When the analysis involves three descriptors only, the simple method of *causal modelling using correlations* may be used (Subsection 4.5.5). For three resemblance matrices, causal modelling may be carried out using the results of Mantel and partial Mantel tests, as described in Subsection 10.5.2 and Section 13.6.

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

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 $x_1, x_2, \ldots, x_p$. 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 set of relationships among variables, whose parameters have to be estimated, or that are to be tested; 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 analysing ecological structures. It is also used as a platform to help introduce the principles of structure analysis. The same principles will apply to more advanced forms, collectively referred to as canonical analysis, that are discussed in Chapter 11.

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

* A random variable is a variable whose values are assumed to result from some random process (p. 1); 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 hypothesis.

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 (i.e. the slope is different from 0), and/or it may state that the intercept is different from zero. The test consists in finding the *two-tailed* probability of observing the slope (b_1) and/or intercept (b_0) values which have been estimated from the sample data, given the null hypothesis (H_0) stating that the slope (β_1) and/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 and/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 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), or textbooks such as those of Ratkowski (1983) or Ross (1990) for nonlinear estimation.

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.

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 mentioned, such as dummy variable regression, 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 offspring. 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

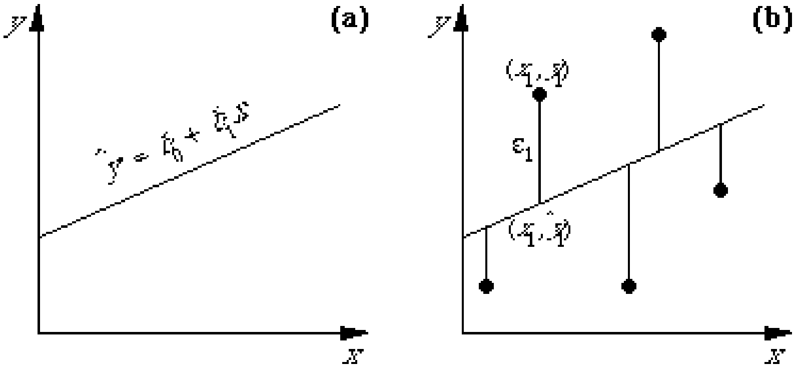

Figure 10.5 (a) Linear regression line, of equation $\hat{y} = b_0 + b_1 x$, fitted to the scatter of points shown in b. (b) Graphical representation of regression residuals ε_i (vertical lines); ε_1 is the residual for point 1 with coordinates (x_1, y_1).

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:

$$\hat{y} = b_0 + b_1 x \tag{10.1}$$

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 $\hat{y}$ (ordinate 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 the Subsection on polynomial regression, a distinction will be made between linearity in parameters and linearity in response to the explanatory variables.

Intercept
Slope

When using this type of regression, one must be aware of the fact that a *linear model* is imposed. 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 may lie above or 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) of the data set; y_i is equal to the value $\hat{y}_i$ predicted by the regression equation plus the residual ε_i:

$$y_i = \hat{y}_i + \varepsilon_i = b_0 + b_1 x_i + \varepsilon_i \qquad \textbf{(10.2)}$$

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 $y = b_0 + b_1 x$ 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 4 and 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 of x, the y's 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.

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 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

Least squares

OLS

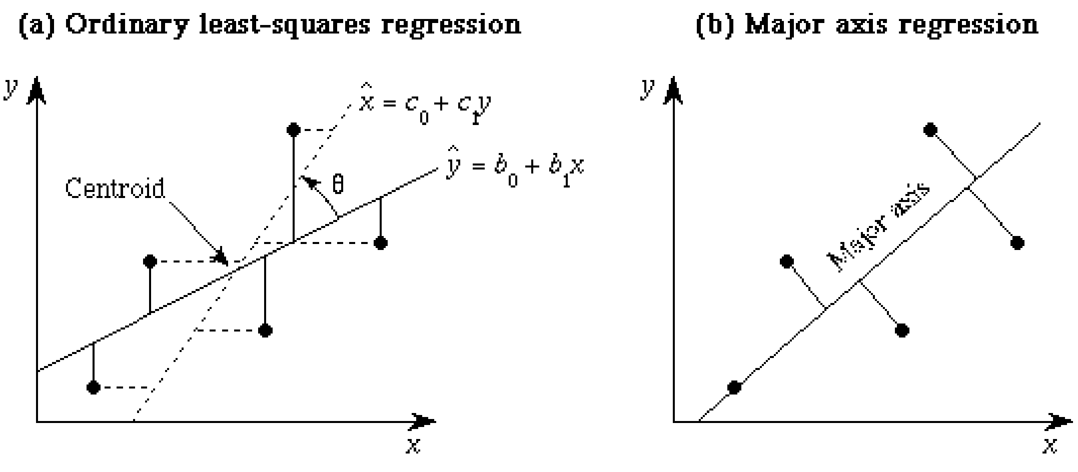

Figure 10.6 (a) Two least-squares regression equations are possible in the case of two random variables (called x and y here, for simplicity). When regressing y on x, the sum of *vertical* squared deviations is minimized (full lines); when regressing x on y, the sum of *horizontal* squared deviations 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.

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:

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

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 may also be computed directly from the **x** and **y** data vectors, using matrix 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; they would thus 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 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 computations, there is a direct relationship between the Pearson correlation coefficient r_{xy} and angle θ between the two regression lines:

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

If $r = 0$, the scatter of points is circular and angle $\theta = 90°$, 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:

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

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 $\hat{y}_i$. It is 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. With two random variables, the regression of y on x makes as much sense as the regression of x on y. In this case, the coefficient of determination may be computed as the product of the two regression coefficients:

$$r^2 = b_{1(y \cdot x)} c_{1(x \cdot y)} \qquad \textbf{(10.7)}$$

In other words, the coefficient of correlation is the geometric mean of the coefficients of linear regression of each variable on the other: $r = (b_{1(y \cdot x)} c_{1(x \cdot y)})^{1/2}$. 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 the one may, indeed, be directly calculated 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 is error associated with the measurements of x and y. Such situations are often referred to as *model II regression*; they are not the regression equivalent of model II ANOVA, though. Examples are:

- 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 allows ecologists to use the second substance as a proxy for the first.

- In aquatic ecology, *in vivo* fluorescence is routinely used to estimate phytoplankton chlorophyll *a*. The two variables must be determined to establish their relationship; they are both random and measured with error.

- In comparative growth studies, one may use length as an indicator for the mass of individuals pertaining to a given taxonomic group.

In all these cases, one may be interested in estimating the parameters of the functional relationship, or to use one variable to forecast the other. In model II regression, different computational procedures may be required for description and inference, as opposed to forecasting. Other applications follow.

- In freshwater sediment, one may be interested to compare the rate of microbial anaerobic methane production to total particulate carbon, in two environments (e.g. two lakes) in which several sites have been studied, but that differ in some other way. 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; the confidence intervals of these slopes may serve to compare the two environments.

- Deterministic models are often used to describe ecological processes. In order to test how good a model is at describing reality, one uses field data about the control variables incorporated in the model, and compares the predictions of the model to the observed field values of the response variable. Since both sets of variables (control,