:

$$\frac{1}{n-1}\mathbf{W'W} = \frac{1}{n-1}\mathbf{R}^{0.5}\mathbf{X'XR}^{0.5} = \mathbf{R}^{0.5}\left(\frac{1}{n-1}\mathbf{X'X}\right)\mathbf{R}^{0.5} = \mathbf{R}^{0.5}\mathbf{IR}^{0.5} = \mathbf{R}$$

For the simulated data, the correlation between $\mathbf{w_1}$ and $\mathbf{w_2}$ turned out to be 0.801, which is very close to 0.8. The test data set was completed by computing a variable $\mathbf{y_2}$ from $[\mathbf{w_1}, \mathbf{w_2}]$ with the same error term ε as in the equation for $\mathbf{y}_1$ above:

$$y_{2i} = w_{i1} + w_{i2} + \varepsilon_i$$

Each data matrix was divided into five independent groups of 20 observations each, and multiple regression equations were computed; the groups were independent of one another since the generated data were not autocorrelated. Results are shown in Table 10.5. Note the high variability of the slope estimates obtained for the test data groups (lower panel, with collinearity in the explanatory variables) compared to the control data groups (upper panel). In two cases, the signs of the regression coefficients were changed: for b_1 in group 5 and for b_2 in group 1.

Parsimony

When trying to find the 'best' possible model describing an ecological process, another important aspect is the principle of *parsimony*, also called *Ockham's razor*.

Table 10.5 Parameters of the multiple regression equations for two data sets, each divided into five groups of 20 objects. Top: control data where variables x_1 and x_2 are uncorrelated. Bottom: test data with $r(\mathbf{x_1}, \mathbf{x_2}) \approx 0.8$. Note how the range and standard deviation statistics indicate higher slope variability among the test groups (lower panel). The intercepts are the same in the two panels.

$\hat{y}_1 = b_0 + b_1x_1 + b_2x_2$ ⇒	b_0	b_1	b_2
Group 1	0.922	1.457	0.247
Group 2	0.002	–0.033	1.032
Group 3	0.494	1.264	1.206
Group 4	0.343	0.614	0.339
Group 5	0.209	0.410	1.410
Mean	0.394	0.742	0.847
Range of slope estimates = Max – Min		1.491	1.163
Standard deviation of slope estimates		0.615	0.524

$\hat{y}_2 = b_0 + b_1w_1 + b_2w_2$ ⇒	b_0	b_1	b_2
Group 1	0.922	1.988	–0.718
Group 2	0.002	–0.819	1.563
Group 3	0.494	0.985	0.855
Group 4	0.343	0.663	0.048
Group 5	0.209	–0.440	1.796
Mean	0.394	0.475	0.709
Range of slope estimates = Max – Min		2.807	2.514
Standard deviation of slope estimates		1.129	1.050

Ockham's razor

This principle, formulated by the English logician and philosopher William Ockham (1290-1349), professor at Oxford University, states that

Pluralitas non est ponenda sine necessitate

which literally translates: "Multiplicity should not be posited without necessity". In other words, unnecessary assumptions should be avoided (i.e. "shaved away") when formulating hypotheses. Following this principle, parameters should be used with parsimony in modelling, so that any parameter that does not significantly contribute to the model (e.g. by increasing the R^2 coefficient in an important way, or by decreasing *AIC*) should be eliminated. Indeed, any model containing as many parameters as the number of data points can be adjusted to perfectly fit the data. The corresponding 'cost' is that there is no degree of freedom left to test its significance, hence the 'model' cannot be extended to any other situation.

When the explanatory variables of the model are orthogonal to one another (no collinearity, for example among the controlled factors of well-planned and balanced factorial experiments), applying Ockham's razor is easy: one can remove from the model any variable whose contribution (slope parameter) is not statistically significant. Tests of significance for the partial regression coefficients (i.e. the individual *b*'s) are described in standard textbooks of statistics. The task is not that simple, however, with observational data, because these often display various degrees of collinearity. The problem is that significance may get 'diluted' among collinear variables contributing in the same way to the explanation of a response variable *y*. Consider a data set where an explanatory variable x_1 makes a significant contribution to a regression model; introducing a highly correlated copy of x_1 in the calculation is usually enough to make the contribution of each copy non-significant, simply as the result of the collinearity that exists between copies (if the second copy is a perfect copy of x_1, the regression coefficients must be computed using a generalized inverse; see Section 2.11, Application 3). Linear dependence (or full collinearity) in a group of explanatory variables is easy to detect; see *Identify collinear variables* in the margin a few pages above. Multicollinearity (without full collinearity) among explanatory variables is measured by *VIF* coefficients (eq. 10.17). Hocking (1976) compared a number of methods proposed for selecting variables in linear regression exhibiting collinearity.

Some statistical programs offer procedures that allow one to compute and compare all possible regression submodels for a small set of *k* explanatory variables. When such a procedure is not available and one does not want to manually test all possible models, heuristic methods that have been developed for selecting the 'best' subset of explanatory variables may be used, although with caution. The explanatory variables with the strongest contributions may be chosen by backward elimination, forward selection, or stepwise procedure. The three strategies do not necessarily lead to the same selection of explanatory variables.

Backward elimination

• The *backward elimination procedure* is easy to understand. All variables are initially included and, at each step, the variable that contributes the least to explaining the response variable (usually that with the smallest partial correlation) is removed, until all explanatory variables remaining in the model have a significant partial regression coefficient. Some programs express the selection criterion in terms of a *F*-to-remove (*F*-statistic for testing the significance of the partial regression coefficient) or a p-to-remove criterion (same, but expressed in terms of probability), instead of the value of the partial correlation, or else in terms of *AIC* or AIC_c (eqs. 10.22 and 10.23, below).

Forward selection

• The *forward selection procedure* starts with no explanatory variable in the model. The variable entered is the one that produces the largest increase in R^2, provided this increase is significantly different from zero using a predetermined significance level. The procedure is iteratively repeated until no more explanatory variable can be found that produces a significant increase in R^2. Calculations may be simplified by computing partial correlations for all variables not yet in the model, and only testing the significance of the largest partial correlation. Again, some programs base the final decision for including an explanatory variable on a *F*-to-enter value, which is equivalent to using the actual probability values, or on AIC or AIC_c (eqs. 10.22 and 10.23, below). The major problem with forward selection is that all variables included at previous steps are kept in the model, even though some of them may finally contribute little to the R^2 after incorporation of some other variables.

Stepwise procedure

• The latter problem may be alleviated by the *stepwise procedure,* which alternates between forward selection and backward elimination. After each step of forward inclusion, the significance of all the variables in the model is tested, and those that are not significant are excluded before the next forward selection step.

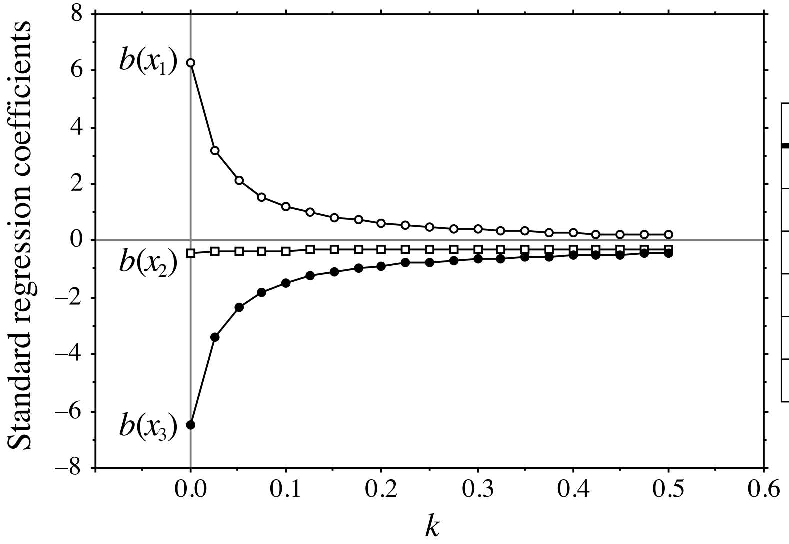
In any case, a problem common to all stepwise inclusion procedures remains: when a model with, say, k explanatory variables has been selected, the procedure offers no guarantee that there does not exist another subset of k explanatory variables, with significant partial correlations, that would explain together more of the variation of y (larger R^2) than the subset selected by stepwise procedure. Furthermore, Sokal & Rohlf (1995) warn users that, after doing repeated tests, the probability of type I error is far greater than the nominal significance value α. The stepwise approach to regression can only be recommended in empirical studies, where one must reduce the number of explanatory variables in order to simplify data collection during the next phase of field study.

There are other ways to counter the effects of multicollinearity in multiple regression. Table 10.5 shows that collinearity has the effect of inflating the variance of regression coefficients, with the exception of the intercept b_0. When the objective is forecasting or prediction, one can use regression on principal components or ridge regression, described below. These methods reduce the variance of the regression coefficients, which leads in turn to better predictions of the response variable. However, the regression coefficients they produce are biased; despite of that, they are still better estimates of the 'true' regression coefficients than those obtained by ordinary multiple regression for collinear variables. In other words, the price to pay for reducing the inflation of variance is some bias in the estimates of the regression coefficients. This may provide better forecasting or prediction than the ordinary multiple regression solution since, as a consequence of the larger variance in the regression coefficients, multicollinearity tends to increase the variance of the forecasted or predicted values (Freund & Minton, 1979).

Regression on principal components

• *Regression on principal components* consists of the following steps: (1) perform a principal component analysis on the matrix of the explanatory variables $\mathbf{X}$, (2) compute the multiple regression of y on the principal components (matrix $\mathbf{F}$, eq. 9.4) of $\mathbf{X}$ instead of the original explanatory variables, and (3) find back the contributions of the explanatory variables by multiplying matrix $\mathbf{U}$ of the eigenvectors with the vector of regression coefficients $\mathbf{c}$ of y regressed on the selected principal components (without including the intercept). One obtains a new vector $\mathbf{b}$ of contributions of the original variables to the regression equation as follows:

$$\mathbf{b}_{(m \times 1)} = \mathbf{U}_{(m \times k)} \mathbf{c}_{(k \times 1)} \tag{10.18}$$



k	R^2
0.0	0.346
0.1	0.275
0.2	0.257
0.3	0.249
0.4	0.245
0.5	0.243

Figure 10.8 'Ridge trace' diagram showing the estimates of the standard regression coefficients $b(x_1)$ to $b(x_3)$ for explanatory variables x_1 to x_3 as a function of k. Table on the right: decrease of R^2 as a function of k.

where m is the number of explanatory variables in the analysis and k is the number of principal components retained for step 3. This procedure does not necessarily resolve the problem of multicollinearity, although it is true that the regression is performed on principal components, which are not correlated to one another by definition. Consider the following case: if all m eigenvectors are kept in matrix **U** for step 3, one obtains exactly the same regression coefficients as in ordinary multiple regression. When **X** contains collinear variables, there is a gain in stability of the regression coefficients only if some of the principal components are eliminated from the computation of eq. 10.18. One may either eliminate the eigenvectors with the smallest eigenvalues or, better, use only in eq. 10.18 the principal components that significantly contribute to explain the variation of y. By doing so, the regression coefficient estimates become biased, of course. In problems involving a small number of explanatory variables, regression on principal components may be difficult to use because the number of principal components is small, so that eliminating one of them from the analysis may result in a large drop in R^2. Ecological application 12.7 provides an example of regression on principal components.

Ridge regression

• *Ridge regression*, developed by Hoerl (1962) and Hoerl & Kennard (1970a, b), approaches the problem in a different way; another important paper on the subject is

Marquardt & Snee (1975). Instead of the usual matrix eq. 2.19 $\mathbf{b} = [\mathbf{X'X}]^{-1}[\mathbf{X'y}]$, the regression coefficients are estimated using a modified equation,

$$\mathbf{b} = [\mathbf{X'X} + k\mathbf{I}]^{-1}[\mathbf{X'y}] \text{ , where } k > 0. \qquad \mathbf{(10.19)}$$

Hence, the method consists in increasing the diagonal terms (variances) of the covariance matrix $[\mathbf{X'X}]$ by a constant positive quantity k. This reduces the variance of the regression coefficients while creating a bias in the resulting estimates. So, users are left with the practical problem of choosing a value for k that is optimal in some sense. This is accomplished by computing regression coefficient estimates for a series of values of k, and plotting them (ordinate) as a function of k (abscissa); this plot is called the 'ridge trace', for historical reasons (Hoerl, 1962). After studying the plot, one chooses a value of k which is as small as possible, but large enough that the regression coefficient estimates change little after it. Since ridge regression is usually computed on standardized variables, no intercept is estimated. A number of criteria have been proposed by Obenchain (1977) to help choose the value of k. These criteria must be used with caution, however, since they often do not select the same value of k as the optimal one.

An example of a 'ridge trace' diagram is presented in Fig. 10.8. The data set consists of a response variable y and three collinear explanatory variables x_1 to x_3; their empirical correlation matrix is the following:

	y	x_1	x_2	x_3
y	1			
x_1	–0.40	1		
x_2	–0.44	0.57	1	
x_3	–0.41	**0.99**	0.56	1

Variables x_1 and x_3 are highly correlated. The leftmost regression coefficient estimates in Fig. 10.8 (for $k = 0$) are the standardized OLS multiple regression coefficients. Going from left to right in the figure, the regression coefficients stabilize after a sharp decrease or increase. One may decide that setting the cut-off point at $k = 0.2$ would be an appropriate compromise between small k and stable regression coefficients. Boudoux & Ung (1979) and Bare & Hann (1981) provide applications of ridge regression to forestry; in both papers, some regression coefficients change signs with increasing k. An application of ridge regression to modelling heterotrophic bacteria in a sewage lagoon ecosystem is presented by Troussellier *et al.* (1986, followed-up by Troussellier & Legendre, 1989).

Coefficient of determination, R^2

The coefficient of multiple determination R^2, also called the *unadjusted* coefficient of multiple determination, is the square of the multiple correlation coefficient R of Section 4.5; it varies between 0 and 1. $R^2_{\mathbf{y}|\mathbf{X}}$ measures the proportion of the variation of variable y about its mean that is explained by the linear model of the variables included in explanatory matrix **X**. As in simple linear regression, where the coefficient of determination is r^2 (eq. 10.7), $R^2_{\mathbf{y}|\mathbf{X}}$ is the regression sum of squares (SS) divided by the total sum of squares (total SS, TSS), or the one-complement of the ratio of the sum of squared residuals (residual sum of squares, RSS) to the total sum of squares (TSS):

$$R^2_{\mathbf{y}|\mathbf{X}} = \frac{\text{regression SS}}{\text{total SS}} = \frac{\Sigma\,(\hat{y}_i - \bar{y})^2}{\Sigma\,(y_i - \bar{y})^2} = 1 - \frac{\text{RSS}}{\text{TSS}} \qquad \textbf{(10.20)}$$

The expected value of R^2 in a regression involving m random predictors is not 0 but $m/(n-1)$, as explained below. As a consequence, if **X** contains $m = (n-1)$ predictors that are linearly unrelated to the response variable y, for example m columns of random numbers, $R^2 = 1$ even though the explanatory variables explain none of the variation of y. For that reason, R^2 cannot be interpreted as a correct (i.e. unbiased) estimate of the proportion of variation of y explained by **X**.

Three useful statistics can, however, be derived from R^2. They serve distinct purposes in regression analysis.

Adjusted R^2

1. The *adjusted coefficient of multiple determination* R^2_a or *adjusted* R^2 (Ezekiel, 1930), provides an unbiased estimate of the proportion of variation of y explained by **X**. The formula takes into account the numbers of degrees of freedom (d.f.) of the numerator and denominator portions of R^2:

$$R^2_a = 1 - \frac{\text{residual mean square}}{\text{total mean square}} = 1 - (1 - R^2_{\mathbf{y}|\mathbf{X}})\left(\frac{\text{total d.f.}}{\text{residual d.f.}}\right) \qquad \textbf{(10.21)}$$

- In ordinary multiple regression, the total degrees of freedom of the F-statistic are $(n-1)$ and the residual d.f. are $(n-m-1)$, where n is the number of observations and m is the number of explanatory variables in the model (eq. 4.40).

- In multiple regression through the origin, where the intercept is forced to zero, the total degrees of freedom of the F-statistic are n and the residual d.f. are $(n-m)$.

These same degrees of freedom are used in eq. 10.21. The logic of this adjustment is the following: in ordinary multiple regression, a random predictor explains on average a proportion $1/(n-1)$ of the response's variation, so that m random predictors explain together, on average, $m/(n-1)$ of the response's variation; in other words, the expected value of R^2 is $E(R^2) = m/(n-1)$. Applying eq. 10.21 to that value, where all predictors are random, gives $R^2_a = 0$. In regression through the origin, a random predictor explains on average a proportion $1/n$ of the response's variation, so that m random predictors explain together, on average, m/n of the response's variation, and $R^2 = m/n$. Applying eq. 10.21 to that case gives, again, $R^2_a = 0$.

R^2_a is a suitable measure of goodness of fit for comparing the success of regression equations fitted to different data sets, with different numbers of objects and explanatory variables. Using simulated data with normal error, Ohtani (2000) has shown that R^2_a is an unbiased estimator of the contribution of a set of random predictors **X** to the explanation of *y*. This adjustment may be too conservative when $m > n/2$ (Borcard *et al*., 2011); this is a rule of thumb rather than a statistical principle.

With real matrices of random variables (defined at the beginning of Section 10.3), when the explanatory variables explain no more of the response's variation than the same number of variables containing random numbers, the value of R_a^2 is near zero; it can be negative on occasion. Contrary to R^2, R_a^2 does not necessarily increase with the addition of explanatory variables to the regression model if these explanatory variables are linearly unrelated to *y*. R_a^2 is a better estimate of the population coefficient of determination ρ^2 than R^2 (Zar, 1999, Section 20.3) because it is unbiased.

Healy (1984) pointed out that Ezekiel's (1930) adjusted R^2 equation (R_a^2, eq. 10.21) makes sense and should be used when **X** contains observed values of random variables. That is not the case for ANOVA fixed factors, which can be used in a multiple regression equation when they are recoded into binary dummy variables or Helmert contrasts (Subsection 1.5.7).

In canonical analysis (Chapter 11), the canonical R^2 is called the *bimultivariate redundancy statistic* (Miller & Farr, 1971), *canonical coefficient of determination*, or *canonical* R^2. Using numerical simulations, Peres-Neto *et al*. (2006) have shown that, in redundancy analysis (RDA, Section 11.1), for normally distributed data or Hellinger-transformed species abundances, the adjusted canonical R^2 (R_a^2, eq. 11.5), obtained by applying eq. 10.21 to the canonical R^2 ($R_{\mathbf{Y}|\mathbf{X}}^2$, eq. 11.4), produces unbiased estimates of the contributions of the variables in **X** to the explanation of a response matrix **Y**, just as in multiple regression. With simulated data, they also showed the artificial increase of R^2 as the number of unrelated explanatory variables in explanatory matrix **X** increases.

AIC, AIC$_c$

2. The *Akaike Information Criterion* (*AIC*) is a measure of the goodness of fit of the data to an estimated statistical model (Akaike, 1974). When comparing linear regression models, *AIC* is computed as follows (RSS, TSS: see eq. 10.20):

$$AIC = n\ \log_e\left(\frac{\text{RSS}}{n}\right) + 2k \qquad \textbf{(10.22)}$$

where k is the number of parameters, including the intercept, in the regression equation. Independence of the observations is assumed in the calculation of *AIC*, as well as normality of the residuals and homogeneity of their variances. The following formula is also found in the literature: $AIC = n\ \log_e((1-R^2)/n) + 2k$. A constant, $n\log_e(\text{TSS})$, must be added to this formula to obtain eq. 10.22. Since *AIC* is used to compare different models of the same response data, either formula will identify the same model as the one that minimizes *AIC*.

The corrected form of *AIC*, abbreviated AIC_c (Hurvich & Tsai, 1993), is *AIC* with a second-order correction for small sample size:

$$AIC_c = AIC + \frac{2k(k+1)}{n-k-1} \qquad \textbf{(10.23)}$$

Burnham & Anderson (2002) strongly recommend using AIC_c rather than AIC when n is small or k is large. Because AIC_c converges towards AIC when n is large, AIC_c should be used with all sample sizes.

The AIC_c statistic is not the basis for a test of significance. It plays a different role than the F-test (below): it is used to compare models. For a given data set, several competing models may be ranked by AIC_c. The model with the *smallest* value of AIC_c is the best-fitting one, i.e. the most likely for the data. For example, in selection of explanatory variables, the model for which AIC_c is minimum is retained.

F-statistic

3. The F-statistic (see eq. 4.40) serves as the basis for the test of significance of the coefficient of multiple determination, R^2. A parametric test can be used if the regression residuals are normal. Otherwise, a permutation test should be used.

F-statistic for nested models

There is another way of comparing models statistically, but it is limited to nested models of the same response data. A model is nested in another if it contains one or several variables less than the reference model. The method consists in calculating the R^2 of the two linear models and computing a F-statistic to test the difference in R^2 between them. The F-statistic is computed as follows for two nested models, the most inclusive containing m_2 variables and the model nested into it containing m_1 variables:

$$F = \frac{(R^2_{y.1\ldots m_2} - R^2_{y.1\ldots m_1}) / (m_2 - m_1)}{(1 - R^2_{y.1\ldots m_2}) / (n - m_2 - 1)}$$

The difference in R^2 is tested for significance parametrically with $\nu_1 = (m_2 - m_1)$ and $\nu_2 = (n - m_2 - 1)$ degrees of freedom, or by permutation. This method can be used in forward selection or backward elimination. It is implemented, for example, in functions ***ordiR2step()*** of VEGAN and ***forward.sel()*** of PACKFOR (Subsection 11.1.10, paragraph 7), which can be used in models involving a single response variable y.

As a final note, it is useful to remember that several types of explanatory variables can be used in multiple regression, besides quantitative variables:

Dummy variable regression

- Binary descriptors can be used as explanatory variables in multiple regression. This means that multistate qualitative variables can also be used, insofar as they are recoded into binary dummy variables, as described in Subsection 1.5.7*. This case is referred to as *dummy variable regression*.

- Geographic information may be used in multiple regression models in different ways. On the one hand, latitude (Y) and longitude (X) information form perfectly valid quantitative descriptors if they are recorded as axes of a Cartesian plane. Geographic data in the form of degrees-minutes-seconds should, however, be recoded to decimal

* In R, qualitative multistate descriptors used as explanatory variables are automatically recoded into dummy variables by function ***lm()*** if they are identified as *factors* in the data frame.

form before they are used as explanatory variables in regression. The X and Y coordinates may be used either alone, or in the form of a polynomial (X, Y, X^2, XY, Y^2, etc.). Regression using such explanatory variables is referred to as *trend surface analysis* in Chapter 13. Spatial eigenfunctions, described in Chapter 14, are more sophisticated descriptions of geographic relationships among study sites; they can also be used as explanatory variables in regression.

- If replicate observations are available for each site, the grouping of observations, which is also a kind of geographic information, may be used in multiple regression as a qualitative multistate descriptor, recoded into a set of dummy variables.

- Finally, any analysis of variance may be reformulated as a linear regression analysis; actually, linear regression and ANOVA both belonging to the General Linear Model. Consider one-way ANOVA for instance: the classification criterion can be written as a multistate qualitative variable and, as such, recoded as a set of dummy variables (Subsection 1.5.7) on which multiple regression may be performed. The analysis of variance table obtained by multiple regression is identical to that produced by ANOVA. This equivalence is discussed in more detail by ter Braak & Looman (1987) in an ecological framework. Draper & Smith (1981) and Searle (1987) discuss in some detail how to apply multiple regression to various analysis of variance configurations. ANOVA by regression can be extended to cross-factor (two-way or multiway) ANOVA. How to carry out these analyses is described in Subsection 11.1.10, point 4, for the more general analysis of multivariate response data **Y** (MANOVA).

4 — Polynomial regression

Several solutions have been proposed to the problem of fitting, to a response variable y, a nonlinear function of a single explanatory variable x. An elegant and easy solution is to use a polynomial of x, whose terms are treated as so many explanatory variables in a multiple regression procedure. In this approach, y is modelled as a polynomial function of x:

Polynomial model

$$\hat{y} = b_0 + b_1 x + b_2 x^2 + \ldots + b_k x^k \qquad \textbf{(10.24)}$$

Such an equation is linear in its parameters (if one considers the terms $x^2, \ldots, x^k$ as so many explanatory variables), although the modelled response of y to the explanatory variable x is nonlinear. The degree of the equation, which is its highest exponent, determines the shape of the curve: each degree above 1 (straight line) and 2 (concave up or down) adds an inflexion point to the curve. Increasing the degree of the equation always increases its adjustment to the data (R^2). If one uses as many parameters b (including the intercept b_0) as there are data points, one can fit the data perfectly ($R^2 = 1$). However, the cost of that perfect fit is that there are no degrees of freedom left to test the relationship and, therefore, the “model” cannot be extended to other situations. Hence, a perfectly fitted model is useless. In any case, a high-degree polynomial would be of little interest in view of the principle of parsimony (Ockham’s razor) discussed in Subsection 10.3.3, which states that the best model is the simplest one that adequately describes the relationship. Each term of a polynomial expression is called a *monomial*.

Monomial

So, the problem left to ecologists is to find the most parsimonious polynomial equation that adequately fits the data. The methods for selecting variables, described above for multiple regression, may be used to profit here.

One can start with a polynomial equation of degree k (e.g. $k = 4$) and use a selection procedure, based on AIC, to determine which subset of the monomials produces the most parsimonious model. Backward, forward or stepwise procedures can be applied. One could add the following constraint: that all monomials in the final model be significant, e.g. at level $\alpha = 0.05$. It may turn out that some higher-degree monomials are retained by the selection procedure, and are significant, whereas some of the lower-order monomials are excluded; this is entirely permissible. Beware: in some statistical packages, selection of monomials in polynomial regression only removes higher-degree monomials; monomials of degrees lower than k cannot be removed if x^k is retained in the model. These procedures do not produce a parsimonious model in cases where some lower-degree terms should be eliminated.

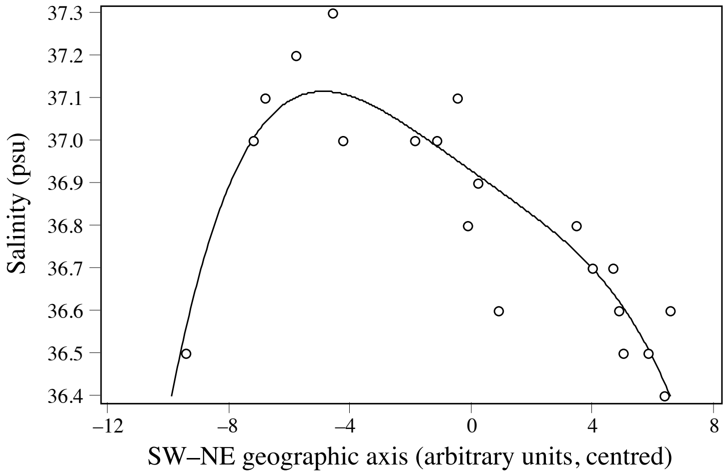
The successive terms of an ordinary polynomial expression are collinear. Starting for instance with a variable x made of the successive integers 1 to 10, variables x^2, x^3, and x^4 computed from it display the following correlations:

	x	x^2	x^3	x^4
x	1			
x^2	0.975	1		
x^3	0.928	0.987	1	
x^4	0.882	0.961	0.993	1

The problem of multicollinearity is severe with such data. Centring variable x on its mean before computing the polynomial is good practice. It reduces the linear dependency of x^2 on x (it actually eliminates it when the x values are at perfectly regular intervals, as in the present example), and somewhat alleviates the problem for the higher terms of the polynomial. This may be enough when the objective is descriptive. If, however, it is important to estimate the exact contribution (standard regression coefficient) of each term of the polynomial in the final equation, the various monomials (x, x^2, etc.) should be made orthogonal to one another before computing the regression equation. Orthogonal monomials may be obtained, for example, through the Gram-Schmidt procedure described in Table 9.5 and in textbooks of linear algebra for instance Lipschutz (2009); see function ***poly()*** in Section 10.7.

Orthogonal monomials

Numerical example. Data from the ECOTHAU program (Ecology of the Thau lagoon, southern France; Amanieu *et al.*, 1989) are used to illustrate polynomial regression. Salinity (response variable y) was measured at 20 sites in the brackish Thau lagoon (Mediterranean Sea) on 25 October 1988. The lagoon is elongated in a SW-NE direction. The explanatory variable x



Regression equation:

	Regr. coeff.	Standard coeff.
b_0	36.929	0
$b_1(x)$	–0.0509	–0.9699
$b_2(x^4)$	–0.0001	–0.7261

Figure 10.9 Polynomial regression line describing the structure of salinity (psu: practical salinity units) in the Thau lagoon (Mediterranean Sea) along its main geographic axis on 25 October 1988.

is the projection of the positions of the sampling sites on the long axis of the lagoon, as determined by principal component analysis of the site coordinates. Being a principal component, variable x is centred. The other terms of an ordinary 6th-degree polynomial were computed from it. After stepwise selection, the model with the lowest AIC_c contained variables x, x^4 and x^5 ($AIC_c = –80.845$, $R_a^2 = 0.815$); the regression parameters for x^4 and x^5 were not significant at the 0.05 level. Then, all possible models involving x, x^2, x^3, x^4 and x^5 were computed. The model with the largest number of significant regression coefficients contained variables x and x^4 ($AIC_c = –80.374$, $R_a^2 = 0.792$). These results indicate that the model with three monomials (x, x^4 and x^5) is slightly better in terms of AIC_c and is thus the best-fitting model for the data. The line fitted to the second model, which is more parsimonious with only two explanatory variables (x and x^4), is shown in Fig. 10.9).

5 — Partial linear regression and variation partitioning

There are situations where two or more complementary sets of hypotheses may be invoked to explain the variation of an ecological variable. For example, the abundance of a species could vary as a function of biotic and abiotic factors. Regression modelling may be used to study one set of factors, or the other, or the two sets together. In most if not all cases involving field data (by opposition to experimental designs), there are correlations among variables across the two (or more) explanatory data sets. Partial regression is a way of estimating how much of the variation of the response variable can be attributed exclusively to one set once the effect of the other has been taken into account and controlled for. The purpose may be to estimate the amount of variation that can be attributed exclusively to one or the other set of explanatory variables and the amount explained jointly by the two explanatory data sets, or else to estimate the vector of fitted values corresponding to the exclusive effect of one set of variables. When the objective is simply to assess the unique contribution of each explanatory variable, there is no need for partial regression analysis: the coefficients of multiple regression of the standardized variables already provide that information since they are *standard partial regression coefficients*.

Consider three data sets. Vector **y** is the response variable whereas matrices **X** and **W** contain the explanatory variables. Assume that one wishes to model the relationship between **y** and **X**, while controlling for the effects of the variables in matrix **W**, which is called the *matrix of covariables*. The roles of **X** and **W** could of course be inverted.

Matrix of covariables

Variation partitioning

Variation partitioning consists in apportioning the variation* of variable **y** among two or more explanatory data sets. This approach was first proposed by Mood (1969, 1971) and further developed by Borcard *et al.* (1992) and Peres-Neto *et al.* (2006). The method is described here for two explanatory data sets, **X** and **W**, but it can be extended to more explanatory matrices. When **X** and **W** contain random variables (defined at the beginning of Section 10.3), adjusted coefficients of determination (R_a^2, eq. 10.21) are used to compute the fractions following the method described below. Ordinary R^2 (eq. 10.20) are used instead of R_a^2 when **X** and **W** represent ANOVA fixed factors coded into binary dummy variables or Helmert contrasts (Subsection 1.5.7).

Figure 10.10 sets a nomenclature, [a] to [d], for the fractions of variation that can be identified in **y**. Kerlinger & Pedhazur (1973) called this form of analysis "commonality analysis" by reference to the common fraction of variation (fraction [b] in Fig. 10.10) that two sets of explanatory variables may explain jointly. Partial regression assumes that the effects are linear and additive. There are two ways of carrying out the partitioning computations, depending on whether one wishes to obtain vectors of fitted values corresponding to fractions of variation, or simply estimate the amounts of variation corresponding to the fractions. In the description that follows, the fractions of variation are computed from R_a^2 statistics.

(1) If one is interested in obtaining a partial regression equation and computing a vector of partial fitted values, one first computes the residuals of **y** on **W** (noted $\mathbf{y}_{\mathrm{res|}\mathbf{W}}$) and the residuals of **X** on **W** (noted $\mathbf{X}_{\mathrm{res|}\mathbf{W}}$):

Residuals of **y** on **W**: $\mathbf{y}_{\mathrm{res|}\mathbf{W}} = \mathbf{y} - \mathbf{W}\,[\mathbf{W'W}]^{-1}\mathbf{W'}\;\mathbf{y}$

Residuals of **X** on **W**: $\mathbf{X}_{\mathrm{res|}\mathbf{W}} = \mathbf{X} - \mathbf{W}\,[\mathbf{W'W}]^{-1}\mathbf{W'}\;\mathbf{X}$

In both cases, the regression coefficients are computed here through eq. 2.19 in which **X** is replaced by **W**. QR decomposition (see Section 10.7), which is also used in some

* The term *variation*, a less technical and looser term than *variance*, is used because one is partitioning the total sum of squared deviations of **y** from its mean (total SS). In variation partitioning, there is no need to divide the total SS of **y** by its degrees of freedom to obtain the variance s_y^2 (eq. 4.3).

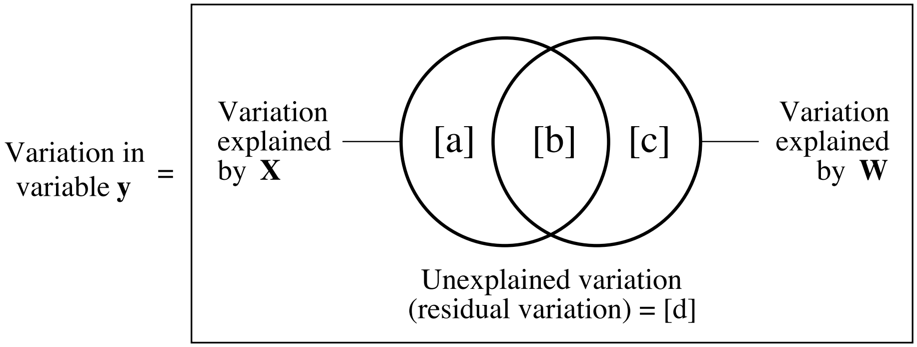


Figure 10.10 Partition of the variation of a response variable **y** among two sets of explanatory variables, **X** and **W**. The rectangle represents 100% of the variation in **y**. Fraction [b] is the intersection (*not* the interaction) of the variation explained by linear models of **X** and **W**. Adapted from Legendre (1993).

situations in Subsection 11.1, e.g. in Table 11.5, offers another way of computing regression equations.

Then, two computation methods are available: one can either

(1.1) regress $\mathbf{y}_{\text{res}|\mathbf{W}}$ on $\mathbf{X}_{\text{res}|\mathbf{W}}$,

(1.2) or regress **y** on $\mathbf{X}_{\text{res}|\mathbf{W}}$. The same partial regression coefficients are obtained in both cases, as will be verified in the numerical example below. Between calculation methods, the vectors of fitted values only differ by the values of the intercepts. The R^2 of analysis 1.1 is the partial R^2 whereas that of analysis 1.2 is the semipartial R^2; their square roots are the partial and semipartial correlation coefficients (Box 4.1).

(2) If one is interested in estimating the fractions resulting from partitioning the variation of vector **y** among the explanatory data sets **X** and **W**, there is a simple way to obtain the information, considering the ease with which multiple regressions can be computed using R or commercial statistical packages:

- Compute the multiple regression of **y** against **X** and **W** together. The corresponding R_a^2 measures the fraction of information [a + b + c], which is the sum of the fractions of variation [a], [b], and [c] defined in Fig. 10.10. For the example data set (below), $R^2 = 0.5835$, so $R_a^2 = 0.3913$ = [a + b + c]. The vector of fitted values corresponding to fraction [a + b + c], which is required to plot Fig. 10.13 (below), is also computed.

- Compute the multiple regression of **y** against **X**. The corresponding R_a^2 measures [a + b], which is the sum of the fractions of variation [a] and [b]. For the example data,

$R^2 = 0.4793$, so $R_a^2 = 0.3817$ = [a + b]. The vector of fitted values corresponding to fraction [a + b], which is required to plot Fig. 10.13, is also computed.

- Compute the multiple regression of **y** against **W**. The corresponding R_a^2 measures [b + c], which is the sum of the fractions of variation [b] and [c]. For the example data, $R^2 = 0.3878$, so $R_a^2 = 0.2731$ = [b + c]. The vector of fitted values corresponding to fraction [b + c], which is required to plot Fig. 10.13, is also computed.

- If needed, fraction [d] may be computed by subtraction. For the example, it is equal to 1 – [a + b + c], or 1 – 0.3913 = 0.6087.

As explained in Subsection 10.3.3, the adjusted *R*-square, R_a^2 (eq. 10.21), is an unbiased estimator of the real contribution of a set of random variables **X** to the explanation of **y**. Following Peres-Neto *et al*. (2006), the values of the individual fractions [a], [b], and [c] must be computed by combining the R_a^2 values obtained from the three multiple regressions that produced fractions [a + b + c], [a + b], and [b + c]:

- fraction [a] is computed by subtraction, using the R_a^2 values: [a] = [a + b + c] – [b + c];

- likewise, fraction [c] is computed by subtraction, using the R_a^2 values: [c] = [a + b + c] – [a + b];

- fraction [b] is also obtained by subtraction, using the R_a^2 values, in the same way as the quantity B used for comparing two qualitative descriptors in Section 6.2:

[b] = [a + b] + [b + c] – [a + b + c] *or* [b] = [a + b] – [a] *or* [b] = [b + c] – [c]

Negative [b]

Fraction [b] may be negative. As such, it is not a rightful measure of variance; this is another reason why it is referred to by the looser term *variation*. A negative fraction [b] indicates that two variables (or groups of variables **X** and **W**), together, explain **y** better than the sum of the individual effects of these variables. This can happen: see Numerical examples 2 and 3. Fraction [b] is the *intersection* of the variation explained by linear models of **X** and **W**. It is *not an interaction* in the ANOVA sense.

Vectors of fitted values corresponding to fractions [a] and [c] can be computed using partial regression, as explained above, while the vector of residuals of the regression equation that uses all predictors corresponds to fraction [d]. No fitted vector can be estimated for fraction [b], however, because no partial regression model can be written for that fraction. No degrees of freedom are attached to fraction [b]; hence [b] cannot be tested for significance.

Selection of explanatory variables

If a selection procedure (backward, forward, stepwise; Subsection 10.3.3) is used, it must be applied to data matrices **X** and **W** separately, before partitioning, in order to preserve fraction [b] of the partition. Applying the selection to matrices **X** and **W** combined could result in the elimination of variables from one or both matrices because they are correlated with variables in the other matrix, thereby reducing or eliminating fraction [b].

Table 10.6 Data collected at 20 sites in the Thau coastal lagoon on 25 October 1988. There are two bacterial response variables (Bna and Ma), three environmental variables (NH_4, phaeopigments, and bacterial production), and three spatial variables (the X and Y geographic coordinates measured with respect to arbitrary axes and centred on their respective means, plus the quadratic monomial X^2). The variables are further described in the text. The code names of these variables in the present section are y, x_1 to x_3, and w_1 to w_3, respectively.

Site	Bna	Ma	NH_4	Phaeo. *a*	Prod.	X	Y	X^2
No.		y	x_1	x_2	x_3	w_1	w_2	w_3
1	4.615	10.003	0.307	0.184	0.274	–8.75	3.7	76.5625
2	5.226	9.999	0.207	0.212	0.213	–6.75	2.7	45.5625
3	5.081	9.636	0.140	0.229	0.134	–5.75	1.7	33.0625
4	5.278	8.331	1.371	0.287	0.177	–5.75	3.7	33.0625
5	5.756	8.929	1.447	0.242	0.091	–3.75	2.7	14.0625
6	5.328	8.839	0.668	0.531	0.272	–2.75	3.7	7.5625
7	4.263	7.784	0.300	0.948	0.460	–1.75	0.7	3.0625
8	5.442	8.023	0.329	1.389	0.253	–0.75	–0.3	0.5625
9	5.328	8.294	0.207	0.765	0.235	0.25	–1.3	0.0625
10	4.663	7.883	0.223	0.737	0.362	0.25	0.7	0.0625
11	6.775	9.741	0.788	0.454	0.824	0.25	2.7	0.0625
12	5.442	8.657	1.112	0.395	0.419	1.25	1.7	1.5625
13	5.421	8.117	1.273	0.247	0.398	3.25	–4.3	10.5625
14	5.602	8.117	0.956	0.449	0.172	3.25	–2.3	10.5625
15	5.442	8.487	0.708	0.457	0.141	3.25	–1.3	10.5625
16	5.303	7.955	0.637	0.386	0.360	4.25	–5.3	18.0625
17	5.602	10.545	0.519	0.481	0.261	4.25	–4.3	18.0625
18	5.505	9.687	0.247	0.468	0.450	4.25	–2.3	18.0625
19	6.019	8.700	1.664	0.321	0.287	5.25	–0.3	27.5625
20	5.464	10.240	0.182	0.380	0.510	6.25	–2.3	39.0625

Numerical example 1. The example data set (Table 10.6) is from the ECOTHAU research program mentioned in the numerical example of Subsection 10.3.4 (Amanieu *et al.*, 1989). It contains two bacterial variables (Bna, the concentration of colony-forming units of aerobic heterotrophs growing on bioMérieux nutrient agar, with low NaCl concentration; and Ma, the concentration of aerobic heterotrophs growing on marine agar with a salt content of 34 gL^{-1}); three environmental variables (NH_4 in the water column, in $\mu molL^{-1}$; phaeopigments from degraded chlorophyll *a*, in μgL^{-1}; and bacterial production, determined by incorporation of tritiated thymidine in bacterial DNA, in $nmolL^{-1}d^{-1}$); and three spatial variables of the sampling sites on the nodes of an arbitrarily located grid (the X and Y geographic coordinates, in km, each centred on its mean, and the quadratic monomial X^2, which was found to be important for explaining the response variables). All bacterial and environmental variables were log-transformed using $\log_e(x + 1)$. One of the bacterial variables, Ma, is used here as the response variable $\mathbf{y}$; the three environmental variables form the matrix of explanatory variables $\mathbf{X}$; the three spatial variables make up matrix $\mathbf{W}$ of the covariables. Table 10.6 will be used again in

Section 13.4. A multiple regression of **y** against **X** and **W** together was computed first as a reference. The regression equation was the following:

$$\hat{y} = 9.64 - 0.90x_1 - 1.34x_2 + 0.54x_3 + 0.10w_1 + 0.14w_2 + 0.02w_3$$
$$(R^2 = 0.5835;\ R_a^2 = 0.3913 = [\mathrm{a + b + c}])$$

The adjusted coefficient of determination (R_a^2) is an unbiased estimate of the proportion of the variation of **y** explained by the regression model containing the 6 explanatory variables; it corresponds to fraction [a+b+c] in the partitioning table below and to the sum of fractions [a], [b] and [c] in Fig. 10.10. The vector of fitted values was also computed; after centring, this vector will be plotted as fraction [a + b + c] in Fig. 10.13. Since the total sum of squares in **y** is 14.9276 [$\mathrm{SS} = s_y^2 \times (n-1)$], the R^2 allowed the computation of the sum of squares corresponding to the vector of fitted values: SS($\hat{\mathbf{y}}$) =14.9276 × 0.5835 = 8.7109. This value can also be obtained by computing directly the sum of squared deviations about the mean of the values in the fitted vector $\hat{\mathbf{y}}$.

For calculation of the partial regression equation using method 1.1, the residuals[*] of the regression of **y** on **W** were computed. One way is to use the following equation, which requires adding a column of “1” to matrix **W** in order to estimate the regression intercept:

$$\mathbf{y}_{\mathrm{res|}\mathbf{W}} = \mathbf{y} - \mathbf{W}\,[\mathbf{W}'\mathbf{W}]^{-1}\mathbf{W}'\ \mathbf{y}$$

The residuals of the regressions of **X** on **W** were computed in the same way:

$$\mathbf{X}_{\mathrm{res|}\mathbf{W}} = \mathbf{X} - \mathbf{W}\,[\mathbf{W}'\mathbf{W}]^{-1}\mathbf{W}'\ \mathbf{X}$$

Then, vector $\mathbf{y}_{\mathrm{res|}\mathbf{W}}$ was regressed on matrix $\mathbf{X}_{\mathrm{res|}\mathbf{W}}$ with the following result:

regression equation: $\hat{y} = 0 - 0.90x_{r(\mathbf{W})1} - 1.34x_{r(\mathbf{W})2} + 0.54x_{r(\mathbf{W})3}$ $(R^2 = 0.3197)$

The value $R^2 = 0.3197$ is the partial R^2. In its calculation, the denominator is the sum of squares corresponding to fractions [a] and [d], as shown for the partial correlation coefficient in Box 4.1.

For calculation through method 1.2, **y** was regressed on matrix $\mathbf{X}_{\mathrm{res|}\mathbf{W}}$ with the following result:

regression equation: $\hat{y} = 8.90 - 0.90x_{r(\mathbf{W})1} - 1.34x_{r(\mathbf{W})2} + 0.54x_{r(\mathbf{W})3}$ $(R^2 = 0.1957)$

The value $R^2 = 0.1957$ is the semipartial R^2. The semipartial R^2 is the square of the semipartial correlation defined in Box 4.1. It represents the fraction of the total variation of **y** explained by the partial regression equation because, in its calculation, the denominator is the total sum of squares of the response variable **y**, [a+b+c+d]. That value is shown in the variation partitioning table below, but it will not be used to compute the individual fractions of variation.

Note that the three regression coefficients for the three *x* variables in the last equation are exactly the same as in the two previous equations; only the intercepts differ. This gives substance to the statement of Subsection 10.3.3 that regression coefficients obtained in multiple

[*] In the R language, regression residuals can be computed using ***residuals(lm())***.

linear regression are *partial regression coefficients* in the sense of the present subsection. Between calculation methods, the vectors of fitted values only differ by the value of the intercept of the regression of **y** on $\mathbf{X}_{\text{res}|\mathbf{W}}$, 8.90, which is also the mean of **y**. The centred vector of fitted values will be plotted as fraction [a] in Fig. 10.13.

The calculation of partial regression can be done in the opposite way, regressing **y** on **W** while controlling for the effects of **X**. First, $\mathbf{y}_{\text{res}|\mathbf{X}}$ and $\mathbf{W}_{\text{res}|\mathbf{X}}$ were computed. Then, for method 1.1, $\mathbf{y}_{\text{res}|\mathbf{X}}$ was regressed on $\mathbf{W}_{\text{res}|\mathbf{X}}$ with the following result:

regression equation: $\hat{y} = 0 + 0.10w_{r(\mathbf{X})1} + 0.14w_{r(\mathbf{X})2} + 0.02w_{r(\mathbf{X})3}$ $(R^2 = 0.2002)$

where $R^2 = 0.2002$ is the partial R^2. For method 1.2, **y** was regressed on $\mathbf{W}_{\text{res}|\mathbf{X}}$ with the following result:

regression equation: $\hat{y} = 8.90 + 0.10w_{r(\mathbf{X})1} + 0.14w_{r(\mathbf{X})2} + 0.02w_{r(\mathbf{X})3}$ $(R^2 = 0.1043)$

where $R^2 = 0.1043$ is the semipartial R^2, shown in the variation partitioning table below, but not used to compute the individual fractions of variation.

Again, the three regression coefficients in these partial regression equations are exactly the same as in the first regression equation of this example; only the intercepts differ. Between calculation methods, the vectors of fitted values only differ by the value of the intercept of the regression of **y** on $\mathbf{X}_{\text{res}|\mathbf{W}}$, 8.90, which is also the mean of **y**. The centred vector of fitted values will be plotted as fraction [c] in Fig. 10.13.

To estimate fraction [a + b] of Fig. 10.10, the multiple regression of **y** on the three original (non-residualized) variables in **X** was computed. The regression equation was:

$$\hat{y} = 10.20 - 0.93x_1 - 2.02x_2 + 0.89x_3 \qquad (R^2 = 0.4793;\ R_a^2 = 0.3817 = [\text{a + b}])$$

The value $R_a^2 = 0.3817$ is an unbiased estimate of the fraction of the variation of **y** accounted for by the linear model of the three explanatory variables **X**. The vector of fitted values was computed; after centring, this vector will be plotted as fraction [a + b] in Fig. 10.13.

To obtain fraction [b + c] of Fig. 10.10, the multiple regression of **y** on the three original (non-residualized) variables in **W** was computed. The regression equation was:

$$\hat{y} = 8.32 + 0.09w_1 + 0.10w_2 + 0.03w_3 \qquad (R^2 = 0.3878;\ R_a^2 = 0.2731 = [\text{b + c}])$$

The value $R_a^2 = 0.2731$ is the unbiased estimation of the fraction of the variation of **y** accounted for by the linear model of the three explanatory variables **W**. The vector of fitted values was computed; after centring, this vector will be plotted as fraction [b + c] in Fig. 10.13.

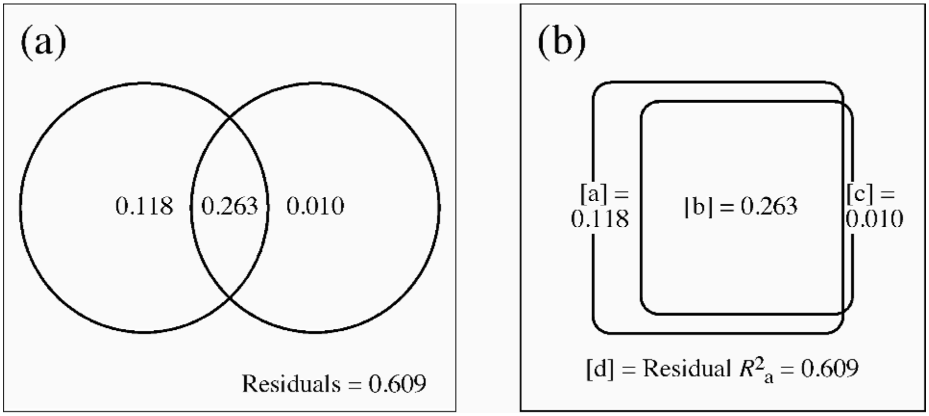


Figure 10.11 Venn diagram illustrating the results of variation partitioning of the numerical example. (a) Diagram drawn by the plotting function ***plot.varpart()*** of the VEGAN package. The circles are of equal sizes despite differences in the corresponding R_a^2. (b) Prior to publication of the partitioning results, the diagram can be redrawn, here using rounded rectangles, to better represent the relative fraction sizes with respect to the size of the outer rectangle, which represents the total variation in the response data. The fractions are identified by letters [a] to [d]; the value next to each identifier is the adjusted R^2 (R_a^2). Rectangle sizes are approximate.

Following the fraction nomenclature convention set in Fig. 10.10, the variation partitioning results were assembled in the following table (rounded values):

Fractions of variation	**Sums of squares (SS)**	**Proportions of variation of y (R^2)**	**Adjusted R^2 (R_a^2)**
[a + b]	7.1547	0.4793	0.3817
[b + c]	5.7895	0.3878	0.2731
[a + b + c]	8.7109	0.5835	0.3913
[a]	2.9213	*0.1957*	0.1183
[b]	4.2333	*0.2836*	0.2634
[c]	1.5562	*0.1043*	0.0097
Residuals = [d]	6.2167	*0.4165*	0.6087
[a + b + c + d]	14.9276	1.0000	1.0000

The partitioning results are illustrated as Venn diagrams in Fig. 10.11*. In Chapter 11, Fig. 11.6 shows partitioning results for multivariate response data involving three explanatory matrices.

* A Venn diagram with proportional circle and intersection sizes can be obtained using function ***venneuler()*** of the same-name package (Section 10.7).

As mentioned at the beginning of this subsection, and following Peres-Neto *et al*. (2006), when **X** and **W** contain random variables, R_a^2 values corresponding to [a + b + c], [a + b], and [b + c] are used to compute, by subtraction, the fractions [a] to [d] shown in column 4 of the table. R_a^2 provides unbiased estimates of the contributions of the explanatory data sets **X** and **W** to **y** when **X** and **W** contain random variables. The adjusted fractions [a], [b], and [c] cannot be directly computed using the non-adjusted fractions computed from non-adjusted R^2 coefficients, shown in italics in the 3*rd* column. When n is small as in this example, the estimated fractions computed from R_a^2 may be very different from the fractions computed from R^2 values.

Ordinary R^2 (3*rd* column) are used to compute the fractions (values in italics) when **X** and **W** represent ANOVA fixed factors coded into dummy variables. When these values are required, they can be calculated by subtraction from the R^2 values in the first three rows of the table: R^2[a] = R^2[a+b+c] – R^2[b+c] = 0.1957 (which is equal to the R^2 of the partial regression equation computed above through method 1.2); R^2[c] = R^2[a+b+c] – R^2[a+b] = 0.1043 (which is equal to the R^2 of the partial regression equation computed above through method 1.2); R^2[b] = R^2[a+b] + R^2[b+c] – R^2[a+b+c] = 0.2836 (this value can only be obtained by subtraction). The sums of squares in the 2nd column of the table are obtained by multiplying these R^2 values by the total sum of squares in **y**, which is 14.9276.

The *partial correlation coefficient* between **y** and matrix **X** while controlling for the effect of **W** can be obtained from the values [a] and [d] in the column "Sums of squares" of the table, as explained in Box 4.1 of Section 4.5:

$$r_{\mathbf{yX.W}} = \sqrt{\frac{[\mathrm{a}]}{[\mathrm{a+d}]}} = \sqrt{\frac{2.9213}{2.9213 + 6.2167}} = 0.5654$$

This value is not the same as the *semipartial* R^2, which is computed as follows (Box 4.1):

$$r_{\mathbf{y(X.W)}} = \sqrt{\frac{[\mathrm{a}]}{[\mathrm{a+b+c+d}]}} = \sqrt{\frac{2.9213}{14.9276}} = 0.4424$$

Tests of significance of the fractions

If the conditions of homoscedasticity and normality of the residuals are satisfied, the fractions (with the exception of [b]) can be tested for significance through parametric tests. For fractions [a + b + c], [a + b], and [b + c], one can use the results of the parametric tests produced by the statistical software. For fractions [a] and [c], one must construct a F-statistic as in eq. 11.22, using the sum of squares corresponding to fraction [a] (symbol: SS[a]) or [c] (symbol: SS[c]) in the numerator, and the residual sum of squares corresponding to [d] (symbol: SS[d]) in the denominator, together with appropriate numbers of degrees of freedom. The test statistic for fraction [a], for example, is constructed as follows:

$$F_{[\mathrm{a}]} = \frac{\mathrm{SS}[\mathrm{a}]/m}{\mathrm{SS}[\mathrm{d}]/(n-m-q-1)}$$

where m is the number of explanatory variables in set **X** and q is the number of covariables in set **W**. In the parametric framework, the statistic is tested against the

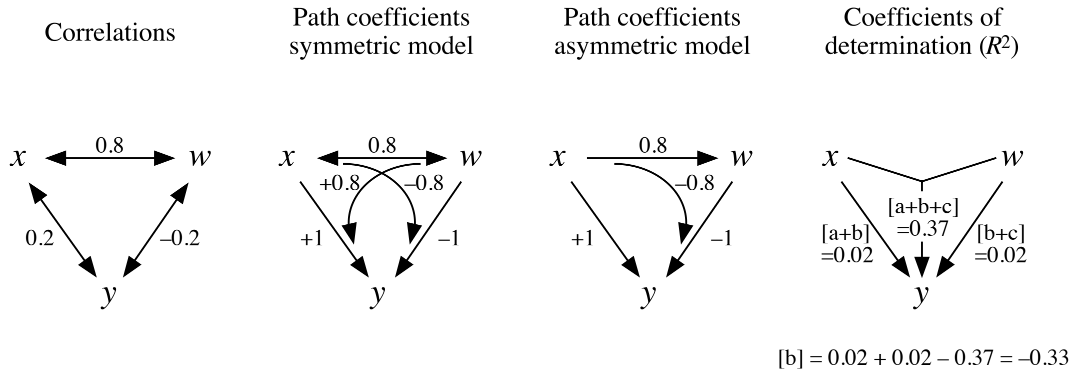

Correlations
Path coefficients symmetric model
Path coefficients asymmetric model
Coefficients of determination (R^2)
x
w
y
0.8
0.2
–0.2
+0.8
–0.8
+1
–1
[a+b+c] =0.37
[a+b] =0.02
[b+c] =0.02
[b] = 0.02 + 0.02 – 0.37 = –0.33

Figure 10.12 Correlations, path coefficients, and coefficients of determination for Numerical example 2.

F–distribution with m and $(n - m - q - 1)$ degrees of freedom. An example of that F-statistic for the test of partial or semipartial correlation coefficients is given in Box 4.1 for the simple case where there is a single variable in **X** and **W**.

If the conditions of homoscedasticity or normality of the residuals are not satisfied, one can use permutation tests to obtain p-values. Permutation of the raw data is used to test fractions [a + b + c], [a + b], and [b + c]. To test fractions [a] and [c], permutation of the residuals of a null or full model should be used (Anderson & Legendre, 1999). These permutation methods are described in Subsection 11.1.8.

Numerical example 2. This example illustrates the appearance of a negative fraction [b] when there are strong direct effects of opposite signs of x and w on y and a strong correlation between x and w (non-orthogonality). For three variables measured over 50 objects, the following correlations are obtained: $r(x, w) = 0.8$, $r(y, x) = 0.2$ and $r(y, w) = –0.2$; y, x, and w have the same meaning as in the previous numerical example. $r(y, x)$ and $r(y, w)$ are not statistically significant at the $\alpha = 0.05$ level. Referring to Section 10.4, one may use path analysis to compute the direct and indirect causal covariation relating the explanatory variables x and w to the response variable y. One can also compute the coefficient of determination of the model $y = f(x, w)$; its value is $R^2 = 0.40$. From these values, the partition of the variation of y can be achieved: R^2 of the whole model = 0.40, R_a^2 = [a + b + c] = 0.37447; $r^2(w, y) = 0.04$, R_a^2 = [a + b] = 0.02; $r^2(x, y) = 0.04$, R_a^2 = [b + c] = 0.02. Hence, [b] = [a + b] + [b + c] – [a + b + c] = –0.33447, [a] = [a + b] – [b] = 0.35447, and [c] = [b + c] – [b] = 0.35447. How is that possible?

Carrying out path analysis (Fig. 10.12), and assuming a symmetric model of relationships (i.e. w affects x and x affects w), the direct effect of x on y, p_{xy} = 1.0, is positive and highly significant, but it is counterbalanced by a strong negative indirect covariation of –0.8 going through w. In the same way, $p_{wy} = –1.0$ (which is highly significant), but this direct effect is counterbalanced by a strong positive indirect covariation of +0.8 going through x. As a result, and although they both have the maximum possible value of 1.0 for direct effects on the response variable y, both w and x turn out to have non-significant total correlations with y. In the present variation partitioning model, this translates into small adjusted amounts of explained variation $[a + b] = 0.02$ and $[b + c] = 0.02$, and a negative value for fraction [b]. If an asymmetric model of relationship had been assumed (e.g. w affects x but x does not affect w), essentially the same conclusion would have been reached from path analysis.

Numerical example 3. Another situation can give rise to a negative fraction [b], i.e. when there is no linear correlation between y and one of the explanatory variables, e.g. $r(y, x) = 0.0$, but the other two correlations differ from 0, e.g. $r(y, w) = 0.5$ and $r(x, w) = 0.5$. For this example, assuming again $n = 50$, we find $[a + b + c] = 0.30497$, $[a + b] = –0.02083$, and $[b + c] = 0.23438$ (computed from the R_a^2 coefficients), so that $[b] = –0.09142$. The partial explanation of the variation of y provided by x, estimated by the partial regression or partial correlation coefficient, is not zero and may be significant in the statistical sense: using path analysis (Section 10.4) for this example, the direct effect of x on y is $p_{xy} = –0.33333$ (p = 0.019, which is significant) and the indirect effect is 0.33333, these two effects summing to zero. The direct effect of w on y is $p_{wy} = 0.66667$ and its indirect effect is –0.16667. The negative [b] fraction indicates that x and w, together, explain the variation of y better than the sum of the individual effects of these variables. The signs of the regression coefficients (path coefficients) actually vary depending on the signs of the correlations $r(y, w)$ and $r(x, w)$.

The above decomposition of the variation of a response vector **y** between two sets of explanatory variables **X** and **W** was described by Whittaker (1984) for the simple case where there is a single regressor in each set **X** and **W**. Whittaker showed that the various fractions of variation may be represented as vectors in space, and that the value of fraction [b] [noted G(12:) by Whittaker, 1984] is related to the angle θ between the two regressors through the following formula:

$$1 - 2\cos^2(\theta/2) \le [b] \le 2\cos^2(\theta/2) - 1 \qquad \textbf{(10.25)}$$

Fraction [b] for orthogonal regressors

θ is related to the coefficient of linear correlation (eq. 10.4). This formula has three interesting properties. (1) If the two regressors are orthogonal ($r = 0$), then $2\cos^2(\theta/2) = 1$, so that $0 \le [b] \le 0$ and consequently $[b] = 0$. Turning the argument around, the presence of a non-zero fraction [b] indicates that the two explanatory variables are not orthogonal. There are also instances where [b] is zero with two non-orthogonal regressors; a simple example is when the two regressors are uncorrelated with **y** and explain none of its variation. (2) If the two regressors are identical, or at least pointing in the same direction ($\theta = 0°$), then $–1 \le [b] \le 1$. It follows that the proportion of variation of **y** that is accounted for by either regressor (fraction [b]) may be, in some cases, as large as 1, i.e. 100%. (3) The formula allows for negative values of [b], as shown in Numerical example 2.

In conclusion, fraction [b] represents the fraction of variation of **y** that may indifferently be attributed to **X** or **W**. The interpretation of a negative [b] is that the two processes, represented in the analysis by data sets **X** and **W**, are competitive; in other words, they have opposite effects, one process hindering the contribution of the other in the joint regression model. One could use eq. 6.15, $S = [b]/[a + b + c]$, to quantify how similar **X** and **W** are in explaining **y**. Whittaker (1984) also suggested that if **X** and **W** represent two factors of an experimental design, [b] may be construed as a

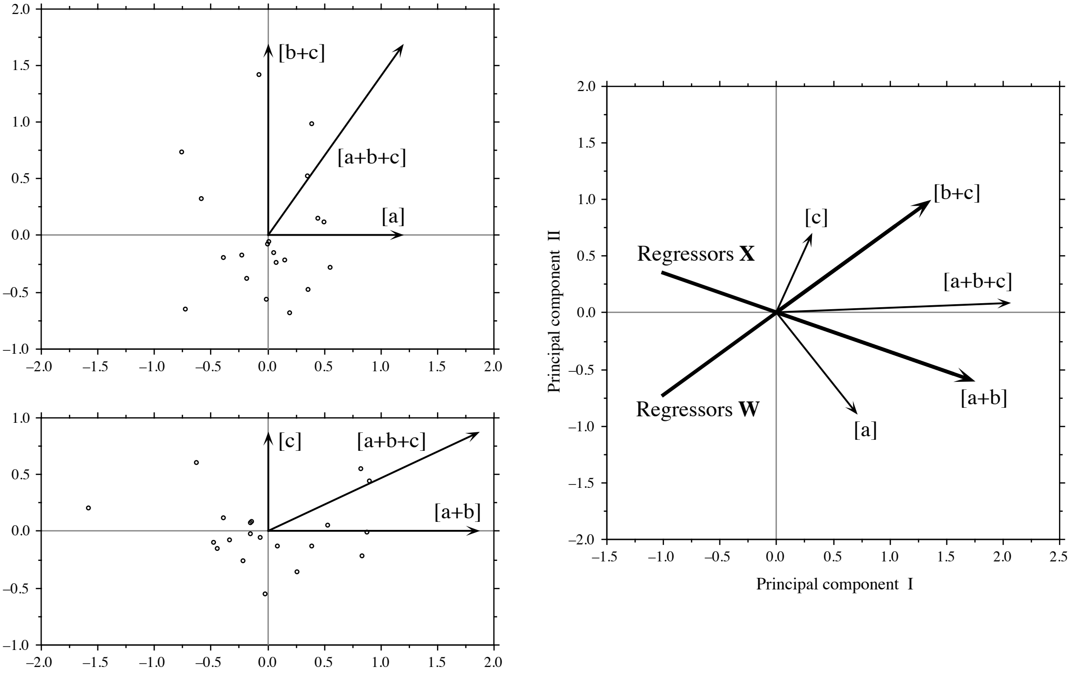


Figure 10.13 Numerical example of partial regression analysis: representation of the fitted vectors in regression space. Vectors are represented with lengths proportional to their standard deviations. Upper left: scatter diagram of objects along orthogonal vectors [a] and [b + c]. Vector [a + b + c], also shown, is obtained by adding vectors [a] and [b + c]. Lower left: same for orthogonal vectors [c] and [a + b]. Right: all five fitted vectors are represented in a compromise plane obtained by principal component analysis (PCA axes I and II, which explain 96.7% of the variation). [a] is still orthogonal to [b + c] in three-dimensional space, and [c] to [a + b]; these orthogonal relationships are slightly deformed by the projection in two dimensions.

measure of the effective balance (i.e. orthogonality) of the design; [b] is 0 in a balanced crossed design.

Whittaker's representation may be used even when regressors **X** and **W** are multivariate data sets. Figure 10.13 illustrates the angular relationships among the fitted vectors corresponding to the fractions of variation of Numerical example 1. One plane is needed for vectors {[a], [b + c], and [a + b + c]} in which [a] is orthogonal and additive to [b + c]; another plane is needed for vectors {[c], [a + b], and [a + b + c]} where [c] is orthogonal and additive to [a + b]. However, the sets {[a], [b + c]} and {[c], [a + b]} belong to different planes, which intersect along vector [a + b + c]; so, the whole set of fitted vectors is embedded in a three-dimensional space when there are two explanatory data sets; this is independent of the number of variables in each set. The vector of residuals corresponding to fraction [d] is orthogonal to all the fitted vectors and lies in a fourth dimension. Whittaker (1984) gives examples involving more than two explanatory data sets. The graphical representation of the partitioned fitted vectors in such cases requires spaces with correspondingly more dimensions.

Ecological application 10.3b

Birks (1996) used partial regressions to analyse the mountain plant species richness in 75 grid squares covering Norway (109 species in total), in order to test whether the nunatak hypothesis was necessary to explain the present distribution of these plants. The nunatak, or refugial hypothesis, holds that apparent anomalies in present-day species distributions are explained by survival through glaciations on ice-free mountain peaks or rocky outcrops, called ‘nunataks’ (from Inuit *nunataq*), projecting above continental glaciers. Implicit in this hypothesis is that the presumed refugial species have poor dispersal ability. According to the nunatak hypothesis, one would expect a concentration of rare plants in the glacial refuges or their vicinity. Hence, a variable describing unglaciated areas (3 abundance classes for occurrence of presumed unglaciated areas) was introduced in the analysis to represent “history”. The alternative hypothesis, called “tabula rasa”, holds that present-day distributions are well-explained by the environmental control model (Whittaker, 1956; Bray & Curtis, 1957). To materialise this hypothesis in the analysis, Birks used 10 explanatory variables that described bedrock geology, geography, topography, and climate. “Geography” was introduced in the analysis in the form of a third-degree polynomial of the geographic coordinates, which allowed a representation of the geographic variation of species richness by a cubic trend surface of latitude and longitude, as explained in Subsection 13.2.1; the terms of the polynomial representing latitude and longitude^2 were retained by a forward selection procedure.

(1) Birks (1996) first used a form of stepwise multiple regression, adding variables in a specified order, to determine the importance of unglaciated areas in explaining mountain plant species richness. In the “ecology first” analysis, history (i.e. variable “unglaciated areas”) was introduced last in the analysis; it added about 0.1% to the explained variation, whereas the environmental variables explained together 84.9% of the variation. In the “history first” analysis, history was entered first; it only explained 7.6% of the variation, which was not a significant contribution. (2) The contribution of “history” did not improve in partial regression analyses, when controlling for either land area per grid square alone, or land area, latitude and longitude. Modern ecological variables such as bedrock geology, climate, topography, and geography were considerably more effective explanatory variables of species richness than “history”. (3) In order to find out whether “history” made a unique statistically significant contribution to the variation of the species richness when the effects of the other variables were controlled for, Birks computed variation partitioning, described above, after partial regression analyses, using non-adjusted R^2 coefficients. Fraction [a], corresponding to the influence of all environmental variables independent of “history”, explained 77.4% of the variation of species richness; fraction [b], in which “environment” covaried with “history”, explained 7.5%; fraction [c], “history” independent of “environment”, explained 0.1%; the unexplained variation, fraction [d], was 15.0%. Fraction [b] is likely to result from the spatial coincidence of unglaciated areas with high elevation, western coastal areas, and certain types of bedrock, all these being included among the environmental variables.

In another paper, Birks (1993) used partial canonical correspondence analysis, instead of partial regression analysis, to carry out the same type of analysis (including variation decomposition) on a matrix of grid cells × species presence/absence. Again, the results suggested that there was no statistically significant contribution from unglaciated areas in explaining present-day distribution patterns when the effects of modern topography, climate, and geology were considered first.

These two papers (Birks, 1993, 1996) show that the hypothesis of survival in glacial nunataks is unnecessary to explain the present-day patterns of species distribution and richness of Norwegian mountain plants. Following Ockham's razor principle (Subsection 10.3.3), this unnecessary assumption should be avoided when formulating hypotheses intended to explain present-day species distributions.

6 — Nonlinear regression

Logistic equation

In some applications, ecologists know from existing theory the algebraic form of the nonlinear relationship between a response variable and one or several explanatory variables. An example is the logistic equation, which describes population growth in population dynamics:

$$N_t = \frac{K}{1 + \mathrm{e}^{(a - rt)}} \qquad \textbf{(10.26)}$$

This equation gives the population size (N_t) of a species at time t as a function of time (t). The equation contains three parameters a, r, and K, which are adjusted to the data; r is the Malthus parameter describing the natural rate of increase of the population, and K is the support capacity of the ecosystem. Nonlinear regression allows one to estimate the parameters (a, r, and K in this example) of the curve that best fits the data, for a user-selected function. This type of modelling does not assume linear relationships among the variables; the equation to be fitted is provided by the user. The algorithm for nonlinear parameter estimation tries to minimize an objective function.

The most usual objective functions to minimize are (1) the usual least-squares criterion $\Sigma(y_i - \hat{y}_i)^2$ and (2) the sum of squared Euclidean distances of the points to the regression function. These two criteria are illustrated in Fig. 10.6. The parameters of the best-fitting equation are found by iterative adjustment; users usually have the choice among a variety of rules for stopping the iterative search process. Common choices are: when the improvement in R^2 becomes smaller than some preselected value, when some preselected maximum number of iterations is reached, or when the change in all parameters becomes smaller than a given value. Useful references on this topic are Hollander & Wolfe (1973), Ratkowsky (1983), Ross (1990), Huet *et al.* (1992), and Bates & Chambers (1992). Nonlinear regression is available in several statistical packages, including R (see Section 10.7).

Consider the Taylor equation relating the means $\bar{y}$ and variances s_y^2 of several groups of data:

$$s_{y_k}^2 = a\bar{y}_k^b \qquad \textbf{(1.17)}$$

One must decide whether the equation should be fitted to the data by nonlinear regression, or to the corresponding logarithmic form (eq. 1.18) by linear regression. Look at the data in the original mean-variance space and in the transformed log(mean)-log(variance) space, and choose the form for which the data are homoscedastic.

Other often-encountered functions are the exponential, hyperbolic, Gaussian, and trigonometric (for periodic phenomena; see Subsection 12.4.5), and other growth models for individuals or populations.

Monotone regression

As an alternative to linear or nonlinear regression, Conover (1980, his Section 5.6) proposed *monotone regression* which may be used when (1) the relationship is monotonic (increasing or decreasing), (2) the purpose is forecasting or prediction rather than parameter estimation, and (3) one does not wish to carefully model the functional relationship; see also Iman & Conover (1983, their Section 12.6). Monotone regression consists in assigning ranks to the x and y observations and computing a linear regression on these ranks. Simple, natural rules are proposed to reassign real-number values to the forecasted/predicted values obtained from the rank-based equation for given values of x. Monotone regression is sometimes called *nonparametric regression*. A specialized form of monotone regression is used in nMDS algorithms (Section 9.4).

7 — Logistic regression

Binary variables form an important category of response variables that ecologists may wish to model. In process studies, one may wonder whether a given effect will be present under a variety of circumstances. Population ecologists are also often interested in determining the factors responsible for the presence or absence of a species. When the explanatory variables of the model are qualitative, modelling may call upon log-linear models computed on multiway contingency tables (Section 6.3). When the explanatory variables are quantitative, or represent a mixture of quantitative and qualitative data, logistic regression is the approach of choice.

In logistic regression, the response variable is binary (presence-absence, or 1-0; see example below). A linear model of quantitative explanatory variables would necessarily produce some forecasted/predicted values larger than 1 and some values smaller than 0. Consider Fig. 10.14, which illustrates the example developed below. A linear regression line fitting the data points would have a positive slope and would span outside the vertical [0, 1] interval, so that the equation would forecast ordinate values smaller than 0 (for small x) and larger than 1 (for large x); these would not make sense since the response variable can only be 0 or 1.

If one tries to predict the *probability* of occurrence of an event (for example the presence of a species), instead of the event itself (0 or 1 response), the model should be able to produce real-number values in the range [0, 1]. The logistic equation (eq. 10.26) described in Subsection 10.3.6 provides a sigmoid model for such a

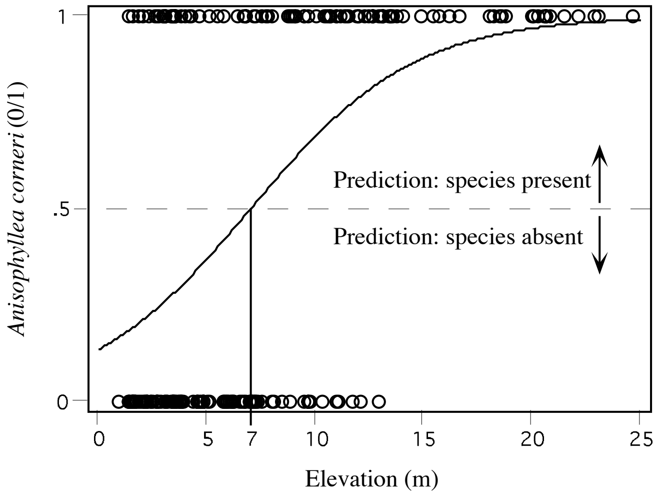


Figure 10.14 Logistic regression equation fitted to presence/absence of *Anisophyllea corneri*, as a function of elevation, in 200 forest quadrats.

response between limit values (Fig. 10.14). It is known to adequately model several ecological, physiological and chemical phenomena. Since the extreme values of the probabilistic response to be modelled are 0 and 1, then $K = 1$, so that eq. 10.26 becomes:

$$\mathrm{p} = \frac{1}{1 + \mathrm{e}^{-z}} \qquad \textbf{(10.27)}$$

where p is the probability of occurrence of the event. z is a linear function of the explanatory variable(s):

$$z = b_0 + b_1 x \qquad \text{for a single predictor } x \qquad \textbf{(10.28a)}$$

or $$z = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p \qquad \text{for several predictors} \qquad \textbf{(10.28b)}$$

Note that there are other, equivalent algebraic forms for the logistic equation. A form equivalent to eq. 10.27 is: $\mathrm{p} = \mathrm{e}^z/(1 + \mathrm{e}^z)$.

For the error part of the model, the ε_i values cannot be assumed to be normally distributed and homoscedastic, as it is the case in linear regression, since the response variable can only take two values (presence or absence). The binomial distribution is the proper model in such a case, or the multinomial distribution for multistate qualitative response variables, as allowed in some computer software (e.g. CATMOD in

SAS). The parameters of the model cannot be estimated by ordinary least-squares since the error term is not normally distributed. This is done instead by maximum likelihood. Logistic regression is a special case of the generalized linear model (GLM: McCullagh & Nelder, 1983; Section 10.7); least-squares regression is another special case of GLM. According to the *maximum likelihood principle*, the best values for the parameters of a model are those for which the likelihood is maximum. The likelihood L of a set of parameter estimates is defined as the probability of observing the values that have actually been observed, given the model and the parameter estimates. This probability, which is not the same as p in eq. 10.27, is expressed as a function of the parameters:

Maximum likelihood

$$L = \mathrm{p}(\text{observed data} \mid \text{model, parameters})$$

So, one iteratively searches for parameter estimates that maximize the likelihood function.

Numerical example. Data describing the structure of a tree community, sampled over a 50–ha plot in the Pasoh forest[*], Malaysia, were studied by He *et al.* (1994, 1996, 1997). The plot was established to monitor long-term changes in a primary tropical forest. The precise locations of the 334 077 individual trees and shrubs at least 1 cm in diameter at breast height (dbh) were determined (825 species in total) and a few environmental variables were recorded at the centres of 20 × 20 m quadrats. The present example uses the presence or absence of a species, *Anisophyllea corneri* Ding Hou (Cucurbitales), in each quadrat. One hundred quadrats were selected at random in the plot among those where *A. corneri* was present, and 100 among the quadrats where it was absent, for a total of 200 quadrats. Results of the logistic regression study presented below were reported by He *et al.* (1997).

Stepwise logistic regression was used to model the presence or absence of the species with respect to *slope* and *elevation* (i.e. altitude in metres measured by reference to the lowest part of the forest plot floor), using the SPSS software package. Following the calculations, *elevation* was included in the model for its significant contribution, whereas *slope* was left out. The linear part of the fitted model (eq. 10.28a) was:

$$z = -1.8532 + 0.2646 \times \textit{elevation}$$

Significance of the regression coefficients was tested using the Wald statistic, which is the square of the ratio of a regression coefficient to its standard error; this statistic is distributed like χ^2. Both the intercept and slope coefficients of the model were significant ($p < 0.001$).

As explained above, the probability of the observed values of the response variable, for given values of the parameters, is called the likelihood. Since a probability is in the range [0, 1], its natural logarithm is a negative number. It is customary to multiply it by –2 to obtain $-2\log_e(L)$, noted –2LL, a positive number that measures of how poorly a model fits the data; –2LL = 0 represents a perfect fit. This value presents the advantage of being distributed like χ^2,

[*] The Pasoh forest is one of the CTFS permanent forest plots. See note on these forest plots in Subsection 6.5.3.

so that it can be tested for significance. The significance of the model was tested using the following table:

	χ^2	ν	p (χ^2)
Intercept only	277.259	199	0.0002
Difference	59.709	1	< 0.0001
Intercept + *elevation*	217.549	198	0.1623
Difference	1.616	1	0.2057
Intercept + *elevation* + *slope*	215.933	197	0.1690
Goodness of fit	183.790	198	0.7575

Parameters were added to the model, one by one, as long as they improved the fit. The procedure is the same as in log-linear models (e.g. Table 6.6).

- For a model with an intercept only, –2LL = 277.259. The hypothesis to be tested was that –2LL = 277.259 was not significantly different from 0, which would be the value of –2LL for a model fitting the data perfectly. Degrees of freedom were computed as the number of observations (200) minus the number of fitted parameters (a single one up this point). The significant χ^2 statistic ($p < 0.05$) indicated that the model did not fit the data well.

- Inclusion of *elevation* added a second parameter to the model; this parameter was fitted iteratively and the resulting value of –2LL was 217.549 at convergence, i.e. when –2LL did not change by more than a small preselected value. Since the probability associated with the χ^2 statistic was large, the null hypothesis that the model fitted the data could not be rejected. The difference in χ^2 between the two models (277.259 – 217.549 = 59.709) was tested with 1 degree of freedom. The significant probability ($p < 0.05$) showed that *elevation* brought a significant contribution to the likelihood of the model.

- Inclusion of *slope* added a third parameter to the model. The resulting model also fitted the data well ($p > 0.05$), but the difference in χ^2 between the two models (217.549 – 215.933 = 1.616) was not significant ($p = 0.2057$), indicating that *slope* did not significantly contribute to increase the likelihood of the model. Hence, *slope* was left out of the final model.

The last row of the table tested a goodness-of-fit statistic that compared the observed values (0 or 1 in logistic regression) to the probabilities forecasted by the model, which included the intercept and *elevation* in this example (Norusis, 1990, p. 52). The statistic (183.790) is distributed like χ^2 and has the same number of degrees of freedom as the χ^2 statistic for the complete model. In the present example, this statistic was not significant ($p > 0.05$), which led to conclude that there was no significant discrepancy between the forecasted values and the data.

Putting back the observed values of the explanatory variable(s) into the model (eq. 10.28a) provided estimates of z. For instance, one of the quadrats in the example data had *elevation* = 9.5 m, so that

$$z = -1.8532 + 0.2646 \times 9.5 = 0.6605$$

Incorporating this value into eq. 10.27 provided the following probability that *A. corneri* would be present in the quadrat:

$$p = \frac{1}{1 + e^{-0.6605}} = 0.659$$

Since $p > 0.5$, the forecast was that the species should be found in this quadrat. In general, if $p < 0.5$, the event is unlikely to occur whereas it is likely to occur if $p > 0.5$. (Flip a coin if a forecasted value is required in a case where $p = 0.5$ exactly.) With the present equation, the breaking point between forecasted values of 0 and 1 (i.e. the point where $p = 0.5$) corresponded to an *elevation* of 7 m. The logistic curve fitted to the *A. corneri* data is shown in Fig. 10.14.

Classification table

Forecasted values may be used to produce a classification (or "confusion") table, as in linear discriminant analysis (Section 11.3), in which the forecasted values are compared to observations. For the example data, the classification table was:

Observed	*Forecasted*		*Percent*
	0	1	*correct*
0	78	22	78%
1	35	65	65%
		Total correct classification	71.5%

Since most values are in the diagonal cells of the table, one concludes that the logistic regression equation based solely on elevation was successful at forecasting the presence of *A. corneri* in the quadrats.

Gaussian logistic model

A Gaussian logistic equation may be used to model the unimodal response of a species to an environmental gradient. Fit the logistic equation with a quadratic response function $z = b_0 + b_1x + b_2x^2$, instead of eq. 10.28a, to obtain a Gaussian logistic model; the response function for several predictors (eq. 10.28b) may be modified in the same way. See ter Braak & Looman (1987) for details.

Linear discriminant analysis (Section 11.3) has often been used by ecologists to study niches of plants or animals, before logistic regression became widely available in computer packages. Williams (1983) gives examples of such works. The problem with discriminant analysis is that it constructs a linear model of the explanatory variables, so that the forecasted values are not limited to the [0, 1] range. Negative values and values higher than 1 can be produced, which are ecologically unrealistic for presence-absence data. This problem does not appear with logistic regression, which is available in major statistical packages as well as in S-PLUS®, MATLAB® and R. This question is further discussed in Section 11.6.

In procedure CATMOD of SAS, the concept of logistic regression is extended to multi-state qualitative response variables. Trexler & Travis (1993) provide an application of logistic regression to an actual ecological problem, including selection of the most parsimonious model; they also discuss the relative merits of various alternatives to the logistic model.

8 — Splines and LOWESS smoothing

There are instances where one is only interested in estimating an empirical relationship between two variables, without formally modelling the relationship in an equation and estimating its parameters. In such instances, smoothing methods may be the most appropriate, since they provide an empirical representation of the relationship, efficiently and at little cost in terms of time spent specifying a model. Since they fit the data locally (i.e. within small windows), smoothing methods are useful when the relationship greatly varies in shape along the abscissa. This is the opposite of the parametric regression methods, where a single set of parameters is used to adjust the same function to all data points (global fit). Smoothing methods are far less sensitive to exceptional values and outliers than regression, including polynomial regression. Several numerical methods are available for smoothing.

Moving average

A simple way to visualize an empirical relationship is the method of moving averages, described in more detail in Section 12.2. Define a 'window' of a given width, position it at one of the margins of the scatter diagram, and compute the mean ordinate value (y) of all the observations in the window. Move the window by small steps along the abscissa, recomputing the mean every time, until the window reaches the opposite margin of the scatter diagram. Plot the window means as a function of the positions of the window centres along the abscissa. Link the mean estimates by line segments. This empirical line may be used to estimate y as a function of x.

Piecewise polynomial fitting by "splines" is a more advanced form of local smoothing. In its basic form, spline estimation consists in dividing the range of the explanatory variable x (which is also the width of the scatter diagram) into a number of intervals, which are generally of equal widths and separated by *knots*, and adjusting a polynomial of order k to the data points within each segments using polynomial regression (Subsection 10.3.4). To make sure that the transitions between spline segments are smooth at the junction points (knots), one imposes two constraints: (1) that the values of the function be equal on the left and right of the knots, and (2) that the (k–1) first derivatives of the curves be also equal on the left and right of the knots. Users of the method have to make arbitrary decisions about (1) the level k of the polynomials to be used for regression (a usual choice is cubic splines) and (2) the number of segments along the abscissa. If a large enough number of intervals is used, the spline function can be made to fit every data point. A smoother curve is obtained by using fewer knots. It is recommended to choose the interval width in such a way as to have at least 5 or 6 data points per segment (Wold, 1974). Knots should be positioned at or near inflexion points, where the behaviour of the curve changes (see example below). A large body of literature exists about splines. Good introductory texts are Chambers (1977), de Boor (1978), Eubank (1988), and Wegman & Wright (1983). The simplest text is Montgomery & Peck (1982, Section 5.2.2); it inspired the explanation of the method that follows.

When the positions of the knots are known (i.e. decided by users), a cubic spline model *with no continuity restriction* is written as:

$$\hat{y} = \sum_{j=0}^{3} b_{0j}x^j + \sum_{k=1}^{h}\sum_{j=0}^{3} b_{kj}(x-t_k)_+^j \tag{10.29}$$

In this equation, the parameters b_{0j} in the first sum correspond to a cubic polynomial equation in x. The parameters b_{kj} in the second sum allow the curve segments to be disconnected at the positions of the knots. There are h knots, and their positions along the abscissa are represented by t_k; the knots are ordered in such a way that $t_1 < t_2 < \ldots < t_h$. This equation, written out in full, is the following for a single knot (i.e. $h = 1$) located at position t:

$$\hat{y} = b_{00} + b_{01}x + b_{02}x^2 + b_{03}x^3 + b_{10}(x-t)_+^0 + b_{11}(x-t)_+^1 + b_{12}(x-t)_+^2 + b_{13}(x-t)_+^3$$

The expression $(x-t_k)_+$ takes the value $(x-t_k)$ when $x-t_k > 0$ (i.e. if the given value x is to the right of the knot), and 0 when $x-t_k \le 0$ (for values of x on the knot or to the left of the knot). The constraint of continuity is implemented by giving the value zero to all terms b_{kj}, except the last one. In eq. 10.29, it is these parameters that allow the relationship to be described by discontinuous curves; by removing them, eq. 10.29 becomes a cubic splines equation with continuity constraint:

Cubic splines

$$\hat{y} = \sum_{j=0}^{3} b_{0j}x^j + \sum_{k=1}^{h} b_k(x-t_k)_+^3 \tag{10.30}$$

which has a single parameter b_k for each knot. Written in full, eq. 10.30 is the following for two knots (i.e. $h = 2$) located at positions $t_1 = -5$ and $t_2 = +4$, as in the numerical example below:

$$\hat{y} = b_{00} + b_{01}x + b_{02}x^2 + b_{03}x^3 + b_1(x+5)_+^3 + b_2(x-4)_+^3$$

This approach is not the one used in advanced spline smoothing packages because it has some numerical drawbacks, especially when the number of knots is large. It is, however, the most didactic, because it shows spline smoothing to be an extension of OLS polynomial regression. Montgomery & Peck (1982) give detailed computational examples and show how to test the significance of the difference in R^2 between models with decreasing numbers of knots, or between a spline model and a simple polynomial regression model. They finally show that *piecewise linear regression* — that is, fitting a continuous series of straight lines through a scatter of points — is a natural extension of the spline eq. 10.30 in which the exponent is limited to 1.

LOWESS

LOWESS refers to ***Lo****cally* ***We****ighted* ***S****catterplot* ***S****moothing* (Cleveland, 1979). This method is an extension of moving averages in the sense that, for each value x_i along the abscissa, a value $\hat{y}_i$ is estimated from the data present in a window around x_i. The number of data points included in the moving window is a proportion f, determined by users, of the total number of observations; a commonly-used first approximation for f is 0.5. The higher this proportion, the smoother the line of fitted values will be. For the end values, all observed points in the window come from the same side of x_i; this prevents the lines from becoming flat near the ends. Estimation proceeds in two steps:

- First, a weighted simple linear regression is computed for the points within the window and an estimate $\hat{y}_i$ is obtained. Weights, given to the observation points by a 'tricube' formula, decrease from the focal point x_i outwards. Points outside the window receive a zero weight. This regression procedure is repeated for all values x_i for which estimates are sought.

- The second step is to make these first estimates more robust, by reducing the influence of exceptional values and outliers. Residuals are computed from the fitted values and, from these, new weights are calculated that give more importance to the points with low residuals. Weighted linear regression is repeated, using as weights the products of the new weights with the original neighbourhood weights. This second step may be repeated until the recomputed weights display no more changes.

Trexler & Travis (1993) give a detailed account of the LOWESS method, together with a full example, and details on two techniques for choosing the most appropriate value for f. The simplest approach is to start with a (low) initial value, and increase it until a non-random pattern along x appears in the residuals; at that point, f is too large. Other important references are Chambers *et al.* (1983) and Cleveland (1985).

Numerical example. Consider again the dependence of salinity on the position along a transect, as modelled in Fig. 10.9. This same relationship may be studied using cubic splines and LOWESS (Fig. 10.15). For splines smoothing, the arbitrary rule stated above (5 or 6 points at least per interval) leads to 3 or 4 intervals. Figure 10.9 indicates, on the other hand, that there are at least three regions in the scatter of points, which can be delimited by knots located at approximately –5 and +4 along the abscissa. The computed spline regression equation which follows has $R^2 = 0.841$:

$$\hat{y} = 37.500 + 0.291x + 0.072x^2 + 0.006x^3 - 0.005\,(x+5)_+^3 - 0.007\,(x-4)_+^3$$

The difference in explained variation between this spline model and a cubic polynomial model ($R^2 = 0.81$, Fig. 10.9) is not significant.

The LOWESS curve also clearly suggests the presence of three distinct physical processes which determine the values of salinity along the long axis of the lagoon, i.e. from abscissa –10 to about –5, the central portion, and the right-hand portion from abscissa 4 and on.

Other smoothing methods are available in computer software, such as negative exponentially weighted smoothing (the influence of neighbouring points decreases exponentially with distance); inverse squared distance smoothing, described in Subsection 13.2.2 (eq. 13.21 with $k = 2$); distance-weighted least-squares smoothing (the surface is allowed to bend locally to fit the data); and step smoothing (a step function is fitted to the data).

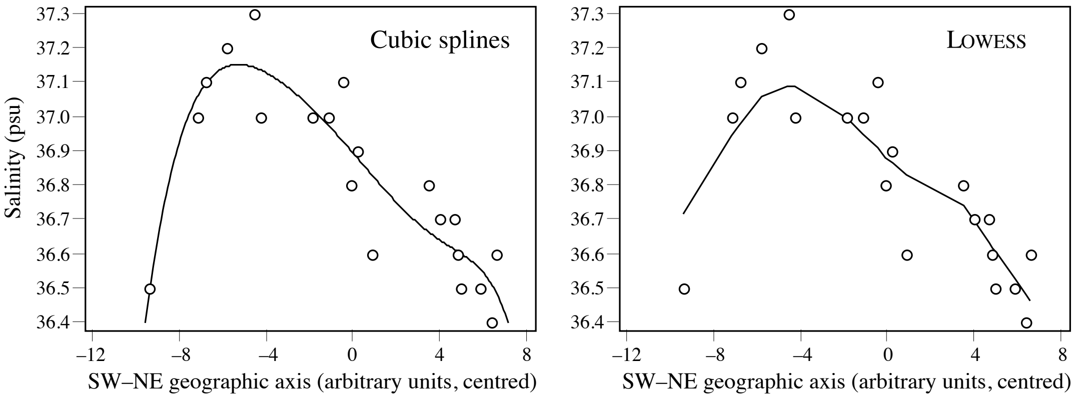


Figure 10.15 Cubic splines and LOWESS scatter diagrams describing the relationship of salinity with the position of the sites along the main geographic axis of the Thau lagoon, on 25 October 1988. Cubic splines were computed with knots at –5 and +4 on abscissa. For LOWESS (computed using SYSTAT), the proportion of the points included in each smoothing window was $f = 0.5$.

10.4 Path analysis

Subsection 4.5.4 showed that causal relationships among descriptors cannot be unambiguously derived from the sole examination of correlation coefficients, whether simple, multiple, or partial. Several causal models may account for the same correlation coefficients. In the case of *prediction* (*versus forecasting*, see Subsection 10.2.2), however, causal (and not only correlative) relationships among descriptors must be established with reasonable certainty. *Path analysis* is an extension of *multiple linear regression* (Subsection 10.3.3) that allows the decomposition and interpretation of *linear* relationships among a (small) number of descriptors. It is thus possible to formally state *a priori hypotheses* concerning the causal relationships among descriptors and, using path analysis, examine their consequences given the coefficients of regression and correlation computed among these descriptors.

Structural equation modelling

Path analysis was developed by Wright (1921, 1960). It is now recognized as a special case of a more general method called *structural equation modelling* (SEM), which includes latent variables (unmeasured, but estimated in the model by several measured variables) in addition to the measured variables. Structural equation models allow both exploratory and confirmatory modelling, meaning that the method is suited to develop as well as test theories. There are many interesting applications of path analysis and SEM in ecology, evolution, population genetics, and the social sciences. An introductory presentation of path analysis is found in Sokal & Rohlf (1995). The present section only provides a summary of path analysis showing its link with linear

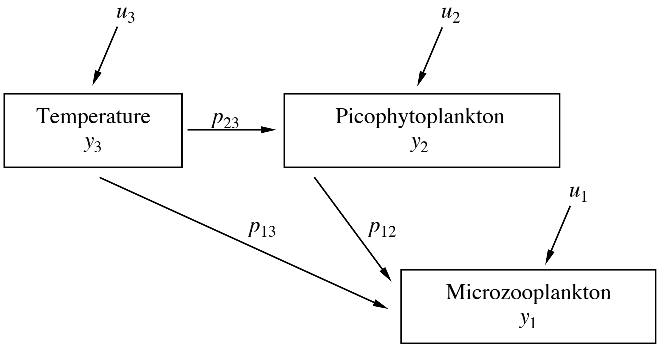


Figure 10.16 Path diagram for three linearly related descriptors. Adapted from Nie *et al.* (1975).

regression, and concludes with an ecological application. More complete and detailed presentations of path analysis and structural equation modelling are found in the books of Shipley (2002), Pugesek *et al.* (2003) and Grace (2006) written for ecologists, as well as books written for the social sciences, e.g. Kaplan (2009) and Kline (2011).

Causal order

Path diagram

Causal closure

As mentioned in Section 10.2, path analysis is based on two fundamental assumptions. (1) There exists a *causal order* among the variables. This causal order, which must be defined by the researchers, may be derived from ecological theory, or established experimentally (for a brief discussion of experiments, see Subsection 10.2.3). The assumption is that of *weak causal* ordering, e.g. y_1 *may* affect y_2 but y_2 *cannot* affect y_1. In *path diagrams* (Figs. 10.16 to 10.18), the causal ordering is represented by arrows, e.g. $y_1 \rightarrow y_2$. (2) No model can account for all the observed variance. Path models thus include *residual variables* u_i, which represent the unknown factors responsible for the residual variance (i.e. the variance not accounted for by the observed descriptors). The assumption of *causal closure* implies the independence of the residual causal variables; in other words, one assumes the existence of residual variables such that $u_1 \rightarrow y_1$ and $u_2 \rightarrow y_2$, whereas $u_1 \rightarrow y_2$ or $u_2 \rightarrow y_1$ is not allowed.

Numeral example. A simple example, with three variables exhibiting causal relationships, is used to illustrate the main features of path analysis. It is adapted from Nie *et al.* (1975, p. 386 *et seq.*). The example considers hypothesized relationships among water temperature, picophytoplankton (algae < 2 μm), and microzooplankton (e.g. ciliates) grazing on the picophytoplankton. In the model, it is assumed that water temperature (y_3) directly affects the growth of microzooplankton (y_1) and picophytoplankton (y_2), whose abundance, in turn, affects that of microzooplankton. Following the terminology of Sokal & Rohlf (1995, Section 16.3), y_2 and y_3 are *predictor* (or explanatory) variables while y_1 is the *criterion* (or response) variable. Figure 10.16 illustrates this hypothetical network of causal relationships in schematic form. Since the three variables probably do not explain all the observed variance, the model also

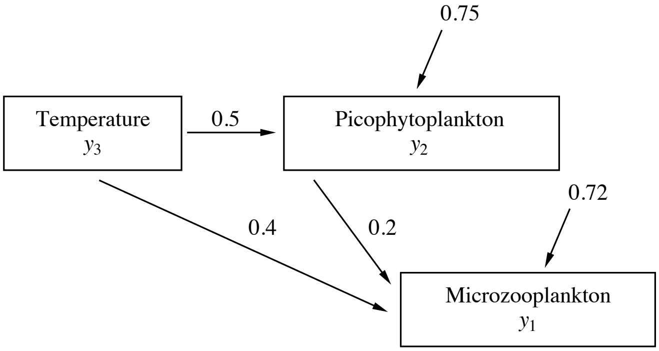


Figure 10.17 Results of path analysis for the example of Fig. 10.16. See text.

includes residual variables u_1 to u_3. The *causal ordering* of Fig. 10.16 is summarized in the following system of linear equations:

$$y_3 = u_3$$

$$y_2 = p_{23}y_3 + u_2$$

$$y_1 = p_{13}y_3 + p_{12}y_2 + u_1$$

Path coefficient

where parameters p_{ij} are the *path coefficients*. All variables are centred on their respective means. The hypothesis of causal closure implies that:

$$s(u_1,u_2) = s(u_1,u_3) = s(u_2,u_3) = 0$$

because the residual causes are independent; s represents covariances.

The path coefficients are estimated using multiple linear regression (Subsection 10.3.3):

$$\hat{y}_2 = p_{23}y_3$$

$$\hat{y}_1 = p_{13}y_3 + p_{12}y_2$$

There are no intercepts (coefficients p_0) in the regression equations because the data are centred. For a model with n descriptors, one can estimate all path coefficients using at most $(n - 1)$ regression equations. Each descriptor is predicted from the descriptors with immediately higher causal order. Two regression equations are needed to calculate the three path coefficients in Fig. 10.16. Let us use the following values for the path coefficients (Fig. 10.17) and coefficients of determination (R^2) of the numerical example:

$$\hat{y}_2 = 0.5y_3 \qquad R^2 = 0.25$$

$$\hat{y}_1 = 0.4y_3 + 0.2y_2 \qquad R^2 = 0.28$$

and the following correlation coefficients among the descriptors:

$$r_{12} = 0.4 \qquad r_{13} = 0.5 \qquad r_{23} = 0.5$$

The correlation r_{13} depends on both the direct relationship between y_1 and y_3 and the indirect relationship *via* y_2 (Figs. 10.16 and 10.17). Path analysis makes it possible to interpret the correlation r_{13} within the framework of the above model of causal relationships. Because the regressions that provide the estimates of the path coefficients are computed using *standardized* variables (eq. 1.12), it follows (Sokal & Rohlf, 1995, eq. 16.6) that

$$\begin{aligned} r_{13} &= p_{13} + r_{23}p_{12} \\ &= 0.4 + 0.5 \times 0.2 \\ &= 0.4 + 0.1 = 0.5 \end{aligned}$$

The correlation between y_3 (predictor variable) and y_1 (criterion variable) includes the direct contribution of y_3 to y_1 (path coefficient p_{13}), and also the common causes behind the correlations between y_3 and y_1. More generally, the correlation between a predictor variable y_i and a criterion variable y_1 includes the direct contribution of y_i to y_1, plus the common causes behind the correlations between y_i and any other variable that has a direct effect on y_1. These various contributions may either increase (as in the present example) or decrease the correlation between the predictor and criterion variables. The correlation coefficient r_{13} thus includes both a direct (0.4) and an indirect component (0.1).

Coefficient of nondetermination

Coefficients of nondetermination[*] are used to estimate the fractions of the variance that are not explained by the models (Fig. 10.17):

$$\begin{aligned} r^2(u_2, y_2) &= 1 - R^2_{2.3} = 1 - 0.25 = 0.75 \\ r^2(u_1, y_1) &= 1 - R^2_{1.23} = 1 - 0.28 = 0.72 \end{aligned}$$

One concludes that 75% of the variance of picophytoplankton (y_2) and 72% of the variance of microzooplankton (y_1) are not explained by the causal relationships stated in the model. The same results are obtained using the following general formula (Sokal & Rohlf, 1995):

$$\begin{aligned} r^2(u_1, y_1) &= 1 - \left[\sum_i p^2_{1i} + 2\sum_{ij} p_{1i}p_{1j}r_{ij} \right] \\ &= 1 - [(p^2_{12} + p^2_{13}) + 2(p_{12}p_{13}r_{23}) \\ &= 1 - [(0.04 + 0.16) + 2(0.2 \times 0.4 \times 0.5)] \\ &= 1 - [0.20 + 2(0.04)] \\ &= 1 - 0.28 = 0.72 \end{aligned}$$

The above results may be summarized in a single table. In the numerical example (Table 10.7), 0.1/0.5 = 20% of the covariation between microzooplankton (y_1) and temperature (y_3) is through picophytoplankton (y_2). In addition, 0.2/0.4 = 50% of the observed relationship between microzooplankton (y_1) and picophytoplankton (y_2) is not causal, and thus spurious

[*] The *coefficient of nondetermination* is $(1 - R^2)$; $\sqrt{1 - R^2}$ is called the *coefficient of alienation*.

Table 10.7 Decomposition of bivariate covariation among the (standardized) variables of Fig. 10.17. Adapted from Nie *et al.* (1975).

Bivariate relationships	Total covariation	Causal covariation Direct	 Indirect	 Total	Noncausal covariation
	(A)	(B)	(C)	(D = B+C)	(A–D)
y_2y_3	$r_{23} = 0.5$	0.5	0.0	0.5	0.0
y_1y_3	$r_{13} = 0.5$	0.4	0.1	0.5	0.0
y_1y_2	$r_{12} = 0.4$	0.2	0.0	0.2	0.2

according to the path model of Figs. 10.16 and 10.17. Such spurious correlations occur when two descriptors are caused by a third one (e.g. Fig. 4.11, Model 2) whose values have not been observed in the study.

Path analysis can be applied to more than three variables. As the number of variables increases, interpretation of the results becomes more complex and the number of possible models increases rapidly. In practice, path analysis is restricted to exploring the causal structure of relatively simple systems. This type of analysis is very useful in many ecological situations, if only because it forces researchers to explicitly state their hypotheses about the causal relationships among descriptors. The method helps assess the consequences of hypotheses, given the observed covariation among descriptors. Other methods, mentioned in Table 10.3, must be used when the descriptors do not exhibit linear relationships, or when they are not quantitative.

The following Ecological application 10.4 concerns freshwater ecology. Other applications of path analysis may be found, for example, in the fields of bacterial ecology (Troussellier *et al.*, 1986), biological oceanography (Gosselin *et al.*, 1986; Legendre *et al.*, 1991), and plant ecology (Hermy, 1987; Kuusipalo, 1987).

Ecological application 10.4

Harris & Charleston (1977) used path analysis to compare the microhabitats of two pulmonate snails, *Lymnaea tomentosa* and *L. columella*. The two species live in freshwater marshes; there are no obvious differences in the physical or chemical features of their respective habitats. Path analysis was used to examine, for each of the two species, the hypothetical model of causal relationships represented in schematic form in Fig. 10.18. In this model, water was assumed to affect snail numbers directly, and also *via* mud and flocculence, since both factors are partly determined by the amount of water present. The amount of mud was also expected to influence

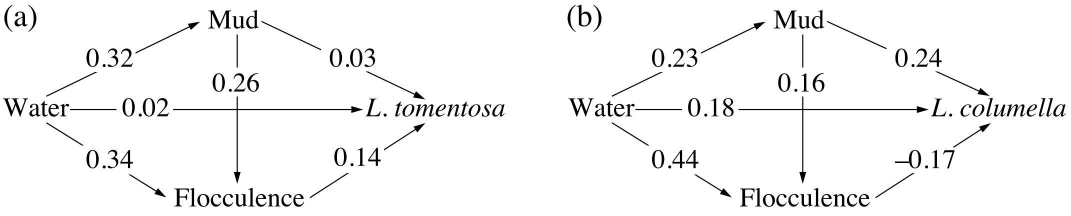


Figure 10.18 Path diagrams of the hypothesized effects of water, mud and flocculence on population densities of two pulmonate snails in marsh microhabitats. After Harris & Charleston (1977).

snails directly; however, larger mud areas are less likely to contain vegetation and are thus more likely to be flocculent, hence the indirect path from mud to snails *via* flocculence.

Results of path analysis (Fig. 10.18) suggest major differences between the microhabitats of the two species. Overall, increasing water cover has a direct (positive) effect on *L. columella*; in addition, flocculent mud appears to favour *L. tomentosa* whereas *L. columella* seem to prefer firm mud. The effects of water and mud on *L. columella* are thus direct, whereas they are indirect on *L. tomentosa* (i.e. *via* flocculence). The tentative hypothesis generated by the path diagrams must be further tested by observations and experiments. However, designing experiments to test the role played by the consistency of mud, while controlling for other (confounding) variables, would require considerable ingenuity.

10.5 Matrix comparisons

Regression and path analysis are restricted to the interpretation of univariate response variables. Other methods are required to perform direct comparison analyses when the descriptors form multivariate data tables. As shown in Fig. 10.4, canonical methods (Chapter 11) analyse the relationship between two rectangular data tables, whereas Mantel tests and derived forms, described in the present section, relate similarity or distance matrices derived from rectangular data tables.

Three main approaches are discussed in the present section. The Mantel test (Subsection 10.5.1) and derived forms (partial Mantel test, multiple regression on distance matrices, Subsection 10.5.2) are used to test relationships between association matrices, not between the rectangular date tables from which they originate. This is also the case of the analysis of similarities (ANOSIM, Subsection 10.5.3). The Procrustes test (Subsection 10.5.4) is different: it assesses the relationship between two rectangular data tables, not between association matrices. That test, derived from Procrustes analysis (Subsection 11.5.2), is presented in the present section to indicate that there are alternatives to the Mantel test to relate data matrices. Chapter 11 describes several other methods for the comparison of raw data matrices.

1 — Two association matrices: Mantel test

The Mantel (1967) test is a method to compare two similarity (**S**) or distance matrices (**D**), computed for the same objects, and test a hypothesis about the relationship between these matrices. For simplicity, the presentation will focus on distance matrices **D**. *Mantel tests should not be used to test hypotheses about the relationships between the original data tables, for reasons explained at the end of this subsection*. The data tables used to compute the two distance matrices must have been obtained independently of each other (i.e. different variables). One of the matrices may actually represent a hypothesis instead of real data, as shown below.

Ecological theory sometimes predicts relationships between resemblance matrices (**S** or **D**). This is the case with neutral theory, which predicts a monotonic relationship to appear in similarity decay plots where community composition similarity is expected to decreases with geographic distance (Nekola & White, 1999; Hubbell, 2001). In genetics, the theory of isolation by distance (Wright, 1943) is based on the fact that in sexually reproducing organisms, individuals tend to find mates in nearby rather than distant populations; for sessile organisms, this theory applies to species with short-range dispersal. As a consequence, populations living near each other tend to be more genetically similar than distant populations. In both cases, the theoretical predictions can be tested by analysing matrices of ecological or genetic distances $\mathbf{D_Y}$ versus geographic distances $\mathbf{D_X}$ using the Mantel test or regression on distance matrices (Subsection 10.5.2). Matrices $\mathbf{D_Y}$ and $\mathbf{D_X}$ must be computed for the same n objects *listed in the same order*. For ecological data, the choice of an appropriate resemblance measure is discussed in Chapter 7. In the two examples of the present paragraph, one of the matrices contains geographic distances among sites and Mantel tests may be used to test the predictions of these theories concerning distances. Other statistical methods can and should be used to test other predictions of these theories.

z_M statistic

The basic form of the Mantel statistic, called z_M, is the scalar product (Section 2.5) of the (*unstandardized*) values in the two resemblance matrices, excluding the main diagonal, which only contains trivial values (1's for similarities, 0's for distances) for which no estimate has been computed (Fig. 10.19). A second approach is to *standardize* the values in each of the two vectors of resemblance before computing the Mantel statistic. The cross-product statistic, divided by the number of distances in each half-matrix minus 1 [i.e. $(n(n-1)/2) - 1$], is bounded between –1 and +1; it behaves like a correlation coefficient and is called r_M. A third approach is to transform the distances into ranks (Dietz, 1983) before computing the standardized Mantel statistic; this is equivalent to computing a Spearman correlation coefficient (Section 5.3) between the corresponding values of matrices $\mathbf{D_Y}$ and $\mathbf{D_X}$.

r_M statistic

Permutation test

Mantel statistics are tested by permutation (Section 1.2). The n objects forming the rows and columns of the similarity or distance matrices are the permutable units, so that the permutations actually concern the n objects, not the $[n(n-1)/2]$ values in each half-matrix of distances. The testing procedure for Mantel statistics is summarized in Box 10.2.

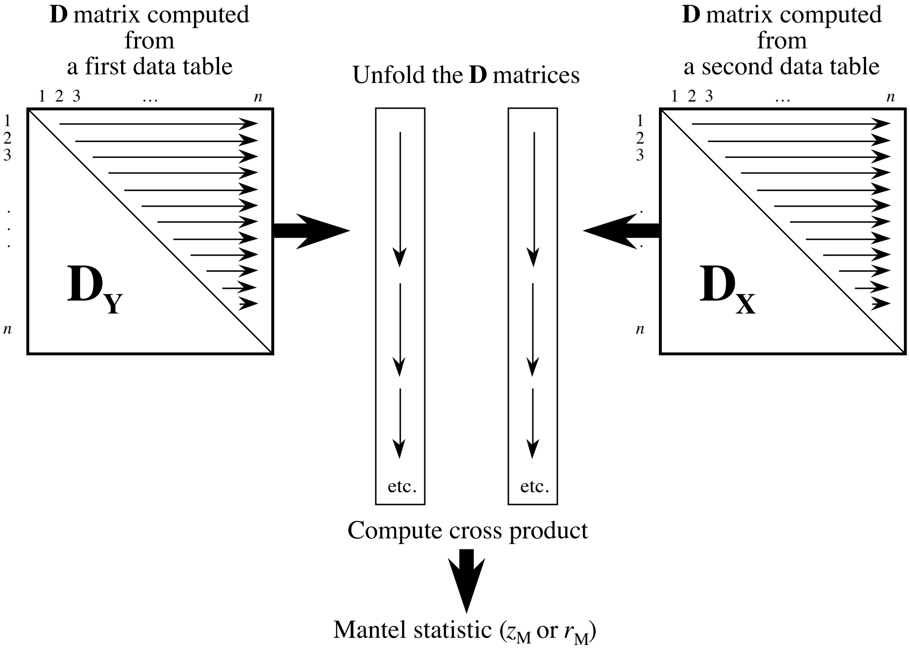

D matrix computed from a first data table
Unfold the **D** matrices
D matrix computed from a second data table
1 2 3 ... n
1 2 3 . . . n
D_Y
D_X
etc.
etc.
Compute cross product
Mantel statistic (z_M or r_M)

Figure 10.19 The Mantel statistic is the scalar product (sum of cross products) of the corresponding values in two distance matrices (**D**). Values in the vectors representing the unfolded matrices (i.e. written out as vectors) may be standardized before computing the statistic (r_M), or not (z_M), or transformed into ranks.

The permutation test leads to the same p-value with statistics z_M or r_M because all cross-product results, permuted or not, are affected in the same way by linear transformations such as standardization (eq. 1.12) of one or both vectors of distances. This is a most important property of the Mantel test. Thanks to it, the arbitrary values used in *model matrices* (below) are not an issue because any pair of chosen contrasting values leads to the same p-value.

Mantel tests are usually one-tailed since, in most cases, ecologists have a strong hypothesis about the sign of the correlation between the two matrices being compared. The hypothesis may be that the two distance matrices are positively related, which leads to a test of significance in the upper tail of the reference distribution. This is certainly the case when testing a hypothesis of isolation by distance in genetics. When comparing a similarity to a distance matrix, as in similarity decay plots, one generally expects a negative relationship to be found, if any; the test is then in the lower tail of the reference distribution.

Theory of the Mantel test Box 10.2

Hypotheses

H_0: The distances among objects in matrix $\mathbf{D_Y}$ are not (linearly or monotonically) related to the corresponding distances in $\mathbf{D_X}$.

H_1: The distances among points in matrix $\mathbf{D_Y}$ are related to the distances in $\mathbf{D_X}$.

Test statistics

- Mantel (1967) statistic: $z_M = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbf{D_Y}_{ij}\mathbf{D_X}_{ij}$ where i and j are row and column indices of the $\mathbf{D}$ matrices.

- Standardized Mantel statistic: $r_M = \frac{1}{d-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \text{stand}\,(\mathbf{D_Y})_{ij}\ \text{stand}\,(\mathbf{D_X})_{ij}$

where stand($\mathbf{D_Y}$) stand($\mathbf{D_X}$) contain standardized distances in their upper-triangular portions and $d = [n(n-1)/2]$ is the number of distances in the upper-triangular portion of each matrix.

Distribution of the test statistic

- According to H_0, the vector of values observed for any one object could have been observed for any other object; in other words, the objects are the permutable units. A realization of H_0 is obtained by permuting the objects (rows) in one of the original data matrices, bringing with them their vectors of values for the observed variables, and recomputing the distance matrix.

- An equivalent result is obtained by permuting at random the rows and corresponding columns of matrix $\mathbf{D_Y}$. Either $\mathbf{D_Y}$ or $\mathbf{D_X}$ can be permuted at random, with the same net effect.

- Repeating the above operation, the different permutations produce a set of values of the Mantel statistic, z_M or r_M, obtained under H_0. These values estimate the sampling distribution of the Mantel statistic under H_0.

Statistical decision

As in any other statistical test, the decision to reject H_0 or not is made by comparing the actual value of the auxiliary variable (z_M or r_M) to the reference distribution obtained under H_0. If the actual value of the Mantel statistic is one likely to have been obtained under the null hypothesis (no relationship between $\mathbf{D_Y}$ and $\mathbf{D_X}$), H_0 is not rejected; if it is too extreme to be considered a likely result under H_0, H_0 is rejected. See Section 1.2 for details.

Remarks

- The z_M or the r_M statistics may be transformed into another statistic, called t by Mantel (1967), which is asymptotically normal. It is tested by referring to a table of the standard normal distribution. It provides a good approximation of the probability **when *n* is large**.

- Like the Pearson correlation coefficient, the Mantel statistic formula is a linear model that brings out the linear component of the relationship between the values in two distance matrices. Strong nonlinearities may prevent the identification of relationships in Mantel tests. This led Mantel (1967) and Dietz (1983) to suggest the use of the Spearman or Kendall nonparametric correlation coefficients, instead of Pearson's r, as the statistic in Mantel tests.

Examples of Mantel tests are found in Upton & Fingleton (1985), Legendre & Fortin (1989), Sokal & Rohlf (1995), and elsewhere. The Mantel test is the statistical basis for the Mantel correlogram described in Subsection 13.1.6.

The Mantel test is only valid if matrix $\mathbf{D_X}$ is independent of the resemblance measures in $\mathbf{D_Y}$, i.e. $\mathbf{D_X}$ should not be derived in any way from $\mathbf{D_Y}$ nor from the data that were used to compute $\mathbf{D_Y}$. The Mantel test has two chief domains of application in community ecology:

1. It may be used to compare two resemblance matrices computed from empirical data and test a hypothesis about the relationship between the distances, as in the similarity decay plot example described above. For the test to be valid, $\mathbf{D_X}$ must be computed from the same objects but a different set of variables than those used to compute $\mathbf{D_Y}$.

Model matrix

2. The Mantel test may also be used to assess the goodness-of-fit of data to an *a priori* distance model. The test compares the empirical distance matrix to a *model matrix* (also called a *pattern* or *design matrix*). This matrix is constructed to represent the model to be tested; in other words, it depicts the alternative hypothesis of the test. For example, in the Mantel correlogram (Subsection 13.1.6), the model is a classification of the distances in two groups, e.g. the distances smaller than a given value of interest and the larger distances. Figure 13.14 (Chapter 13) shows two matrices, **X**(1) and **X**(2), representing such models. Other examples are given by Sokal & Rohlf (1995, Section 18.3).

The Mantel test cannot be used to check the conformity to a matrix $\mathbf{D_Y}$ of a model derived from the same data, e.g. to test the conformity of $\mathbf{D_Y}$ to a group structure obtained by clustering matrix $\mathbf{D_Y}$. In such a case, the model matrix $\mathbf{D_X}$, which depicts the *alternative hypothesis* of the test, would describe a structure made to fit the very data that would now be used to test the null hypothesis. The hypothesis ($\mathbf{D_X}$) would not be independent of the data ($\mathbf{D_Y}$) used to test it. Such a test would be incorrect; it would almost always reject the null hypothesis and support the conformity of $\mathbf{D_Y}$ to $\mathbf{D_X}$. This point has been mentioned in Subsection 8.12.2.

Goodness-of-fit Mantel tests have been used in vegetation studies to investigate hypotheses related to questions like the concept of climax (McCune & Allen, 1985) and the environmental control model (Burgman, 1987, 1988). Hypotheses of niche segregation have been tested for trees by Legendre & Fortin (1989), and for animals by Hudon & Lamarche (1989). Somers & Green (1993) used Mantel tests based on Spearman correlation coefficients (see Box 10.2, Remarks) to assess the relationship between crayfish catches in six Ontario lakes and five model matrices corresponding to different ecological hypotheses. Considering what is now known about the properties of the Mantel test (see Box 10.3), in all these applications, the Mantel test should be replaced by canonical analysis (Chapter 11), which provides more powerful tests of significance.

Further developments, Mantel test Box 10.3

Mantel tests should be restricted to test hypotheses concerning distances. A Mantel test between two distance matrices $\mathbf{D_Y}$ and $\mathbf{D_X}$ derived from raw data tables **Y** and **X** is not equivalent to (1) the test of a correlation coefficient computed between two vectors of raw data, (2) a test computed by linear regression between a response vector **y** and an explanatory matrix **X**, or (3) a test in canonical analysis between a multivariate response matrix **Y** and an explanatory matrix **X**. This statement is supported by the following observations.

1. The sum of squares of the distances *is not* the sum of squares of the raw data (Legendre *et al*., 2005; Legendre & Fortin, 2010). On the one hand,

$$\text{SS}(\mathbf{Y}) = \sum_{j=1}^{p}\sum_{i=1}^{n}(y_{ij}-\bar{y}_j)^2 = \left(\sum_{i \neq h} D_{ih}^2\right) / n$$

as shown in Box 6.1, eqs. 6.55 and 6.56. On the other hand, the sum of squares of the distances in a distance matrix **D** is computed as follows:

$$\text{SS}(\mathbf{D}) = \sum_{i \neq h}(D_{ih}-\bar{D})^2 = \sum_{i \neq h} D_{ih}^2 - \frac{\left(\sum_{i \neq h} D_{ih}\right)^2}{n\,(n-1)/2}$$

The last equation is for symmetric distance matrices, where only the D_{ih} values in the upper or lower-triangular portion of **D** are used. The right-hand parts of the two equations above are irreductible to one another. Consider the numbers 1 to 10 for example: their total sum of squares SS(**Y**) in the first equation is 82.5 and the sum of squares of the Euclidean distances among these values (SS(**D**), second equation) is 220. SS(**Y**) is the denominator of R^2 in multiple regression (eq. 10.20) and canonical analysis (eq. 11.4) whereas SS(**D**) is the denominator of the R^2 in regression on distance matrices (Subsection 10.5.2); in simple Mantel tests, the square root of this R^2 is the Mantel statistic r_{M}. As a consequence, the R^2 computed by regressing $\mathbf{D_Y}$ on $\mathbf{D_X}$ has nothing to do with the canonical R^2 obtained by analysing the variation of **Y** with respect to the variation of **X**.

2. An example given by Legendre *et al*. (2005) concerns a group of four sites that have one species in common; in addition, each site harbours one species that is not present in any of the three other sites (Fig. 10.20). This group of sites clearly displays spatial variation in community structure, or beta diversity (Subsection 6.5.3). The total sum of squares of the species data, SS(**Y**), is 3.0; it is positive, as expected for a group of sites showing beta diversity (Box 6.1). However, the sum of squares of the distances in the upper (or lower) triangular portion of matrix **D**, SS(**D**), is zero. Because SS(**D**) is the denominator of the R^2 in regression on distance matrices and the square root of this R^2 is the Mantel statistic r_{M}, the variation in the data shown in Fig. 10.20a cannot be analysed by a Mantel test because the Mantel correlation r_{M} would be indeterminate. This example also shows that SS(**D**) is not a measure of beta diversity.

(a) Data

	Sp.1	Sp.2	Sp.3	Sp.4	Sp.5
Site 1	1	1	0	0	0
Site 2	1	0	1	0	0
Site 3	1	0	0	1	0
Site 4	1	0	0	0	1

(b) **D** = [1 – Jaccard similarity]

	Site 1	Site 2	Site 3	Site 4
Site 1	0	0.667	0.667	0.667
Site 2	0.667	0	0.667	0.667
Site 3	0.667	0.667	0	0.667
Site 4	0.667	0.667	0.667	0

Figure 10.20 Illustrative example. (a) Community composition data table and (b) derived distance matrix, $D_{ij} = (1 - S_{ij})$, based on the Jaccard similarity index (S_7). Redrawn from Legendre *et al.* (2005).

Box 10.3 (*continued*)

3. Numerical simulations involving two variables were carried out by Legendre & Fortin (2010, Table 2) to demonstrate the difference in power between tests of significance of correlation coefficients between two variables and Mantel tests carried out between distance matrices computed from these same variables. A population correlation value was imposed between two vectors of random variables, as in Table 10.5 of Subsection 10.3.3. When the correlation value was 0, all tests (the parametric and permutation test of the correlation coefficient, as well as the Mantel test) had correct levels of type I error, i.e. all tests rejected H_0 at the α level in a proportion of the cases approximately equal to α. When the population correlation was $\rho = 0.5$, the mean of the Pearson correlations computed on samples of $n = 10$ to 100 data (10000 repetitions for each value of n) was approximately 0.5; the mean of the Mantel r_M statistics was near 0.2. Tests of the Pearson correlations increased in power as n increased, from a rejection rate of H_0 of 0.455 for $n = 10$ to 1.000 for $n = 100$; Mantel tests had a rejection rate of 0.279 for $n = 10$ to 0.968 for $n = 100$. When the population correlation was negative ($\rho = -0.5$), the mean of the Pearson correlations was approximately –0.5; the mean of the Mantel r_M statistics was near 0.2; note the positive sign. Powers for the two statistics were the same as when the population correlation was $\rho = 0.5$. To summarize, these simulations showed the following: when it detects a correlation in the original data, the Mantel test may not correctly estimate the sign of the correlation coefficient, and it produces tests with lower power than the test of Pearson's r. Conclusion: the Mantel test is inappropriate to test hypotheses concerning correlations in raw data.

4. Dutilleul *et al.* (2000) described cases where the values of the Mantel statistics were negative whereas the Pearson correlation was strictly 0; their Table 4 also showed cases, for real bivariate data, where the signs of the Mantel statistics varied but were unrelated to the signs of the Pearson correlations. Again, the Mantel test seemed inappropriate to test hypotheses concerning correlations in raw data.

2 — More than two association matrices

Smouse *et al*. (1986) proposed to compute partial correlations involving similarity or distance matrices. Consider distance matrices $\mathbf{D}_1$, $\mathbf{D}_2$, and $\mathbf{D}_3$ computed from three multivariate data tables, using a distance measure appropriate to each case. The partial Mantel statistic, $r_M(\mathbf{D}_1\mathbf{D}_2.\mathbf{D}_3)$, estimating the correlation between matrices $\mathbf{D}_1$ and $\mathbf{D}_2$ while controlling for the effect of $\mathbf{D}_3$, is computed in the same way as a partial correlation coefficient (eq. 4.36), except that the calculation is based here on standardized Mantel statistics r_M (Box 10.2) instead of Pearson correlations r. For symmetric distance matrices, only the upper (or lower) triangular portions are used in the calculations. The tests of significance applicable to partial Mantel statistics (permutation tests) are described in Legendre (2000) and in Appendix 4 of Legendre & Fortin (2010).

Partial Mantel test

Like Mantel tests, partial Mantel tests are only applicable to questions that concern relationships among three distance matrices, not raw data. Another method described in the present section, multiple regression on distance matrices, is applicable to questions involving more than three distance matrices,

Ecological application 10.5

This application analyses the microgeographic morphological differentiation of muskrats (*Ondatra zibethicus*) in the upper basin of River La Houille in southern Belgium. Muskrats were introduced into Bohemia (now part of the Czech Republic) from North America in 1905 for breeding and fur production. In later years, the species was introduced into other European countries, including Belgium, where individuals were released to the wild in 1928 (Le Boulengé, 1972). After their release from breeding farms, muskrats colonized ponds and waterways throughout Europe.

Muskrats were captured during a government-sponsored trapping campaign conducted in 1971-1972 to eradicate rats from the ponds of the upper La Houille basin (approximately 150 km^2) where the river forms a broad, 15 km long loop, before flowing towards the Ardennes Department of France where it becomes a tributary of River Meuse (Fig. 10.21a). Muskrats were captured in nine local population zones, seven of which are included in the part of the study of Le Boulengé *et al*. (1996) reported here. Age and sex of the captured specimens were determined and measurements of the skull were taken. Mahalanobis distances based on 10 age-adjusted skull measurements were computed among the muskrat population zones.

Despite the absence of environmental heterogeneity across the study region, significant skull morphological differences were identified among the local populations by ANOVA and MANOVA. These differences were possibly due to founder effects and/or colonization of the tributaries by animals from different origins, coupled with a spatial pattern of genetic relatedness among the zones. The question addressed by the authors was: how can the relatedness of the populations in the different zones be explained? Are geographically closer populations more similar in their skull morphology? And then, what is it to be “geographically closer” for muskrats, which are semi-aquatic mammals? The populations are genetically interconnected by the migration of the young which, after weaning, may disperse to other population zones. In this study, the relationships to be tested clearly concerned distances (morphological and geographic), so Mantel tests were appropriate.

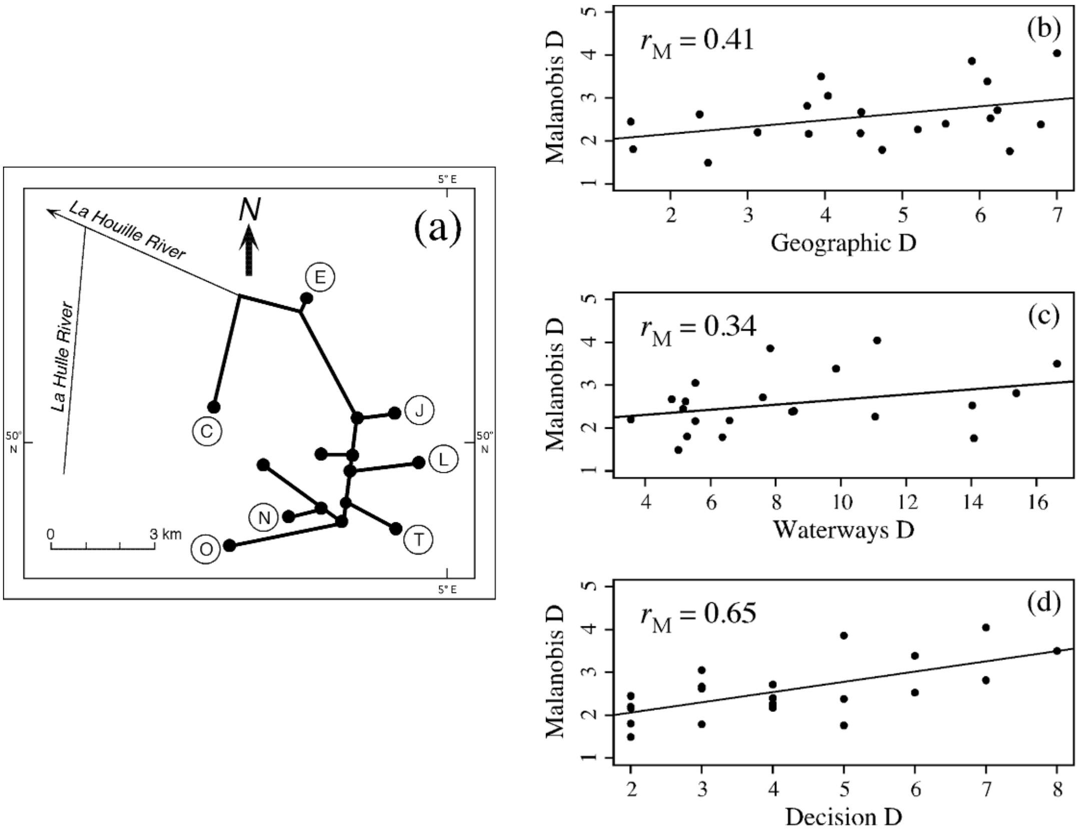


Figure 10.21 (a) Schematic representation of the upper La Houille River network in Belgium, showing the seven muskrat local population zones, identified by letters C to T. (b) to (d) Distance comparison diagrams and simple Mantel statistics (r_M) for three types of geographic separation distances. An OLS regression line indicates the trend in each graph.

The authors measured the geographic distance between the zones in two ways: straight-line geographic distance and distance along the waterways. Muskrats are fast-moving animals when they travel, so perhaps the actual distance is not really the determining factor. For that reason, the authors also devised a “Decision distance”, which is the number of furcations of the river network separating two zones. At these points, a migrating muskrat must decide to go either left or right along the river network. The relationship between the three types of geographic separation distances and the morphometric (Mahalanobis) distances are shown in Fig. 10.21b-d, where simple Mantel correlations (r_M) are shown. The graphs also show that the different types of geographic distances are linearly related to the morphometric distances, so that the Mantel statistic based on the Pearson correlation coefficient was appropriate in this study.

The authors formulated a sociobiological hypothesis called "isolation by distance along corridors". This hypothesis states that the decision distances explain the morphometric differentiation of the local populations best. Partial Mantel tests were computed to compare the ability of different distances to explain the morphological distances. Comparisons of the geographic and decision distances produced the following results:

r_M(Mahalanobis Geographic.Decision) = 0.03, p = 0.455

r_M(Mahalanobis Decision.Geographic) = 0.55, p = 0.010

showing that the decision distance matrix explained the morphometric distances significantly better than the geographic distances. Likewise, the decision distance matrix explained the morphometric distances significantly better than the waterways distances; there was no significant difference in explanation of the morphometric distances between the geographic and waterways distances (results computed from the distance matrices found in the paper). Hence the results were consistent with the hypothesis of "isolation by distance along corridors".

Causal modelling

Partial Mantel tests are not always easy to interpret. Legendre & Troussellier (1988) have shown the consequences of all possible three-matrix causal models on the significance of Mantel and partial Mantel statistics. The models (and their predictions) are the same as those illustrated in Fig. 4.11 for three simple variables. This approach leads to a form of *causal modelling on resemblance matrices* (Legendre, 1993). It should only be used to analyse questions that require the modelling of distances, not raw data.

In ecology, this type of analysis has been used mostly to study the distribution of organisms (matrix $\mathbf{D}_1$) with respect to environmental variables (matrix $\mathbf{D}_2$) while considering the spatial locations of the sampling sites (matrix $\mathbf{D}_3$). We now know that in all these applications, including Legendre & Trousselllier (1988), the Mantel test should be replaced by canonical analysis (Chapter 11), which provides much more powerful tests of significance (see Box 10.4).

Regression on distance matrices

One may also want to model the variation in a first distance matrix as a function of the variation in other distance matrices about the same objects. *Multiple regression on resemblance matrices* has been suggested by several authors (Hubert & Golledge, 1981; Smouse *et al.*, 1986; Manly, 1986; Krackhardt, 1988) to address research questions formulated in terms of distances. Legendre *et al.* (1994) described appropriate testing procedures for evolutionary studies where the response data was a dendrogram or an evolutionary tree. The parameters of the multiple regression model are obtained using a procedure similar to that of the Mantel test (Fig. 10.22). The response distance matrix $\mathbf{D_Y}$, which represents the evolutionary tree, is unfolded into a vector $\mathbf{y}$; likewise, each explanatory distance matrix $\mathbf{D_X}$ is unfolded into a vector $\mathbf{x}$. A multiple regression is computed in which $\mathbf{y}$ is a function of vectors $\mathbf{x}_j$. The parameters of that regression (the coefficient of multiple determination R^2 and the partial regression coefficients) are tested by permutations, as follows. When the response distance matrix $\mathbf{D_Y}$ is an ordinary distance or similarity matrix, the permutations of the corresponding vector $\mathbf{y}$ are carried out in the way of the Mantel permutational test (Subsection 10.5.1). When it is an ultrametric matrix representing a dendrogram (Subsection 8.3.1), the double-permutation method of Lapointe and Legendre (1990,

Permutation test

Further developments, partial Mantel test Box 10.4

During the past 15 years, partial Mantel tests and regression on distance matrices have been used in many ecological papers that had for objective to analyse the spatial variation of community composition (raw data, not distances) among sites, i.e. beta diversity (Subsection 6.5.3). Some of these papers were listed as examples by Legendre *et al*. (2005). To demonstrate that the Mantel test should not be used for that type of objective, that paper presented simulation results involving multivariate, spatially correlated data. The simulations compared canonical analysis (RDA, Section 11.1) to Mantel tests to detect the effect of environmental variables **X** on species-like response data **Y**, as well as the presence of spatial structures in the species-like data (10 simulated species, $n = 100$). The results found in Table 1 and Fig. 3 of Legendre *et al*. (2005) showed the following:

- The two testing methods had correct levels of type I error. They were thus statistically valid.

- When **Y** was related to the environmental variables **X** (plus random error in **Y**), RDA detected a significant relationship in 97% of the simulations whereas the Mantel test detected it in 49% of the cases.

- Using the distance-based Moran's eigenvector map method of spatial analysis (dbMEM, Section 14.1) in RDA, significant spatial structures were detected in the simulated data in 99% of the cases, compared to 8 to 22% of the cases detected by Mantel tests.

These findings support the conclusion that the Mantel test is inappropriate to test hypotheses concerning correlations in raw data. Other simulation results, where community composition data were simulated according to Hubbell's (2001) neutral model, led to the same conclusions about the difference in power between the two types of tests when applied to raw data (Legendre *et al*., 2008).

Not everyone agrees about the questions that can be answered by Mantel tests. See the controversy raised by Tuomisto & Ruokolainen (2006) and the exchanges that followed in the ecological literature (Pélissier *et al*., 2008; Laliberté, 2008; Legendre *et al*., 2008; Tuomisto & Ruokolainen, 2008). Everyone now seems to agree, however, that Mantel tests should be limited to questions about relationships between distance matrices.

1991) is used. When it is a path-length matrix representing an additive tree (i.e. a cladogram in phylogenetic studies), a triple-permutation method (Lapointe and Legendre, 1992a) is used. Vectors $\mathbf{x}_j$ representing the explanatory matrices are kept fixed with respect to one another during the permutations. Selection of explanatory matrices may be done by forward selection, backward elimination, or a stepwise procedure, which are described in Legendre *et al*. (1994). For research questions that do not strictly concern distances, the method of multiple factor analysis (MFA, briefly described at the end of Subsection 11.5.1) should be used for analysis.

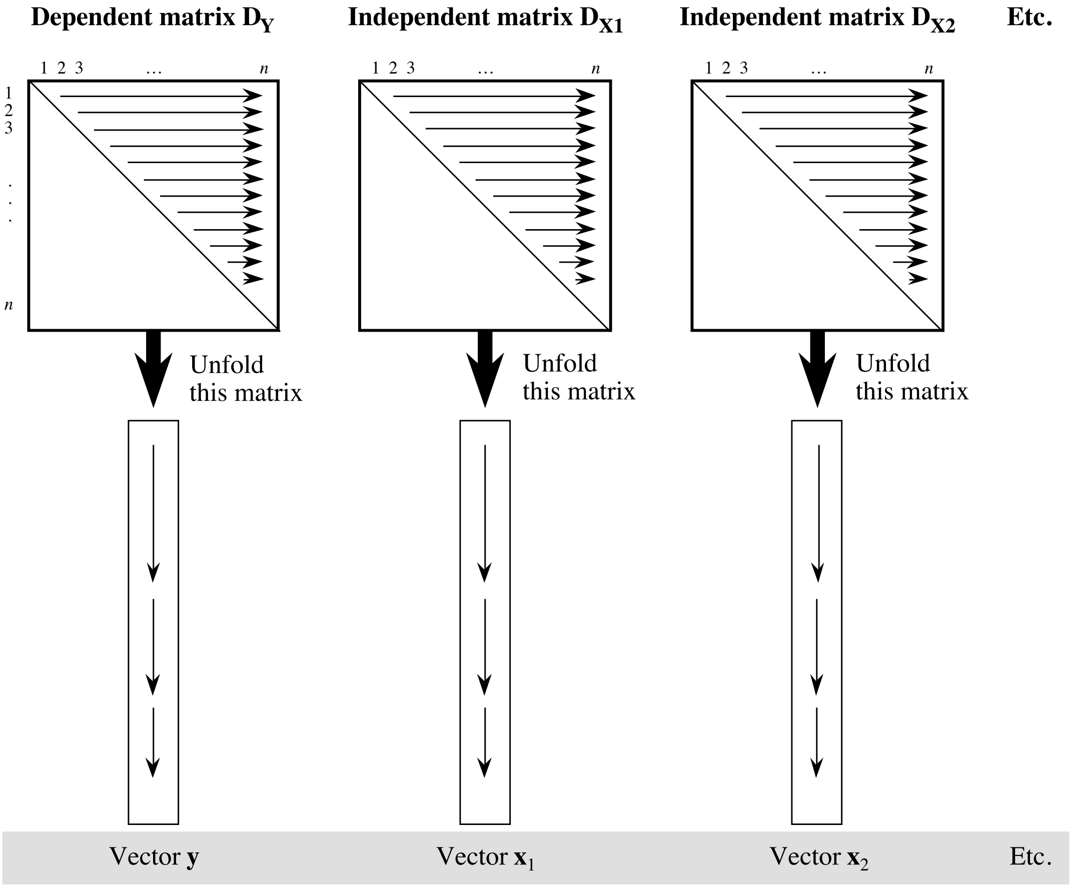
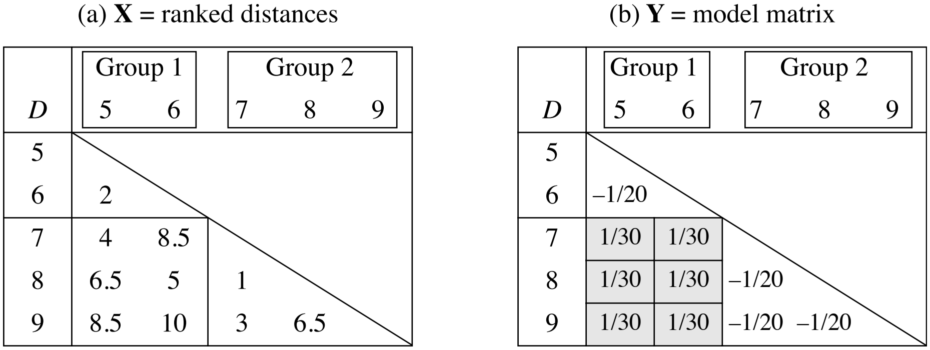

Dependent matrix $\mathbf{D_Y}$
Independent matrix $\mathbf{D_{X1}}$
Independent matrix $\mathbf{D_{X2}}$
Etc.
1 2 3 … n
1
2
3
n
Unfold this matrix
Vector **y**
Vector $\mathbf{x}_1$
Vector $\mathbf{x}_2$
Etc.

Figure 10.22 Multiple regression is computed on the vectors resulting from unfolding matrices $\mathbf{D_Y}$ (response) and $\mathbf{D_{X1}}$, $\mathbf{D_{X2}}$, etc. (explanatory).

The CADM method to test the congruence among distance matrices, described in Subsection 5.4.3, is another extension of the Mantel test to several distance matrices.

3 — ANOSIM test

Focusing on problems of analysis of variance that involved community composition data, Clarke (1988, 1993) developed a parallel approach to the goodness-of-fit Mantel tests. Clarke's method, called ANOSIM (*ANalysis Of SIMilarities*), is implemented in the PRIMER package, referred to in Section 9.4, and in R. In PRIMER, program ANOSIM includes one-way and two-way analyses (crossed or nested) for replicated data, whereas program ANOSIM2 covers two-way analyses without replication (Clarke &



(a) **X** = ranked distances

	Group 1		Group 2		
D	5	6	7	8	9
5					
6	2				
7	4	8.5			
8	6.5	5	1		
9	8.5	10	3	6.5	

(b) **Y** = model matrix

	Group 1		Group 2		
D	5	6	7	8	9
5					
6	–1/20				
7	1/30	1/30			
8	1/30	1/30	–1/20		
9	1/30	1/30	–1/20	–1/20	

Figure 10.23 (a) Distances from the numerical example in Fig. 12.22a are transformed into ranks, the most similar pair receiving rank 1. (b) Weighting required to compute the ANOSIM statistic as a Mantel statistic.

Warwick, 1994). After a brief presentation of Clarke's statistic, below, the similarities and differences between the ANOSIM and Mantel approaches will be shown.

Consider the situation illustrated in Fig. 10.23a. The distances shown in Fig. 12.22a were transformed into ranks, the least dissimilar pair (i.e. the most similar) receiving rank 1. Tied values in Fig. 12.22a were given mean rank values, as usual in nonparametric statistics. Objects are arbitrarily numbered 5, 6, 7, 8, 9. The objects are assumed to form two groups, defined here on *a priori* bases; the two groups are not supposed to result from clustering as in Fig. 12.22a. The two *a priori* groups are (5, 6) and (7, 8, 9). The null hypothesis is of the ANOVA type:

H_0: There is no difference between the two (or more) groups.

In Fig. 10.23a, does one find the kind of variation among distance values that one might expect if the data corresponded to the null hypothesis? Clarke (1988, 1993) proposed the following statistic to assess the differences among groups:

$$R = \frac{\bar{r}_B - \bar{r}_W}{n(n-1)/4} \qquad \textbf{(10.31)}$$

where $\bar{r}_B$ is the mean of the ranks in the *between*-group submatrix (i.e. in Fig. 10.23a, the rectangle crossing groups 1 and 2), $\bar{r}_W$ is the mean of the ranks in all *within*-group submatrices (i.e. the two triangles in the figure), and n is the total number of objects. In the present example, $\bar{r}_B = 7.083$ and $\bar{r}_W = 3.125$, so that $R = 0.79167$ (eq. 10.31).

Using ranks instead of the original distances is not a fundamental requirement of the method. It comes from a (reasonable) recommendation, by Clarke and co-authors, that the test statistic should reflect the patterns formed among sites represented by multidimensional scaling plots (nMDS, Section 9.4), which preserve rank-transformations of distances. The R statistic is tested by permutations of the objects, as explained in Box 10.2. The denominator of eq. 10.31 is chosen in such a way that $R = 1$ if all the lowest ranks are in the "within-group" submatrices, and $R = 0$ if the high and low ranks are perfectly mixed across the "within" and "between" submatrices. R is unlikely to be substantially smaller than 0; this would indicate that the similarities within groups are systematically lower than among groups.

Permutation test

Clarke (1988, 1993) actually applied the method to the analysis of several groups. This is also the case in the nonparametric ANOVA-like example of the Mantel test in the Sokal & Rohlf (1995) book. The statistic (eq. 10.31) can readily handle the more-than-two-group case: $\bar{r}_B$ is then the mean of the ranks in *all* between-group submatrices, whereas $\bar{r}_W$ is the mean of the ranks in *all* within-group submatrices.

Equation 10.31 may be reformulated as a Mantel cross-product statistic z_M (Box 10.2). To achieve this, define a model matrix containing positive constants in the "between-group" portion and negative constants in the "within-group" parts:

- the "between" values (shaded area in Fig. 10.23b) are chosen to be the inverse of the number of between-group distances (1/6 in this example), divided by the denominator of eq. 10.31, i.e. $[n(n-1)/4]$ (which is 5 in the present example);

- similarly, the "within" values in Fig. 10.23b are chosen to be the inverse, with negative signs, of the number of distances in all within-group submatrices (–1/4 in the example), also divided by $[n(n-1)/4]$ (= 5 in the present example).

The coding is such that the sum of values in the half-matrix is zero. The unstandardized Mantel statistic (Box 10.2), computed between matrices $\mathbf{D_X}$ and $\mathbf{D_Y}$ of Fig. 10.23, is $z_M = 0.79167$. This result is identical to Clarke's ANOSIM statistic.

Since the permutation method is the same in the Mantel and ANOSIM procedures, the tests should produce similar p-values. They may differ slightly in practice because different programs, and even different runs of the same program, may produce different sequences of permutations of the objects. As shown in Subsection 10.5.1, *any binary coding* of the "within" and "between" submatrices of the model matrix leads to the same probabilities. Of course, interchanging the small and large values produces a change of sign of the statistic and turns an upper-tail test into a lower-tail test. The only substantial difference between the Mantel goodness-of-fit and ANOSIM tests is one of tradition: Clarke (1988, 1993) and the ANOSIM function in the PRIMER package (Clarke & Warwick, 1994) and in R (Section 10.7) transform the distances into ranks before computing eq. 10.31. Since Clarke's R is equivalent to a Mantel statistic computed on ranked distances, it is thus analogous to a Spearman correlation coefficient (eqs. 5.1 and 5.3).

The Mann-Whitney U statistic could also be used for analysis-of-variance-like tests of significance performed on distance matrices. This has been suggested by

Gordon (1994) in a different context, i.e. as a way of measuring the differentiation of clusters produced by clustering procedures (internal validation criterion), as reported in Section 8.13. In Gordon's method, distances are divided in two subsets, i.e. the within-group (*W*) and between-group (*B*) distances — just like in Clarke's method. A *U* statistic is computed between the two subsets. *U* is closely related to the Spearman rank correlation coefficient (eqs. 5.1 and 5.3); a *U* test of a variable against a dummy variable representing a classification in two groups is equivalent to a Spearman correlation test (same probability). Since Clarke's statistic is also equivalent to a Spearman correlation coefficient, the Mann-Whitney *U* statistic should lead to the exact same probability as the Clarke or Mantel statistics, if *U* was used as the statistic in a Mantel-like permutation test. [Using the *U* statistic as an internal validation criterion, as proposed by Gordon (1994), is different. On the one hand, the grouping of data into clusters is obtained from the distance matrix that is also used for testing; this is not authorized in an analysis-of-variance approach. On the other hand, Gordon's Monte Carlo testing procedure differs from the Mantel permutation test.]

4 — Procrustes test

In Greek mythology, Procrustes was a son of Poseidon and a rogue. He invited travellers to spend the night with him, then tied them down to an iron bed and either cut off their limbs if they were taller than the bed, or stretched the victims if they were too short, till they fitted in.

Procrustes analysis, proposed by Gower (1971b, 1975, 1987), is primarily a canonical ordination method; it is described in Subsection 11.5.2. The Procrustes test (PROTEST) is presented here as a statistical method for comparing two rectangular data matrices about the same objects. It is appropriate to answer questions about the relationship between the original data sets (i.e. raw data), which is not the case of the Mantel test. Another statistic that can be used in the same situation is the *RV* coefficient (eqs. 11.65 and 11.66) described with co-inertia analysis (Subsection 11.5.1).

The purpose of Procrustes analysis is to find a compromise ordination for two raw data matrices with the same objects in rows, using a rotational-fit algorithm that minimizes the sum of squared distances between corresponding points of the two matrices in a joint ordination. In that ordination, each object has two representations, one from each matrix, so that the scatter diagram allows one to visualize the differences between the two original matrices. In *orthogonal Procrustes*, two matrices are considered and fitted using rigid-body motions (translation, rotation, and mirror reflection). *Generalized Procrustes analysis* is the extension of the method to more than two matrices. Details are found in the references given above.

The present subsection focuses on the residual sum-of-squares statistic of orthogonal Procrustes analysis, which is a goodness-of-fit statistic. It was called m^2 by Gower and is computed as follows:

Procrustes statistic m^2

$$m_{12}^2 = \mathrm{Trace}(\mathbf{Y}_1\mathbf{Y}_1') - \frac{(\mathrm{Trace}\mathbf{W})^2}{\mathrm{Trace}(\mathbf{Y}_2\mathbf{Y}_2')} \qquad \textbf{(10.32)}$$

where $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are the two rectangular matrices of raw data to be analysed, with column vectors centred on their respective means, and $\mathbf{W}$ is a diagonal matrix of singular values found by the singular value decomposition $\mathbf{Y}_1'\mathbf{Y}_2 = \mathbf{VWU}'$ (SVD, eq. 2.31).

Equation 10.32 is not symmetric; indeed, the m_{12}^2 value resulting from fitting $\mathbf{Y}_2$ to $\mathbf{Y}_1$ differs from m_{21}^2 obtained by fitting $\mathbf{Y}_1$ to $\mathbf{Y}_2$. To solve that problem, transform the column-centred matrix $\mathbf{Y}_1$ to $\mathbf{Y}_{1.\mathrm{tr}}$ by dividing each value of $\mathbf{Y}_1$ by the square root of the trace of $\mathbf{Y}_1\mathbf{Y}'_1$, which is the same as the trace of $\mathbf{Y}_1'\mathbf{Y}_1$; that trace is easily computed as the sum of squares of all values in $\mathbf{Y}_1$. Using the same method, transform the centred matrix $\mathbf{Y}_2$ to $\mathbf{Y}_{2.\mathrm{tr}}$. For $\mathbf{Y}_{1.\mathrm{tr}}$ and $\mathbf{Y}_{2.\mathrm{tr}}$, the two Procrustes statistics are now identical:

$$m_{12}^2 = m_{21}^2 = 1 - (\mathrm{Trace}\mathbf{W})^2 \qquad \textbf{(10.33)}$$

Jackson (1995) suggested using the symmetric orthogonal Procrustes statistic m_{12}^2 (eq. 10.33) as a measure of concordance, or similarity, between two data matrices representing, in particular, species abundances and environmental variables. The statistic is tested by permutation. Jackson (1995) called this procedure the *Procrustean randomization test* (PROTEST). He provided examples of applications to ecological data: benthic invertebrates, lake morphometry, lake water chemistry, and geographic coordinates, for 19 lakes in Ontario, Canada. What Jackson actually compared in that paper were, for each data set, the first two ordination axes from correspondence analysis (CA, for benthic invertebrates) or principal component analysis (PCA, for lake morphometry and chemistry); the geographic coordinates were left untransformed. The PROTEST method, as re-described by Peres-Neto and Jackson (2001), can actually be used to test the significance of the relationship between data matrices in all situations where co-inertia and orthogonal Procrustes analyses are applicable (Section 11.5). In R, function ***protest()*** of VEGAN uses Trace**W** instead of m_{12}^2 or m_{21}^2 as the test statistic in the Procrustean permutation test.

PROTEST

Numerical simulations carried out by Peres-Neto & Jackson (2001) showed that PROTEST was more powerful than the Mantel test to identify correlations generated between raw data matrices. This finding is in accordance with the conclusions of other authors, reported in Box 10.3, that the Mantel test should not be used to test hypotheses concerning correlations between raw data matrices.

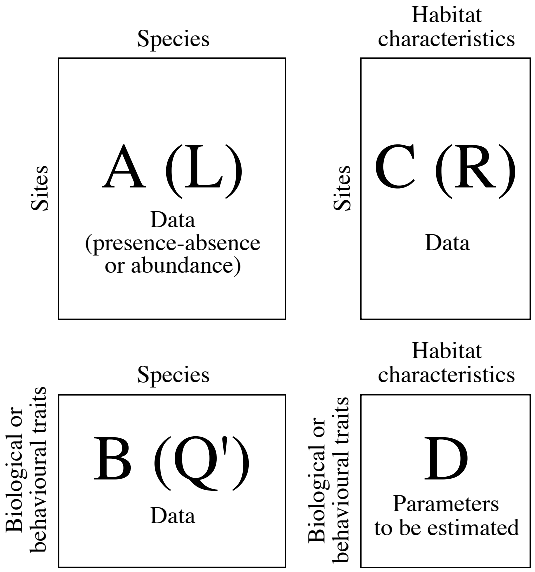

Species
Habitat characteristics
Sites
A (L)
Data (presence-absence or abundance)
Sites
C (R)
Data
Species
Habitat characteristics
Biological or behavioural traits
B (Q')
Data
Biological or behavioural traits
D
Parameters to be estimated

Figure 10.24 Given the information in matrices **A**, **B**, and **C**, the fourth-corner problem is to estimate the parameters in the fourth-corner matrix **D** that crosses the habitat characteristics with the biological or behavioural traits of the species. In Dray & Legendre (2008), matrix **A** is called **L**, **B** is called **Q'**, and **C** is called **R**.

10.6 The fourth-corner problem

How do the biological and behavioural characteristics of species determine the niches they occupy or their geographic locations in an ecosystem?

This question, which stems from niche theory, has long been neglected by ecologists because they lacked an appropriate method of analysis. Observation of species in nature helps ecologists formulate hypotheses in that respect. Testing such hypotheses requires (1) a way of detecting relationships between species traits and habitat characteristics, and (2) of testing the significance of these relationships. A method of analysis for this problem was proposed by Legendre *et al*. (1997a) and the statistical theory was completed by Dray & Legendre (2008) and by ter Braak *et al*. (2012).

Consider a matrix **A** ($n \times p$) containing data on the presence-absence or abundance of p species at n sites (Fig. 10.24)[*]. A second matrix **B** ($q \times p$) describes q biological or behavioural traits of the same p species. A third matrix **C** ($n \times m$) contains information about m habitat characteristics (environmental variables) at the n sites. How does one

[*] Matrices **A** to **D** are transposed compared to the presentation in Legendre *et al*. (1997).

go about associating the q biological and behavioural traits to the m habitat characteristics? To help find a solution, let us translate the problem into matrix form:

$$\begin{bmatrix} \mathbf{A}_{(n \times p)} & \mathbf{C}_{(n \times m)} \\ \mathbf{B}_{(q \times p)} & \mathbf{D}_{(q \times m)} \end{bmatrix} \qquad \textbf{(10.34)}$$

Using this representation, the problem may now be stated as follows:

- How does one go about estimating the parameters in matrix **D** ($q \times m$) where the q biological and behavioural traits are related to the m habitat characteristics?

- Are these parameters significant in some sense, i.e. are they different from 0 (no relationship) or from the value they could take in a randomly organized environment?

The statistical problem of estimating the parameters in matrix **D** is referred to as the *fourth-corner problem* because matrix **D** lies in the fourth corner of the matrix arrangement shown in eq. 10.34. Data in matrix **A** belong to the presence/absence or abundance types (only presence-absence data were considered by Legendre *et al.*, 1997a). Matrices **B** and **C** may contain quantitative or qualitative (nominal) data. The papers referenced at the beginning of the section describe solutions to accommodate the different types of variables. The relationship between **B** and **C** mediated by **A** can also be analysed by a related method called RLQ analysis (Dolédec *et al.*, 1996).

1 — Comparing two qualitative variables

The first situation considered here concerns two qualitative variables, one from matrix **B** (behaviour), the other from matrix **C** (habitat). Any qualitative variable can be expanded into a series of binary variables, one for each state (Subsection 1.5.7).

Numerical example. In test cases 1 and 2 (Table 10.8), **A** is a matrix of presence-absence of species at two sites; **B** and **C** contain supplementary variables (qualitative, two states) for the rows and columns of **A**, respectively. To fix ideas, let us assume that the variable in **B** describes two feeding habits (herbivorous, carnivorous) and **C** is the nature of the substrate at the study sites on a coral reef (live coral, turf). This example describes the approach for qualitative variables (Subsection 10.6.1) and introduces the method for significance testing (Subsection 10.6.2).

Matrices **A**, **B** and **C** (or **L**, **Q** and **R)** are all needed to estimate the parameters in the fourth-corner matrix **D**. The three matrices can be combined by multiplication around the set of four matrices while preserving matrix compatibility:

clockwise: $$\mathbf{D} = \mathbf{B}\,\mathbf{A}'\,\mathbf{C} \quad \text{or} \quad \mathbf{D} = \mathbf{Q}'\,\mathbf{L}'\,\mathbf{R} \qquad \textbf{(10.35)}$$

or counter-clockwise: $$\mathbf{D}' = \mathbf{C}'\,\mathbf{A}\,\mathbf{B}' \quad \text{or} \quad \mathbf{D}' = \mathbf{R}'\,\mathbf{L}\,\mathbf{Q} \qquad \textbf{(10.36)}$$

For the two test cases of the numerical example, matrix **D** is shown in Table 10.8. Equations 10.35 and 10.36 have an equivalent in traditional statistics. If the data in **A**,

Table 10.8 Test cases for qualitative variables. Matrices are transposed to reduce their widths in the page. **A'** is (10 species × 2 sites), **B'** is (10 species × 2 feeding habits), and **C'** is (2 habitat types × 2 sites). So, **D'** is (2 habitat types × 2 feeding habits). Probabilities (p) are one-tailed, assuming that H_1 states the sign of the relationship. H_1 is indicated by a sign in each cell of **D'**, + meaning that the actual value is larger than the expected value and is tested in the upper tail, and – in the opposite case. Probabilities computed after 9999 permutations. *E* = exact probabilities; see text.

Test case 1

A':	Site 1	Site 2	**B'**: Herbiv.	Carniv.
Sp. 1	1	0	0	1
Sp. 2	0	1	0	1
Sp. 3	1	0	0	1
Sp. 4	1	0	0	1
Sp. 5	1	0	0	1
Sp. 6	0	1	1	0
Sp. 7	0	1	1	0
Sp. 8	0	1	1	0
Sp. 9	0	1	1	0
Sp. 10	0	1	1	0

C':	Site 1	Site 2	**D'**: Herbiv.	Carniv.
Live coral	1	0	0 – p = 0.029 *E* = 0.031	4 + p = 0.189 *E* = 0.188
Turf	0	1	5 + p = 0.029 *E* = 0.031	1 – p = 0.189 *E* = 0.188

Contingency statistic:

G = 8.4562, p (9999 permutations) = 0.021

Test case 2

A':	Site 1	Site 2	**B'**: Herbiv.	Carniv.
Sp. 1	1	1	0	1
Sp. 2	1	1	0	1
Sp. 3	1	1	0	1
Sp. 4	1	1	0	1
Sp. 5	1	1	0	1
Sp. 6	1	1	1	0
Sp. 7	1	1	1	0
Sp. 8	1	1	1	0
Sp. 9	1	1	1	0
Sp. 10	1	1	1	0

C':	Site 1	Site 2	**D'**: Herbiv.	Carniv.
Live coral	1	0	5 p = 1.000 *E* = 1.000	5 p = 1.000 *E* = 1.000
Turf	0	1	5 p = 1.000 *E* = 1.000	5 p = 1.000 *E* = 1.000

Contingency statistic:

G = 0.0000, p (9999 permutations) = 1.000

Inflated data matrix

B, and **C** are frequencies, they can be combined to form an "inflated data matrix". Matrix **D**, which results from crossing the two columns of the inflated matrix, is a contingency table as shown in Table 10.9; values d_{ij} in matrix **D** are frequencies or pseudo-frequencies (see Ecological application 10.6). So, a solution that naturally comes to mind for significance testing is to compute a χ^2 statistic, using either Pearson's (eq. 6.5) or Wilks' formula (eq. 6.6, also called the *G* statistic). The *G* statistic is used here; it is the first type of fourth-corner statistic.

Table 10.9 Inflated data matrix (left); there is one row in this matrix for each species "presence" (value 1) in matrix **A'** of test case 1 (Table 10.8). The contingency table (matrix **D'**, right) is constructed from the inflated matrix.

Inflated data matrix

Occurrences in test case 1	Feeding habits from **B**	Habitat types from **C**
Sp. 1 @ Site 1	Carnivorous	Live coral
Sp. 2 @ Site 2	Carnivorous	Turf
Sp. 3 @ Site 1	Carnivorous	Live coral
Sp. 4 @ Site 1	Carnivorous	Live coral
Sp. 5 @ Site 1	Carnivorous	Live coral
Sp. 6 @ Site 2	Herbivorous	Turf
Sp. 7 @ Site 2	Herbivorous	Turf
Sp. 8 @ Site 2	Herbivorous	Turf
Sp. 9 @ Site 2	Herbivorous	Turf
Sp. 10 @ Site 2	Herbivorous	Turf

Contingency table

D':	Herbivorous	Carnivorous
Live coral	0	4
Turf	5	1

Dray & Legendre (2008) have shown that species abundance data can be used as well as presence-absence data in the calculation of fourth-corner statistics and in the permutation tests described in the next two subsections.

For large contingency tables **D**, relationships among descriptor states could be visualized in a correspondence analysis (CA) biplot (Subsection 9.2.1). Consider matrix **D** shown in Table 10.10 (below) as an example. It may be simplified as follows before CA: for the cells where d_{ij} is significant, code those that are above the expected value (sign + in the matrix) with +1 and those that are below the expected value (sign – in the matrix) with –1. Code the non-significant cells with 0. After coding, add 1 to all cells because CA requires that the values in the matrix subjected to the analysis be non-negative. Carry out CA of the coded matrix and use scaling type 4 for the biplot. A CA biplot remains a simplified summary; it contains less precise information than the original matrix **D**.

2 — Test of statistical significance

In fourth-corner problems, one cannot test the G statistics in the usual manner because, in the general case (although not in test case 1 of Table 10.8), several species are observed at any one sampling site so that the rows of the inflated matrix are not independent of one another; several rows of that matrix result from observations at a single site. To solve the problem, G is tested by permutations (Section 1.2). The procedure is as follows.

Permutation test

Hypotheses

- H_0: the species traits (matrix **B**) are unrelated to the characteristics of the sites (matrix **C**), their relationships (links) being mediated by the species presence-absence or abundance data (matrix **A**). Different permutation null models are detailed in the next subsection.

- H_1: the species traits are related to the characteristics of the sites.

Test statistic

Compute a χ^2 statistic (G here) on the contingency table (matrix **D**) and use it as the reference value for the remainder of the test.

Distribution of the test statistic

Under H_0, the species found at any one site could have been observed at any other site. Where the species have actually been observed is due to chance alone. So, a realization of H_0 is obtained by permuting at random the values in matrix **A**, using one of the methods described in the next subsection. After each permutation of matrix **A**, recompute the χ^2 statistic on **D**.

- Repeat the permutation a large number of times (say, 999 or 9999 times). The different permutations produce a set of values of the χ^2 statistic, obtained under H_0.

- Add to this set the reference value of the statistic, computed for the unpermuted data matrix. Together, the unpermuted and permuted values (for a total of 1000 values, 10000 values, etc.) form an estimate of the sampling distribution of χ^2 under H_0.

Statistical decision

As in any other statistical test, the decision is made by comparing the reference value of the χ^2 statistic to the distribution obtained under H_0. If the reference value of χ^2 is one likely to have been obtained under the null hypothesis, H_0 is not rejected. If it is too extreme (i.e. located out in a tail) to be considered a likely result under H_0, then H_0 is rejected.

Individual values d_{ij} in matrix **D** can also be tested for significance, as shown below in the numerical example and the ecological application.

In addition, a global test of significance can be carried out for the fourth-corner relationship involving all variables in matrices **A** and **C** in the analysis. The global test uses statistic S_{RLQ}, which is the trace of a cross-product matrix computed from the fourth-corner matrix **D**. See Dray & Legendre (2008, eq. 8). This quantity is equal to the total inertia of an RLQ analysis (Dolédec *et al*., 1996).

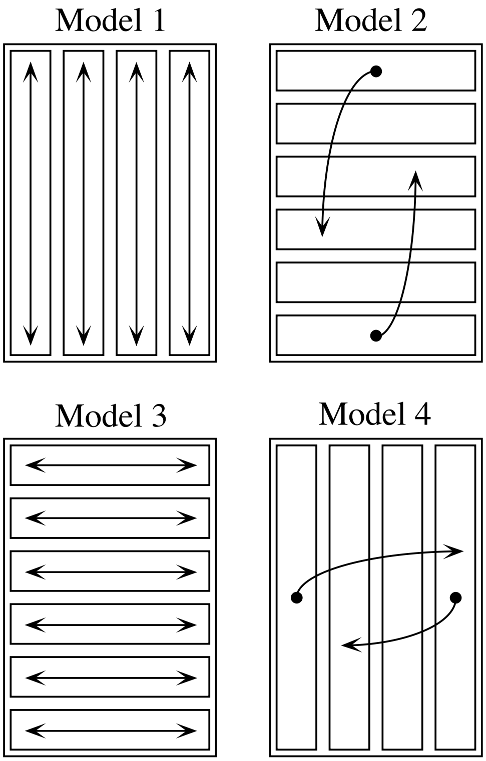


Figure 10.25 Permutations of matrix **A** may be performed in different ways which correspond to different null ecological models.

(1) The occurrence of a species in the study area is constant, but positions are random; permute at random within columns.

(2) Positions of species assemblages are random; permute whole rows (assemblages).

(3) Lottery hypothesis: the species that arrived first occupied a site; permute at random within rows.

(4) Species have random attributes; permute whole columns.

3 — Permutational models

Permutations may be conducted in different ways, depending on the ecological hypotheses to be tested against observations. Technically, the fourth-corner statistical method can accommodate any of the permutation models described below. The random component is clearly the field information about the species found at the sampling sites, i.e. matrix **A**. It is thus matrix **A** that should be permuted (randomized) for the purpose of hypothesis testing. This may be done in various ways (Fig. 10.25). Models 1 to 4 were described by Legendre *et al*. (1997a), model 5 by Dray & Legendre (2008).

Model 1: Environmental control over individual species. — Permute the species presence-absence or abundance data within each column of matrix **A**, independently from column to column. This not only destroys the link between **A** and **C**, but also the relationship between **A** and **B**, as shown by Dray & Legendre (2008, Appendix A). The null hypothesis (H_0) states that individuals of a species are randomly distributed with respect to the site characteristics. The corresponding alternative hypothesis (H_1) states that individuals of a species are distributed according to their preferences for site conditions. Under this permutation model, the number of sites occupied by each species is kept constant, as in permutation method 2b of the Raup & Crick similarity coefficient (S_{27}, Chapter 7).

Model 2: Environmental control over species assemblages. — Permute entire rows of matrix **A** at random. This method destroys the link between **A** and **C** but keeps **A** linked to **B**; it is equivalent to permuting the rows of **C**. H_0 states that species assemblages are randomly attributed to sites, irrespective of the site characteristics. The corresponding alternative hypothesis (H_1) states that species assemblages are dependent upon the physical characteristics of the locations where they are found. This method preserves the covariances among the species throughout the permutations, as well as the number of sites occupied by each species.

Model 3: Lottery. — Permute the species data within each row of matrix **A**, independently from row to row. This not only destroys the link between **A** and **B**, but also the relationship between **A** and **C**. H_0 states that the distribution of the presences of various species at a site is the result of a random allocation process (the lottery for space model advocated by Sale, 1978); it is not due to the adaptation of the species traits to the sites. The alternative hypothesis (H_1) states that due to their traits, species have some competitive advantages over chance settlers in the habitats where they are found. Under this model, the number of species present in a given site (i.e. species richness) is kept constant.

Model 4: Random species attributes. — Permute entire columns of matrix **A** at random. This destroys the link between **A** and **B** but keeps **A** linked to **C**; it is equivalent to permuting the columns of **B**. H_0 states that species are distributed according to their preferences for site conditions, but irrespective of their traits or other characteristics included in **B**. The alternative hypothesis (H_1) states that the distributions of the species among the sites, which are related to their preferences for site conditions, depend on the adaptations (traits) of the species. Under this model, the number of species present at each site (i.e. the species richness) is kept constant.

Model 5: Permute rows and columns. — Permute entire rows of **A** at random, then (or before) permute entire columns at random. The links between **A** and **B** and between **A** and **C** are destroyed. An alternative, equivalent method is to permute at random the rows of **C** and the columns of **B** while keeping **A** fixed, which was the method used by Dolédec *et al.* (1996). H_0 states that the species distributions among the sites are not related to the site conditions nor to the traits of the species. The alternative hypothesis (H_1) states that the species distributions across the sites are related to species traits, and/or that species assemblages are dependent upon the environmental conditions.

The type I error rate and power of these permutation models were studied by Dray & Legendre (2008) under six data generation scenarios. All permutation models have nearly equal power to detect a relationship between **B** and **C** mediated by **A** when such a relationship is present in the data. In the opposite situation, when there is no relationship between **A** and **B** nor between **A** and **C**, all permutation models have correct rates of type I error and reject H_0 at the α significance level. However, in some simulation situations, the permutation models differ in type I error rates, which can be very high with some models (Dray & Legendre, 2008). This makes it difficult, with some models, to interpret a rejection of H_0: does it mean that there is a relationship between **B** and **C** via **A**, or is it a type I error?

- When one can assume that a relationship exists between **A** and **B** (i.e. the species have fixed trait values) and the test concerns the relationship between **A** and **C**, (i.e. test the species-environment relationship), permutation model 2 has a correct type I error rate; so this permutation model can be used in that situation.

- Similarly, in the opposite situation, when one can assume that a relationship exists between **A** and **C** and the test concerns the relationship between **A** and **B**, permutation model 4 can be used.

- When no *a priori* assumption can be made about existing relationships (the data in **B** and **C** are considered random instead of fixed), the best strategy is to carry out two tests in sequence using permutation models 2 and 4 and take the maximum of the two p-values as the probability of the data under the combined null hypothesis (ter Braak *et al*., 2012).

Numerical example. Let us examine how the fourth-corner method behaves when applied to the data sets introduced in the numerical example of Subsection 10.6.1. The first test case (Table 10.8, left) was constructed to suggest that herbivores are found on turf while carnivores are more ubiquitously distributed. Globally, the G statistic indicates a significant relationship ($\alpha = 0.05$) between behavioural states and types of habitat ($p = 0.0207$ after 9999 random permutations under model 1 above). The expected values in the various cells of matrix **D** determine the tail in which each frequency d_{ij} of the contingency table is to be tested for significance; this value is taken to be the mean frequency expected from all possible permutations of matrix **A**, given the permutation model that has been selected. Looking at individual values d_{ij}, herbivores are clearly positively associated with turf and negatively with coral ($p = 0.0287$, computed from the random permutation results), while carnivores are not significantly associated with either live coral or turf ($p = 0.1890$). These probabilities are very close to the exact probabilities calculated for the same data, which are the values obtained from a complete permutation procedure (E in Table 10.8). Values of exact probabilities E are computed as follows: consider all possible permutations that result from independently permuting the rows of matrix **A** (permutation model 1); count how many of these would produce values equal to, or more extreme than the observed value in each given cell of matrix **D**. This value may differ slightly from the random permutational probability. Globally, the testing procedure for the relationship between behaviour and habitat behaved as expected in this example, and the random permutation procedure produced values quite close to the exact probabilities.

The second test case (Table 10.8, right) illustrates a situation where the null hypothesis is true in all cases, matrix **A** indicating all 10 species to be present everywhere. Indeed, the testing procedure finds all permutation statistics to be equal to the unpermuted ones, so that the probability of the data under the null hypothesis is 1 everywhere. The procedure once more behaved correctly.

4 — Other types of comparisons among variables

Variables in matrices **B** and **C** are not always qualitative. Through lines of reasoning similar to that of Subsection 10.6.1, involving inflated data matrices (as in Table 10.9), fourth-corner statistics can be formulated to accommodate other types of variable comparisons.

- To compare a quantitative variable in **B** to a quantitative variable in **C**, a Pearson correlation coefficient may be computed between the columns of the inflated matrix. A correlation coefficient is directly obtained from the fourth-corner equation **D = BA'C** if the columns of the inflated data matrix are first standardized and the scalar product is divided by the number of rows of the inflated matrix minus 1.

- When comparing a quantitative variable in **B** to a qualitative variable coded into dummy variables (Subsection 1.5.7) in **C**, or the converse, the fourth-corner matrix product (eq. 10.35) is equivalent to computing an overall F-statistic for the pair of variables, as explained in Legendre *et al.* (1997a); the cells d_{ij} of matrix **D** contain measures of within-group homogeneity. Correlations may also be computed between the quantitative variable on the one hand, and each of the dummy variables coding for the qualitative variable on the other hand.

Each of these statistics can be tested for significance using the permutational procedure described in Subsection 10.6.2.

The fourth-corner method offers a way of analysing the relationships between *supplementary variables* associated with the rows and columns of a community composition data matrix. Other types of problems could be studied using this method. Here are two examples.

- In biogeography, consider a matrix **A** of presence/absence or abundance of species; a matrix **B** describing the extensiveness of the species' distributions, their migratory behaviour, etc.; and a matrix **C** of habitat characteristics (environmental variables), as above. The question is again to relate habitat to species characteristics.

- In the study of feeding behaviour, consider a matrix **A** with columns that are *individuals* while rows correspond to sites. The prey ingested by each individual are found in matrix **B**. Matrix **C** may contain either microhabitat environmental variables, or prey availability variables. The question is to determine feeding preferences: choice of prey *versus* availability, or choice of prey *versus* microhabitat conditions. Problems of the same type are found in such fields as sociology, marketing, political science, and the like.

- In studies involving spatial data, matrix **C** may contain spatial eigenfunctions (Chapter 14) representing the spatial relationships among the study sites. A global test of significance can be carried our between the characteristics of the species in **B** and the spatial eigenfunctions in **C** using the global statistic S_{RLQ}.

Ecological application 10.6

Development of the fourth-corner method was motivated by the study of a fish assemblage (280 species) surveyed along a one-km transect across the coral reef of Moorea Island, French Polynesia (Legendre *et al*., 1997a). Biological and behavioural characteristics of the species were used as descriptors (supplementary variables) for the rows, and characteristics of the environment for the columns of the fish presence-absence data matrix **A**. Parameters of the relationship between habitat characteristics (distance from the beach, water depth, and substrate variables) and biological and behavioural traits of the species (feeding habits, ecological niche categories, size classes, egg types, activity rhythms) were estimated and tested for significance. Results were compared to predictions made independently by reef fish ecologists, in order to assess the method as well as the pertinence of the variables subjected to the analysis.

Table 10.10 summarizes the comparison of reef bottom materials to feeding habits. This is an interesting case: the eight "reef bottom materials" variables are relative frequencies; each one represents the proportion of the habitat covered by a category of substrate material, so that non-integer pseudo-frequencies are obtained in the contingency table where the variables are crossed (Table 10.10). The permutation testing procedure allows data in matrices **B** and **C** to be relative or absolute frequencies. Probabilities remain the same under any linear transformation of the frequency values, even though the value of the G statistic is changed. This would not be allowed by a standard χ^2 test.

The relationship is globally significant (G = 15.426, p(G) = 0.0001 after 9999 random permutations following model 1 of Subsection 10.6.3 above); 20 of the 56 fourth-corner statistics d_{ij} were significant (*) after applying Holm's correction for multiple testing (Box 1.3). Compared to the null hypothesis, fish are under-represented on sand and large algae, and are unrelated to stone slab. In addition, herbivores are over-represented on live coral and calcareous algae. Grazers of sessile invertebrates and carnivores of types 1 and 2 are over-represented on coral debris, turf and dead coral, live coral, calcareous algae, and other types of substrate (large echinoderms, sponges, anemones, alcyonarians); this includes all areas where herbivores are found. Copepod eaters are over-represented on live coral and calcareous algae. Omnivores and specialist piscivores (fish-only diet) do not exhibit significant relationships with substrate.

Distance from the beach and size of fish species (adult individuals) are quantitative variables. The fourth-corner statistic that crosses these two variables is thus correlation-like; its value is $r = 0.0504$, with a probability of 0.001 after 999 random permutations. There is thus a weak but significant correlation, indicating that larger fish are found farther away from the beach than smaller ones. Other comparisons between biological-behavioural and habitat variables are presented in the published paper.

10.7 Software

Functions in the R language are available to carry out all analyses described in this chapter.

1. Linear regression. — In package STATS, function ***lm()*** computes simple or multiple linear regression. Function ***step()*** used in conjunction with ***lm()*** offers model selection by *AIC* using a backward, forward, or stepwise strategy.

Table 10.10 Contingency table comparing feeding habits (7 states) to materials covering reef bottom (8 proportions). From Legendre *et al*. (1997a, Table 6). First row in each cell: pseudo-frequency resulting from the matrix operation **D** = **BA'C**; lower row, probability adjusted using Holm's procedure; *: p ≤ 0.05. Probabilities before correction resulted from 9999 random permutations. Sign indicates whether a statistic is above (+) or below (–) the expected value, estimated as the mean of the permutation results.

	Herbivorous	Omnivorous	Sessile invertebrates	Carniv. 1 diurnal	Carniv. 2 nocturnal	Fish only	Copepod eater
Stone slab	6.20–	5.84+	3.72–	8.42–	5.18+	0.96+	2.40–
p	0.429	0.232	1.535	2.650	2.650	2.650	2.650
Sand	81.22–	54.26–	43.34–	94.38–	35.90–	8.94–	26.26–
p	0.039*	0.799	0.006*	0.006*	0.006*	0.799	0.039*
Coral debris	34.96+	20.22–	24.32+	46.74+	25.60+	4.48+	12.08–
p	1.976	1.976	0.006*	0.009*	0.645	2.650	2.650
Turf, dead cor.	45.46+	27.88+	28.28+	57.58+	33.58+	6.20+	15.76+
p	0.207	2.650	0.081	0.013*	0.029*	1.976	2.650
Live coral	49.86+	28.50+	29.20+	58.28+	40.82+	6.22+	21.06+
p	0.006*	1.976	0.006*	0.006*	0.006*	1.976	0.006*
Large algae	44.66–	37.50+	28.12–	59.68–	32.26–	6.34–	19.20–
p	0.006*	2.650	0.105	0.048*	0.140	2.650	2.650
Calcar. algae	29.12+	16.32+	16.08+	31.00+	26.02+	4.50+	11.32+
p	0.006*	1.030	0.079	0.122	0.006*	0.207	0.036*
Other substrate	2.52+	1.48+	1.94+	2.92+	1.64+	0.36+	0.92+
p	0.105	2.650	0.006*	0.795	1.734	1.976	1.976

Functions ***lmodel2()*** of LMODEL2 and ***sma()*** of SMATR compute model II simple linear regressions. Function ***lmorigin()*** in APE computes regression through the origin with permutation test. Variance inflation factors are computed by function ***vif()*** of packages CAR and DAAG, applied to models computed by ***lm()***.

QR decomposition, carried out by function ***qr()*** of BASE, is an efficient method to compute coefficients in univariate or multivariate linear regression. Multivariate linear regression can be computed using either ***lm()***, which takes either a single variable **y** or a whole matrix **Y** as the response data, or ***qr()*** after incrementing the explanatory matrix **X** with a column of 1's to estimate the intercept, producing matrix $\mathbf{X}_{+1}$. For example, the matrix of fitted values in multivariate regression can be computed as follows: fitted(lm(as.matrix(**Y**) ~ ., data=**X**)), or qr.fitted(qr($\mathbf{X}_{+1}$), as.matrix(**Y**)).

Ridge regression is available in functions ***lm.ridge()*** of MASS, ***ridge()*** of SURVIVAL, and ***penalized()*** of PENALIZED. Generalized linear models are computed by function ***glm()*** of STATS. Among the generalized linear models, only logistic regression is discussed in detail in the present chapter; it is computed by ***glm****(y~x, family=binomial(logit))*. In STATS, function ***nls()*** computes nonlinear weighted least-squares estimates of the parameters of a nonlinear statistical model; ***optim()*** is a general-purpose nonlinear optimization function offering a variety of optimization algorithms.

2. Partial regression and variation partitioning. — Partial linear regression can be computed by function ***rda()*** of VEGAN. ***varpart()*** of VEGAN is used for variation partitioning; ***plot.varpart()*** plots a Venn diagram with fixed circle and intersection sizes. A Venn diagram with proportional circle and intersection sizes can be obtained with function ***venneuler()*** of package VENNEULER[*].

3. Path analysis. — Structural equation modelling, which is a generalized form of analysis encompassing path analysis, is available in package SEM.

4. Matrix comparisons. — Simple Mantel tests are found in functions ***mantel.test()*** of APE and ***mantel.rtest()*** of ADE4. For simple and partial Mantel tests, use ***mantel()*** of VEGAN, ***mantel()*** of ECODIST, ***mantel.test()*** and ***partial.mantel.test()*** of NCF. ***protest()*** in VEGAN computes the Procrustes permutation test. ***anosim()*** in VEGAN computes the ANOSIM test. The ***MRM()*** function in ECODIST carries out multiple regression on distance matrices.

5. Fourth-corner problem. — Functions ***fourthcorner()*** and ***fourthcorner2()*** of ADE4 compute fourth-corner analysis; function ***rlq()*** of ADE4 carries out RLQ analysis.

6. Miscellaneous methods. — Function ***poly()*** of STATS computes ordinary or orthogonal polynomials, the latter of the degree specified by the user, from a data vector. The resulting monomial vectors are normalized (i.e. scaled to length 1, eq. 2.7) and made to be orthogonal to one another. Several packages contain functions for spline and LOWESS smoothing, e.g. STATS, SPLINES and DIERCKXSPLINE.

[*] *Beware*: the fraction names in the *combinations* option of function ***venneuler()*** follow a different convention than in ***varpart()***. For two explanatory matrices for example, the first element mentioned, e.g. A, is fraction [c] of Fig. 10.10; the second element, e.g. B, is fraction [a]; the intersection [b] is called "A&B". See the examples in the documentation file.