Country		Dr A.'s results		Dr B.'s results	
		Calluna	*Deschampsia*	*Calluna*	*Deschampsia*
Netherlands	mean	44.3	39.0	6.8	4.8
	s.d.	42.1	41.8	2.1	4.4
Scotland	mean	78.8	9.8	8.2	5.3
	s.d.	6.3	7.3	1.0	1.3

Both investigators agree about *Calluna*: it is more dominant in Scotland than in the Netherlands. However they disagree about *Deschampsia*, since Dr B. concludes that *Deschampsia* is more important in Scotland than in the Netherlands. Note also the very high standard deviations in the Dutch samples, which might be caused by a skewed or bimodal distribution.

3 Regression

C.J.F. ter Braak and C.W.N. Looman

3.1 Introduction

3.1.1 Aim and use

Regression analysis is a statistical method that can be used to explore relations between species and environment, on the basis of observations on species and environmental variables at a series of sites. Species may be recorded in the form of abundances, or merely as being present. In contrast with ordination and cluster analysis, we cannot analyse data on all species simultaneously; in regression analysis, we must analyse data on each species separately. Each regression focuses on a particular species and on how this particular species is related to environmental variables. In the terminology of regression analysis, the species abundance or presence is the response variable and the environmental variables are explanatory variables. The term 'response variable' stems from the idea that the species react or respond to the environmental variables in a causal way; however, causality cannot be inferred from a regression analysis. The goal of regression analysis is more modest, namely to describe the response variable as a function of one or more explanatory variables. This function, termed the response function, usually cannot be chosen such that the function predicts responses without errors. By using regression analysis, we attempt to make the errors small and to average them to zero. The value predicted by the response function is then the expected response: the response with the error averaged out.

Regression analysis is well suited for what Whittaker (1967) termed 'direct gradient analysis'. In ecology, regression analysis has been used mainly for the following:

- estimating parameters of ecological interest, for example the optimum and ecological amplitude of a species
- assessing which environmental variables contribute most to the species' response and which environmental variables appear to be unimportant. Such assessment proceeds through tests of statistical significance
- predicting the species' responses (abundance or presence–absence) at sites from the observed values of one or more environmental variables
- predicting the values of environmental variables at sites from observed values of one or more species. Such prediction is termed calibration and is treated separately in Chapter 4.

3.1.2 Response model and types of response variables

Regression analysis is based on a response model that consists of two parts: a systematic part that describes the way in which the expected response depends on the explanatory variables; and an error part that describes the way in which the observed response deviates from the expected response.

The systematic part is specified by a regression equation. The error part can be described by the statistical distribution of the error. For example, when fitting a straight line to data, the response model (Figure 3.1) is

$$y = b_0 + b_1\,x + \varepsilon \qquad \text{Equation 3.1}$$

with
y the response variable
x the explanatory variable
ε the error
b_0 and b_1 fixed but unknown coefficients; they are the intercept and slope parameter, respectively.

The expected response, denoted by Ey, is equal to $b_0 + b_1\,x$. The systematic part of the model is thus a straight line and is specified by the regression equation

$$\mathrm{E}y = b_0 + b_1\,x.$$

The error part is the distribution of ε, i.e. the random variation of the observed response around the expected response. The aim of regression analysis can now be specified more precisely. The aim is to estimate the systematic part from data while taking account of the error part of the model. In fitting a straight line, the systematic part is simply estimated by estimating the parameters b_0 and b_1.

In the most common type of regression analysis, least-squares regression, the distribution of the error is assumed to be the normal distribution (Subsection 2.4.3). Abundance values of a species commonly show a skew distribution that looks like a log-normal distribution (Subsection 2.4.3), with many small to moderate values and a few extremely large values. Abundance values often show this type of distribution even among sites whose environmental conditions are apparently identical. By transforming the abundance values to logarithms, their distribution becomes more like a normal distribution (Williamson 1972). To analyse abundance values by least-squares regression, it is therefore often more appropriate to use log-abundance values. A problem then arises when the species is absent, because the abundance is then zero and the logarithm of zero is undefined.

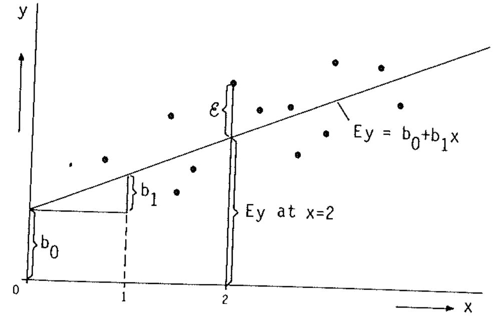

Figure 3.1 Response model used in fitting a straight line to data points (●) by least squares regression. For explanation see Subsection 3.1.2.

A regression technique appropriate for presence–absence data is logit regression. Logit regression attempts to express the probability that a species is present as a function of the explanatory variables.

3.1.3 Types of explanatory variables and types of response curves

The explanatory variables can be nominal, ordinal or quantitative (Subsection 2.4.2). Regression techniques can easily cope with nominal and quantitative environmental variables, but not with ordinal ones. We suggest treating an ordinal variable as nominal when the number of possible values is small, and as quantitative when the number of possible values is large.

Regression with a single quantitative explanatory variable consists of fitting a curve through the data. The user must choose in advance how complicated the fitted curve is allowed to be. The choice may be guided by looking at a scatter

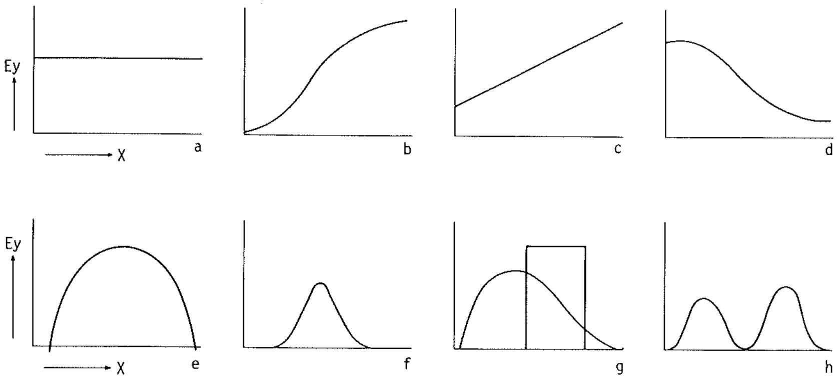

Figure 3.2 Shapes of response curves. The expected response (Ey) is plotted against the environmental variable (x). The curves can be constant (a: horizontal line), monotonic increasing (b: sigmoid curve. c: straight line), monotonic decreasing (d: sigmoid curve), unimodal (e: parabola. f: symmetric, Gaussian curve. g: asymmetric curve and a block function) or bimodal (h).

plot of the response variable against the explanatory variable or can be guided by available knowledge and theory about the relation. We denote the environmental variable by the letter x and the expected response by Ey, the expected value of the response y. We distinguish the following types of curves, often referred to as response curves (Figure 3.2):

- constant: Ey is equal to a constant; the expected response does not depend on x (Figure 3.2a).
- monotonically increasing (or decreasing): Ey increases (or decreases) with increasing values of x. Examples are straight lines and sigmoid curves (Figure 3.2b,c,d).
- unimodal (single-peaked): Ey first increases with x, reaches a maximum and after that decreases. Examples are the parabola with a maximum (Figure 3.2e), and a bell-shaped curve like the Gaussian curve (Figure 3.2f). The value of x where Ey reaches its maximum is termed the mode or optimum. The optimum does not need to be unique when the curve has a 'plateau' (Figure 3.2g). A unimodal curve can be symmetric (with the optimum as point of symmetry) or asymmetric (Figure 3.2g).
- bimodal: Ey first increases with x, reaches a maximum, then decreases to a minimum, after which Ey increases to a new maximum, from which Ey finally decreases again (Figure 3.2h).
- other: Ey has another shape.

The types of curves are listed in order of their complexity. Only in the simplest case of a constant response curve does the environmental variable have no effect on the response of the species. Monotonic curves can be thought of as special cases of unimodal curves; when the optimum lies outside the interval that is actually sampled, then the unimodal curve is monotonically increasing or decreasing within that interval (Figure 3.3). Similarly, unimodal curves can be special cases of bimodal curves (Figure 3.3). Curves with a single minimum fall in our classification in the category 'other', but can also be thought of as special cases of bimodal curves (Figure 3.3).

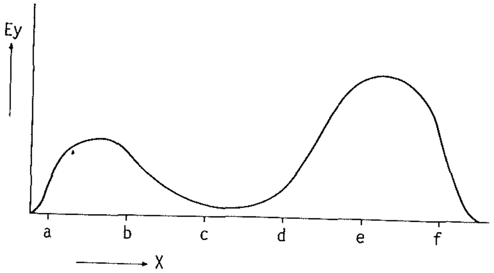

Figure 3.3 Response curves derived from a bimodal curve by restricting the sampling interval. The curve is bimodal in the interval a-f, unimodal in a-c and in d-f, monotonic in b-c and c-e and almost constant in c-d. In the interval b-e, the curve has a single minimum. (Ey, expected response; x, environmental variable).

3.1.4 *Outline of the chapter*

In this chapter, we introduce regression techniques for analysing quantitative abundance data (least-squares regression, Section 3.2) and presence–absence data (logit regression, Section 3.3). In both sections, we first present a model in which the explanatory variable is nominal, and then models in which the explanatory variable is quantitative, in particular models that are based on straight lines and parabolas.

From the straight line and the parabola, we derive curves that are more useful in ecological data analysis. For abundance data, we derive from them the exponential curve and the Gaussian curve, respectively, and for the analysis of presence–absence data, the sigmoid curve and the Gaussian logit curve. The curves based on a parabola allow estimation of the indicator value (optimum) and the ecological amplitude (tolerance) of the species. Problems involved in analysing quantitative data containing many zero values are dealt with in Section 3.4. In Section 3.5, both least-squares regression and logit regression are extended to multiple regression. Multiple regression can be used to study the effect of many environmental variables on the response by the species, be it quantitative or of presence–absence type. The topic of Section 3.6 is model choice and regression diagnostics. Finally, we leave regression and introduce the method of weighted averaging, which is a simple method for estimating indicator values of species. This method has a long tradition in ecology; it has been used by Gause (1930). We compare the weighted averaging method with the regression method to estimate indicator values.

3.2 Regression for quantitative abundance data: least-squares regression

3.2.1 *Nominal explanatory variables: analysis of variance*

The principles of regression are explained here using a fictitious example, in which we investigate whether the cover proportion of a particular plant species at sites systematically depends on the soil type of the sites. We distinguish three soil types, namely clay, peat and sand. The observed cover proportions showed a skew distribution within each soil type and therefore we decided to transform them by taking logarithms. Before taking logarithms, we added the value 1 to the cover expressed in percentages, to avoid problems with the two zero values in the data. Figure 3.4 displays the resulting response values for each soil type.

Our response model for this kind of data is as follows. The systematic part simply consists of three expected responses, one for each soil type, and the error part is the way in which the observed responses within each soil type vary around the expected responses in each soil type. On the basis of Figure 3.4, it appears not unrealistic to assume for the error part of the response model that the transformed relative covers within each soil type follow a normal distribution and that the variance of this distribution is the same for each of the three soil types. We further assume that the responses are independent. These assumptions constitute the response model of the analysis of variance (ANOVA), which is

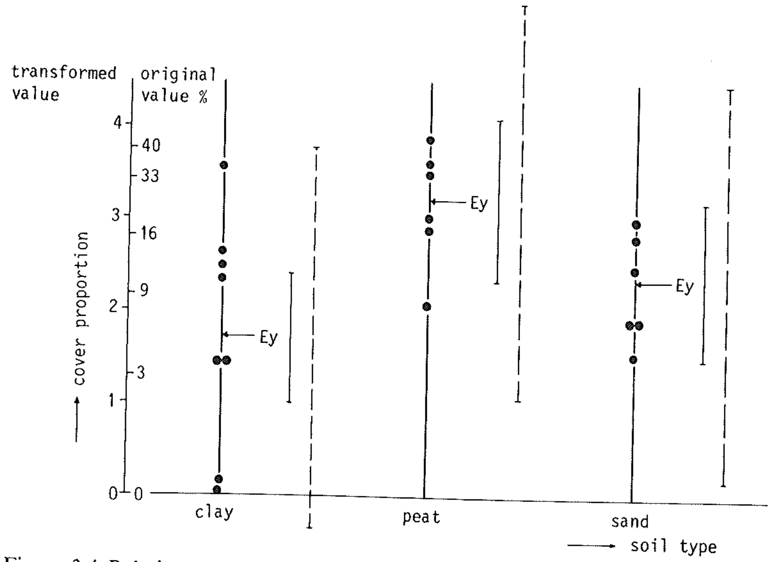

Figure 3.4 Relative cover (log-transformed) of a plant species (●) in relation to the soil types clay, peat and sand. The horizontal arrows indicate the mean value in each type (Table 3.1). The solid vertical bars show the 95% confidence interval for the expected values in each type and the dashed vertical bars the 95% prediction interval for the log-transformed cover in each type. (fictitious data).

one particular form of least-squares regression.

The first step in regression analysis is to estimate the parameters of the model. The parameters are here the expected responses in the three soil types. We estimate them by using the least-squares principle. We choose values for the parameters such that the sum (over all sites) of the squared differences between observed and expected responses is minimal. The parameter values that minimize this sum of squares, are simply the mean values of the transformed relative covers for each soil type. The expected response as fitted (estimated) by regression is, therefore just the mean of the response in each soil type. The fitted values are indicated by arrows in Figure 3.4. The difference between an observed response (a dot in Figure 3.4) and the fitted value is termed a residual, which in Figure 3.4 is a vertical distance. Least-squares thus minimizes a sum of squared vertical distances; the minimum obtained is called the residual sum of squares.

Regression analysis by computer normally gives not only parameter estimates but also an analysis-of-variance table (ANOVA table). From the ANOVA table (Table 3.1), we can derive how well the regression equation explains the response variable. In the example, the fraction of variance accounted for (R^2_{adj}) is 0.25, which means that only a quarter of the variance in the responses is explained by the differences between soil types. The ANOVA table can further be used for testing statistically whether the expected responses differ among soil types; that is, whether the mean values for each soil type differ more than could be expected by chance if soil type did not affect the relative cover. For this test, the variance ratio F (Table 3.1) must be compared with the critical value of an F distribution with 2 and 17 degrees of freedom in the numerator and denominator, respectively (2 and 17 are the degrees of freedom for the regression and residual in the ANOVA table). The critical value (at the 5% significance level) is 3.59. (Consult for this a table of the F distribution, for instance in Snedecor & Cochran 1980.) In the example, the variance ratio (4.248) is larger than 3.59. Under the null hypothesis of equal expected responses, this happens in 5% of the cases only. So it is unlikely that the expected responses are equal. From the ANOVA table, we can thus conclude that the expected responses do differ and we say that the cover proportions differ significantly between soil types at the 5% level ($P < 0.05$, F test).

Table 3.1 Means and ANOVA table of the transformed relative cover of Figure 3.4.

Term	mean	s.e.	95% confidence interval
Clay	1.70	0.33	(1.00, 2.40)
Peat	3.17	0.38	(2.37, 3.97)
Sand	2.33	0.38	(1.53, 3.13)

Overall mean 2.33

ANOVA table

	d.f.	s.s	m.s.	F
Regression	2	7.409	3.704	4.248
Residual	17	14.826	0.872	
Total	19	22.235	1.170	

$R^2_{\text{adj}} = 0.25$

How precisely have we estimated the expected responses? An indication for this is the standard error of the estimates (Table 3.1). The standard error can be used to construct a confidence interval for the expected response. The endpoints of a 95% confidence interval for a parameter (and the expected response is a parameter in this example) are given by

$$\text{(estimate)} \pm t_{0.05}(\nu) \times \text{(standard error of estimate)} \qquad \text{Equation 3.2}$$

The symbol ± is used to indicate addition or subtraction in order to obtain upper and lower limits. The symbol $t_{0.05}(\nu)$ denotes the 5% critical value of a two-tailed t test. The value of $t_{0.05}(\nu)$ depends on the number of degrees of freedom (ν) of the residual and can be obtained from a t table (e.g. Snedecor & Cochran 1980). In our example, $\nu = 17$ and $t_{0.05}(17) = 2.11$, which gives the intervals shown in Figure 3.4 and Table 3.1.

We may also want to predict what responses are likely to occur at new sites of a particular soil type. A prediction interval for new responses is much wider than the confidence interval for the expected response. To construct a prediction interval, we need to know the residual standard deviation. This is the standard deviation of the residuals and is obtained from the ANOVA table by taking the square root of the 'mean square of the residual'. We obtain from Table 3.1 the residual standard deviation of $\sqrt{(0.872)} = 0.93$. In the example, the residual standard deviation is simply an estimate of the standard deviation within soil types. The prediction interval within which 95% of the new responses fall is now given by

$$\text{(estimated response)} \pm t_{0.05}(\nu)\sqrt{(\text{s.d.}^2 + \text{s.e.}^2)}$$ Equation 3.3

where s.d. is the residual standard deviation and s.e. the standard error of the estimated response. Equation 3.3 yields for clay the interval

$$1.70 \pm 2.11\sqrt{(0.93^2 + 0.33^2)} = (-0.38, 3.78).$$

If we had done many observations, the estimated response would be precisely the expected response. Then s.e. $= 0$ and $t_{0.05}(\infty) = 1.96$, so that Equation 3.3 reduces to: expected response $\pm$ 1.96 $\times$ s.d. Figure 3.4 also displays the prediction intervals for the three soil types.

That procedure is sufficient for ANOVA by computer and for interpretation of the results. But for a better understanding of the ANOVA table, we now show how it is calculated. After we have estimated the parameters of the model, we can write each response as

$$\text{observed value} = \text{fitted value} + \text{residual}.$$ Equation 3.4

For example, one of the observed responses on peat is 3.89 (corresponding to a cover of 48%). Its fitted value is 3.17, the mean response on peat, and the residual is thus $3.89 - 3.17 = 0.72$. We therefore write this response as $3.89 = 3.17 + 0.72$.

Each term in Equation 3.4 leads to a sum of squares; we first subtract the overall mean (2.33) from the observed and fitted values and then calculate (over all sites) sums of squares of the observed values and of the fitted values so corrected and of the residuals. These sums of squares are the total sum of squares, the regression sum of squares and the residual sum of squares, respectively, and are given in Table 3.1 in the column labelled with s.s. (sum of squares). The total sum of squares is always equal to the regression sum of squares and the residual sum of squares added together. Each sum of squares is associated with several degrees of freedom (d.f. in Table 3.1). The number of degrees of freedom equals $n - 1$ for the total sum of squares (n being the number of sites), $q - 1$ for the regression sum of squares (q being the number of estimated parameters, the value 1 is subtracted because of the correction for the overall mean) and $n - q$ for the residual sum of squares.

In the example, $n = 20$ and $q = 3$. The column labelled m.s. (mean square) is obtained by dividing the sum of squares by its number of degrees of freedom. The mean square of the residual is a measure of the difference between the observed and the fitted values. It is the variance of the residuals; hence its usual name residual variance. Similarly, the total variance is obtained; this is just the sample variance of the responses, ignoring soil type. The fraction of variance accounted for by the explanatory variable can now be defined as

$$R^2_{\text{adj}} = 1 - (\text{residual variance}/\text{total variance}),$$

which is also termed the adjusted coefficient of determination. In the example, $R^2_{\text{adj}} = 1 - (0.872/1.170) = 0.25$. The original, unadjusted coefficient of determination (R^2) does not take into account how many parameters are fitted as compared to the number of observations, its definition being

$$R^2 = 1 - (\text{residual sum of squares}/\text{total sum of squares}).$$

When a large number of parameters is fitted, R^2 may yield a value close to 1, even when the expected response does not depend on the explanatory variables. The multiple correlation coefficient, which is the product–moment correlation between the observed values and the fitted values, is just the square root of the coefficient of determination. Finally, the ratio of the mean squares of the regression and the residual is the variance ratio (F). If the expected responses are all equal, the variance ratio randomly fluctuates around the value 1, whereas it is systematically greater than 1, if the expected values differ; hence its use in statistical testing.

3.2.2 Straight lines

In Figure 3.5a, the explanatory variable is mean water-table, a quantitative variable that enables us to fit a curve through the data. A simple model for these data is a straight line with some scatter around the line. The systematic part of the response model is then

$$\text{E}y = b_0 + b_1 x$$ Equation 3.5

in which
Ey denotes the expected value of the response y
x denotes the explanatory variable, the mean water-table
b_0 and b_1 are the parameters that must be estimated
b_0 is the intercept (the value at which the line crosses the vertical axis)
b_1 is the slope parameter or the regression coefficient of the straight line (Figure 3.1)
b_1 is the expected change in y divided by the change in x.

The error part of the model is the same as for ANOVA (Subsection 3.2.1), i.e.

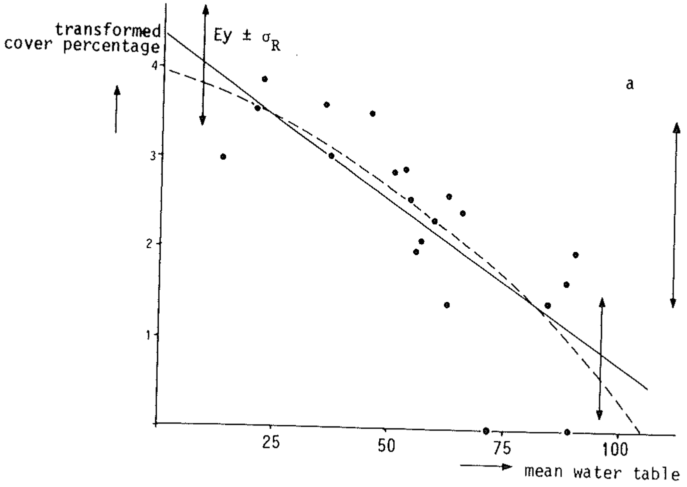

Figure 3.5a Straight line fitted by least-squares regression of log-transformed relative cover on mean water-table. The vertical bar on the far right has a length equal to twice the sample standard deviation σ_T, the other two smaller vertical bars are twice the length of the residual standard deviation (σ_R). The dashed line is a parabola fitted to the same data (●).

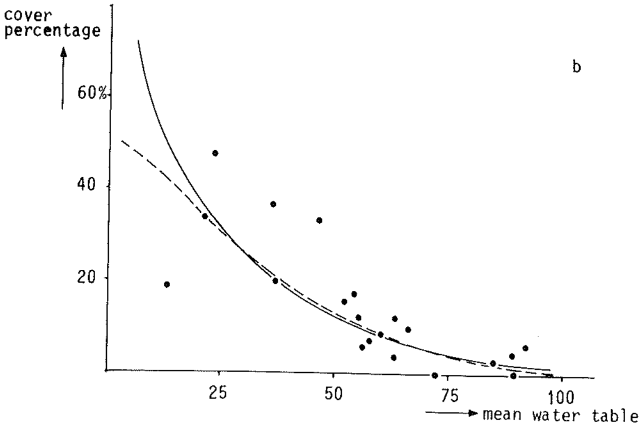

Figure 3.5b Relative cover in relation to water-table with curves obtained by back transformation of the straight line and parabola of Figure 3.5a.

the responses are taken to be mutually independent and are normally distributed around their expected values (E*y*) as specified by the straight line (Equation 3.5). The errors are thus taken to follow a normal distribution and the variance of the errors to be independent of the value of *x*.

We again use the least-squares principle to estimate the parameters. That is, we choose arbitrary values for b_0 and b_1, calculate with these values the expected responses at the sites by Equation 3.5, calculate the sum of squared differences between observed and expected responses, and stop when we cannot find values for b_0 and b_1 that give a smaller sum of squared differences. In Figure 3.5a, this procedure means that we choose the line such that the sum of squares of the vertical distances between the data points and the line is least. (Clearly, any line with a positive slope is inadequate!) For many regression models, the estimates can be obtained by more direct methods than by the trial and error method just described. But, the estimates are usually obtained by using a computer program for regression analysis so that we do not need to bother about the numerical methods used to obtain the least-squares estimates. For the straight-line model, we need, for later reference, the equations for estimating b_0 and b_1

$$b_0 = \bar{y} - b_1 \bar{x} \qquad \text{Equation 3.6a}$$

$$b_1 = \Sigma_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) / \Sigma_{i=1}^{n} (x_i - \bar{x})^2 \qquad \text{Equation 3.6b}$$

where

y_i and x_i are the values of y and x at the *i*-th site
$\bar{y}$ and $\bar{x}$ are the mean values of y and x, respectively.

Table 3.2 shows standard output of a computer program for regression analysis, in which b_0 is estimated at 4.411 and b_1 at −0.0370. The ANOVA table can in

Table 3.2 Straight line fitted by least-squares: parameter estimates and ANOVA table for the transformed relative cover of Figure 3.5.

Term	Parameter	estimate	s.e.	*t*
Constant	b_0	4.411	0.426	10.35
Water-table	b_1	−0.0370	0.00705	−5.25

ANOVA table

	d.f.	s.s	m.s.	*F*
Regression	1	13.45	13.45	27.56
Residual	18	8.78	0.488	
Total	19	22.23	1.170	

$R^2_{\text{adj}} = 0.58$

principle be obtained by the rules of Subsection 3.2.1 below Equation 3.4, by using fitted values as calculated from Equation 3.5. The following statistics are derived from the ANOVA table as in Subsection 3.2.1. The residual standard deviation is $\sqrt{0.488} = 0.70$, which is much smaller than the standard deviation of the observed response, $\sqrt{1.170} = 1.08$.

The fraction of variance accounted for by the straight line is 0.58. The multiple correlation coefficient reduces in straight line regression to the absolute value of the product-moment correlation between x and y (0.78 in Figure 3.5a). The variance ratio can again be used for statistical testing, here for testing whether the expected responses depend on the mean water-table. The critical F at the 5% significance level is now 4.41, because there is only 1 degree of freedom for the regression (Snedecor & Cochran 1980). Because the variance ratio (27.56) exceeds this F, the expected response does depend on the mean water-table. An alternative for this F test is to use a two-tailed t test of whether b_1 equals 0; if b_1 were 0, the straight line would be horizontal, so that the expected response would not depend on x. This t test uses the t of b_1, which is the estimate of b_1 divided by its standard error (Table 3.2). This value (–5.25) is greater (in absolute value) than the critical value of a two-tailed t test at the 5% level obtained from a t table: $t_{0.05}(18) = 2.10$, and so b_1 is not equal to zero; thus the relative cover of our species does significantly depend on the mean water-table. Yet another way of testing whether $b_1 = 0$ is by constructing a 95% confidence interval for b_1 with Equation 3.2. The result is the interval $-0.037 \pm 2.10 \times 0.00705 = (-0.052, -0.022)$.

The value 0 does not lie in this interval and 0 is therefore an unlikely value for b_1. Which of the three tests to use (F test, t test or test through the confidence interval) is a matter of convenience; they are equivalent in straight-line regression.

After regression analysis, we should make sure that the assumptions of the response model have not been grossly violated. In particular, it is useful to check whether the variance of the errors depends on x or not, either by inspecting Figure 3.5a or by plotting the residuals themselves against x. Figure 3.5a does not give much reason to suspect such a dependence.

In the analysis, we used transformed relative covers. The data and the fitted straight line of Figure 3.5a are back-transformed to relative covers in Figure 3.5b. The fitted line is curved on the original scale: it is an exponential curve. Note that in Figure 3.5b, the assumption that the error variance is independent of x does not hold; this could have been a reason for using a transformation in the first place. It may be instructive now to do Exercise 3.1.

3.2.3 *Parabolas and Gaussian curves*

In Subsection 3.2.2, we fitted a straight line to the responses in Figure 3.5a. But wouldn't a concave curve have been better? We therefore extend Equation 3.5 with a quadratic term in x and obtain the parabola (Figure 3.2e)

$$Ey = b_0 + b_1 x + b_2 x^2 \qquad \text{Equation 3.7}$$

Table 3.3 Parabola fitted by least-squares regression: parameter estimates and ANOVA table for the transformed relative cover of Figure 3.5.

Term	Parameter	estimate	s.e.	t
Constant	b_0	3.988	0.819	4.88
Water-table	b_1	–0.0187	0.0317	–0.59
(Water-table)^2	b_2	–0.000169	0.000284	–0.59

ANOVA table

	d.f.	s.s	m.s.	F
Regression	2	13.63	6.815	13.97
Residual	17	8.61	0.506	
Total	19	22.23	1.170	

$R^2_{adj} = 0.57$

We again use the least-squares principle to obtain estimates. The estimates are given in Table 3.3. The estimates for b_0 and b_1 change somewhat from Table 3.2; the estimate for b_2 is slightly negative. The parabola fitted (Figure 3.5, dashed line) gives a slightly smaller residual sum of squares than the straight line. But, with the change in the number of degrees of freedom of the residual (from 18 to 17), the residual variance is greater and the fraction of variance accounted for is lower. In the example, the advantage of the parabola over the straight line is therefore doubtful. A formal way to decide whether the parabola significantly improves the fit over the straight line is by testing whether the extra parameter b_2 is equal to 0. Here we use the t test (Subsection 3.2.2). The t of b_2 (Table 3.3) is much smaller in absolute value than the critical value, 2.11; hence the data provide no evidence against b_2 being equal to 0. We conclude that a straight line is sufficient to describe the relation between the transformed cover proportions and mean water-table; a parabola is not needed.

Generally the values of t for b_0 and b_1 in Table 3.3 are not used, because they do not test any useful hypothesis. For example, the t test of whether b_1 is equal to 0 in Equation 3.7 would test a particular kind of parabola against the general parabola of Equation 3.7, quite different from the meaning of the t test of the slope parameter b_1 in Table 3.2.

In principle, we can extend the response function of Equation 3.7 to higher-order polynomials in x by adding terms in x^3, x^4, ... There is no advantage in doing so for the data in Figure 3.5a. Polynomial regression of species data has limited use except for one special case. *When we fit a parabola to log-transformed abundances, we actually fit a Gaussian response curve to the original abundance data.* The Gaussian response curve has the formula

$$z = c \exp\left[-0.5(x - u)^2 / t^2\right] \qquad \text{Equation 3.8}$$

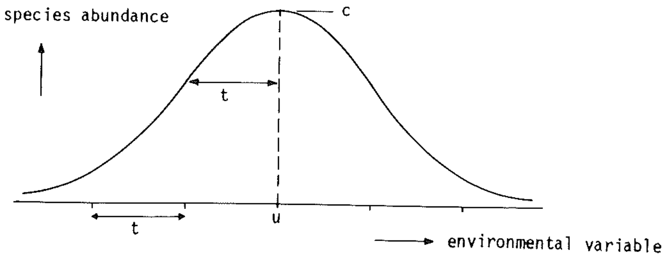

Figure 3.6 Gaussian response curve with its three ecologically important parameters: maximum (c), optimum (u) and tolerance (t). Vertical axis: species abundance. Horizontal axis: environmental variable. The range of occurrence of the species is seen to be about $4t$.

where
z is the original abundance value
c is the species' maximum abundance
u is its optimum (the value of x that gives maximum abundance)
t is its tolerance (a measure of ecological amplitude).

Note that in this chapter the symbol t is used in two ways: the t of a regression coefficient (Subsection 3.2.2) and the t of a Gaussian curve. Which t is intended should be clear from the context of the passages concerned.

Figure 3.6 displays the Gaussian curve and its parameters. The curve is seen to rise and fall over a length of about $4t$. If we take the logarithm on both sides of Equation 3.8, we obtain

$$\log_e z = \log_e (c) - 0.5\,(x-u)^2/t^2 = b_0 + b_1\,x + b_2\,x^2 \qquad \text{Equation 3.9}$$

where the third form follows by expanding

$$(x-u)^2 = x^2 - 2\,u\,x + u^2$$

and by setting:

$$b_0 = \log_e (c) - u^2/(2t^2);\ b_1 = u/t^2;\ b_2 = -1/(2t^2). \qquad \text{Equation 3.10}$$

By fitting a parabola to log-abundances, we obtain least-squares estimates for b_0, b_1 and b_2, from which we can obtain estimates of

the optimum, $u = -b_1/(2b_2)$ Equation 3.11a

the tolerance, $t = 1/\sqrt{(-2b_2)}$ Equation 3.11b

the maximum, $c = \exp\,(b_0 + b_1\,u + b_2\,u^2)$. Equation 3.11c

These equations are derived from Equation 3.10 where $b_2 < 0$. If the estimate of b_2 is positive, the fitted curve has a minimum instead of a maximum. Approximate standard errors of the estimated optimum and tolerance can be derived from the variances and covariances of b_1 and b_2 that are provided as options by statistical packages. A confidence interval for the optimum can also be calculated. Details of these calculations are given in Section 3.9.

It may be instructive now to do Exercise 3.2 (except Part 3.2.8).

3.3 Regression for presence-absence data: logit regression

3.3.1 Nominal explanatory variables: chi-square test

Table 3.4 shows the numbers of dune meadow fields in which the plant species *Achillea ptarmica* was present and in which it was absent. The fields are divided into four classes depending on agricultural use. The relevant question for these data is whether the frequency of occurrence of *Achillea ptarmica* depends systematically on agricultural use. This question is analogous to the question that was studied in Subsection 3.2.1, although here the response of the species is not relative cover, but merely presence or absence. The usual thing to do is to calculate the relative frequency in each class, i.e. the number of fields of a given class in which the species is present divided by the total number of fields of that class (Table 3.4). But relative frequency of occurrence is simply the mean value when we score presence as 1 and absence as 0. Calculating means was what we did in Subsection 3.2.1. The response is thus $y = 1$ or $y = 0$, and the expected response, Ey, is the expected frequency, i.e. the probability of occurrence of the species in a field randomly drawn from all fields that belong to the class. Relative frequency is therefore an estimate of probability of occurrence.

Table 3.4 Numbers of fields in which *Achillea ptarmica* is present and absent in meadows with different types of agricultural use and frequency of occurrence of each type (unpublished data from Kruijne et al. 1967). The types are pure hayfield (ph), hay pastures (hp), alternate pasture (ap) and pure pasture (pp).

Achillea ptarmica	Agricultural use				
	ph	hp	ap	pp	total
present	37	40	27	9	113
absent	109	356	402	558	1425
total	146	396	429	567	1538
frequency	0.254	0.101	0.063	0.016	0.073

If the probabilities of occurrence of *Achillea ptarmica* were the same for all four classes, then we could say that its occurrence did not depend on agricultural use. We shall test this null hypothesis by the chi-square test. This test proceeds as follows. Overall, the relative frequency of occurrence is $113/1538 = 0.073$ (Table 3.4). Under the null hypothesis, the expected number of fields with *Achillea ptarmica* is in pure hayfield $0.073 \times 146 = 10.7$ and in hay pasture $0.073 \times 396 = 29.1$ and so on for the remaining types. The expected number of fields in which *Achillea ptarmica* is absent is therefore in pure hayfield $146 - 10.7 = 135.3$ and in hay pasture $396 - 29.1 = 366.9$.

We now measure the deviation of the observed values (o) and the expected values (e) by the chi-square statistic, that is the sum of $(o - e)^2/e$ over all cells of Table 3.4. We get $(37 - 10.7)^2/10.7 + (109 - 135.3)^2/135.3 + (40 - 29.1)^2/29.1 + ... = 102.1$. This value must be compared with the critical value, $\chi^2_\alpha(\nu)$, of a chi-square distribution with ν degrees of freedom, where $\nu = (r - 1)(c - 1)$, r is the number of rows and c the number of columns in the table. In the example, $\nu = 3$ and the critical value at the 5% level is $\chi^2_{0.05}(3) = 7.81$. Consult a χ^2 table, for instance Snedecor & Cochran (1980). The chi-square calculated, 102.1, is much greater than 7.81, and we conclude therefore that the probability of occurrence of *Achillea ptarmica* strongly depends on agricultural use. Notice that the chi-square statistic is a variant of the residual sum of squares: it is a weighted sum of squares with weights $1/e$.

The chi-square test is an approximate test, valid only for large collections of data. The test should not be used when the expected values in the table are small. A rule of thumb is that the test is accurate enough when the smallest expected value is at least 1. A remedy when some expected numbers are too small is to aggregate classes of the explanatory variable.

3.3.2 *Sigmoid curves*

We now look at the situation in which we have a presence–absence response variable (y) and a quantitative explanatory variable (x). Data of this kind are shown in Figure 3.7. Just as in Subsection 3.3.1, the expected response is the probability of occurrence of the species in a site with a particular value of the environmental variable. This probability will be described by a curve. Probabilities always have values between 0 and 1. So a straight-line equation

$$\mathrm{E}y = b_0 + b_1 x \qquad \text{Equation 3.12}$$

is not acceptable, because $b_0 + b_1 x$ can also be negative. This difficulty could be solved by taking the exponential curve

$$\mathrm{E}y = \exp(b_0 + b_1 x) \qquad \text{Equation 3.13}$$

However the right side of Equation 3.13 can be greater than 1, so we adapt the curve once more to

$$\mathrm{E}y = p = [\exp(b_0 + b_1 x)]/[1 + \exp(b_0 + b_1 x)] \qquad \text{Equation 3.14}$$

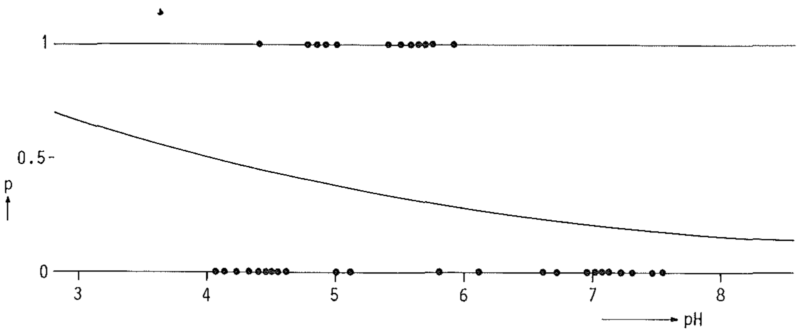

Figure 3.7 Sigmoid curve fitted by logit regression of the presences • at $p = 1$) and absences (• at $p = 0$) of a species on acidity (pH). In the display, the sigmoid curve looks like a straight line but it is not. The curve expresses the probability (p) of occurrence of the species in relation to pH.

This curve satisfies the requirement that its values are all between 0 and 1. The only further reason to take this curve, and not another one, is mathematical convenience. The curves representing Equations 3.12-3.14 are shown in Figure 3.8; Equation 3.14 represents a sigmoid curve. All three curves are monotonic and have two parameters, namely b_0 and b_1. The part $b_0 + b_1 x$ is termed the linear predictor. For probabilities, we use the symbol p instead of Ey (Equation 3.14).

The systematic part of the response model is now defined. Next, we deal with the error part. The response can only have two values, hence, the error distribution is the Binomial distribution with total 1 (Subsection 2.4.3). So the variance of y is $p(1 - p)$. We have now completed the description of the model.

To estimate the parameters from data, we cannot use ordinary least-squares regression because the errors are not normally distributed and have no constant variance. Instead we use logit regression. This is a special case of the generalized linear model (GLM, McCullagh & Nelder 1983). The term logit stems from logit transformation, that is the transformation of p

$$\log_e [p/(1 - p)] = \text{linear predictor} \qquad \text{Equation 3.15}$$

which is just another way of writing

$$p = [\exp(\text{linear predictor})]/[1 + \exp(\text{linear predictor})] \qquad \text{Equation 3.16}$$

The solution to Exercise 3.3 shows that Equations 3.15 and 3.16 are equivalent. The left side of Equation 3.15 is termed the link function of the GLM. Logit regression is sometimes called logistic regression.

In GLM, the parameters are estimated by the maximum likelihood principle. The likelihood of a set of parameter values is defined as the probability of the responses actually observed when that set of values were the true set of parameter

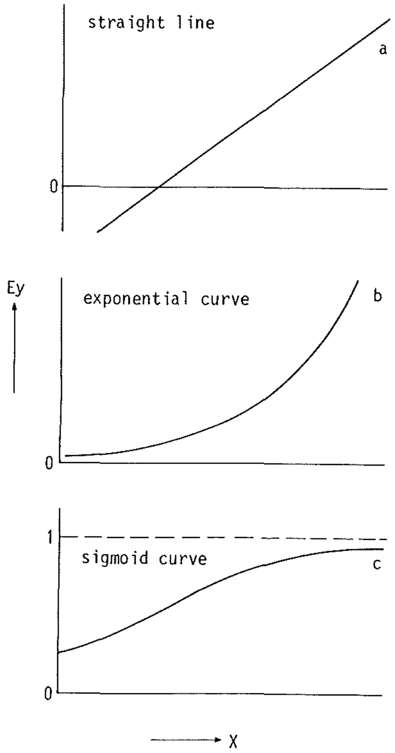

Figure 3.8 Straight line (a), exponental curve (b) and sigmoid curve (c) representing Equations 3.12, 3.13 and 3.14, respectively.

values. The maximum likelihood principle says that we must choose that set of parameter values for which the likelihood is maximum. A measure for the deviation of the observed responses from the fitted responses is the residual deviance, which is $-2 \log_e L$, where L is the maximized likelihood. The residual deviance takes the place of the residual sum of squares in least-squares regression. The least-square principle (Subsection 3.2.1) is equivalent to the maximum likelihood principle, if the errors are independent and follow a normal distribution. Least-squares regression is thus also a special case of GLM. In general, the parameters of a GLM must be calculated in an iterative fashion; provisional estimates of parameters are updated several times by applying repeatedly a weighted least-squares regression, in which responses with a small variance receive a larger weight in the residual sum of squares than responses with a large variance. In logit regression, the variance of the response was $p(1 - p)$. So the weight depends

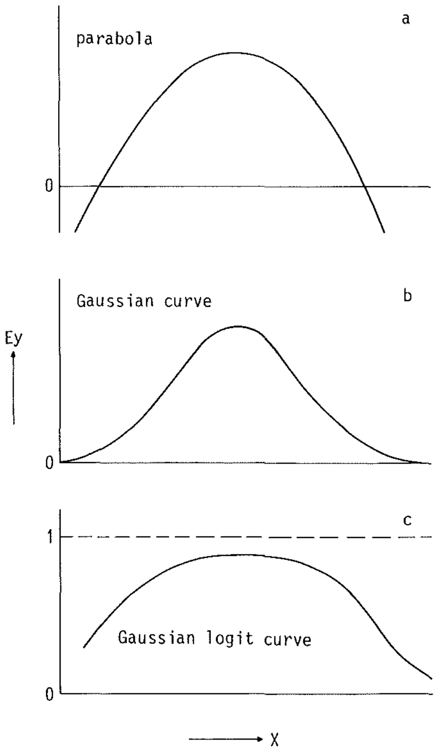

Figure 3.9 Parabola (a), Gaussian curve (b) and Gaussian logit curve (c) representing Equations 3.7, 3.8 and 3.17, respectively.

on the fitted value of p and hence on the parameter estimates; calculations must therefore be iterative. Computer programs for logit regression are available in statistical packages including GLIM (Baker & Nelder 1978), GENSTAT (Alvey et al. 1977), BMDP (Dixon 1981, subprogram PLR) and SAS (Harrell 1980). Ter Braak & Looman (1986) give an example of a program in GLIM.

We fitted the sigmoid curve of Equation 3.14 to the data of Figure 3.7 by logit regression. Table 3.5 shows the estimated parameter and the residual deviance; its number of degrees of freedom is $n - q$, where q is the number of estimated parameters (Subsection 3.2.1). The resulting curve (Figure 3.7) does not differ significantly ($P > 0.05$) from a horizontal line, as judged by a t test of whether b_1 equals 0. All tests in logit regression are approximate, because the error distribution is not normal (cf. the chi-square test of Subsection 3.3.1). Apart from this, there is no difference from the t test described in Subsection 3.2.2.

Table 3.5 Sigmoid curve fitted by logit regression: parameter estimates and deviance table for the presence–absense data of Figure 3.7.

Term	Parameter	estimate	s.e.	t
Constant	b_0	2.03	1.98	1.03
pH	b_1	-0.484	0.357	-1.36
		d.f.	deviance	mean deviance
Residual		33	43.02	1.304

3.3.3 *Gaussian logit curves*

When we take for the linear predictor in Equation 3.16 a parabola, we obtain the Gaussian logit curve

$$\begin{aligned} p &= [\exp{(b_0 + b_1\,x + b_2\,x^2)}]/[1 + \exp{(b_0 + b_1\,x + b_2\,x^2)}] \\ &= c\exp{[-0.5\,(x - u)^2/t^2]}/\;[1 + c\exp{(-0.5\,(x - u)^2/t^2)}] \end{aligned} \qquad \text{Equation 3.17}$$

The third form of the equation follows from Equations 3.8-3.10 and shows the relation to the Gaussian curve (Equation 3.8). The relation between parabola, Gaussian curve and Gaussian logit curve is shown graphically in Figure 3.9 (contrast Figure 3.8). The Gaussian logit curve has a flatter top than the Gaussian curve but the difference is negligible when the maximum of the Gaussian logit curve is small (< 0.5). The Gaussian logit curve was fitted to the data in Figure 3.7 by using GENSTAT and the result is shown in Figure 3.10. Table 3.6 gives the parameter estimates of b_0, b_1 and b_2, from which we obtain estimates for the optimum and the tolerance by using Equations 3.11a,b. The result is $u = 5.28$ and $t = 0.327$. The maximum of the fitted curve in Figure 3.10 is the (estimated) maximum probability of occurrence of the species ($p_{\max}$) and can be calculated from the second form of Equation 3.17 by inserting the value of u (5.28) for x and the values of b_0, b_1 and b_2 from Table 3.6; we obtain $p_{\max} = 0.858$.

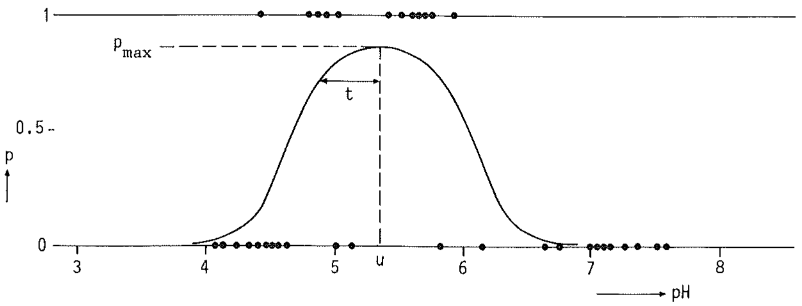

Figure 3.10 Gaussian logit curve fitted by logit regression of the presences (● at $p = 1$) and absences (● at $p = 0$) of a species on acidity (pH). Same data as in Figure 3.7. u = optimum; t = tolerance; $p_{\max}$ = maximum probability of occurrence.

Table 3.6 Gaussian logit curve fitted by logit regression: parameter estimates and deviance table for the presence–absence data of Figure 3.10. The data are the same as in Figure 3.7.

Term		Estimate	s.e.	t
Constant	b_0	-128.8	51.1	-2.52
pH	b_1	49.4	19.8	2.50
pH^2	b_2	4.68	1.90	-2.47
		d.f.	deviance	mean deviance
Residual		32	23.17	0.724

We can decide whether the Gaussian logit curve significantly improves the fit over the sigmoid curve by testing whether b_2 equals 0. Here we use the t test again (Subsection 3.2.3). The t of b_2 is -2.47 (Table 3.6) and we conclude that the fitted curve differs significantly from a sigmoid curve. It is justified to use a one-tailed t test here, if we only want to detect unimodal curves, i.e. curves with $b_2 < 0$ (Snedecor & Cochran 1980, Section 5.5). If b_2 is significantly smaller than 0, then the optimum is said to be significant. An approximate 95% confidence interval for u is (5.0, 5.8), obtained from Section 3.9.

A more general method of statistical testing in GLM is by the deviance test, in which the residual deviance of a model is compared with that of an extended model. The additional parameters in the latter model are significant when the drop in residual deviance is larger than the critical value of a chi-square distribution with k degrees of freedom, k being the number of additional parameters. As an example, the drop in deviance going from the sigmoid curve to the Gaussian logit curve (Tables 3.5 and 3.6) is 43.02 – 23.17 = 19.85. This drop is larger than $\chi^2_{0.05}(1) = 3.84$. Hence the single additional parameter b_2 is significant. The deviance test replaces the F test of least-squares regression.

An example of analysing presence–absence data is provided in Exercises 3.4 and 3.5.

3.4 Regression for abundance data with many zero values

Abundance data with many zero values (i.e. absence) always show a skew distribution. So one should transform them before analysing them by least-squares regression. But the logarithmic transformation does not work, because the logarithm of zero is undefined. The value 0 might be caused by rounding error, but even then one often does not know whether the original value was 0.1, 0.01 or even smaller. On a log scale the difference between these values is large, and one does not know which value to choose. A common practice is to add a small value to the abundance data before logs are taken, as was done in Subsection 3.2.1, but this is somewhat arbitrary; different values may lead to different results of analysis if there are many zeros among the data. An additional problem is that in the model abundance values may be negative, which does not make sense (e.g. the prediction interval for clay in Figure 3.4). Other transformations do not work either.

In least-squares regression after logarithmic transformation, the implicit assumption is that the abundance data follow a log-normal distribution. The probability of observing the value 0 from a log-normal distribution is, however, zero. A distribution that allows zero values is the Poisson distribution (Subsection 2.4.3). Observations arising from a Poisson distribution can take the integer values 0, 1, 2, 3, ... and have a variance that is equal to the mean. Counts of the number of animals in a region, for example, take integer values only. We assume for a moment that the data follow a Poisson distribution and seek appropriate response curves. The curves must not be negative, but may rise above the value 1. The exponential transformation used in Equation 3.13 is therefore sufficient. The exponential curve can be fitted to data by log-linear regression, which is again a special case of GLM (Subsection 3.3.2). The regression is termed log-linear because another way of writing Equation 3.13 is

$$\log_e \mathrm{E}y = \text{linear predictor} \qquad \text{Equation 3.18}$$

By using $b_0 + b_1\, x + b_2\, x^2$ in the linear predictor, we again obtain the Gaussian curve provided $b_2 < 0$ (Equations 3.8-3.11). The Gaussian curve can thus be fitted to abundance data with zero values by carrying out a log-linear regression. In this way we circumvent the problem of having to take logarithms of zeros. The optimum, tolerance and maximum are derived from the estimates of b_0, b_1 and b_2 as in Subsection 3.2.3.

The assumption that abundance data follow a Poisson distribution is often false (Subsection 2.4.3). Fortunately, the assumptions of log-linear regression can be relaxed. It is sufficient that the variance in the data is proportional to the mean (McCullagh & Nelder 1983). When this weaker assumption is also inappropriate, a possible ad-hoc method is to transform the species data to presence–absence. This method sacrifices all the quantitative information. The quantitative information can be retained partly by also analysing 'pseudo-species' (Hill et al. 1975). A pseudo-species is a presence–absence variable that is defined, for instance, by a cut-level value y_c. The pseudo-species at cut-level value y_c is present if the abundance of the species exceeds the cut-level value y_c, and is absent if the abundance is less. By choosing a set of cut levels, we get a set of pseudo-species, each of which can be analysed separately by logit regression. An attractive property of the method of pseudo-species is that the response curve of each pseudo-species is unimodal whenever the response curve for the original abundances is unimodal. Then, the tolerances of the response curves of the pseudo-species decrease with increasing value of the cut level; their optima may shift when the response curve for abundance is asymmetric. A disadvantage of the method is that the choice of cut levels is arbitrary and that the results of the separate analyses cannot be combined easily into a simple description of the relation between the abundance of the species and the environmental variable under consideration.

In some ecological applications the quantitative information on abundance is of the type 'absent, a few, many'. We suggest transforming such data to 'Is the species present?' and 'Is the species abundant?', and to analyse each variable separately by logit regression. The second variable is a pseudo-species.

3.5 Multiple regression

3.5.1 *Introduction*

In the previous sections, the response variable was expressed in various ways as a function of a single environmental variable. A species may, however, respond to more than one environmental variable. To investigate such a situation, we need multiple regression. In multiple regression, the response variable is expressed as a function of two or more explanatory variables (response-surface analysis). Separate analyses of the response for each of the environmental variables cannot replace multiple regression if the environmental variables show some correlation with one another and if there are interaction effects, i.e. if the effect of one variable depends on the value of another variable.

We will show how least-squares regression and logit regression can be extended to study the effect of two environmental variables. The extension to more than two variables will then be obvious. Typical cases of multiple regression will be illustrated in the section on multiple logit regression, although they occur equally in multiple least-squares regression. In separate subsections, we will discuss the analysis of interaction effects and the inclusion of nominal explanatory variables in multiple regression.

3.5.2 *Multiple least-squares regression: planes and other surfaces*

An extension of the straight line to two explanatory variables is a plane (Figure 3.11). A plane has the formula

$$\mathrm{E}y = b_0 + b_1\, x_1 + b_2\, x_2 \qquad \text{Equation 3.19}$$

where

x_1 and x_2 are two explanatory variables
b_0, b_1 and b_2 are parameters or regression coefficients.

b_0 is the expected response when $x_1 = 0$ and $x_2 = 0$. b_1 and b_2 are the rates of change in the expected response along the x_1 and x_2 axes, respectively. b_1 thus measures the change in Ey with x_1 for a fixed value of x_2, and b_2 the change in Ey with x_2 for a fixed value of x_1.

The parameters are, again, estimated by the least-squares method, i.e. by minimizing the sum of squares of the differences between the observed and expected response. This means in geometric terms (Figure 3.11) that the regression plane is chosen in such a way that the sum of squares of the vertical distances between the observed responses and the plane is minimum.

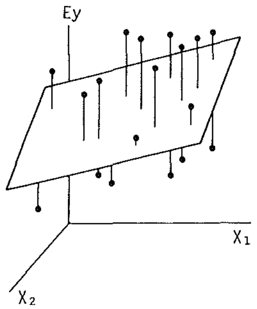

Figure 3.11 Three-dimensional view of a plane fitted by least-squares regression of responses (●) on two explanatory variables x_1 and x_2. The residuals, i.e. the vertical distances between the responses and the fitted plane are shown. Least-squares regression determines the plane by minimization of the sum of these squared vertical distances.

A multiple regression analysis carried out by computer not only gives estimates for b_0, b_1 and b_2, but also standard errors of the estimates and associated values of t (Table 3.3). Fitting a parabola is a special case of multiple regression analysis where $x_1 = x$ and $x_2 = x^2$. The values of t can be used to test whether a coefficient is zero (Subsection 3.2.1), i.e. whether the corresponding variable contributes to the fit of the model in addition to the fit already provided by the other explanatory variable(s).

By extending the parabola we obtain the quadratic surface

$$Ey = b_0 + b_1 x_1 + b_2 x_1^2 + b_3 x_2 + b_4 x_2^2 \qquad \text{Equation 3.20}$$

which has five parameters. When y in this model is the logarithm of abundance, we are fitting through multiple regression a bivariate Gaussian response surface to the observed abundances, provided b_2 and b_4 are both negative. With t tests, we can see whether one of the parameters is equal to zero. In particular, to detect whether the surface is unimodal in the direction of x_1, we test the null hypothesis ($b_2 \geqslant 0$) against the alternative hypothesis ($b_2 < 0$) through the t corresponding to the coefficient b_2, as in Subsection 3.3.3. Similarly, we use the t corresponding to b_4 to test whether the surface is unimodal in x_2.

The optimum and tolerance of the species with respect to x_1 are calculated as in Subsection 3.2.3 by inserting in Equation 3.11a,b the values of b_1 and b_2 obtained from fitting Equation 3.20. Standard errors and the confidence interval for the optimum can still be obtained by using the equations given in Section 3.9. The optimum and tolerance with respect to x_2 are obtained analogously by replacing b_1 by b_3 and b_2 by b_4.

To investigate whether x_2 in this model influences the abundance of a species in addition to x_1, we need to test whether both b_3 and b_4 equal 0. This test requires simultaneous testing of two parameters, which cannot be done with two separate t tests. For this, we need the F test. For an F test, we must fit two regression equations, a simple one with only x_1 and x_1^2 and an extended one, in which x_2 and x_2^2 are added, and compare the residual sum of squares, RSS_1 and RSS_2, respectively, by calculating

$$F = [(\mathrm{RSS}_1 - \mathrm{RSS}_2)/(\mathrm{df}_1 - \mathrm{df}_2)]/\ (\mathrm{RSS}_2/\mathrm{df}_2) \qquad \text{Equation 3.21}$$

where df_1 and df_2 are the degrees of freedom of RSS_1 and RSS_2, respectively.

Under the null hypothesis that the additional parameters b_3 and b_4 equal 0, F follows an F distribution with $\mathrm{df}_1 - \mathrm{df}_2$ and df_2 degrees of freedom (Subsection 3.2.1). The null hypothesis is rejected if the calculated F exceeds the critical value of this distribution. This test can be used whenever simple and extended models are to be compared in multiple least-squares regression. Our previous applications of the F test were special cases of Equation 3.21, in which the simple model was the no-effect model 'Ey is constant'.

3.5.3 Multiple logit regression: logit planes and Gaussian logit surfaces

In Subsection 3.3.2, logit regression was obtained from least-squares regression by replacing Ey by $\log_e [p/(1-p)]$ and there is no reason not to do so in multiple regression. This replacement transforms the plane of Equation 3.19 into a logit plane defined by the equation

$$\log_e [p/(1-p)] = b_0 + b_1 x_1 + b_2 x_2 \qquad \text{Equation 3.22}$$

We will now show what multiple regression can add to the information provided by separate regressions with one explanatory variable. Figure 3.12 displays the values of x_1 and x_2 in a sample of 35 sites and also shows which sites an imaginary species is present at. Fitting Equation 3.22 to the data by using GLM (Subsection 3.3.2) gives the results shown in the first line of Table 3.7. Judged by t tests, both b_1 and b_2 differ significantly from 0, and we conclude that the presence of the species depends both on x_1 and x_2. By fitting a model with x_1 only (Equation 3.14), we obtain the second line of Table 3.7. The estimated probability of occurrence increases somewhat with x_1 ($b_1 = 0.16$), but not significantly (the t of b_1 is 1.33). We would thus have concluded wrongly that the presence of the species did not depend on x_1. From fitting a model with x_2 only, we would also have concluded wrongly that x_2 was irrelevant for predicting presence of the species. By comparing the residual deviances of the models fitted (Table 3.7), we see that x_1 and x_2 are good explanatory variables only when taken together. Such variables are said to be complementary in explanatory power (Whittaker 1984).

The values of b_1 and b_2 in the multiple regression clearly describe the pattern of species occurrence in Figure 3.12. In words, for any given value of x_2, the probability of occurrence strongly increases with x_1, and for any given value of x_1, it strongly decreases with x_2. A line drawn at about 45° in Figure 3.12 actually separates most of the species presences from the absences.

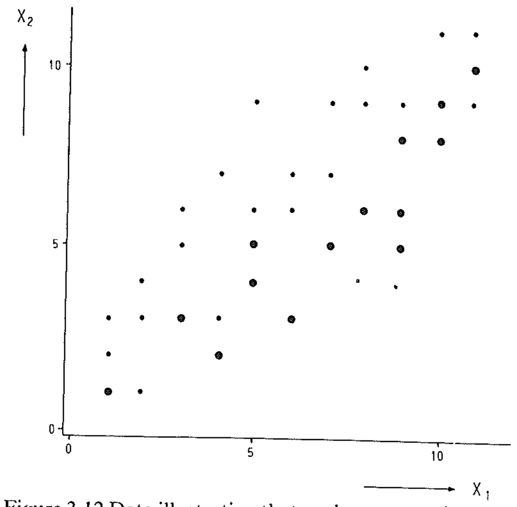

Figure 3.12 Data illustrating that explanatory variables can be complementary in explanatory power. The scatter diagram of x_1 and x_2 shows the sites where a particular species is present (●) and absent (•).

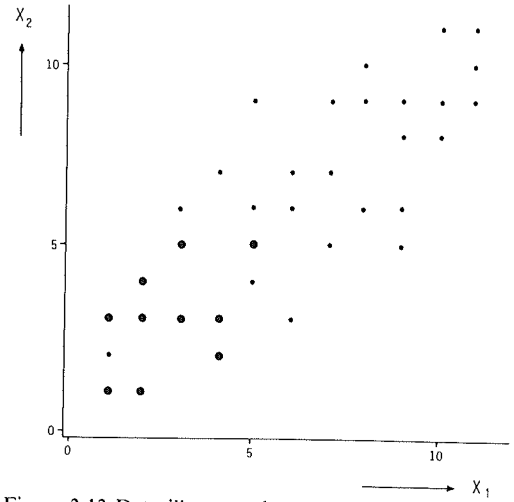

Figure 3.13 Data illustrate that explanatatory variables can replace each other in multiple regression equations. The scatter diagram of x_1 and x_2 shows the sites where a particular species is present (●) and absent (•).

Table. 3.7 Multiple logit regressions of the data of Figure 3.12 to illustrate that explanatory variables can be complementary in explanatory power (d(res) = residual deviance).

Terms in model	b_1	b_2	t value of b_1	t value of b_2	d(res)	d.f.
x_1, x_2	1.53	-1.66	2.98	-2.96	23.99	32
x_1	0.16	-	1.33	-	45.25	33
x_2	-	-0.15	-	-1.17	45.69	33
none	-	-	-	-	47.11	34

Table 3.8 Multiple logit regressions of the data of Figure 3.13 to illustrate that explanatory variables can substitute each other in a model (d(res) = residual deviance).

Terms in model	b_1	b_2	t value of b_1	t value of b_2	d(res)	d.f.
x_1, x_2	-0.61	-0.625	-1.63	-1.59	17.47	32
x_1	-0.94	-	-2.85	-	20.57	33
x_2	-	-1.016	-	-2.88	20.82	33
none	-	-	-	-	41.88	34

Figure 3.13 shows the occurrence of another species. When x_1 and x_2 are used to explain this species' occurrence, the t's (first line of Table 3.8) show that neither b_1 nor b_2 differs significantly from 0. It should not be concluded now that neither x_1 nor x_2 has an effect on the species' presence. These t tests only say that we do not need both x_1 and x_2 in the model. The fits with x_1 only and with x_2 only show that, taken singly x_1 and x_2 have both an effect. Moreover, these fits give about the same deviance; hence, x_1 can substitute x_2 in the model (Whittaker 1984). We observe in Figure 3.13 that the species occurs at low values of x_1 and x_2, but cannot say which variable this is caused by, because there were too few sites where x_1 was low and x_2 high or vice versa. We cannot distinguish their effects. This problem often arises when explanatory variables are highly correlated in the sample. This problem is known as the multicollinearity problem. For example, we may wish to know whether the probability of occurrence of a certain rare meadow plant decreases with potassium or with phosphate. But, in a survey potassium and phosphate will be strongly correlated, because they are usually applied simultaneously; so the question cannot be answered by a survey. Multicollinearity also arises when the number of explanatory variables is only slightly less than the number of sites.

Figures 3.12 and 3.13 illustrate the cases that create the most surprise at first. Less surprising are the cases in which neither multiple regression nor separate regressions show up any effects, or in which the techniques demonstrate the same

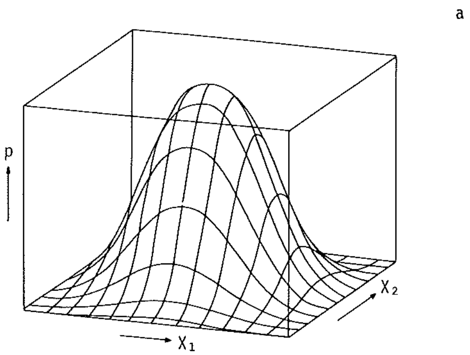

Figure 3.14a Three-dimensional view of a bivariate Gaussian logit surface with the probability of occurrence (p) plotted vertically and the two explanatory variables x_1 and x_2 plotted in the horizontal plane.

effects. Finally, it may also happen that both x_1 and x_2 show an effect on the species in the separate regressions, whereas in multiple regression only one of them shows an effect. This happens, for example, when x_1 is the only effective variable, and x_2 is correlated with x_1. The possible effect of x_2 in the regression with x_2 only is then due to its correlation with x_1, as multiple regression may show.

In multiple regression with more than two explanatory variables, all the previous cases may occur together in one analysis. Further, instead of pairs of variables that are substitutable or complementary, we may have triplets, quadruplets, etc. (Whittaker 1984). These concepts are important when one wants to select the best set of explanatory variables in a regression equation (Montgomery & Peck 1982; Whittaker 1984).

We now proceed to quadratic models. By inserting the quadratic surface of Equation 3.20 in Equation 3.15, we obtain a bivariate Gaussian logit surface, provided both b_2 and b_4 are negative (Figure 3.14a). This surface has ellipses as contour lines (lines of equal probability) with main axes parallel to the x_1 and x_2 axis (Figure 3.14b). The parameters of this models can again be estimated by GLM. Further analysis proceeds as from Equation 3.20, except that the F test must be replaced by the deviance test (Subsection 3.3.3).

3.5.4 *Interaction between explanatory variables*

Two explanatory variables show interaction of effects if the effect of the one variable depends on the value of the other. We can test for interaction by extending regression equations with product terms, like $x_1 \, x_2$.

By extending Equation 3.19 in this way, we obtain

$$\mathrm{E}y = b_0 + b_1 \, x_1 + b_2 \, x_2 + b_3 \, x_1 \, x_2 = (b_0 + b_2 \, x_2) + (b_1 + b_3 \, x_2) x_1 \qquad \text{Equation 3.23}$$

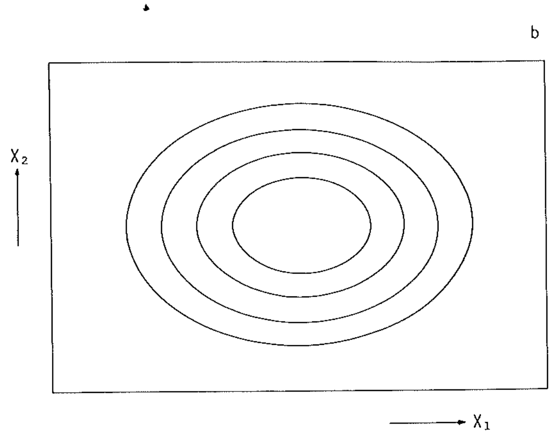

Figure 3.14b Elliptical contours of the probability of occurrence p plotted in the plane of the explanatory variables x_1 and x_2. One main axis of the ellipses is parallel to the x_1 axis and the other to the x_2 axis.

The final expression in Equation 3.23, obtained by simple algebra, shows that the relation between Ey and x_1 in this model is still a straight line, but that the intercept and slope and hence the effect of x_1 depend on the value of x_2. Conversely, the effect of x_2 depends on the value of x_1. The parameters b_1, b_2 and b_3 in Equation 3.23 can be estimated by using any multiple regression program and calculating the new variable $x_3 = x_1 \, x_2$ and specifying x_1, x_2 and x_3 as the explanatory variables. The interaction can be tested by a t test whether b_3 equals 0.

By extending the Gaussian model of Equation 3.20 with a product term, we obtain in the logit case

$$\log_e [p/(1-p)] = b_0 + b_1 \, x_1 + b_2 \, x_1^2 + b_3 \, x_2 + b_4 \, x_2^2 + b_5 \, x_1 \, x_2 \qquad \text{Equation 3.24}$$

If $b_2 + b_4 < 0$ and $4 \, b_2 \, b_4 - b_5^2 > 0$, Equation 3.24 describes a unimodal surface with ellipsoidal contours as in Figure 3.14b, but without the restriction that the main axes are horizontal or vertical. If one of these conditions is not satisfied, it describes a surface with a single minimum or one with a saddle point (e.g. Carroll 1972). When the surface is unimodal, the overall optimum (u_1, u_2) can be calculated from the coefficients in Equation 3.24 by

$$u_1 = (b_5 \, b_3 - 2 \, b_1 \, b_4)/d \qquad \text{Equation 3.25a}$$

$$u_2 = (b_5 \, b_1 - 2 \, b_3 \, b_2)/d \qquad \text{Equation 3.25b}$$

where $d = 4 \, b_2 \, b_4 - b_5^2$.

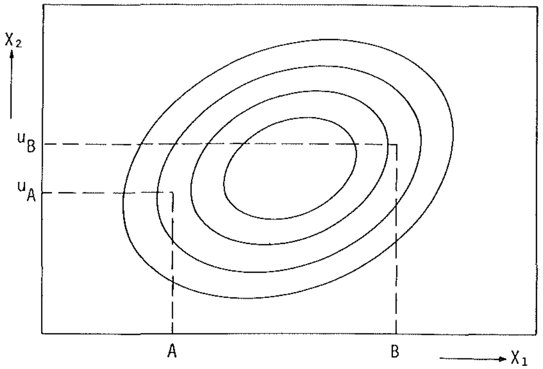

Figure 3.15 Interaction in the Gaussian logit model. The elliptical contours of the probability of occurrence p with respect to the explanatory variables x_1 and x_2 are shown. The main axes of the ellipses are not parallel to either the x_1 axis or the x_2 axis. u_A and u_B are the optima with respect to x_2 that correspond to levels A and B of x_1.

The optimum with respect to x_1 for a given value of x_2 is $-(b_1 + b_5\, x_2)/(2\, b_2)$ and thus depends on the value of x_2 if $b_5 \neq 0$. The expression is obtained by rearranging Equation 3.25 in the form of a parabola and using Equations 3.10 and 3.11. Figure 3.15 clearly shows this interaction. We can test this interaction by using a t test whether b_5 equals 0.

3.5.5 Nominal explanatory variables

Multiple regression can also be used to study the simultaneous effect of nominal environmental variables or of both quantitative and nominal environmental variables. To show how nominal variables may enter the multiple regression equation, we express the ANOVA model of Subsection 3.2.1 as a regression equation. In the example of Subsection 3.2.1, the nominal variable soil type had three classes: clay; peat; sand. We take clay as the reference class and define for peat and sand two dummy variables, x_2 and x_3, the values of which are either 0 or 1. The dummy variable for peat x_2 takes the value 1 when the site is on peat and the value 0 when the site is on clay or sand. The dummy variable for sand x_3 takes the value 1 when the site is on sand and the value 0 when the site is on clay or peat. A site on clay thus scores the value 0 for both dummy variables, a site on peat scores the value 1 for x_2 and 0 for x_3, etc. The systematic part of the model of Subsection 3.2.1 can be written as

$$Ey = b_1 + b_2\, x_2 + b_3\, x_3 \qquad \text{Equation 3.26}$$

The coefficient b_1 gives the expected response on the reference class clay, the coefficient b_2 the difference in expected response between peat and clay, and coefficient b_3 the difference between sand and clay. The coefficients b_1, b_2 and b_3 can be estimated by multiple least-squares regression. For the data of Figure 3.4, we obtain $b_1 = 1.70$, $b_2 = 1.47$, $b_3 = 0.63$. The mean is then on clay $b_1 = 1.70$, on peat $b_1 + b_2 = 3.17$ and on sand $b_1 + b_3 = 2.33$, as can be checked with Table 3.1. The ANOVA table of this multiple regression analysis is precisely that of Table 3.1. When a nominal variable has k classes, we simply specify $k - 1$ dummy variables (Montgomery & Peck 1982, Chapter 6).

The next example concerns the presence–absence of the plant species *Equisetum fluviatile* in fresh water ditches in the Netherlands. We will investigate the effect of electrical conductivity (mS m^{-1}) and of soil type (clay, peat, sand) on the species by logit regression, using the model

$$\log_e [p/(1 - p)] = b_0 + b_1\, x_1 + b_2\, x_1^2 + b_3\, x_2 + b_4\, x_3 \qquad \text{Equation 3.27}$$

where x_1 is the logarithm of electrical conductivity and x_2 and x_3 are the dummy variables defined in the previous example. Here b_3 and b_4 represent the effect of the nominal variable soil type. Figure 3.16 shows that this model consists of three curves with different maxima but with identical optima and tolerances. The coefficient b_3 is the difference between the logits of the maxima of the curves for peat and the reference class clay; the coefficient b_4 is the analogous difference between the curves for sand and clay. We can test whether the maxima of these curves are different by comparing the residual deviance of the model with x_1 and x_1^2 with the residual deviance of Equation 3.27. The difference is a chi-square with two degrees of freedom if soil type has no effect. This is another example of the deviance test.

To calculate the optimum and tolerance in Equation 3.27, we simply use Equation 3.11; to calculate standard errors and a confidence interval for the optimum, we can use the equations of Section 3.9. Exercise 3.6 may serve as an example.

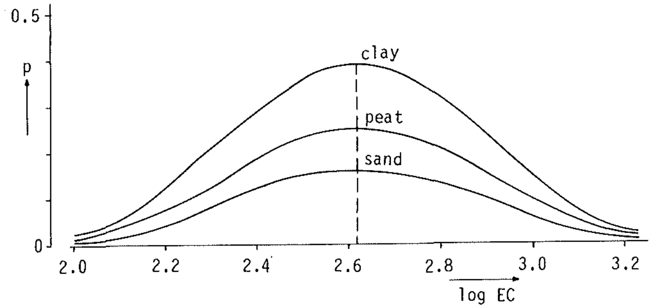

Figure 3.16 Response curves for *Equisetum fluviatile* fitted by multiple logit regression of the occurrence of *E. fluviatile* in freshwater ditches on the logarithm of electrical conductivity (EC) and soil type surrounding the ditch (clay, peat, sand). Data from de Lange (1972).

3.6 Model choice and regression diagnostics

Many things can go wrong in regression analysis. The type of response curve or the error distribution may have been chosen incorrectly and there may be outliers that unduly influence the regression. To detect such faults is the purpose of regression diagnostics (Belsley et al. 1980; Cook & Weisberg 1982; Hocking & Pendleton 1983). What we can do, for instance, is to plot the residuals of a regression against the fitted values or against each of the explanatory variables and look for outliers and systematic patterns in these plots. The references just given deal mainly with regression diagnostics for quantitative response variables. Here we focus on presence–absence data and response curves of species.

One would like to base the shape of a response curve of a species on physiological and ecological theory. But there is no generally accepted theory (Austin 1980) and therefore no ubiquitously applicable response curve. In the absence of theory, one can still proceed by empirical methods and decide upon an applicable curve on the basis of many empirical results. Early studies by Gause (1930), Curtis & Mcintosh (1951) and Whittaker (1956) showed that monotonic response curves are too simple as an ecological response model and that a unimodal model is more appropriate. Simple ecological reasoning shows that also bimodal curves are a realistic option: a species can be outcompeted near its physiological optimum by more competitive species whereas the species may be able to cope with less favourable environmental conditions when competition is less. The response curve applicable to field conditions is then the result of the physiological response curve and competition between species (Fresco 1982). Hill (1977) suggested, however, that a good ecological variable, minimizes the occurrence of bimodal species distributions.

When there are no ideas a priori of the shape of the response curve, one can best divide the quantitative environmental variable into classes and calculate the frequency of occurrence for each class as in Subsection 3.3.1 (Gounot 1969; Guillerm 1971). By inspection of the profiles of the frequencies for several species, one may get an idea which type of response curve is appropriate.

Curves have several advantages over frequency profiles for quantitative environmental variables:

- curves when kept simple, provide through their parameters a more compact description than frequency profiles
- there is no need to choose arbitrary class boundaries
- there is no loss of information because the environmental variable is not divided into classes
- when the Gaussian model applies, statistical tests based on curves have greater power to detect that the environmental variable influences the species than the chi-square test based on the frequency profile. This is because the chi-square test of Subsection 3.3.1 is an omnibus test that is able to detect many types of deviations from the null hypothesis, whereas the *t* tests and deviance tests of Subsections 3.3.2, 3.3.3 and Section 3.5 test the null hypothesis against a specified alternative hypothesis.

A clear disadvantage of curves is that one is forced to choose a model which

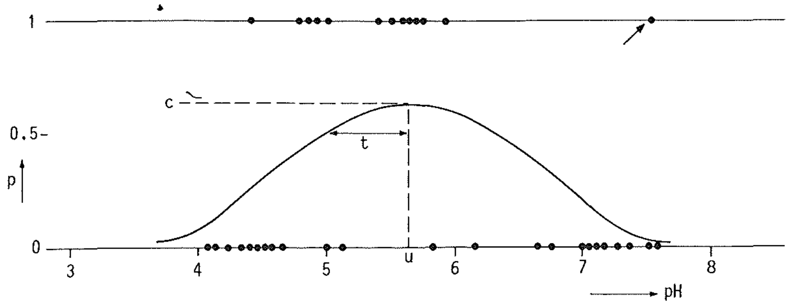

Figure 3.17 The change in a fitted Gaussian logit curve by adding an influential point. Adding a single presence at pH = 7.6 (indicated by an arrow) to Figure 3.10 considerably decreases the estimated maximum and increases the estimated tolerance and optimum.

may be wrong for the data at hand. For example, is the true response curve symmetric? When asymmetry is suspected, one can transform the explanatory variable, for example by taking logarithms, and one can compare the residual deviances before and after transformation. The detection of a deviation from a supposed response curve may aid our understanding of the relation of the species with the environment and in general triggers off a new cycle in the process of model building.

Data points that unduly influence the regression require special attention with presence–absence data. For example, adding a presence to Figure 3.10 at pH 7.6 drastically changes the fitted response curve (Figure 3.17). When there are two or more explanatory variables, we suggest you plot the variables in pairs as in Figures 3.12 and 3.13 and inspect the plots for outlying presences. When such an outlier is spotted, you must attempt to find out whether it is a recording error or whether the site was atypical for the conditions you intended to sample, and decide after such attempts whether or not to retain the outlier in the data. We also suggest that you always try to remove the lowest or highest *x* where the species is present to check that the fitted response stays roughly the same (cf. the jackknife technique, Efron 1982).

3.7 The method of weighted averaging

This section is devoted to estimation of species indicator values (Ellenberg 1982). In terms of response curves, there are two possible definitions of species indicator value: it is either the optimum or the centroid of the species response curve. These definitions coincide only if the response curve is symmetric. In Subsections 3.2.3 and 3.5.2, we have shown how an optimum can be estimated by fitting a curve or a surface to the species data by regression. In the regression method, we have to assume a particular response curve. Ecologists have long used a simpler method for estimating indicator values (Ellenberg 1948; 1979). This is the method of weighted averaging, which circumvents the problem of having to fit a response curve. When a species shows a unimodal curve against a particular environmental variable, the presences of the species will most frequently occur near the optimum of the curve. An intuitively reasonable estimate of the indicator value is therefore obtained by taking the average of the values of the environmental variable over those sites where the species is present. For abundance data, a weighted average may be taken in which values are weighted proportional to the species' abundance, i.e.

$$u^* = (y_1 x_1 + y_2 x_2 + \ldots + y_n x_n)/(y_1 + y_2 + \ldots + y_n) \qquad \text{Equation 3.28}$$

where
u^* is the weighted average
$y_1, y_2, \ldots, y_n$ are the abundances of the species
$x_1, x_2, \ldots, x_n$ the values of the environmental variable at the Sites 1, 2 ... n.

The weighted average disregards species absences. An unpleasant consequence of this is that the weighted average depends on the distribution of the environmental variable in the sample (Figure 3.18). Highly uneven distributions can even scramble the order of the weighted averages for different species (Figure 3.18).

Ter Braak & Looman (1986) compared the performance of the methods of weighted averaging and of Gaussian logit regression to estimate the optimum of a Gaussian logit curve from presence–absence data. Through simulation and practical examples, they showed that the weighted average is about as efficient as the regression method for estimating the optimum:

- when a species is rare and has a narrow ecological amplitude
- when the distribution of the environmental variable among the sites is reasonably homogeneous over the whole range of occurrence of the species along the environmental variable.

In other situations, weighted averaging may give misleading results (Exercise 3.2.8). Similar conclusions also hold for quantitative abundance data; for quantitative abundance data, the weighted average efficiently estimates the optimum of the Gaussian response curve, if the abundances are Poisson-distributed and the sites are homogeneously distributed over the whole range of the species.

Despite its deficiencies, the method of weighted averaging is a simple and useful method to show up structure in a data table such as Table 0.1 by rearranging species and sites on the basis of an explanatory variable. As an example, we shall demonstrate this by rearranging the Dune Meadow Data in Table 0.1 on the basis of the moisture value of the sites (relevés). For each species, we calculate its weighted average for moisture, e.g. for *Aira praecox*

$$u^* = (2 \times 2 + 3 \times 5)/(2 + 3) = 3.8$$

and arrange the species in order of the values so obtained and the sites in order of their moisture value (sites with equal moisture are arranged in arbitrary order). The result is shown in Table 3.9. *Plantago lanceolata* is clearly restricted to the driest sites, *Ranunculus flammula* to the wettest sites, and *Alopecurus geniculatus* to sites with intermediate moisture. In Table 3.9, most of the abundance values (> 0) are arranged in a band along the 'diagonal'. The method of weighted averaging tends to show up such a diagonal structure in a table, when species show unimodal curves for the environmental variable. This idea is extended in Section 5.2.

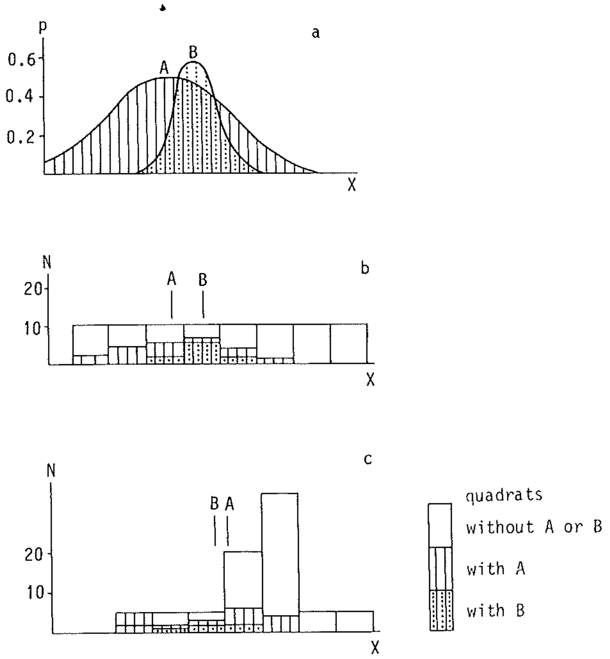

Figure 3.18 The response curves of imaginary species A and B (a); the occurrence of these species in two samples of 80 sites, in which the environmental variable is distributed evenly (b) or unevenly (c). The weighted averages are indicated with lines. The two sampling designs yield weighted averages that are in reverse order. p = probability of occurence; N = number of sites; x = environmental variable.

3.8 Bibliographic notes

The least-squares technique dates back to the early nineteenth century with the work of K.F. Gauss. The principle of maximum likelihood was founded by R.A. Fisher in the 1920s. The generalized linear model (GLM) was introduced by Nelder & Wedderburn (1972) and made it easy to fit a major class of non-linear models to data. Among the many statistical textbooks on least-squares regression are Draper & Smith (1981), Seber (1977), Montgomery & Peck (1982) and Mosteller & Tukey (1977). Useful more general statistical texts for biologists are Parker (1979), Sokal & Rohlf (1981) and Snedecor & Cochran (1980). Dobson

Table 3.9 Weighted averaging used for rearranging species and sites in Table 0.1. The sites (columns) are arranged in order of moisture and the species (rows) in order of their weighted average (u^*) with respect to moisture. Species abundance is printed as a one-digit number, a blank denoting absence. Site identification numbers are printed vertically. For abbrevations of species names see Table 0.1.

```
      species       sites
                         11  11 1 111112
                    12567183407928345690  u*
26  Tri  pra          252                 1.0
18  Pla  lan          55533  32           1.2
28  Vic  lat             21  1            1.3
 1  Ach  mil        13222    42           1.4
 6  Bel  per         32   2222            1.5
 7  Bro  hor         42 2   34            1.5
23  Rum  ace          563      22         1.7
17  Lol  per        7526672656 2 4        1.7
 9  Cir  arv                2             2.0
11  Ely  rep        444    44  6          2.0
19  Poa  pra        442344354414 42       2.0
 5  Ant  odo          432    44        4  2.1
20  Poa  tri        27645  654 5449  2    2.6
16  Leo  aut         5333552232223222 62  2.6
27  Tri  rep         525232216 332261 2   2.7
29  Bra  rut          26246222 242  4434  2.9
13  Hyp  rad             2    2        5  3.4
24  Sag  pro             2  5  2422   3   3.5
 4  Alo  gen         2     72  3855  4    3.7
15  Jun  buf            2      44 3       3.8
 3  Air  pra                  2       3   3.8
25  Sal  rep              3           35  3.9
 2  Agr  sto               48  3445447 5  4.1
14  Jun  art                   4 4  33 4  4.8
 8  Che  alb                      1       5.0
10  Ele  pal                     4 458 4  5.0
12  Emp  nig                          2   5.0
21  Pot  pal                       22     5.0
22  Ran  fla                     22222 4  5.0
30  Cal  cus                       4 3 3  5.0

    MOISTURE        11111112222445555555
```

(1983) and McCullagh & Nelder (1983) provide an introduction to GLM.

A major contribution to the analysis of species-environment relations was made by Whittaker (1956; 1967). His direct gradient analysis focused on response curves and surfaces of species with respect to a complex of environmental variables, that changed gradually in geographic space. The term 'gradient' therefore then had a geographical meaning, but in recent use the term is equivalent to 'environmental variable'. Whittaker used simple smoothing methods to fit the curves and surfaces. Following Gleason (1926), Ramensky (1930) and Gause (1930), he stressed that species react 'individualistically' to environmental variables and that response surfaces of species are often unimodal. Whittaker's view opposed the 'integrated-community hypothesis' of Clements (1928), which viewed communities of species as organisms of a higher scale. The integrated-community hypothesis stimulated much work on succession and on the interrelations between species, disregarding environmental variables. Conversely, the individualistic concept (in its most extreme form, at least) disregards direct relations between species. McIntosh (1981) discussed these apparently contrasting views. Fresco (1982) attempted to incorporate species-environment and inter-species relations into a single regression equation.

Whittaker (1956; 1967) dealt with gradients, i.e. ordinal or quantitative environmental variables. Gounot (1969) and Guillerm (1971) proposed methods similar to that of Subsection 3.3.1, which are applicable for presence-absence species data and nominal environmental variables. They divided environmental variables into classes when the variables were quantitative. Our approach of using logit regression makes it possible to deal with quantitative and nominal variables in a single analysis.

An early ecological example of fitting sigmoid curves to presence-absence data is found in Jowett & Scurfield (1949). They applied probit analysis (Finney 1964), an alternative for logit regression that usually gives similar results. Polynomial least-squares regression was advocated by Yarranton (1969; 1970). He noticed the problem of absences of species (zero abundance values). Austin (1971) stressed the power of regression analysis and gave several examples from plant ecology where abundance data were first transformed logarithmically and then analysed by least-squares regression using parabolas and second-order response surfaces. Alderdice (1972) explained and applied second-order response surfaces in marine ecology. Gauch & Chase (1974) provided a computer program to fit the Gaussian response curve by least squares to ecological data that might include zero abundances. Their approach has become outdated with the advent of GLM. Austin et al. (1984) showed the usefulness of GLM in direct gradient analysis, using log-linear regression and logit regression, with second-order polynomials as linear predictors. We believe that GLM (Section 3.5) should become a standard tool in applied ecology. Response surfaces fitted by GLM are particularly useful in models simulating the impact of various options in environmental management.

3.9 Standard errors of estimated optimum and tolerance; confidence interval for the optimum

We denote the variance of the estimates of b_1 and b_2 in Equations 3.9, 3.17, 3.20 or 3.24 by v_{11} and v_{22} and their covariance by v_{12}. Using Taylor expansion, we calculate the approximate variance of the estimated optimum and tolerance:

$$\mathrm{var}\,(\hat{u}) = (v_{11} + 4\,u\,v_{12} + 4\,u^2\,v_{22})/(4\,b_2^2) \qquad \text{Equation 3.28}$$

$$\mathrm{var}\,(\hat{t}) = v_{22}/(-8\,b_2^3) \qquad \text{Equation 3.29}$$

An approximate 100(1 - α)% confidence interval for the optimum is derived from Fiellers theorem (Finney 1964, p.27-29). Let t_α be the critical value of a two-sided t test at chosen probability level α with n - 3 degrees of freedom, where n is the number of sites. For example, t = 2.00 for a 95% confidence interval and 63 sites. Calculate

$$g = t_\alpha^2 \, v_{22}/b_2^2 \qquad \text{Equation 3.30a}$$

and

$$D = 4\, b_2^2 \text{ var } (\hat{u}) - g(v_{11} - v_{12}^2/v_{22}). \qquad \text{Equation 3.30b}$$

$$u_{\text{lower}}, u_{\text{upper}} = [\hat{u} + 0.5\, g\, v_{12}/v_{22} \pm 0.5\, t_\alpha\, (\sqrt{D})/b_2]/(1 - g) \qquad \text{Equation 3.31}$$

where the symbol ± indicates addition and subtraction in order to obtain the lower and upper limits of the confidence interval, respectively. If b_2 is not significantly different from zero ($g > 1$), then the confidence interval is of infinite length and, taken alone, the data must be regarded as valueless for estimating the optimum.

3.10 Exercises

Exercise 3.1 *Straight line regression*

In a study of the impact of acid rain on diatoms, van Dam et al. (1981) collected data on diatom composition and water chemistry in Dutch moorland pools. For each sample, a total of 400 diatomaceous frustules were identified under a microscope. The numbers of frustules of the species *Frustulia rhomboides* var. *saxonica* and the relative sulphate concentrations $S_{\text{rel}} = [SO_4^{2-}]/([Cl^-] + [SO_4^{2-}] + [HCO_3^-])$ in the 16 samples taken in 1977 and 1978 were as follows (van Dam et al. 1981, Tables 2 and 5):

pool	V2	B6	B3	B4	V1	B5B	B8	B1	D6	B7	B2	D3	D2	D1	D5	D6
Frustulia count	0	0	14	3	0	5	6	21	62	26	14	48	97	99	28	202
S_{rel}	0.78	0.64	0.69	0.70	0.64	0.77	0.73	0.77	0.58	0.44	0.44	0.37	0.23	0.19	0.31	0.23

Exercise 3.1.1 Construct a graph of the data, plotting $\log_e$ [(*Frustulia* count) + 1] on the vertical axis. Note that the relation looks linear.

Exercise 3.1.2 Fit a straight line to the data taking $\log_e$ (*Frustulia* count + 1) as the response variable and the relative sulphate concentration as the explanatory variable. Use a pocket calculator or a computer for least-squares regression to verify the following results.

		estimate	s.e.	t
constant	b_0	5.848	0.806	7.26
S_{rel}	b_1	-5.96	1.41	-4.22
ANOVA table				
	d.f.	s.s.	m.s.	
regression	1	24.34	24.340	
residual	14	19.11	1.365	
total	15	43.45	2.897	

Exercise 3.1.3 What are the systematic part and the error part of the response model fitted in Exercise 3.1.2? What are the fitted value and the residual for Pool B2?

Exercise 3.1.4 What are the residual sum of squares, the residual variance, the residual standard deviation and the fraction of variance accounted for? How many degrees of freedom are there for the residual sum of squares?

Exercise 3.1.5 Calculate a 95% confidence interval for the regression coefficient b_1. Is the estimate of b_1 significantly ($P < 0.05$) different from 0?

Exercise 3.1.6 Estimate the expected responses when the relative concentrations of sulphate equal 0.25, 0.50 and 0.75. Calculate the 95% confidence interval of each of these expected responses. The standard errors of the estimates are 0.49, 0.30 and 0.42, respectively. Back-transform the estimates obtained to counts of *Frustulia*.

Exercise 3.1.7 Calculate 95% prediction intervals when the relative sulphate concentrations are equal to 0.25, 0.50 and 0.75.

Exercise 3.2 Parabola, Gaussian response curve and weighted averaging

In a study aimed at reconstructing past temperatures of the sea-surface from fossil distributions of *Radiolaria*, Lozano & Hays (1976) investigated the relation between different taxa of *Radiolaria* and sea-surface temperature in present-day samples. The following data extracted from their Figure 11 concern the abundance (%) of *Spongotrochus glacialis* and February sea-surface temperature (temp., °C) at 34 sites in the Atlantic and Antarctic Oceans.

site	1	2	3	4	5	6	7	8	9	10	11	12
abundance	12	14	13	22	18	19	7	8	11	15	12	14
temp	0.8	1.1	1.6	1.8	1.7	2.0	1.6	1.9	2.0	2.5	3.7	4.2

site	13	14	15	16	17	18	19	20	21	22	23	24
abundance	16	21	35	30	34	48	47	63	54	62	56	52
temp.	4.1	5.8	6.1	6.6	7.9	10.2	11.0	11.9	12.8	14.8	15.9	18.1

site	25	26	27	28	29	30	31	32	33	34
abundance	41	38	30	18	25	35	37	38	42	41
temp.	16.9	17.1	18.0	18.5	20.0	21.0	19.4	19.8	19.0	21.6

Exercise 3.2.1 Construct a graph of the data, plotting the abundance on the vertical axis. Note that the relation looks unimodal. Plot also the logarithm of abundance against temperature.

Exercise 3.2.2 Use a computer program for least-squares regression to verify the following results. Fitting a parabola to the logarithm of the abundances gives:

		estimate	s.e.	*t*
constant	b_0	2.119	0.133	15.95
temp.	b_1	0.2497	0.0356	7.01
temp. squared	b_2	-0.00894	0.00164	-5.46

ANOVA table

	d.f.	s.s.	m.s.
regression	2	9.42	4.7101
residual	31	3.06	0.0988
total	33	12.48	0.3783

Exercise 3.2.3 Estimate the expected responses when the temperatures are 5, 10, 15 and 20 °C, calculate the optimum, tolerance and maximum of the fitted parabola and use the results to sketch the fitted parabola.

Exercise 3.2.4 What is the residual standard deviation and the fraction of variance accounted for?

Exercise 3.2.5 Calculate a 95% confidence interval for the regression coefficient b_2. Would a straight line be statistically acceptable for these data?

Exercise 3.2.6 Calculate a 95% confidence interval for the optimum using Equation 3.31. Here one needs to know also that covariance between the estimates of b_1 and b_2 equals -0.00005704; the variances required can be obtained from the table of regression coefficients. Hint: write a computer program for the calculations required in order to avoid lengthy hand-calculation.

Exercise 3.2.7 Back-transform the expected responses of Exercise 3.2.3 to abundance and sketch the fitted curve.

Exercise 3.2.8 Calculate (after reading Section 3.7) the weighted average of *Spongotrochus* with respect to temperature, using the abundances and, a second time, using log abundances. Explain the difference from the optimum estimated above. Is the difference large?

Exercise 3.3 Logit link function

Verify the equivalence of Equations 3.15 and 3.16 by showing that $\log_e [p/(1-p)] = c$ if and only if $p = (\exp c)/(1 + \exp c)$.

Exercise 3.4 Chi-square test and logit regression

A sample of 160 fields of meadow is taken to investigate the occurrence of the grass species *Elymus repens* in relation to agricultural use (hayfield or pasture). The data, based on the study of Kruijne et al. (1967), are summarized in the following 2 × 2 table of number of fields.

E. repens	agricultural use		
	hayfield	pasture	total
present	12	96	108
absent	16	36	52
total	28	132	160

Exercise 3.4.1 Estimate the probability of occurrence of *E. repens* in hayfield and in pasture.

Exercise 3.4.2 Is there evidence that the probability of occurrence in hayfield differs from that in pasture? Apply here the chi-square test of Subsection 3.3.1, using a significance level of 5%.

Exercise 3.4.3 Instead of the chi-square test we can use logit regression of the presences and absences of *E. repens* in the 160 fields on the nominal explanatory variable agricultural use. Agricultural use has two classes in this problem and therefore we define a single dummy variable USE, which takes the value 1 if the field is a pasture and the value 0 if the field is a hayfield. A computer program for logit regression gave the following output with the response variable presence–absence of *E. repens*:

		estimate	s.e.	t
constant	c_0	-0.28	0.38	-0.74
USE	c_1	1.27	0.42	3.02
		d.f.	deviance	mean deviance
residual		158	192.9	1.221

The model corresponding to this output is $\log_e [p/(1 - p)] = c_0 + c_1 \times \text{USE}$.

Exercise 3.4.3.1 Calculate from the output the estimates for the probability of occurrence of *E. repens* in hayfield and in pasture. Hint: use Exercise 3.3. Compare the estimates with those of Exercise 3.4.1.

Exercise 3.4.3.2 Show by t test whether the probability of occurrence in hayfield differs from that in pasture. Compare the conclusion with that of Exercise 3.4.2.

Exercise 3.4.3.3 The deviance corresponding to the model $\log_e [p/(1 - p)] = c$ equals 201.7 with 159 degrees of freedom. Apply the deviance test instead of the t test of the previous exercise.

Exercise 3.5 Gaussian logit regression

The acidity (pH) of the fields was recorded also for the sample of the previous exercise. Spatial heterogeneity in acidity was disregarded; pH was the mean of several systematically located points in the field. To investigate the effect of acidity on the occurrence of *E. repens*, a Gaussian logit regression was carried out. The results were:

		estimate	s.e.	t
constant	b_0	-57.26	15.4	-3.72
pH	b_1	19.11	5.3	3.61
pH^2	b_2	-1.55	0.44	-3.52
		d.f.	deviance	mean deviance
residual		157	176.3	1.123

Exercise 3.5.1 At what pH did *E. repens* occur with the highest probability? Calculate also the tolerance and the maximum probability of occurrence.

Exercise 3.5.2 Calculate from the output the estimated probabilities of occurrence of *E. repens* at pH 4.5, 5.0, 5.5, 6.0, 6.5, 7.0 and 7.5 and use the results to sketch the response curve of *E. repens* against pH.

Exercise 3.5.3 Is the estimated Gaussian logit response curve significantly different ($P < 0.05$) from a sigmoid response curve; hence, is the optimum significant? Hint: use a one-tailed t test.

Exercise 3.6 Multiple logit regression

When considered separately, agricultural use and acidity appear to influence the occurrence of *E. repens* in fields (Exercises 3.4 and 3.5). Hayfield and pasture differ, however, in acidity; hayfields tend to be more acid than pastures. It is therefore of interest to investigate whether this difference in acidity between hayfields and pastures can explain the difference in probability of occurrence of *E. repens* between hayfields and pastures. This problem can be attacked by multiple (logit) regression. We fitted the model

$$\log_e [p/(1 - p)] = c_0 + c_1 \text{ USE} + b_1 \text{ pH} + b_2 \text{ pH}^2$$

to the data and obtained the following results:

		estimate	s.e.	t
constant	c_0	-57.82	17.10	-3.38
USE	c_1	-0.04	0.57	-0.07
pH	b_1	19.30	5.81	3.32
pH^2	b_2	-1.56	0.49	-3.18
		d.f.	deviance	mean deviance
residual		156	176.2	1.129

Exercise 3.6.1 Calculate the estimated probabilities of occurrence in hayfields and pastures for pH 5 and for pH 6. Calculate also the optimum pH and the maximum probabilities of occurrence in hayfields and pastures, and the tolerance. Compare the results with those of Exercise 3.5.1 and 3.5.2, and sketch the response curves.

Exercise 3.6.2 Show by a t test whether the probability of occurrence in hayfields differs from that in pastures after correction for the effect of acidity. Can acidity account for the difference found in Exercise 3.4.2

Exercise 3.6.3 Use the deviance test instead of the *t* test in Exercise 3.6.2. Does the conclusion change?

Exercise 3.6.4 Show by a deviance test whether acidity has an effect on the probability of occurrence of *E. repens* after correction for the effect of agricultural use. Are the variables acidity and agricultural use substitutable in the sense of Subsection 3.5.3?

3.11 Solutions to exercises

Exercise 3.1 Straight-line regression

Exercise 3.1.3 The systematic part is $Ey = b_0 + b_1 S_{rel}$ and the error part is that the error $(y - Ey)$ follows a normal distribution with mean at zero and a variance that does not depend on S_{rel}. Pool B2 has a count of 14 (hence, y = 2.71) and S_{rel} = 0.44; hence, the fitted value is 5.848 – 5.96 × 0.44 = 3.23 and the residual is 2.71 – 3.23 = –0.52. The fitted number of *Frustulia* frustules is thus exp (3.23) – 1 = 25 – 1 = 24.

Exercise 3.1.4 From the ANOVA table, we obtain the residual sum of squares 19.11, the residual variance 1.365, the residual standard deviation $\sqrt{1.365} = 1.17$ and the fraction of variance accounted for is 1 – (1.365/2.897) = 0.529. The residual sum of squares has 14 degrees of freedom.

Exercise 3.1.5 In Equation 3.2 with $t_{0.05}(14) = 2.145$, we insert the estimate for b_1 and its standard error and obtain a lower bound of –5.96 – (2.145 × 1.41) = –8.98 and an upper bound of –5.96 + (2.145 × 1.41) = –2.94. The 95% confidence interval for b_1 is therefore (–8.98, –2.94). The value 0 does not lie in this interval. Alternatively, the *t* for b_1 (-4.22) is greater in absolute value than the critical *t* (2.145); hence, the estimate of b_1 is significantly ($P < 0.05$) different from 0.

Exercise 3.1.6 In a pool with S_{rel} = 0.25 the expected response is estimated by 5.848 – 5.96 × 0.25 = 4.36. The standard error of this estimate is 0.49 and the 95% confidence interval is therefore (4.36 – 2.145 × 0.49, 4.36 + 2.145 × 0.49) = (3.31, 5.41). For S_{rel} = 0.50 and 0.75 the estimates are 2.87 and 1.38, with confidence intervals of (2.23, 3.50) and (0.47, 2.29), respectively. Notice that the interval is shortest near the middle of the interval of the relative sulphate values actually sampled. For S_{rel} = 0.25, 0.50, 0.75 back-transformation to counts gives the estimates exp (4.36) – 1 = 77, 17 and 3, respectively.

The latter values estimate the median number of frustules at the respective relative sulphate concentrations, and not the expected number of frustules. We assumed that the log-transformed data do follow a normal distribution. In the normal distribution, the mean is equal to the median (50-percentile) and transformations do not change percentiles of a distribution. Back-transforming the limits of the 95% confidence intervals gives 95% confidence intervals for the median counts. For S_{rel} = 0.25 this interval is (26, 223).

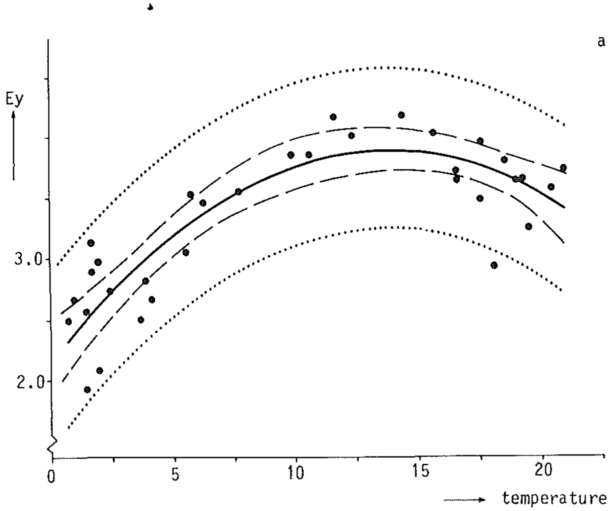

Figure 3.19a Parabola (solid line) fitted by least-squares regression of log-transformed relative abundance of *Spongotrochus glacialis* (●) on February sea-surface temperature (temp). 95% confidence intervals (dashed curve) and 95% prediction intervals (dotted line) are shown. Data from Lozano & Hays (1976).

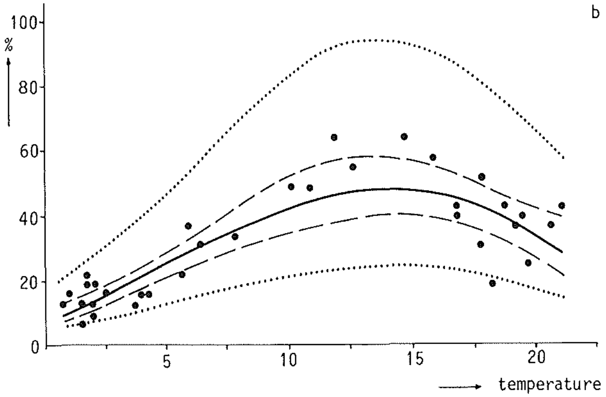

Figure 3.19b Gaussian response curve with 95% confidence and 95% prediction intervals obtained by back-transforming the curves of Figure 3.19a. Vertical axis: abundance (%) of *Spongotrochus glacialis*. Horizontal axis: February sea-surface temperature.

Exercise 3.1.7 With Equation 3.3 and $S_{\mathrm{rel}} = 0.25$ we obtain the interval $4.36 \pm 2.145 \times \sqrt{(1.17^2 + 0.49^2)} = 4.36 \pm 2.145 \times 1.27 = (1.63, 7.08)$.
Back-transforming to counts shows that 95% of the counts are expected to lie between 4 and 1187. The latter value is nonsensical as the maximum count is 400.

For $S_{\mathrm{rel}} = 0.50$ and 0.75 we obtain 95% prediction intervals for the transformed counts of (0.28, 5.46) and (-1.28, 4.05), respectively.

Exercise 3.2 Parabola, Gaussian response curve and weighted averaging

Exercise 3.2.1 See Figure 3.19a,b.

Exercise 3.2.3 The expected response at temp. = 5 is estimated by
$2.119 + 0.2497 \times 5 - 0.00894 \times 5^2 = 3.14$. For temp. = 10, 15 and 20 the estimates are 3.72, 3.85 and 3.54, respectively. It is of interest to note that the standard errors of the estimates are 0.07, 0.10, 0.09 and 0.11 at temp. 5, 10, 15 and 20, respectively.

With Equations 3.11a and 3.11b, the optimum is estimated by $\hat{u} = -\hat{b}_1/(2\ \hat{b}_2) = -0.2497/(-2 \times 0.00894) = 14.0$, so that the optimum temperature is 14.0 °C and the tolerance by $\hat{t} = 1/\sqrt{(-2\ \hat{b}_2)} = 7.48$, so that the tolerance of temperature is 7.48 °C. The maximum of the parabola (Figure 3.9a) is estimated by $2.119 + 0.2497 \times 14.0 - 0.00894 \times 14.0^2 = 3.86$.

Exercise 3.2.4 The residual standard deviation is $\sqrt{0.0988} = 0.314$ and the fraction of variance accounted for is $1 - (0.0988/0.3783) = 0.739$, using the results of the ANOVA table.

Exercise 3.2.5 With Equation 3.2 and $t_{0.05}(31) = 2.04$, a 95% confidence interval for b_2 is $(-0.00894 - 2.04 \times 0.00164,\ 0.00894 + 2.04 \times 0.00164) = (-0.0122, -0.0056)$. The estimate for b_2 is thus significantly ($P < 0.05$) different from 0, in agreement with the t of -5.46; hence, the null hypothesis thast the relation is a straight line ($b_2 = 0$) is rejected in favour of a parabola ($b_2 \neq 0$). A straight line is thus statistically unacceptable for these data.

Exercise 3.2.6 A 95% confidence interval for the optimum temperature is (12.8 °C, 16.2 °C).

Exercise 3.2.7 The median abundances of *Spongotrochus* at temp. = 5, 10, 15 and 20 are exp (3.14) = 23, 41, 47 and 34, respectively. The fitted Gaussian curve with the data points and 95% confidence and 95% prediction intervals (obtained also by back-transformation) is plotted in Figure 3.19b.

Exercise 3.2.8 The weighted average is $(12 \times 0.8 + 14 \times 1.1 + \ldots + 41 \times 21.6)/(12 + 14 + \ldots + 41) = 12.7$, so that the weighted average temperature is 12.7 °C.

With log-transformed abundance data the weighted average temperature is smaller, namely 11.0 °C. Both values are smaller than the optimum (14.0 °C) estimated by regression, because the temperatures are not homogeneously distributed over the range of the species; in particular, the lower temperatures are over-represented and the optimum lies at the higher end of the temperature interval that was actually sampled. So the weighted average estimator is biased. The difference is large in a statistical sense: the weighted averages fall outside the 95% confidence interval for the optimum calculated in Exercise 3.2.6.

Exercise 3.3 Logit link function

$\log_e [p/(1-p)] = c \rightarrow p/(1-p) = \exp c$
$\rightarrow p = (\exp c)(1-p) = \exp c - p \exp c \rightarrow p + p \exp c = \exp c$.
$\rightarrow p(1 + \exp c) = \exp c \rightarrow p = (\exp c)/(1 + \exp c)$.

The arrows hold true also in the reverse direction; hence, the equivalence.

Exercise 3.4 Chi-square test and logit regression

Exercise 3.4.1 The estimated probability of occurrence is: in hayfield 12/28 = 0.43; in pasture 96/132 = 0.73.

Exercise 3.4.2 When the probability of occurrence in hayfield equals that in pasture, this probability is estimated by 108/160 = 0.675. Then, we expect that out of 28 fields $0.675 \times 28 = 18.9$ fields contain *E. repens*, and $28 - 18.9 = 9.1$ fields do not contain *E. repens*.

With 132 fields (pastures) the expected numbers are: 89.1 with *E. repens* and 42.9 without *E. repens*. Inserting the observed and expected numbers in the equation for chi-square gives $(12 - 18.9)^2/18.9 + \ldots + (36 - 42.9)^2/42.9 = 9.39$ which is much greater than the critical value at the 5% significance level of a chi-square distribution with $(2 - 1) \times (2 - 1) = 1$ degree of freedom: $\chi^2_{0.05}(1) = 3.841$. The conclusion is that there is strong evidence ($P < 0.01$) that the probability of occurrence in hayfield differs from that in pasture.

Exercise 3.4.3.1 For hayfield the model reads: $\log_e [p/(1 - p)] = c_0$ because USE = 0 for hayfields. c_0 is estimated by -0.28; hence the estimated probability of occurrence is $\hat{p} = \exp(-0.28)/[1 + \exp(-0.28)] = 0.43$. For pastures, the model reads: $\log_e [p/(1 - p)] = c_0 + c_1$ because USE = 1 for pastures. $c_0 + c_1$ is estimated as $-0.28 + 1.27 = 0.99$, which gives $\hat{p} = \exp 0.99/(1 + \exp 0.99) = 0.73$. The estimates equal those of Exercise 3.4.1, because the regression model simply specifies two probabilities, one for hayfields and one for pastures.

Exercise 3.4.3.2 The estimate of the coefficient c_1 of USE differs significantly ($P < 0.05$) from 0, $t = 3.02$ being greater than $t_{0.05}(158) = 1.98$; hence, the estimated probabilities differ significantly. The conclusion is identical to that of Exercise 3.4.2; we applied a different test for the same purpose.

Exercise 3.4.3.3 The difference in deviance between the model with and without the variable USE is 201.7 - 192.9 = 8.8, which is to be compared with a chi-square distribution with one degree of freedom.

Exercise 3.5 Gaussian logit regression

Exercise 3.5.1 From Equation 3.11a, the estimated optimum of pH is $\hat{u} = -19.11/(-2 \times 1.55) = 6.16$.

With $u = 6.16$ in Equation 3.17, the maximum probability of occurrence is estimated by $\hat{p} = (\exp 1.641)/(1 + \exp 1.641) = 0.84$, because $-57.26 + 19.11 \times 6.16 - 1.55 \times 6.16^2 = 1.641$. The tolerance is $\hat{t} = 0.57$ (Equation 3.11b).

Exercise 3.5.2 Inserting pH = 4.5, 5.0, 5.5, 6.0, 6.5, 7.0 and 7.5 in Equation 3.17, we obtain estimated probabilities of 0.07, 0.39, 0.72, 0.83, 0.81, 0.64 and 0.25.

Exercise 3.5.3 The estimate of b_2 is significantly ($P < 0.05$) smaller than 0, because the t (-3.52) is much greater in absolute value than the critical value of a one-tailed t test (1.65 at $P = 0.05$, one-tailed); hence, the estimated Gaussian logit response curve differs significantly from a sigmoid response curve, so that the optimum is significant.

Exercise 3.6 Multiple logit regression

Exercise 3.6.1 In hayfield (USE = 0) with pH = 5: $\log_e [p/(1 - p)] = -57.82 + (19.30 \times 5) - (1.56 \times 5^2) = -0.32$, which gives $\hat{p} = 0.421$. In pasture (USE = 1) with pH = 5: $\log_e [p/(1 - p)] = -0.32 - 0.04 = -0.36$, which gives $\hat{p} = 0.411$.

For pH = 6, the estimated probabilities of occurrence are 0.860 and 0.855 in hayfield and pasture, respectively. The optimum pH is now estimated as $-19.30/(-2 \times 1.56) = 6.18$ and the tolerance as 0.57, identical for hayfields and pastures. The maximum probabilities of occurrence are 0.867 and 0.862 in hayfield and pasture, respectively. The difference between the estimated curves is small (Figure 3.20), the difference from the curve estimated in Exercise 3.5 is small too.

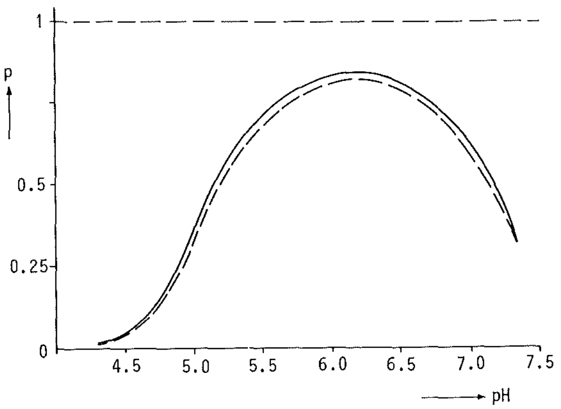

Figure 3.20 Gaussian logit curves of probability of occurrence of *Elymus repens* in hayfield (solid line) and pasture (broken line) against acidity (pH), as fitted by multiple logit regression. The probability of occurrence of *Elymus repens* at pH = 5 is estimated at 0.421 in hayfield and 0.411 in pasture; the difference is not statistically significant. Data from Kruijne et al. (1967).

Exercise 3.6.2 The t of the coefficient c_1 of USE is much smaller than the critical t at 5%. Therefore there is no evidence from these data that the probability of occurrence in fields with the same pH differs between hayfields and pastures. Acidity can therefore account for the overall difference between hayfields and pastures found in Exercise 3.4. The test result is not surprising after our observation in the previous exercise that the difference between the estimated response curves is small.

Exercise 3.6.3 The deviance of the model with acidity and agricultural use is 176.2; dropping agricultural use (variable USE) gives us the model with acidity only (Exercise 3.5), whose deviance is 176.3. The change in deviance (0.1) is much smaller than the critical value of a chi-square distribution with one degree of freedom, the change in the number of parameters between the models being one. The conclusion is the same as in Exercise 3.6.2.

Exercise 3.6.4 The deviance of the model with acidity and agricultural use is 176.2; dropping acidity (pH and pH^2) gives us the model with agricultural use only (Exercise 3.4), whose deviance is 192.9. The change in deviance is 16.7, which must be compared with a chi-square distribution with two degrees of freedom: $\chi^2_{0.05}(2) = 5.99$.

The conclusion is that acidity has an effect after correction for the effect of agricultural use. Acidity and agricultural use are not substitutable in the sense of Subsection 3.5.3; agricultural use cannot replace acidity in explanatory power, as judged by the deviance tests.