Chapter 5

Correlation and regression

Biologists commonly record more than one variable from each sampling or experimental unit. For example, a physiologist may record blood pressure and body weight from experimental animals, or an ecologist may record the abundance of a particular species of shrub and soil pH from a series of plots during vegetation sampling. Such data are termed bivariate when we have two random variables recorded from each unit or multivariate when we have more than two random variables recorded from each unit. There are a number of relevant questions that might prompt us to collect such data, based on the nature of the biological and statistical relationship between the variables. The next two chapters consider statistical procedures for describing the relationship(s) between two or more continuous variables, and using that relationship for prediction. Techniques for detecting patterns and structure in complex multivariate data sets, and simplifying such data sets for further analyses, will be covered in Chapters 15–18.

5.1 Correlation analysis

Consider a situation where we are interested in the statistical relationship between two random variables, designated Y_1 and Y_2, in a population. Both variables are continuous and each sampling or experimental unit (i) in the population has a value for each variable, designated y_{i1} and y_{i2}.

Land crabs on Christmas Island

Christmas Island in the northeast Indian Ocean is famous for its endemic red land crabs, *Gecarcoidea natalis*, which undergo a spectacular mass migration back to the ocean each year to release their eggs. The crabs inhabit the rain forest on the island where they consume tree seedlings. In a study on the ecology of the crabs, Green (1997) tested whether there was a relationship between the total biomass of red land crabs and the density of their burrows within 25 m^2 quadrats (sampling units) at five forested sites on the island. The full analyses of these data are provided in Box 5.1.

5.1.1 Parametric correlation model

The most common statistical procedure for measuring the 'strength' of the relationship between two continuous variables is based on distributional assumptions, i.e. it is a parametric procedure. Rather than assuming specific distributions for the individual variables, however, we need to think of our data as a population of y_{i1} and y_{i2} pairs. We now have a joint distribution of two variables (a bivariate distribution) and, analogous to the parametric tests we described in Chapter 3, the bivariate normal distribution (Figure 5.1) underlies the most commonly used measure of the strength of a bivariate relationship. The bivariate normal distribution is defined by the mean and standard deviation of each variable and a parameter called the correlation coefficient, which measures the strength of the relationship between the two variables. A bivariate normal distribution implies that the individual variables are also normally distributed and also implies that any relationship between the two variables, i.e. any lack of independence between the variables, is a linear one (straight-line; see Box 5.2; Hays 1994). Nonlinear relationships between two variables indicate that the bivariate normal distribution does not apply and we must use other procedures that do not assume this distribution for quantifying the strength of such relationships (Section 5.1.2).

Box 5.1 Worked example: crab and burrow density on Christmas Island

Green (1997) studied the ecology of red land crabs on Christmas Island and examined the relationship between the total biomass of red land crabs and the density of their burrows within 25 m^2 quadrats (sampling units) at five forested sites on the island. We will look at two of these sites: there were ten quadrats at Lower Site (LS) and eight quadrats at Drumsite (DS). Scatterplots and boxplots are presented in Figure 5.3. There was slight negative skewness for biomass and burrow density for LS, and an outlier for burrow density for DS, but no evidence of nonlinearity. Pearson's correlation coefficient was considered appropriate for these data although more robust correlations were calculated for comparison.

Site	Correlation type	Statistic	*P* value
DS ($n=8$)	Pearson	0.392	0.337
	Spearman	0.168	0.691
	Kendall	0.036	0.901
LS ($n=10$)	Pearson	0.882	0.001
	Spearman	0.851	0.002
	Kendall	0.719	0.004

The H_0 of no linear relationship between total crab biomass and number of burrows at DS could not be rejected. The same conclusion applies for monotonic relationships measured by Spearman and Kendall's coefficients. So there was no evidence for any linear or more general monotonic relationship between burrow density and total crab biomass at site DS.

The H_0 of no linear relationship between total crab biomass and number of burrows at LS was rejected. The same conclusion applies for monotonic relationships measured by Spearman and Kendall's coefficients. There was strong evidence of a linear and more general monotonic relationship between burrow density and total crab biomass at site LS.

Covariance and correlation

One measure of the strength of a linear relationship between two continuous random variables is to determine how much the two variables covary, i.e. vary together. If one variable increases (or decreases) as the other increases (or decreases), then the two variables covary; if one variable does not change as the other variable increases (or decreases), then the variables do not covary. We can measure how much two variables covary in a

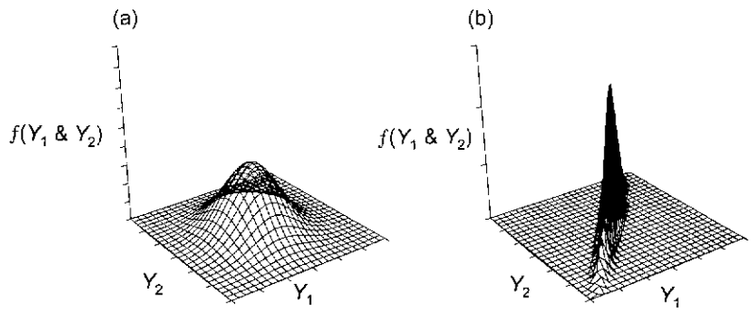

Figure 5.1 Bivariate normal distribution for (a) two variables with little correlation and (b) two variables with strong positive correlation.

Table 5.1 Parameters used for parametric correlation analysis and their estimates, with standard error for correlation coefficient. Note that y_{i1} and y_{i2} are the values of the two variables for observation i, $\bar{y}_1$ and $\bar{y}_2$ are the sample means for the two variables and n is the number of observations

Parameter	Estimate	Standard error
Covariance: $\sigma_{Y_1Y_2}$	$s_{Y_1Y_2} = \frac{\sum_{i=1}^{n}(y_{i1}-\bar{y}_1)(y_{i2}-\bar{y}_2)}{n-1}$	n/a
Correlation: $\rho_{Y_1Y_2}$	$r_{Y_1Y_2} = \frac{\sum_{i=1}^{n}[(y_{i1}-\bar{y}_1)(y_{i2}-\bar{y}_2)]}{\sqrt{\sum_{i=1}^{n}(y_{i1}-\bar{y}_1)^2\sum_{i=1}^{n}(y_{i2}-\bar{y}_2)^2}}$	$s_r = \sqrt{\frac{(1-r^2)}{(n-2)}}$

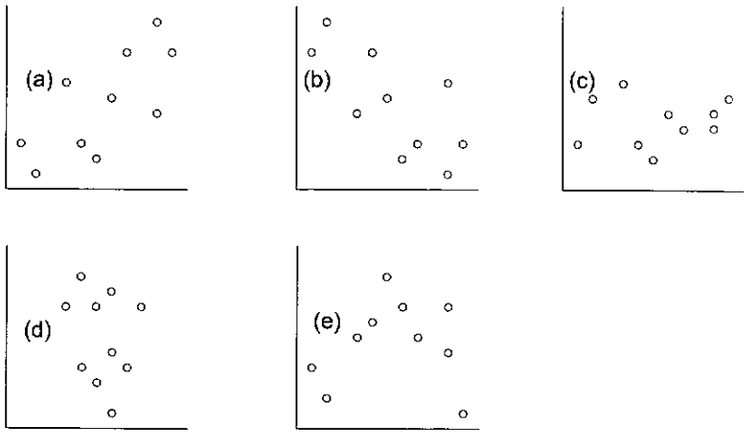

Figure 5.2 Scatterplots illustrating (a) a positive linear relationship ($r = 0.72$), (b) a negative linear relationship ($r = -0.72$), (c) and (d) no relationship ($r = 0.10$ and -0.17), respectively, and (e) a nonlinear relationship ($r = 0.08$).

sample of observations by the covariance (Table 5.1). The numerator is the sum of cross-products (SSCP), the bivariate analogue of the sum of squares (SS). The covariance ranges from $-\infty$ to $+\infty$. Note that a special case of the covariance is the sample variance (see Chapter 2), the covariance of a variable with itself.

One limitation of the covariance as a measure of the strength of a linear relationship is that its absolute magnitude depends on the units of the two variables. For example, the covariance between crab biomass and number of burrows in the study of Green (1996) would be larger by a factor of 10^3 if we measured biomass in grams rather than kilograms. We can standardize the covariance by dividing by the standard deviations of the two variables so that our measure of the strength of the linear relationship lies between -1 and $+1$. This is called the Pearson (product–moment) correlation (Table 5.1) and it measures the "strength" of the linear (straight-line) relationship between Y_1 and Y_2. If our sample data comprise a random sample from a population of (y_{i1}, y_{i2}) pairs then the sample correlation coefficient r is the maximum likelihood (ML) estimator of the population correlation coefficient ρ; r actually slightly under-estimates ρ, although the bias is small (Sokal & Rohlf 1995). Along with the means and standard deviations of the two variables, the population correlation coefficient (ρ) is the parameter that defines a bivariate normal distribution. The sample correlation coefficient is also the sample covariance of two variables that are both standardized to zero mean and unit variance (Rodgers & Nicewander 1988; see Chapter 4 for details on standardized variables). Note that r can be positive or negative (Figure 5.2) with $+1$ or -1 indicating that the observations fall along a straight line and zero indicating no correlation. The correlation coefficient measures linear

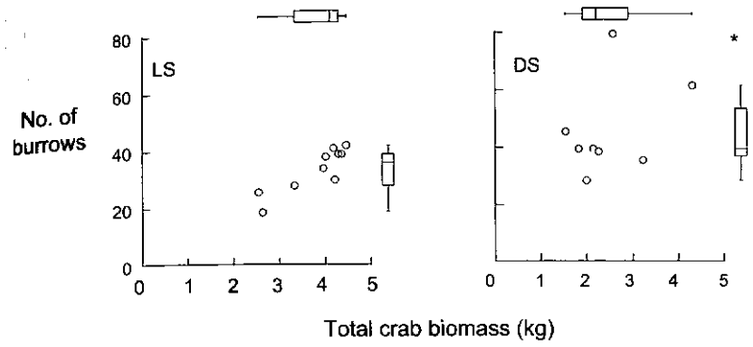

Figure 5.3 Scatterplots showing the relationship between number of burrows of red land crabs and total crab biomass in 25 m^2 quadrats at two sites (LS, DS) on Christmas Island (Green 1997). Each plot includes bordered boxplots for each variable separately.

relationships; two variables may have a strong nonlinear relationship but not have a large correlation coefficient (Figure 5.2(e)).

Since the sample correlation coefficient is a statistic, it has a sampling distribution (probability distribution of the sample correlation coefficient based on repeated samples of size n from a population). When ρ equals zero, the distribution of r is close to normal and the sample standard error of r can be calculated (Table 5.1). When ρ does not equal zero, the distribution of r is skewed and complex (Neter *et al.* 1996) and, therefore, the standard error cannot be easily determined (although resampling methods such as the bootstrap could be used; see Chapter 2). Approximate confidence intervals for ρ can be calculated using one of the versions of Fisher's z transformation (see Sokal & Rohlf 1995) that convert the distribution of r to an approximately normal distribution.

Hypothesis tests for ρ

The null hypothesis most commonly tested with Pearson's correlation coefficient is that ρ equals zero, i.e. the population correlation coefficient equals zero and there is no linear relationship between the two variables in the population. Because the sampling distribution of r is normal when ρ equals zero, we can easily test this H_0 with a t statistic:

$$t = \frac{r}{s_r} \qquad (5.1)$$

We compare t with the sampling distribution of t (the probability distribution of t when H_0 is true) with $n-2$ df. This is simply a t test that a single population parameter equals zero (where t equals the sample statistic divided by the standard error of the statistic) as described for the population mean in Chapter 3. The value of r can also be compared to the sampling distribution for r under the H_0 (see tables in Rohlf & Sokal 1969, Zar 1996). The results of testing the H_0 using the sampling distribution of t or r will be the same; statistical software usually does not provide a t statistic for testing correlation coefficients.

Tests of null hypotheses that ρ equals some value other than zero or that two population correlation coefficients are equal cannot use the above approach because of the complex sampling distribution of r when ρ does not equal zero. Tests based on Fisher's z transformation are available (Sokal & Rohlf 1995).

Assumptions

Besides the usual assumptions of random sampling and independence of observations, the Pearson correlation coefficient assumes that the joint probability distribution of Y_1 and Y_2 is bivariate normal. If either or both variables have non-normal distributions, then their joint distribution cannot be bivariate normal and any relationship between the two variables might not be linear. Nonlinear relationships can even arise if both variables have normal distributions. Remembering that the Pearson correlation coefficient measures the strength of the linear relationship between two variables, checking for a nonlinear relationship with a simple scatterplot and for asymmetrical distributions of the variables with boxplots is important. Modern statistical software produces these plots very easily (see Figure 5.3).

If the assumption of bivariate normality is suspect, based on either of the two variables having non-normal distributions and/or apparent nonlinearity in the relationship between the two

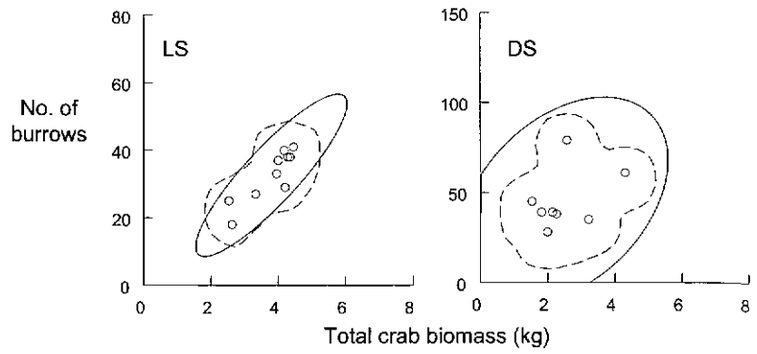

Figure 5.4 Comparison of 95% confidence ellipses (——) and kernel density estimators (-------) for the relationship between total crab biomass and number of burrows at sites LS and DS on Christmas Island (Green 1997).

variables, we have two options. First, we can transform one or both variables if they are skewed and their nature suggests an alternative scale of measurement might linearize their relationship (see Chapter 4 and Section 5.3.11). Second, we can use more robust measures of correlation that do not assume bivariate normality and linear relationships (Section 5.1.2).

5.1.2 Robust correlation

We may have a situation where the joint distribution of our two variables is not bivariate normal, as evidenced by non-normality in either variable, and transformations do not help or are inappropriate (e.g. the log of a variable does not make much theoretical sense). We may also be interested in testing hypotheses about monotonic relationships or more general associations between two variables, i.e. one variable increases (or decreases) as the other increases (or decreases) but not necessarily in a linear (straight-line) manner. One general approach for testing monotonic relationships between variables that does not assume bivariate normality is to examine the association of the ranks of the variables; statistical tests based on rank transformations were described in Chapter 3.

Spearman's rank correlation coefficient (r_s) is simply the Pearson correlation coefficient after the two variables have been separately transformed to ranks but the (y_{i1},y_{i2}) pairing is retained after ranking. An equivalent computation that uses the ranked data directly is also available (e.g. Hollander & Wolfe 1999, Sokal & Rohlf 1995, Sprent 1993). The null hypothesis being tested is that there is no monotonic relationship between Y_1 and Y_2 in the population. An alternative measure is Kendall's rank correlation coefficient, sometimes termed Kendall's tau (τ). The value of Spearman's r_s will be slightly greater than τ for a given data set (Box 5.1), and both are more conservative measures than Pearson's correlation when distribution assumptions hold. Note that these non-parametric correlation analyses do not detect all nonlinear associations between variables, just monotonic relationships.

5.1.3 Parametric and non-parametric confidence regions

When representing a bivariate relationship with a scatterplot, it is often useful to include confidence regions (Figure 5.4, left). The 95% confidence region, for example, is the region within which we would expect the observation represented by the population mean of the two variables to occur 95% of the time under repeated sampling from this population. Assuming our two variables follow a bivariate normal distribution, the confidence band will always be an ellipse centered on the sample means of Y_1 and Y_2 and the orientation of the ellipse is determined by the covariance (or the Pearson correlation coefficient). The two major axes (length and width) of these ellipses are determined from the variances (or standard deviations) of Y_1 and Y_2. These axes are used for some forms of regression analysis (Section 5.3.14) and also for some statistical procedures that deal with multivariate data sets, such as principal components analysis (Chapters 15 and 16). Note that if the linear relationship between Y_1 and Y_2 is weak, then the bounds of the ellipse may exceed the actual and theoretical range of our data, e.g. include impossible values such as negatives (Figure 5.4, right).

Sometimes we are not interested in the population mean of Y_1 and Y_2 but simply want a confidence region for the observations themselves. In Chapter 4, we introduced kernel density estimators for univariate data (Silverman 1986). The estimated density for a value of Y is the sum of the estimates from a series of symmetrical distributions (e.g. normal, although others are often used) fitted to groups of local observations. In the bivariate case, we determine contours that surround regions of high bivariate density where these contours are formed from summing a series of symmetrical bivariate distributions fitted to groups of local paired observations. Note that the kernel estimators are not constrained to a specific ellipsoid shape and will often better represent the pattern of density of observations in our sample (Figure 5.4, right).

Box 5.2 What does "linear" mean?

The term linear model has been used in two distinct ways. First, it means a model of a straight-line relationship between two variables. This is the interpretation most biologists are familiar with. A second, more correct, definition is that a linear model is simply one in which any value of the variable of interest (y_i) is described by a linear combination of a series of parameters (regression slopes, intercept), and "no parameter appears as an exponent or is multiplied or divided by another parameter" (Neter *et al.* 1996, p. 10). Now the term "linear" refers to the combination of parameters, not the shape of the relationship. Under this definition, linear models with a single predictor variable can represent not only straight-line relationships such as Equation 5.3, but also curvilinear relationships, such as the models with polynomial terms described in Chapter 6.

5.2 Linear models

Most of the analyses in the following chapters are concerned with fitting statistical models. These are used in situations where we can clearly specify a response variable, also termed the dependent variable and designated Y, and one or more predictor variables, also termed the independent variables or covariates and designated X_1, X_2, etc. A value for each response and predictor variable is recorded from sampling or experimental units in a population. We expect that the predictor variables may provide some biological explanation for the pattern we see in the response variable. The statistical models we will use take the following general form:

response variable = model + error (5.2)

The model component incorporates the predictor variables and parameters relating the predictors to the response. In most cases, the predictor variables, and their parameters, are included as a linear combination (Box 5.2), although nonlinear terms are also possible. The predictor variables can be continuous or categorical or a combination of both. The error component represents the part of the response variable not explained by the model, i.e. uncertainty in our response variable. We have to assume some form of probability distribution for the error component, and hence for the response variable, in our model.

Our primary aim is to fit our model to our observed data, i.e. confront our model with the data (Hilborn & Mangel 1997). This fitting is basically an estimation procedure and can be done with ordinary least squares or maximum likelihood (Chapter 2). We will emphasize OLS for most of our models, although we will be assuming normality of the error terms for interval estimation and hypothesis testing. Such models are called general linear models, the term "general" referring to the fact that both continuous and categorical predictors are allowed. If other distributions are applicable, especially when there is a relationship between the mean and the variance of the response variable, then ML must be used for estimation. These models are called generalized

linear models, generalized meaning that other distributions besides normal and relationships between the mean and the variance can be accommodated.

We nearly always have more than one statistical model to consider. For example, we might have the simplest model under a null hypothesis versus a more complex model under some alternative hypothesis. When we have many possible predictor variables, we may be comparing a large number of possible models. In all cases, however, the set of models will be nested whereby we have a full model with all predictors of interest included and the other models are all subsets of this full model. Testing hypotheses about predictors and their parameters involves comparing the fit of models with and without specific terms in this nested hierarchy. Non-nested models can also be envisaged but they cannot be easily compared using the estimation and testing framework we will describe, although some measures of fit are possible (Hilborn & Mangel 1997; Chapter 6).

Finally, it is important to remember that there will not usually be any best or correct model in an absolute sense. We will only have sample data with which to assess the fit of the model and estimate parameters. We may also not have chosen all the relevant predictors nor considered combinations of predictors, such as interactions, that might affect the response variable. All the procedure for analyzing linear models can do is help us decide which of the models we have available is the best fit to our observed sample data and enable us to test hypotheses about the parameters of the model.

5.3 Linear regression analysis

In this chapter, we consider statistical models that assume a linear relationship between a continuous response variable and a single, usually continuous, predictor variable. Such models are termed simple linear regression models (Box 5.2) and their analysis has three major purposes:

1. to describe the linear relationship between Y and X,
2. to determine how much of the variation (uncertainty) in Y can be explained by the linear relationship with X and how much of this variation remains unexplained, and
3. to predict new values of Y from new values of X.

Our experience is that biologists, especially ecologists, mainly use linear regression analysis to describe the relationship between Y and X and to explain the variability in Y. They less commonly use it for prediction (see discussion in Ford 2000, Peters 1991).

5.3.1 Simple (bivariate) linear regression

Simple linear regression analysis is one of the most widely applied statistical techniques in biology and we will use two recent examples from the literature to illustrate the issues associated with the analysis.

Coarse woody debris in lakes

The impact of humans on freshwater environments is an issue of great concern to both scientists and resource managers. Coarse woody debris (CWD) is detached woody material that provides habitat for freshwater organisms and affects hydrological processes and transport of organic materials within freshwater systems. Land use by humans has altered the input of CWD into freshwater lakes in North America, and Christensen *et al.* (1996) studied the relationships between CWD and shoreline vegetation and lake development in a sample of 16 lakes. They defined CWD as debris greater than 5 cm in diameter and recorded, for a number of plots on each lake, the density (no. km^{-1}) and basal area ($m^2\ km^{-1}$) of CWD in the nearshore water, and the density (no. km^{-1}) and basal area ($m^2\ km^{-1}$) of riparian trees along the shore. They also recorded density of cabins along the shoreline. Weighted averages of these values were determined for each lake, the weighting based on the relative proportion of lake shore with forest and with cabins. We will use their data to model the relationships between CWD basal area and two predictor variables separately, riparian tree density and cabin density. These analyses are presented in Box 5.3.

Species–area relationships

Ecologists have long been interested in how abundance and diversity of organisms relate to the area of habitat in which those organisms are found.

Box 5.3 Worked example of linear regression analysis: coarse woody debris in lakes

Christensen *et al.* (1996) studied the relationships between coarse woody debris (CWD) and shoreline vegetation and lake development in a sample of 16 lakes in North America. The main variables of interest are the density of cabins (no. km^{-1}), density of riparian trees (trees km^{-1}), the basal area of riparian trees ($m^2\ km^{-1}$), density of coarse woody debris (no. km^{-1}), basal area of coarse woody debris ($m^2\ km^{-1}$).

CWD basal area against riparian tree density

A scatterplot of CWD basal area against riparian tree density, with a Loess smoother fitted, showed no evidence of a nonlinear relationship (Figure 5.13(a)). The boxplots of each variable were slightly skewed but the residuals from fitting the linear regression model were evenly spread and there were no obvious outliers (Figure 5.13(b)). One lake (Tenderfoot) had a higher Cook's D_i than the others that was due mainly to a slightly higher leverage value because this lake had the greatest riparian density (X-variable). Omitting this lake from the analysis did not alter the conclusions so it was retained and the variables were not transformed.

The results of the OLS fit of a linear regression model to CWD basal area against riparian tree density were as follows.

	Coefficient	Standard error	Standardized coefficient	t	P
Intercept	−77.099	30.608	0	−2.519	0.025
Slope	0.116	0.023	0.797	4.929	<0.001

Correlation coefficient (r) = 0.797, r^2 = 0.634

Source	df	MS	F	P
Regression	1	3.205×10^4	24.303	<0.001
Residual	14	1318.969		

The t test and the ANOVA F test cause us to reject the H_0 that β_1 equals zero. Note that $F\ (24.307) = t^2\ (4.929)$, allowing for rounding errors. We would also reject the H_0 that β_0 equals zero, although this test is of little biological interest. The r^2 value (0.634) indicates that we can explain about 63% of the total variation in CWD basal area by the linear regression with riparian tree density.

We can predict CWD basal area for a new lake with 1500 trees km^{-1} in the riparian zone. Plugging 1500 into our fitted regression model:

$$\text{CWD basal area} = -77.099 + 0.116 \times 1500$$

the predicted basal area of CWD is 96.901 $m^2\ km^{-1}$. The standard error of this predicted value (from Equation 5.10) is 37.900, resulting in a 95% confidence interval for true mean CWD basal area of lakes with a riparian density of 1500 trees km^{-1} of ± 81.296.

CWD basal area against cabin density

A scatterplot of CWD basal area against cabin density, with a Loess smoother fitted, showed some evidence of a nonlinear relationship (Figure 5.14(a)). The boxplot of

cabin density was highly skewed, with a number of zero values. The residuals from fitting the linear regression model to untransformed data suggested increasing spread of residuals with an unusual value (Arrowhead Lake) with a low (negative) predicted value and a much higher Cook's D_i than the others (Figure 5.14(b)). Following Christensen *et al.* (1996), we transformed cabin density to $\log_{10}$ and refitted the linear model. The scatterplot of CWD basal area against $\log_{10}$ cabin density suggested a much better linear relationship (Figure 5.15(a)). The boxplot of $\log_{10}$ cabin density was less skewed but the residuals from fitting the linear regression model still showed increasing spread with increasing predicted values. Lake Arrowhead was no longer influential but Lake Bergner was an outlier with a moderate Cook's D_i. Finally, we fitted a linear model when both variables were $\log_{10}$ transformed. The scatterplot of $\log_{10}$ CWD basal area against $\log_{10}$ cabin density suggested a slightly less linear relationship (Figure 5.16(a)) and the boxplot of $\log_{10}$ CWD basal area was now negatively skewed. The residuals from fitting the linear regression model were much improved with constant spread and no observations were particularly influential.

Overall, transforming both variables seems to result in a linear model that fits best to these data, although we will present the analysis with just cabin density transformed as per Christensen *et al.* (1996). The results of the OLS fit of a linear regression model to CWD basal area against $\log_{10}$ cabin density were as follows.

	Coefficient	Standard error	Standardized coefficient	t	P
Intercept	121.969	13.969	0	8.732	<0.001
Slope	−93.301	18.296	−0.806	−5.099	<0.001
Correlation coefficient (r) = −0.806, r^2 = 0.650					
Source	df	MS	F	P	
Regression	1	3.284×10^4	26.004	<0.001	
Residual	14	1262.870			

The t test and the ANOVA F test cause us to reject the H_0 that β_1 equals zero. We would also reject the H_0 that β_0 equals zero, although this test is of little biological interest, especially as the slope of the relationship is negative.

For example, it has been shown that as the area of islands increases, so does the number of species of a variety of taxa (Begon *et al.* 1996). On rocky intertidal shores, beds of mussels are common and many species of invertebrates use these mussel beds as habitat. These beds are usually patchy and isolated clumps of mussels mimic islands of habitat on these shores. Peake & Quinn (1993) investigated the relationship between the number of species of macroinvertebrates, and the total abundance of macroinvertebrates, and area of clumps of mussels on a rocky shore in southern Australia. They collected a sample of 25 clumps of mussels in June 1989 and all organisms found within each clump were identified and counted. We will use their data to model the relationship between two separate response variables, the total number of species and the total number of individuals, and one predictor variable, clump area in dm^2. These analyses are presented in Box 5.4.

5.3.2 Linear model for regression

Consider a set of $i = 1$ to n observations where each observation was selected because of its specific X-value, i.e. the X-values were fixed by the investigator, whereas the Y-value for each observation is sampled from a population of possible Y-values. The simple linear regression model is:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad (5.3)$$

Box 5.4 Worked example of linear regression analysis: species richness of macroinvertebrates in mussel clumps

Peake & Quinn (1993) investigated the relationship between the number of species of macroinvertebrates, and the total abundance of macroinvertebrates, and area of clumps of mussels on a rocky shore in southern Australia. The variables of interest are clump area (dm^2), number of species, and number of individuals.

Number of species against clump area

A scatterplot of number of species against clump area, and the plot of residuals against predicted number of species from a linear regression analysis, both suggested a nonlinear relationship (Figure 5.17(a,b)). Although only clump area was positively skewed, Peake & Quinn (1993) transformed both variables because of the nature of the species–area relationships for other seasons in their study plus the convention in species–area studies to transform both variables.

The scatterplot of log number of species against log clump area (Figure 5.18) linearized the relationship effectively except for one of the small clumps. The residual plot also showed no evidence of nonlinearity but that same clump had a larger residual and was relatively influential (Cook's $D_i = 1.02$). Reexamination of the raw data did not indicate any problems with this observation and omitting it did not alter the conclusions from the analysis (b_1 changed from 0.386 to 0.339, r^2 from 0.819 to 0.850, all tests still $P<0.001$) so it was not excluded from the analysis. In fact, just transforming clump area produced the best linearizing of the relationship with no unusually large residuals or Cook's D_i statistics but, for the reasons outlined above, both variables were transformed.

The results of the OLS fit of a linear regression model to log number of species and log clump area were as follows.

	Coefficient	Standard error	Standardized coefficient	t	P
Intercept	1.270	0.024	0	52.237	<0.001
Slope	0.386	0.038	0.905	10.215	<0.001
Correlation coefficient (r) = 0.905, r^2 = 0.819					
Source	df	MS	F	P	
Regression	1	1.027	104.353	<0.001	
Residual	23	0.010			

The t test and the ANOVA F test cause us to reject the H_0 that β_1 equals zero. We would also reject the H_0 that β_0 equals zero, indicating that the relationship between species number and clump area must be nonlinear for small clump sizes since the model must theoretically go through the origin. The r^2 value (0.819) indicates that we can explain about 82% of the total variation in log number of species by the linear regression with log clump area.

Number of individuals against clump area

A scatterplot of number of individuals against clump area, with a Loess smoother fitted, suggested an approximately linear relationship (Figure 5.19(a)). The plot of residuals against predicted number of individuals from a linear regression model

fitted to number of individuals against clump area (Figure 5.19(b)) showed a clear pattern of increasing spread of residuals against increasing predicted number of individuals (or, equivalently, clump area); the pattern in the residuals was wedge-shaped. The boxplots in Figure 5.19(a) indicated that both variables were positively skewed so we transformed both variables to logs to correct for variance heterogeneity.

The scatterplot of log number of individuals against log clump area (Figure 5.20(a)) showed an apparent reasonable fit of a linear regression model, with symmetrical boxplots for both variables. The residual plot showed a more even spread of residuals with little wedge-shaped pattern (Figure 5.20(b)).

The results of the OLS fit of a linear regression model to log number of individuals and log clump area were as follows.

	Coefficient	Standard error	Standardized coefficient	t	P
Intercept	2.764	0.045	0	60.766	<0.001
Slope	0.835	0.071	0.927	11.816	<0.001
Correlation coefficient (r) = 0.927, r^2 = 0.859					
Source	df	MS	F	P	
Regression	1	4.809	139.615	<0.001	
Residual	23	0.034			

The t test and the ANOVA F test cause us to reject the H_0 that β_1 equals zero. We would also reject the H_0 that β_0 equals zero, although this test is of little biological interest. The r^2 value (0.859) indicates that we can explain about 86% of the total variation in log number of individuals by the linear regression with log clump area.

The details of the linear regression model, including estimation of its parameters, are provided in Box 5.5.

For the CWD data from Christensen *et al.* (1996), we would fit:

$$(\text{CWD basal area})_i = \beta_0 + \beta_1(\text{riparian tree density})_i + \varepsilon_i \quad (5.4)$$

where $n = 16$ lakes.

For the species–area data from Peake & Quinn (1993), we would fit:

$$(\text{number of species})_i = \beta_0 + \beta_1(\text{mussel clump area})_i + \varepsilon_i \quad (5.5)$$

where $n = 25$ mussel clumps.

In models 5.3 and 5.4:

y_i is the value of Y for the ith observation when the predictor variable $X = x_i$. For example, this is the basal area of CWD for the ith lake when the riparian tree density is x_i;

β_0 is the population intercept, the mean value of the probability distribution of Y when $x_i = 0$, e.g. mean basal area of CWD for lakes with no riparian trees;

β_1 is the population slope and measures the change in Y per unit change in X, e.g. the change in basal area of CWD for a unit (one tree km^{-1}) change in riparian tree density; and

ε_i is random or unexplained error associated with the ith observation, e.g. the error terms for a linear model relating basal area of CWD to riparian tree density in lakes are the differences between each observed value for CWD basal area and the true mean CWD basal area at each possible riparian tree density.

In this model, the response variable Y is a random variable whereas the predictor variable X represents fixed values chosen by the researcher. This means that repeated sampling from the population of possible sampling units would use the same values of X; this restriction on X has important ramifications for the use of

Box 5.5 The linear regression model and its parameters

Consider a set of $i = 1$ to n observations with fixed X-values and random Y-values. The simple linear regression model is:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (5.3)$$

In model 5.3 we have the following.

y_i is the value of Y for the ith observation when the predictor variable $X = x_i$.

β_0 is the population intercept, the mean value of the probability distribution of Y when x_i equals zero.

β_1 is the population slope and measures the change in Y per unit change in X.

ε_i is random or unexplained error associated with the ith observation. Each ε_i measures, for each x_i, the difference between each observed y_i and the mean of y_i; the latter is the value of y_i predicted by the population regression model, which we never know. We must make certain assumptions about these error terms for the regression model to be valid and to allow interval estimation of parameters and hypothesis tests. We assume that these error terms are normally distributed at each x_i, their mean at each x_i is zero [$E(\varepsilon_i)$ equals zero] and their variance is the same at each x_i and is designated σ_ε^2. This assumption is the same as the homogeneity of variances of y_i described in Section 5.3.8. We also assume that these ε_i terms are independent of, and therefore uncorrelated with, each other. Since the ε_i terms are the only random ones in our regression model, then these assumptions (normality, homogeneity of variances and independence) also apply to the response variable y_i at each x_i. We will examine these assumptions and their implications in more detail in Section 5.3.8.

Figure 5.5 illustrates the population linear regression model and shows some important features:

1. For any particular value of X (x_i), there is a population of Y-values with a probability distribution. For most regression applications, we assume that the population of Y-values at each x_i has a normal distribution. While not necessary to obtain point estimates of the parameters in the model, this normality assumption is necessary for determining confidence intervals on these parameters and for hypothesis tests.

2. These populations of Y-values at each x_i are assumed to have the same variance (σ^2); this is termed the homogeneity of variance assumption.

3. The true population regression line joins the means of these populations of Y-values.

4. The overall mean value of Y, also termed the expected value of Y [$E(Y)$], equals $\beta_0 + \beta_1 X$. This implies that we can re-express the linear regression model in terms of means of the response variable Y at each x_i:

$$y_i = \mu_i + \varepsilon_i$$

where μ_i is the population mean of Y-values at each x_i. This type of linear model is particularly useful when the predictor variable is categorical and the effects of the predictor on the response variable are usually expressed in terms of mean values.

As we described in Chapter 2, we can use either of two methods for estimating parameters, (ordinary) least squares (OLS) and maximum likelihood (ML). If we assume normality of the ε_i, it turns out that the OLS and ML estimates of β_0 and β_1 are identical, although, as is usual for variance estimation, the ML estimate of the variance (σ_ε^2) is slightly biased whereas the OLS estimate of σ_ε^2 is not. In this book, we will focus on OLS estimates of these parameters; details of the calculations for ML estimation of regression parameters can be found in Neter *et al.* (1996).

The OLS estimates of β_0 and β_1 are the values that produce a sample regression line ($\hat{y}_i = b_0 + b_1 x_i$) that minimize $\Sigma(y_i - \hat{y}_i)^2$. These are the sum of the squared deviations (SS) between each observed y_i and the value of y_i predicted by the sample regression line for each x_i. This is the sum of squared vertical distances between each observation and the fitted regression line (Figure 5.6). Note that for any x_i, $\hat{y}_i$ is our best estimate of the mean of y_i in the usual case of only a single y_i at each x_i. In practice, the values of b_0 and b_1 that minimize $\Sigma(y_i - \hat{y}_i)^2$ are found by using a little calculus to derive two new equations, termed normal equations, that are solved simultaneously for b_0 and b_1 (see Neter *et al.* 1996, Rawlings *et al.* 1998 for details).

Because we have different populations of Y for each x_i, the estimate of the common variance of ε_i and y_i (σ_ε^2) must be based on deviations of each observed Y-value from the estimated value of the mean Y-value at each x_i. As stated above, our best estimate of the mean of y_i is $\hat{y}_i$. This difference between each observed Y-value and each predicted $\hat{y}_i$ is called a residual:

$$e_i = y_i - \hat{y}_i$$

These residuals are very important in the analysis of linear models. They provide the basis of the OLS estimate of σ_ε^2 and they are valuable diagnostic tools for checking assumptions and fit of our model. The OLS estimate of σ_ε^2 is the sample variance of these residuals and is termed the Residual (or Error) Mean Square (Table 5.2). Remember from Chapter 2 that a variance is also termed a mean square. The numerator of the $MS_{Residual}$ is the sum-of-squares (SS) of the residuals and the quantity that OLS estimation minimizes when determining estimates of the regression model parameters. The degrees of freedom (the denominator) are $n-2$ because we must estimate both β_0 and β_1 to estimate σ_ε^2. The $SS_{Residual}$ and $MS_{Residual}$ measure the variation in Y around the fitted regression line. Two other attributes of residuals are important: their sum equals zero ($\Sigma_{i=1}^{n} e_i = 0$) and, therefore, their mean must also equal zero ($\bar{e} = 0$). Note that the residuals ($e_i = y_i - \hat{y}_i$) are related to the model error terms ($\varepsilon_i = y_i - \mu_i$) because our best estimate of the mean of Y at each x_i is the predicted value from the fitted regression model.

regression analysis in biology because usually both Y and X are random variables with a joint probability distribution. For example, the predictor variable in the study by Peake & Quinn (1993) was the area of randomly chosen clumps of mussels, clearly a random variable. Some aspects of classical regression analysis, like prediction and tests of hypotheses, might not be affected by X being a random variable whereas the estimates of regression coefficients can be inaccurate. We will discuss this issue in some detail in Section 5.3.14.

From the characteristics of the regression model summarized in Box 5.5, we assume that (a) there is a population of lakes with a normal distribution of CWD basal areas, (b) the variances of CWD basal area (σ_i^2) are the same for all of these populations and (c) the CWD basal areas in different lakes are independent of each other. These assumptions also apply to the error terms of the model, so the common variance of the error terms is σ_ε^2. We will examine these assumptions and their implications in more detail in Section 5.3.8.

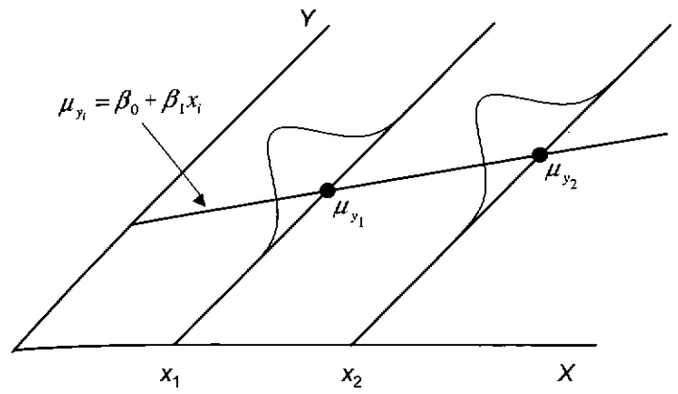

Figure 5.5 Diagrammatic representation of a linear regression model showing the population of y_i at two values of x_i. Note that the population regression model relates the mean of Y at each X-value (x_i) to $\beta_0 + \beta_1 x_i$.

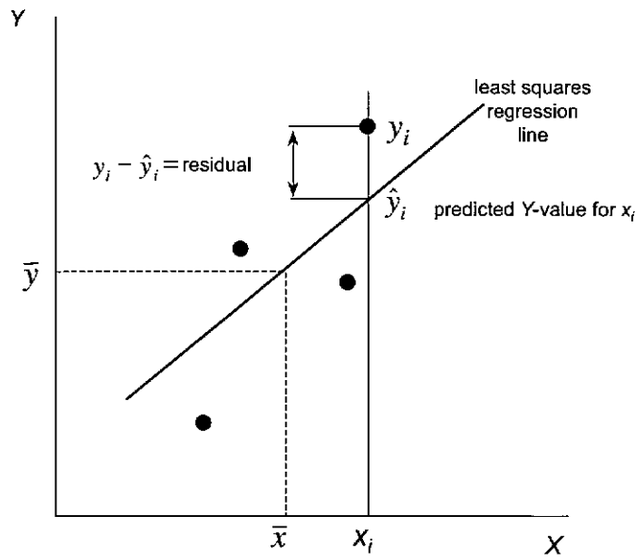

Figure 5.6 Illustration of the least squares regression line and residual values.

5.3.3 Estimating model parameters

The main aim of regression analysis is to estimate the parameters (β_0 and β_1) of the linear regression model based on our sample of n observations with fixed X-values and random Y-values. Actually, there are three parameters we need to estimate: β_0, β_1 and σ_ε^2 (the common variance of ε_i and therefore of y_i). Once we have estimates of these parameters (Box 5.5), we can determine the sample regression line:

$$\hat{y}_i = b_0 + b_1 x_i \qquad (5.6)$$

In model 5.6:

$\hat{y}_i$ is the value of y_i predicted by the fitted regression line for each x_i, e.g. the predicted basal area of CWD for lake i.

b_0 is the sample estimate of β_0, the Y-intercept, e.g. the predicted basal area of CWD for a lake with no riparian trees; and

b_1 is the sample estimate of β_1, the regression slope, e.g. the estimated change in basal area of CWD for a unit (one tree km^{-1}) change in riparian tree density.

The OLS estimates of β_0 and β_1 are the values that minimize the sum of squared deviations (SS) between each observed value of CWD basal area and the CWD basal area predicted by the fitted regression model against density of riparian trees. The estimates of the linear regression model are summarized in Table 5.2.

Regression slope

The parameter of most interest is the slope of the regression line β_1 because this measures the strength of the relationship between Y and X. The estimated slope (b_1) of the linear regression

Table 5.2 Parameters of the linear regression model and their OLS estimates with standard errors

Parameter	OLS estimate	Standard error
β_1	$b_1 = \frac{\sum_{i=1}^{n}[(x_i-\bar{x})(y_i-\bar{y})]}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$	$s_{b_1} = \sqrt{\frac{MS_{Residual}}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$
β_0	$b_0 = \bar{y} - b_1\bar{x}$	$s_{b_0} = \sqrt{MS_{Residual}\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]}$
ε_i	$e_i = y_i - \hat{y}_i$	$\sqrt{MS_{Residual}}$ (approx.)

model derived from the solution of the normal equations is the covariance between Y and X divided by the sum of squares (SS) of X (Table 5.2). The sample regression slope can be positive or negative (or zero) with no constraints on upper and lower limits.

The estimate of the β_1 is based on X being fixed so in the common case where X is random, we need a different approach to estimating the regression slope (Section 5.3.14). Nonetheless, there is also a close mathematical relationship between linear regression and bivariate correlation that we will discuss in Section 5.4. For now, note that we can also calculate b_1 from the sample correlation coefficient between Y and X as:

$$b_1 = r\frac{s_Y}{s_X} \tag{5.7}$$

where s_X and s_Y are the sample standard deviations of X and Y and r is the sample correlation coefficient between X and Y.

Standardized regression slope

Note that the value of the regression slope depends on the units in which X and Y are measured. For example, if CWD basal area was measured per 10 km rather than per kilometer, then the slope would be greater by a factor of ten. This makes it difficult to compare estimated regression slopes between different data sets. We can calculate a standardized regression slope b_1^*, termed a beta coefficient:

$$b_1^* = b_1\frac{s_X}{s_Y} \tag{5.8}$$

This is simply the sample regression slope multiplied by the ratio of the standard deviation of X and the standard deviation of Y. It is also the sample correlation coefficient. The same result can be achieved by first standardizing X and Y (each to a mean of zero and a standard deviation of one) and then calculating the usual sample regression slope. The value of b_1^* provides an estimate of the slope of the regression model that is independent of the units of X and Y and is useful for comparing regression slopes between data sets. For example, the estimated slopes for regression models of CWD basal area and CWD density against riparian tree density were 0.116 and 0.652 respectively, suggesting a much steeper relationship for basal area. The standardized slopes were 0.797 and 0.874, indicating that when the units of measurement were taken into account, the strength of the relationship of riparian tree density on CWD basal area and CWD density were similar. Note that the linear regression model for standardized variables does not include an intercept because its OLS (or ML) estimate would always be zero. Standardized regression slopes are produced by most statistical software.

Intercept

The OLS regression line must pass through $\bar{y}$ and $\bar{x}$. Therefore, the estimate (b_0) of the intercept of our regression model is derived from a simple rearrangement of the sample regression equation, substituting b_1, $\bar{y}$ and $\bar{x}$. The intercept might not be of much practical interest in regression analysis because the range of our observations rarely includes X equals zero and we should not usually extrapolate beyond the range of our sample observations. A related issue that we will discuss below is whether the linear regression line should be forced through the origin (Y equals zero and X equals zero) if we know theoretically that Y must be zero if X equals zero.

Confidence intervals

Now we have a point estimate for both σ_ε^2 and β_1, we can look at the sampling distribution and standard error of b_1 and confidence intervals for β_1. It turns out that the Central Limit Theorem applies to b_1 so its sampling distribution is normal with an expected value (mean) of β_1. The standard error of b_1, the standard deviation of its sampling distribution, is the square root of the residual mean square divided by the SS_X (Table 5.2). Confidence intervals for β_1 are calculated in the usual manner when we know the standard error of a statistic and use the t distribution. The 95% confidence interval for β_1 is:

$$b_1 \pm t_{0.05,n-2}s_{b_1} \tag{5.9}$$

Note that we use $n-2$ degrees of freedom (df) for the t statistic. The interpretation of confidence intervals for regression slopes is as described for means in Chapter 2. To illustrate using 95% confidence interval, under repeated sampling, we would expect 95% of these intervals to contain the fixed, but unknown, true slope of our linear regression model. The standard error (Table 5.2) and confidence intervals for β_0 can also be determined (Neter *et al.* 1996, Sokal & Rohlf 1995) and are standard output from statistical software.

We can also determine a confidence band (e.g. 95%) for the regression line (Neter *et al.* 1996, Sokal & Rohlf 1995). The 95% confidence band is a biconcave band that will contain the true population regression line 95% of the time. To illustrate with the data relating number of individuals of macroinvertebrates to mussel clump area from Peake & Quinn (1993), Figure 5.20(a) shows the confidence bands that would include the true population regression line 95% of the time under repeated sampling of mussel clumps. Note that the bands are wider further away from $\bar{x}$, indicating we are less confident about our estimate of the true regression line at the extremes of the range of observations.

Predicted values and residuals

Prediction from the OLS regression equation is straightforward by substituting an X-value into the regression equation and calculating the predicted Y-value. Be wary of extrapolating when making such predictions, i.e. do not predict from X-values outside the range of your data. The predicted Y-values have a sampling distribution that is normal and we provide the equation for the standard error of a new predicted Y-value because these standard errors are not always produced by statistical software:

$$s_{\hat{y}} = \sqrt{MS_{Residual}\left[1 + \frac{1}{n} + \frac{(x_p-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]} \tag{5.10}$$

where x_p is the new value of X from which we are predicting and the other terms have already been used in previous calculations. This predicted Y-value is an estimate of the true mean of Y for the new X-value from which we are predicting. Confidence intervals (also called prediction intervals) for this mean of Y can be calculated in the usual manner using this standard error and the t distribution with $n-2$ df.

This difference between each observed y_i and each predicted $\hat{y}_i$ is called a residual (e_i):

$$e_i = y_i - \hat{y}_i \tag{5.11}$$

For example, the residuals from the model relating CWD basal area to riparian tree density are the differences between each observed value of CWD basal area and the value predicted by the fitted regression model. We will use the residuals for checking the fit of the model to our data in Section 5.3.9.

Table 5.3 Analysis of variance (ANOVA) table for simple linear regression of Y on X

Source of variation	SS	df	MS	Expected mean square
Regression	$\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$	1	$\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{1}$	$\sigma_\varepsilon^2+\beta_1^2\sum_{i=1}^{n}(x_i-\bar{x})^2$
Residual	$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$	$n-2$	$\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-2}$	σ_ε^2
Total	$\sum_{i=1}^{n}(y_i-\bar{y})^2$	$n-1$		

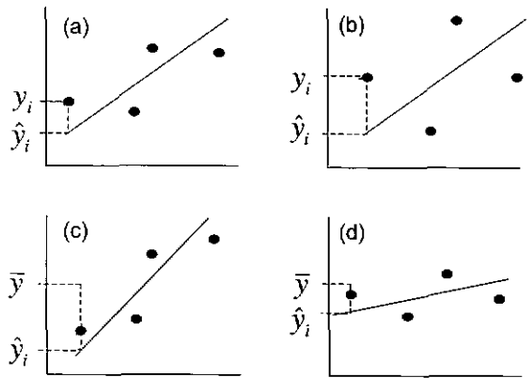

Figure 5.7 Illustration of explained and residual variation in regression analysis. Residual variation: (a) and (b) have identical regression lines but the differences between observed and predicted observations in (b) are greater than in (a) so the $MS_{Residual}$ in (b) is greater than in (a). Explained variation: (c) and (d) have identical $MS_{Residual}$ (the differences between the observed and predicted values are the same) but the total variation in Y is greater in (c) than (d) and the differences between the predicted values and the mean of Y are greater in (c) than in (d) so $MS_{Regression}$ would be greater in (c) than in (d).

5.3.4 Analysis of variance

A fundamental component of the analysis of linear models is partitioning the total variability in the response variable Y into the part due to the relationship with X (or X_1, X_2, etc. – see Chapter 6) and the part not explained by the relationship. This partitioning of variation is usually presented in the form of an analysis of variance (ANOVA) table (Table 5.3). The total variation in Y is expressed as a sum of squared deviations of each observation from the sample mean. This SS_{Total} has $n-1$ df and can be partitioned into two additive components. First is the variation in Y explained by the linear regression with X, which is measured as the difference between $\hat{y}_i$ and $\bar{y}$ (Figure 5.7). This is a measure of how well the estimated regression model predicts $\bar{y}$. The number of degrees of freedom associated with a linear model is usually the number of parameters minus one. For a simple linear regression model, there are two parameters (β_0 and β_1) so $df_{Regression}=1$.

Second is the variation in Y not explained by the regression with X, which is measured as the difference between each observed Y-value and the value of Y predicted by the model ($\hat{y}_i$) (Figure 5.7). This is a measure of how far the Y-values are from the fitted regression line and is termed the residual (or error) variation (see Section 5.3.3). The $df_{Residual}=n-2$, because we have already estimated two parameters (β_0 and β_1) to determine the $\hat{y}_i$.

The SS and df are additive (Table 5.3):

$$SS_{Regression}+SS_{Residual}=SS_{Total}$$
$$df_{Regression}+df_{Residual}=df_{Total}$$

Although the SS is a measure of variation, it is dependent on the number of observations that contribute to it, e.g. SS_{Total} will always get bigger as more observations with different values are included. In contrast to the SS, the variance (mean square, MS) is a measure of variability that does not depend on sample size because it is an average of the squared deviations and also has a known probability distribution (Chapter 2). So the next step in the analysis of variance is to convert the SS into MS by dividing them by their df:

The MS are not additive:

$$MS_{Regression}+MS_{Residual}\neq MS_{Total}$$

and the "MS_{Total}" does not play a role in analyses of variance.

These MS are sample variances and, as such, they estimate parameters. But unlike the situation where we have a single sample, and therefore a single variance (Chapter 2), we now have two variances. Statisticians have determined the expected values of these mean squares, i.e. the average of all possible values of these mean squares or what population values these mean squares actually estimate (Table 5.3).

The $MS_{Residual}$ estimates σ_ε^2, the common variance of the error terms (ε_i), and therefore of the Y-values at each x_i. The implicit assumption here, that we mentioned in Section 5.3.2 and will detail in Section 5.3.8, is that the variance of ε_i (and therefore of y_i) is the same for all x_i (homogeneity of variance), and therefore can be summarized by a single variance (σ_ε^2). If this assumption is not met, then $MS_{Residual}$ does not estimate a common variance σ_ε^2 and interval estimation and hypothesis tests associated with linear regression will be unreliable. The $MS_{Regression}$ also estimates σ_ε^2 plus an additional source of variation determined by the strength of the absolute relationship between Y and X (i.e. β_1^2 multiplied by the SS_X).

Sometimes the total variation in Y is expressed as an "uncorrected" total sum-of-squares ($SS_{Total\ uncorrected}$; see Neter *et al.* 1996, Rawlings *et al.* 1998). This is simply $\sum_{i=1}^{n}y_i^2$ and can be "corrected" by subtracting $n\bar{y}^2$ (termed "correcting for the mean") to convert $SS_{Total\ uncorrected}$ into the SS_{Total} we have used. The uncorrected total SS is occasionally used when regression models are forced through the origin (Section 5.3.12) and in nonlinear regression (Chapter 6).

5.3.5 Null hypotheses in regression

The null hypothesis commonly tested in linear regression analysis is that β_1 equals zero, i.e. the slope of the population regression model equals zero and there is no linear relationship between Y and X. For example, the population slope of the regression model relating CWD basal area to riparian tree density is zero or there is no linear relationship between number of species and mussel clump area in the population of all possible mussel clumps. There are two equivalent ways of testing this H_0.

The first uses the ANOVA we have described in Section 5.3.4. If H_0 is true and β_1 equals zero, then it is apparent from Table 5.3 that $MS_{Regression}$ and $MS_{Residual}$ both estimate σ_ε^2 because the term $\beta_1^2\sum_{i=1}^{n}(x_i-\bar{x})^2$ becomes zero. Therefore, the ratio of $MS_{Regression}$ to $MS_{Residual}$ should be less than or equal to one. If H_0 is not true and β_1 does not equal zero, then the expected value of $MS_{Regression}$ is larger than that of $MS_{Residual}$ and their ratio should be greater than one.

If certain assumptions hold (Section 5.3.8), the ratio of two sample variances (the F-ratio) follows a well-defined probability distribution called the F distribution (Chapter 2). A central F distribution is a probability distribution of the F-ratio[1] when the two sample variances come from populations with the same expected values. There are different central F distributions depending on the df of the two sample variances. Therefore, we can use the appropriate probability distribution of F (defined by numerator and denominator df) to determine whether the probability of obtaining our sample F-ratio or one more extreme (the usual hypothesis testing logic; see Chapter 3), is less than some specified significance level (e.g. 0.05) and therefore whether we reject H_0. This F test basically compares the fit to the data of a model that includes a slope term to the fit of a model that does not.

We can also test the H_0 that β_1 equals zero using a single parameter t test, as described in Chapter 3. We calculate a t statistic from our data:

$$t=\frac{b_1-\theta}{s_{b_1}} \qquad (5.12)$$

In Equation 5.12, θ is the value of β_1 specified in the H_0. We compare the observed t statistic to a t distribution with $(n-2)$ df with the usual logic of

1 *F-ratio versus F.* Hypothesis tests that involve comparisons of variance (ANOVA, ANCOVA, etc.) use an F-ratio, which is the ratio of two variances. This ratio follows an F distribution. Strictly speaking, any test statistic that we calculated as part of an ANOVA or ANCOVA is an F-ratio, but in much of the biological literature, there is reference to the less cumbersome F. We will often use this abbreviation.

a t test. Note that the F test of the H_0 that β_1 equals zero is mathematically identical to the t test; in fact, the F-ratio equals t^2 for a given sample. So, in practice, it does not matter which we use and both are standard output from statistical software. We offer some suggestions about presenting results from linear regression analyses in Chapter 19.

While the test of the H_0 that β_1 equals zero is most common, a test whether β_1 equals some other value may also be relevant, especially when variables have been log transformed. Examples include increases in metabolic rate with body size, an allometric relationship with a predicted slope of 0.75, and the self-thinning rule, that argues that the relationship between log plant size and log plant density would have a slope of $-3/2$ (Begon *et al.* 1996).

We can also test the H_0 that β_0 equals zero, i.e. the intercept of the population regression model is zero. Just as with the test that β_1 equals zero, the H_0 that β_0 equals zero can be tested with a t test, where the t statistic is the sample intercept divided by the standard error of the sample intercept. Alternatively, we can calculate an F test by comparing the fit of a model with an intercept term to the fit of a model without an intercept term (Section 5.3.6). The conclusions will be identical as the F equals t^2 and the t test version is standard output from statistical software. This H_0 is not usually of much biological interest unless we are considering excluding an intercept from our final model and forcing the regression line through the origin (Section 5.3.12).

Finally, we can test the H_0 that two regression lines come from populations with the same slope using a t test, similar to a test of equality of means (Chapter 3). A more general approach to comparing regression slopes is as part of analysis of covariance (ANCOVA, Chapter 12).

5.3.6 Comparing regression models

Methods for measuring the fit of a linear model to sample data fall into two broad categories based on the way the parameters of the models are estimated (see also Chapter 2).

1. Using OLS, the fit of a model is determined by the amount of variation in Y explained by the model or conversely, the lack of fit of a model is determined by the unexplained (residual) variation. This approach leads to the analysis of variance described above and F tests of null hypotheses about regression model parameters.

2. Using maximum likelihood (ML), the fit of a model is determined by the size of likelihood or log-likelihood. This approach leads to likelihood ratio tests of null hypotheses about regression model parameters and is most commonly used when fitting generalized linear models (GLMs) with non-normal error terms (Chapter 13).

The logic of comparing the fit of different models is the same whichever approach is used to measure fit. We will illustrate this logic based on the OLS estimation we have been using throughout this chapter. We can measure the fit of different models to the data and then compare their fits to test hypotheses about the model parameters. For example, smaller unexplained (residual) variation when a full model that includes β_1 is fitted compared with when a reduced model is fitted that omits β_1 is evidence against the H_0 that β_1 equals zero. Including a slope term in the model results in a better fit to the observed data than omitting a slope term. If there is no difference in the explanatory power of these two models, then there is no evidence against the H_0 that β_1 equals zero.

Let's explore this process more formally by comparing the unexplained, or residual, SS (the variation due to the difference between the observed and predicted Y-values) for full and reduced models (Box 5.6). To test the H_0 that β_1 equals zero, we fit the full model with both an intercept and a slope term (Equation 5.3):

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

We have already identified the unexplained SS from the full model as $\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2$. This is the $SS_{Residual}$ from our standard regression ANOVA in Table 5.3.

We then fit a reduced model that omits the slope term, i.e. the model expected if the H_0 that β_1 equals zero is true:

$$y_i = \beta_0 + \varepsilon_i \quad (5.13)$$

This is a model with zero slope (i.e. a flat line). The predicted Y-value for each x_i from this model is the intercept, which equals $\bar{y}$. Therefore, the unexplained SS from this reduced model is the sum of squared differences between the observed Y-values and $\bar{y}$ (i.e. $\Sigma_{i=1}^{n}(y_i - \bar{y})^2$), which is the SS_{Total} from our standard regression ANOVA.

Box 5.6 | Model comparisons in simple linear regression

We can use the model relating CWD basal area to riparian tree density to illustrate comparing the fit of full and reduced models to test null hypotheses about population parameters.

Test H_0: β_1 equals zero:

Full model:

$$(\text{CWD basal area})_i = \beta_0 + \beta_1(\text{riparian tree density})_i + \varepsilon_i$$

$SS_{Residual}$ = 18 465.56 (14 df).

Reduced model:

$$(\text{CWD basal area})_i = \beta_0 + \varepsilon_i$$

$SS_{Residual}$ = 50 520.00 (15 df).

Reduced $SS_{Residual}$ − Full $SS_{Residual}$ = 32 054.44 (1 df). This is identical to $MS_{Regression}$ from the ANOVA from fitting the original full model (Box 5.3).

Test H_0: β_0 equals zero:

Full model:

$$(\text{CWD basal area})_i = \beta_0 + \beta_1(\text{riparian tree density})_i + \varepsilon_i$$

$SS_{Residual}$ = 18 465.56 (14 df).

Reduced model:

$$(\text{CWD basal area})_i = \beta_1(\text{riparian tree density})_i + \varepsilon_i$$

$SS_{Residual}$ = 26 834.35 (15 df).

Reduced $SS_{Residual}$ − Full $SS_{Residual}$ = 8368.79 (1 df).

The difference between the unexplained variation of the full model ($SS_{Residual}$) and the unexplained variation from the reduced model (SS_{Total}) is simply the $SS_{Regression}$. It measures how much more variation in Y is explained by the full model than by the reduced model. It is, therefore, the relative magnitude of the $SS_{Regression}$ (which equals $MS_{Regression}$ with one df) that we use to evaluate the H_0 that β_1 equals zero (Box 5.6). So describing the $SS_{Regression}$ or $MS_{Regression}$ as the variation explained by the regression model is really describing the $SS_{Regression}$ or $MS_{Regression}$ as how much more variation in Y the full model explains over the reduced model.

The same logic can be used to test H_0 that β_0 equals zero by comparing the fit of the full model and the fit of a reduced model that omits the intercept:

$$y_i = \beta_1 x_i + \varepsilon_i \quad (5.14)$$

This is the model expected if the H_0 that β_0 equals zero is true and therefore, when x_i equals zero then y_i equals zero (Box 5.6).

For most regression models, we don't have to worry about comparing full and reduced models because our statistical software will do it automatically and provide us with the familiar ANOVA table and F tests and/or t tests. While comparisons of full and reduced models are trivial for linear models with a single predictor variable, the model comparison approach has broad applicability for testing null hypotheses about particular parameters in more complex linear (Chapter 6) and generalized linear models (Chapter 13).

5.3.7 Variance explained

A descriptive measure of association between Y and X is r^2 (also termed R^2 or the coefficient of

determination), which measures the proportion of the total variation in Y that is explained by its linear relationship with X. When we fit the full model, it is usually calculated as (Kvalseth 1985, Neter *et al.* 1996):

$$r^2 = \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{SS_{Residual}}{SS_{Total}} \quad (5.15)$$

Anderson-Sprecher (1994) argued that r^2 is better explained in terms of the comparison between the full model and a reduced (no slope parameter) model:

$$r^2 = 1 - \frac{SS_{Residual(Full)}}{SS_{Residual(Reduced)}} \quad (5.16)$$

Equations 5.15 and 5.16 are identical for models with an intercept (see below for no intercept models) but the latter version emphasizes that r^2 is a measure of how much the fit is improved by the full model compared with the reduced model. We can also relate explained variance back to the bivariate correlation model because r^2 is the square of the correlation coefficient r. Values of r^2 range between zero (no relationship between Y and X) and one (all points fall on fitted regression line). Therefore, r^2 is not an absolute measure of how well a linear model fits the data, only a measure of how much a model with a slope parameter fits better than one without (Anderson-Sprecher 1994).

Great care should be taken in using r^2 values for comparing the fit of different models. It is inappropriate for comparing models with different numbers of parameters (Chapter 6) and can be problematical for comparing models based on different transformations of Y (Scott & Wild 1991). If we must compare the fit of a linear model based on Y with the equivalent model based on, say log(Y), using r^2, we should calculate r^2 as above after re-expressing the two models so that Y is on the same original scale in both models (see also Anderson-Sprecher 1994).

5.3.8 Assumptions of regression analysis

The assumptions of the linear regression model strictly concern the error terms (ε_i) in the model, as described in Section 5.3.2. Since these error terms are the only random ones in the model, then the assumptions also apply to observations of the response variable y_i. Note that these assumptions are not required for the OLS estimation of model parameters but are necessary for reliable confidence intervals and hypothesis tests based on t distributions or F distributions.

Table 5.4 | Types of residual for linear regression models, where h_i is the leverage for observation i

Residual	$e_i = y_i - \hat{y}_i$
Standardized residual	$\frac{e_i}{\sqrt{MS_{Residual}}}$
Studentized residual	$\frac{e_i}{\sqrt{MS_{Residual}(1-h_i)}}$
Studentized deleted residual	$e_i \sqrt{\frac{n-1}{SS_{Residual}(1-h_i) - e_i^2}}$

The residuals from the fitted model (Table 5.4) are important for checking whether the assumptions of linear regression analysis are met. Residuals indicate how far each observation is from the fitted OLS regression line, in Y-variable space (i.e. vertically). Observations with larger residuals are further from the fitted line that those with smaller residuals. Patterns of residuals represent patterns in the error terms from the linear model and can be used to check assumptions and also the influence each observation has on the fitted model.

Normality

This assumption is that the populations of Y-values and the error terms (ε_i) are normally distributed for each level of the predictor variable x_i. Confidence intervals and hypothesis tests based on OLS estimates of regression parameters are robust to this assumption unless the lack of normality results in violations of other assumptions. In particular, skewed distributions of y_i can cause problems with homogeneity of variance and linearity, as discussed below.

Without replicate Y-values for each x_i, this assumption is difficult to verify. However, reasonable checks can be based on the residuals from the fitted model (Bowerman & O'Connell 1990). The methods we described in Chapter 4 for checking normality, including formal tests or graphical methods such as boxplots and probability plots, can be applied to these residuals. If the assumption is not met, then there are at least two options. First, a transformation of Y (Chapter 4 and Section 5.3.11) may be appropriate if the distribution is positively skewed. Second, we can fit a linear model using techniques that allow other distributions of error terms other than normal. These generalized linear models (GLMs) will be described in Chapter 13. Note that non-normality of Y is very commonly associated with heterogeneity of variance and/or nonlinearity.

Homogeneity of variance

This assumption is that the populations of Y-values, and the error terms (ε_i), have the same variance for each x_i:

$$\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_i^2 = \ldots = \sigma_\varepsilon^2 \text{ for } i = 1 \text{ to } n \quad (5.17)$$

The homogeneity of variance assumption is important, its violation having a bigger effect on the reliability of interval estimates of, and tests of hypotheses about, regression parameters (and parameters of other linear models) than non-normality. Heterogeneous variances are often a result of our observations coming from populations with skewed distributions of Y-values at each x_i and can also be due to a small number of extreme observations or outliers (Section 5.3.9).

Although without replicate Y-values for each x_i, the homogeneity of variance assumption cannot be strictly tested, the general pattern of the residuals for the different x_i can be very informative. The most useful check is a plot of residuals against x_i or $\hat{y}_i$ (Section 5.3.10). There are a couple of options for dealing with heterogeneous variances. If the unequal variances are due to skewed distributions of Y-values at each x_i, then appropriate transformations will always help (Chapter 4 and Section 5.3.11) and generalized linear models (GLMs) are always an option (Chapter 13). Alternatively, weighted least squares (Section 5.3.13) can be applied if there is a consistent pattern of unequal variance, e.g. increasing variance in Y with increasing X.

Independence

There is also the assumption that the Y-values and the ε_i are independent of each other, i.e. the Y-value for any x_i does not influence the Y-values for any other x_i. The most common situation in which this assumption might not be met is when the observations represent repeated measurements on sampling or experimental units. Such data are often termed longitudinal, and arise from longitudinal studies (Diggle *et al.* 1994, Ware & Liang 1996). A related situation is when we have a longer time series from one or a few units and we wish to fit a model where the predictor variable is related to a temporal sequence, i.e. a time series study (Diggle 1990). Error terms and Y-values that are non-independent through time are described as autocorrelated. A common occurrence in biology is positive first-order autocorrelation, where there is a positive relationship between error terms from adjacent observations through time, i.e. a positive error term at one time follows from a positive error term at the previous time and the same for negative error terms. The degree of autocorrelation is measured by the autocorrelation parameter, which is the correlation coefficient between successive error terms. More formal descriptions of autocorrelation structures can be found in many textbooks on linear regression models (e.g Bowerman & O'Connell 1990, Neter *et al.* 1996). Positive autocorrelation can result in underestimation of the true residual variance and seriously inflated Type I error rates for hypothesis tests on regression parameters. Note that autocorrelation can also be spatial rather than temporal, where observations closer together in space are more similar than those further apart (Diggle 1996).

If our Y-values come from populations in which the error terms are autocorrelated between adjacent x_i, then we would expect the residuals from the fitted regression line also to be correlated. An estimate of the autocorrelation parameter is the correlation coefficient between adjacent residuals, although some statistical software calculates this as the correlation coefficient between adjacent Y-values. Autocorrelation can therefore be detected in plots of residuals against x_i by an obvious positive, negative or cyclical trend in the residuals. Some statistical software also provides the Durbin–Watson test of the H_0 that the autocorrelation parameter equals zero. Because we might expect positive autocorrelation, this test is often one-tailed against the alternative hypothesis that the autocorrelation parameter is greater than zero. Note that the Durbin–Watson test is specifically designed for first-order autocorrelations and may not detect other patterns of non-independence (Neter *et al.* 1996).

There are a number of approaches to modeling a repeated series of observations on sampling or experimental units. These approaches can be used with both continuous (this chapter and Chapter 6) and categorical (Chapters 8–12) predictor variables and some are applicable even when the response variable is not continuous. Commonly, repeated measurements on individual units occur in studies that also incorporate a treatment structure across units, i.e. sampling or experimental units are allocated to a number of treatments (representing one or more categorical predictor variables or factors) and each unit is also recorded repeatedly through time or is subject to different treatments through time. Such "repeated measures" data are usually modeled with analysis of variance type models (partly nested models incorporating a random term representing units; see Chapter 11). Alternative approaches, including unified mixed linear models (Laird & Ware 1982, see also Diggle *et al.* 1994, Ware & Liang 1996) and generalized estimating equations (GEEs; see Liang & Zeger 1986, Ware & Liang 1996), based on the generalized linear model, will be described briefly in Chapter 13.

When the data represent a time series, usually on one or a small number of sampling units, one approach is to adjust the usual OLS regression analysis depending on the level of autocorrelation. Bence (1995) discussed options for this adjustment, pointing out that the usual estimates of the autocorrelation parameter are biased and recommending bias-correction estimates. Usually, however, data forming a long time series require more sophisticated modeling procedures, such as formal time-series analyses. These can be linear, as described by Neter *et al.* (1996) but more commonly nonlinear as discussed in Chatfield (1989) and Diggle (1990), the latter with a biological emphasis.

Fixed X

Linear regression analysis assumes that the x_i are known constants, i.e. they are fixed values controlled or set by the investigator with no variance associated with them. A linear model in which the predictor variables are fixed is known as Model I or a fixed effects model. This will often be the case in designed experiments where the levels of X are treatments chosen specifically. In these circumstances, we would commonly have replicate Y-values for each x_i and X may well be a qualitative variable, so analyses that compare mean values of treatment groups might be more appropriate (Chapters 8–12). The fixed X assumption is probably not met for most regression analyses in biology because X and Y are usually both random variables recorded from a bivariate distribution. For example, Peake & Quinn (1993) did not choose mussel clumps of fixed areas but took a haphazard sample of clumps from the shore; any repeat of this study would use clumps with different areas. We will discuss the case of X being random (Model II or random effects model) in Section 5.3.14 but it turns out that prediction and hypothesis tests from the Model I regression are still applicable even when X is not fixed.

5.3.9 Regression diagnostics

So far we have emphasized the underlying assumptions for estimation and hypothesis testing with the linear regression model and provided some guidelines on how to check whether these assumptions are met for a given bivariate data set. A proper interpretation of a linear regression analysis should also include checks of how well the model fits the observed data. We will focus on two aspects in this section. First, is a straight-line model appropriate or should we investigate curvilinear models? Second, are there any unusual observations that might be outliers and could have undue influence on the parameter estimates and the fitted regression model? Influence can come from at least two sources – think of a regression line as a see-saw, balanced on the mean of X. An observation can influence, or tip, the regression line more easily if it is further from the mean (i.e. at the ends of the range of X-values) or if it is far from the fitted regression line (i.e. has a large residual, analogous to a heavy person on the see-saw). We emphasized in Chapter 4 that it is really important to identify if the conclusions from any statistical analysis are influenced greatly by one or a few extreme observations. A variety of "diagnostic measures" can be calculated as part of the analysis that identify extreme or influential points and detect nonlinearity. These diagnostics also provide additional ways of checking the underlying assumptions of normality, homogeneity of variance and independence. We will illustrate some of the more common regression diagnostics that are standard outputs from most statistical software but others are available. Belsley *et al.* (1980) and Cook & Weisberg (1982) are the standard references, and other good discussions and illustrations include Bollen & Jackman (1990), Chatterjee & Price (1991) and Neter *et al.* (1996).

Leverage

Leverage is a measure of how extreme an observation is for the X-variable, so an observation with high leverage is an outlier in the X-space (Figure 5.8). Leverage basically measures how much each x_i influences $\hat{y}_i$ (Neter *et al.* 1996). X-values further from $\bar{x}$ influence the predicted Y-values more than those close to $\bar{x}$. Leverage is often given the symbol h_i because the values for each observation come from a matrix termed the hat matrix (**H**) that relates the y_i to the $\hat{y}_i$ (see Box 6.1). The hat matrix is determined solely from the X-variable(s) so Y doesn't enter into the calculation of leverage at all.

Leverage values normally range between $1/n$ and 1 and a useful criterion is that any observation with a leverage value greater than $2(p/n)$ (where p is the number of parameters in the model including the intercept; $p=2$ for simple linear regression) should be checked (Hoaglin & Welsch 1978). Statistical software may use other criteria for warning about observations with high leverage. The main use of leverage values is when they are incorporated in Cook's D_i statistic, a measure of influence described below.

Residuals

We indicated in Section 5.3.8 that patterns in residuals are an important way of checking regression assumptions and we will expand on this in Section 5.3.10. One problem with sample residuals is that their variance may not be constant for different x_i, in contrast to the model error terms that we assume do have constant variance. If we could modify the residuals so they had constant variance, we could more validly compare residuals to one another and check if any seemed unusually large, suggesting an outlying observation from the fitted model. There are a number of modifications that try to make residuals more useful for detecting outliers (Table 5.4).

Standardized residuals use the $\sqrt{MS_{Residual}}$ as an approximate standard error for the residuals. These are also called semistudentized residuals by Neter *et al.* (1996). Unfortunately, this standard error doesn't solve the problem of the variances of the residuals not being constant so a more sophisticated modification is needed. Studentized residuals incorporate leverage (h_i) as defined earlier. These studentized residuals do have constant variance so different studentized residuals can be validly compared. Large (studentized) residuals for a particular observation indicate that it is an outlier from the fitted model compared to the other observations. Studentized residuals also follow a t distribution with $(n-1)$ df if the regression assumptions hold. We can determine the probability of getting a specific studentized residual, or one more extreme, by comparing the studentized residual to a t distribution. Note that we would usually test all residuals in this way, which will result in very high family-wise Type I error rates (the multiple testing problem; see Chapter 3) so some type of P value adjustment might be required, e.g. sequential Bonferroni.

The deleted residual for observation i, also called the PRESS residual, is defined as the difference between the observed Y-values and those predicted by the regression model fitted to all the observations except i. These deleted residuals are usually calculated for studentized residuals. These studentized deleted residuals can detect outliers that might be missed by usual residuals (Neter *et al.* 1996). They can also be compared to a t distribution as we described above for the usual studentized residual.

Influence

A measure of the influence each observation has on the fitted regression line and the estimates of the regression parameters is Cook's distance statistic, denoted D_i. It takes into account both the size of leverage and the residual for each observation and basically measures the influence of each observation on the estimate of the regression slope (Figure 5.8). A large D_i indicates that removal of that observation would change the estimates of the regression parameters considerably. Cook's D_i can be used in two ways. First, informally by scanning the D_is of all observations and noting if any values are much larger than the rest. Second, by comparing D_i to an $F_{1,n}$ distribution; an approximate guideline is that

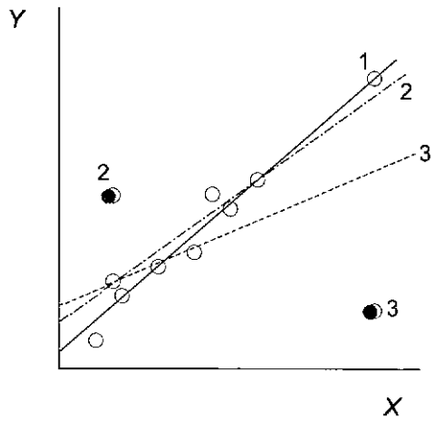

Figure 5.8 Residuals, leverage, and influence. The solid regression line is fitted through the observations with open symbols. Observation 1 is an outlier for both Y and X (large leverage) but not from the fitted model and is not influential. Observation 2 is not an outlier for either Y or X but is an outlier from the fitted model (large residual). Regression line 2 includes this observation and its slope is only slightly less than the original regression line so observation 2 is not particularly influential (small Cook's D_i). Observation 3 is not an outlier for Y but it does have large leverage and it is an outlier from the fitted model (large residual). Regression line 3 includes this observation and its slope is markedly different from the original regression line so observation 3 is very influential (large Cook's D_i, combining leverage and residual).

an observation with a D_i greater than one is particularly influential (Bollen & Jackman 1990). An alternative measure of influence that also incorporates both the size of leverage and the residual for each observation is $DFITS_i$, which measures the influence of each observation (i) on its predicted value ($\hat{y}_i$).

We illustrate leverage and influence in Figure 5.8. Note that observations one and three have large leverage and observations two and three have large residuals. However, only observation three is very influential, because omitting observations one or two would not change the fitted regression line much.

Transformations of Y that overcome problems of non-normality or heterogeneity of variance might also reduce the influence of outliers from the fitted model. If not, then the strategies for dealing with outliers discussed in Chapter 4 should be considered.

5.3.10 Diagnostic graphics

We cannot over-emphasize the importance of preliminary inspection of your data. The diagnostics and checks of assumptions we have just described are best used in graphical explorations of your data before you do any formal analyses. We will describe the two most useful graphs for linear regression analysis, the scatterplot and the residual plot.

Scatterplots

A scatterplot of Y against X, just as we used in simple correlation analysis, should always be the first step in any regression analysis. Scatterplots can indicate unequal variances, nonlinearity and outlying observations, as well as being used in conjunction with smoothing functions (Section 5.5) to explore the relationship between Y and X without being constrained by a specific linear model. For example, the scatterplot of number of species of invertebrates against area of mussel clump from Peake & Quinn (1993) clearly indicates nonlinearity (Figure 5.17(a)), while the plot of number of individuals against area of mussel clump indicates increasing variance in number of individuals with increasing clump area (Figure 5.19(a)). While we could write numerous paragraphs on the value of scatterplots as a preliminary check of the data before a linear regression analysis, the wonderful and oft-used example data from Anscombe (1973) emphasize how easily linear regression models can be fitted to inappropriate data and why preliminary scatterplots are so important (Figure 5.9).

Residual plots

The most informative way of examining residuals (raw or studentized) is to plot them against x_i or, equivalently in terms of the observed pattern, $\hat{y}_i$ (Figure 5.10). These plots can tell us whether the assumptions of the model are met and whether there are unusual observations that do not match the model very well.

If the distribution of Y-values for each x_i is positively skewed (e.g. lognormal, Poisson), we would expect larger $\hat{y}_i$ (an estimate of the population mean of y_i) to be associated with larger residuals. A wedge-shaped pattern of residuals, with a larger spread of residuals for larger x_i or $\hat{y}_i$ as shown for the model relating number of individuals of macroinvertebrates to mussel clump area

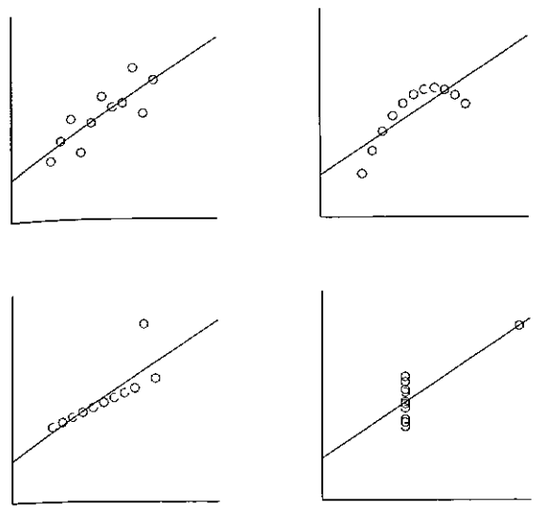

Figure 5.9 Scatterplots of four data sets provided in Anscombe (1973). Note that despite the marked differences in the nature of the relationships between Y and X, the OLS regression line, the r^2 and the test of the H_0 that β_1 equals zero are identical in all four cases: $y_i = 3.0 + 0.5x_i$, $n = 11$, $r^2 = 0.68$, H_0: $\beta_1 = 0$, $t = 4.24$, $P = 0.002$.

in our worked example (Box 5.4 and Figure 5.19(b)), indicates increasing variance in ε_i and y_i with increasing x_i associated with non-normality in Y-values and a violation of the assumption of homogeneity of variance. Transformation of Y (Section 5.3.11) will usually help. The ideal pattern in the residual plot is a scatter of points with no obvious pattern of increasing or decreasing variance in the residuals. Nonlinearity can be detected by a curved pattern in the residuals (Figure 5.17b) and outliers also stand out as having large residuals. These outliers might be different from the outliers identified in simple boxplots of Y, with no regard for X (Chapter 4). The latter are Y-values very different from the rest of the sample, whereas the former are observations with Y-values very different from that predicted by the fitted model.

Searle (1988) pointed out a commonly observed pattern in residual plots where points fall along parallel lines each with a slope of minus one (Figure 5.11). This results from a number of observations having similar values for one of the variables (e.g. a number of zeros). These parallel lines are not a problem, they just look a little unusual. If the response variable is binary (dichotomous),

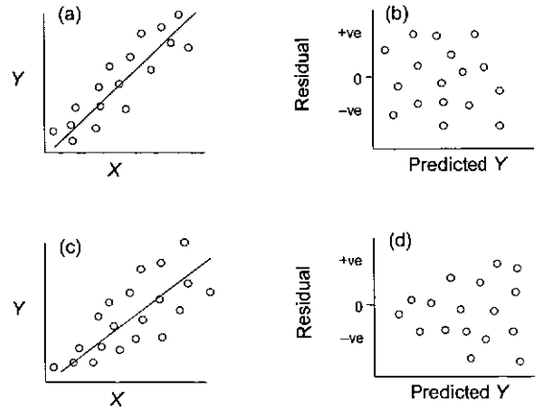

Figure 5.10 Diagrammatic representation of residual plots from linear regression: (a) regression showing even spread around line, (b) associated residual plot, (c) regression showing increasing spread around line, and (d) associated residual plot showing characteristic wedge-shape typical of skewed distribution.

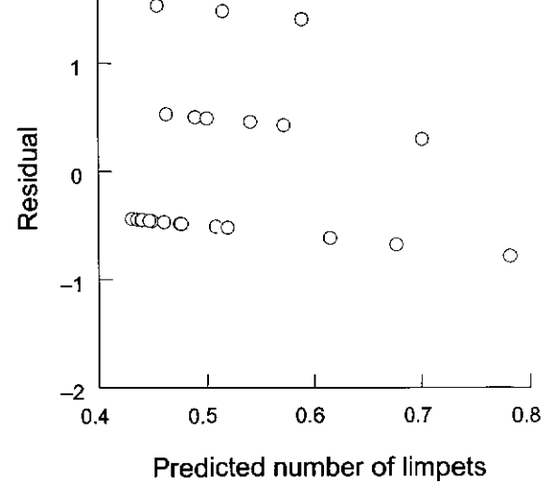

Figure 5.11 Example of parallel lines in a residual plot. Data from Peake & Quinn (1993), where the abundance of the limpets (*Cellana tramoserica*) was the response variable, area of mussel clump was the predictor variable and there were $n = 25$ clumps.

then the points in the residual plot will fall along two such parallel lines although OLS regression is probably an inappropriate technique for these data and a generalized linear model with a binomial error term (e.g. logistic regression) should be used (Chapter 13). The example in Figure 5.11 is from Peake & Quinn (1993), where the response variable (number of limpets per mussel clump) only takes three values: zero, one or two.

5.3.11 Transformations

When continuous variables have particular skewed distributions, such as lognormal or Poisson, transformations of those variables to a different scale will often render their distributions closer to normal (Chapter 4). When fitting linear regression models, the assumptions underlying OLS interval estimation and hypothesis testing of model parameters refer to the error terms from the model and, therefore, the response variable (Y). Transformations of Y can often be effective if the distribution of Y is non-normal and the variance of y_i differs for each x_i, especially when variance clearly increases as x_i increases. For example, variance heterogeneity for the linear model relating number of individuals of macroinvertebrates to mussel clump area was greatly reduced after transformation of Y (and also X – see below and compare Figure 5.19 and Figure 5.20). Our comments in Chapter 4 about the choice of transformations and the interpretation of analyses based on transformed data are then relevant to the response variable.

The assumption that the x_i are fixed values chosen by the investigator suggests that transformations of the predictor variable would not be warranted. However, regression analyses in biology are nearly always based on both Y and X being random variables, with our conclusions conditional on the x_i observed in our sample or we use a Model II analysis (Section 5.3.14). Additionally, our discussion of regression diagnostics shows us that unusual X-values determine leverage and can cause an observation to have undue influence on the estimated regression coefficient. Transformations of X should also be considered to improve the fit of the model and transforming both Y and X is sometimes more effective than just transforming Y.

The other use of transformations in linear regression analysis is to linearize a nonlinear relationship between Y and X (Chapter 4). When we have a clear nonlinear relationship, we can use nonlinear regression models or we can approximate the nonlinearity by including polynomial terms in a linear model (Chapter 6). An alternative approach that works for some nonlinear relationships is to transform one or both variables to make a simple linear model an appropriate fit to the data. Nonlinear relationships that can be made linear by simple transformations of the variables are sometimes termed "intrinsically linear" (Rawlings *et al.* 1998); for example, the relationship between the number of species and area of an island can be modeled with a nonlinear power function or a simple linear model after log transformation of both variables (Figure 5.17 and Figure 5.18). If there is no evidence of variance heterogeneity, then it is best just to transform X to try and linearize the relationship (Neter *et al.* 1996). Transforming Y in this case might actually upset error terms that are already normally distributed with similar variances. The relationship between number of species and area of mussel clump from Peake & Quinn (1993) illustrates this point, as a log transformation of just clump area (X) results in a linear model that best fits the data although both variables were transformed in the analysis (Box 5.4). However, nonlinearity is often associated with non-normality of the response variable and transformations of Y and/or Y and X might be required.

Remember that the interpretation of our regression model based on transformed variables, and any predictions from it, must be in terms of transformed Y and/or X, e.g. predicting log number of species from log clump area, although predictions can be back-transformed to the original scale of measurement if required.

5.3.12 Regression through the origin

There are numerous situations when we know that Y must equal zero when X equals zero. For example, the number of species of macroinvertebrates per clump of mussels on a rocky shore must be zero if that clump has no area (Peake & Quinn 1993), the weight of an organism must be zero when the length of that organism is zero etc. It might be tempting in these circumstances to force our regression line through the origin (Y equals zero, X equals zero) by fitting a linear model without an intercept term:

$$y_i = \beta_1 x_i + \varepsilon_i \qquad (5.14)$$

There are several difficulties when trying to interpret the results of fitting such a no-intercept model. First, our minimum observed x_i rarely extends to zero, and forcing our regression line through the origin not only involves extrapolating the regression line outside our data range but also assuming the relationship is linear outside this range (Cade & Terrell 1997, Neter *et al.* 1996). If we know biologically that Y must be zero when X is zero, yet our fitted regression line has an intercept different to zero, it suggests that the relationship between Y and X is nonlinear, at least for small values of X. We recommend that it is better to have a model that fits the observed data well than one that goes through the origin but provides a worse fit to the observed data.

Second, although residuals from the no-intercept model are $(y_i - \hat{y}_i)$ as usual, they no longer sum to zero, and the usual partition of SS_{Total} into $SS_{Regression}$ and $SS_{Residual}$ doesn't work. In fact, the $SS_{Residual}$ can be greater than SS_{Total} (Neter *et al.* 1996). For this reason, most statistical software presents the partitioning of the variance in terms of $SS_{Total\ uncorrected}$ (Section 5.3.4) that will always be larger than $SS_{Residual}$. However, the value of r^2 for a no-intercept model determined from $SS_{Total\ uncorrected}$ will not be comparable to r^2 from the full model calculated using SS_{Total} (Cade & Terrell 1997, Kvalseth 1985). The residuals are still comparable and the $MS_{Residual}$ is probably better for comparing the fit of models with and without an intercept (Chatterjee & Price 1991).

If a model with an intercept is fitted first and the test of the H_0 that β_0 equals zero is not rejected, there may be some justification for fitting a no-intercept model. For example, Caley & Schluter (1997) examined the relationship between local species richness (response variable) and regional species richness (predictor variable) for a number of taxa and geographic regions at two spatial scales of sampling (1% of region and 10% of region). They argued that local species richness must be zero when regional richness was zero and that no-intercept models were appropriate. Re-analysis of their data showed that when a model with an intercept was fitted to each combination of region and spatial scale, the test of the H_0 that β_0 equals zero was not rejected and the $MS_{Residual}$ was always less for a no-intercept model than a model with an intercept. This indicates that the no-intercept model was probably a better fit to the observed data. So no-intercept models were justified in this case, although we note that the estimates of β_1 were similar whether or not an intercept was included in the models.

Generally, however, we recommend against fitting a model without an intercept. The interpretation is more difficult and we must assume linearity of the relationship between Y and X beyond the range of our observed data.

5.3.13 Weighted least squares

The usual OLS approach for linear regression assumes that the variances of ε_i (and therefore the y_i) are equal, i.e. the homogeneity of variance assumption discussed in Section 5.3.8. If the variance of y_i varies for each x_i, we can weight each observation by the reciprocal of an estimate of its variance (σ_i^2):

$$w_i = \frac{1}{s_i^2} \qquad (5.18)$$

We then fit our linear regression model using generalized least squares which minimizes $\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2$. This is the principle of weighted least squares (Chaterjee & Price 1991, Myers 1990, Neter *et al.* 1996). The difficulty is calculating the w_i because we can't calculate s_i^2 unless we have replicate Y-values at each x_i. One approach is to group nearby observations and calculate s_i^2 (Rawlings *et al.* 1998), although there are no clear guidelines for how many observations to include in each group. A second approach uses the absolute value of each residual ($|e_i|$) from the OLS regression as an estimate of σ_i. Neter *et al.* (1996) suggested that the predicted values from an OLS regression of $|e_i|$ against x_i could be used to calculate the weights for each observation, where w_i is the inverse of the square of this predicted value. These weights can be used in statistical software with a weighted least squares option or, equivalently, OLS regression used once y_i and x_i in each pair has been multiplied (i.e. weighted) by w_i.

Box 5.7 | Model II regression.

For the data from Christensen et *al.* (1996), both the response variable (CWD basal area) and the predictor variable (riparian tree density) are random. These variables are measured in different units, so reduced major axis (RMA; also called standard major axis) and ranged MA regression are appropriate. We used the program "Model II regression" from Pierre Legendre at the University of Montreal.

Statistic	RMA	Ranged MA	OLS
b_1	0.145	0.164	0.116
95% CI	0.103 to 0.204	0.109 to 0.275	0.065 to 0.166
b_0	−113.904	−137.108	−77.099
95% CI	−187.152 to −61.767	−275.514 to −70.160	−142.747 to −11.451

The correlation coefficient was nearly 0.8, so we would not expect much difference in the estimates of the regression slope. The estimated regression slope from the RMA model and the ranged MA model were both larger than the OLS estimate, and, not surprisingly, the estimates of the intercept also differed. Note that the width of the confidence interval for β_1 was the same for RMA and OLS, but wider for ranged MA. A randomization test of the H_0 that β_1 equals zero for ranged MA resulted in a P value of 0.001. The test for the OLS regression is the same as the test for the correlation coefficient and provides a test for the RMA slope, with a P value less than 0.001.

Weighted least squares seems to have been rarely applied in the biological literature, most biologists including us preferring to transform one or both variables to meet the assumption of homogeneity of variance or else use generalized linear models (Chapter 13).

5.3.14 *X* random (Model II regression)

The linear regression model we have been using in this chapter is sometimes called Model I regression because X is a fixed variable, i.e. the x_i are fixed values set by the investigator and a new sample of observations from the population would use the same x_i. As we have previously discussed, most applications of linear regression in biology are unlikely to involve fixed X-values. Although we can usually conceptually distinguish a response variable (Y) from a predictor variable (X), the (x_i,y_i) pairs are commonly a sample from a bivariate distribution of two random variables, X and Y. For example, number of species per clump and area of mussel clump were clearly both random variables in the study by Peake & Quinn (1993) because clumps were chosen haphazardly from the shore and both variables recorded from each clump. Fitting a linear regression model for Y on X to data where both variables are random, and assumed to be jointly distributed with a bivariate normal distribution has been termed Model II regression (Legendre & Legendre 1998, Sokal & Rohlf 1995). It is a topic of some controversy and there are several ways of looking at the problem.

If the main aim of our regression analysis is prediction, then we can use the usual OLS regression model when Y and X are random as long as the probability distributions of y_i at each x_i are normal and independent. We must constrain our inferences about Y to be conditional given particular values of X (Neter *et al.* 1996).

If the main aim of our regression analysis is not prediction but to describe the true nature of the relationship between Y and X (i.e. estimate β_1), then OLS regression might not be appropriate. There is error variability associated with both Y (σ_ε^2) and X (σ_δ^2) and the OLS estimate of β_1 is biased towards zero (Box 5.7). The extent of the bias depends on the ratio of these error variances

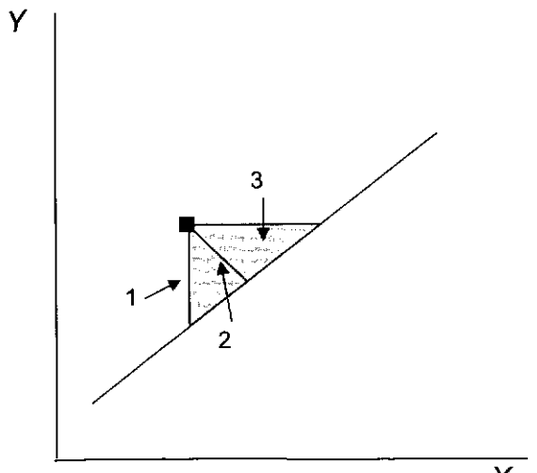

Figure 5.12 Distances or areas minimized by OLS (1), MA (2) and RMA (shaded area 3) linear regressions of Y on X.

(Legendre & Legendre 1998, Prairie *et al.* 1995, Snedecor & Cochran 1989):

$$\lambda = \frac{\sigma_\varepsilon^2}{\sigma_\delta^2} \tag{5.19}$$

If X is fixed then σ_δ^2 equals zero and the usual OLS estimate of β_1 is unbiased; the greater the error variability in X relative to Y, the greater the downward bias in the OLS estimate of β_1. Remember that the usual OLS regression line is fitted by minimizing the sum of squared vertical distances from each observation to the fitted line (Figure 5.12). Here, σ_δ^2 equals zero (fixed X) and λ equals ∞. The choice of method for estimating a linear regression model when both Y and X are random variables depends on our best guess of the value of λ, which will come from our knowledge of the two variables, the scales on which they are measured and their sample variances.

Major axis (MA) regression is estimated by minimizing the sum of squared perpendicular distances from each observation to the fitted line (Figure 5.12). For MA regression, σ_ε^2 is assumed to equal σ_δ^2 so λ equals one. The calculation of the estimate of the slope of the regression model is a little tedious, although it can be calculated using the estimate of the slope of the Model I regression and the correlation coefficients:

$$b_{1(MA)} = \frac{d \pm \sqrt{d^2+4}}{2} \tag{5.20}$$

If r is +ve, use the +ve square root and vice versa. In Equation 5.20:

$$d = \frac{b_{1(OLS)}^2 - r^2}{r^2 b_{1(OLS)}} \tag{5.21}$$

Standard errors and confidence intervals are best estimated by bootstrapping and a randomization test used for testing the H_0 of zero slope. Legendre & Legendre (1988) argued that MA regression was appropriate when both variables are measured on the same scales with the same units, or are dimensionless. They described a modification of MA regression, termed ranged MA regression. The variables are standardized by their ranges, the MA regression calculated, and then the regression slope is back-transformed to the original scale. The advantage of ranged MA regression is that the variables don't need to be in comparable units and a test of the H_0 of zero slope is possible (see below).

Reduced major axis (RMA) regression, also called the standard major axis (SMA) regression by Legendre & Legendre (1998), is fitted by minimizing the sum of areas of the triangles formed by vertical and horizontal lines from each observation to the fitted line (Figure 5.12). For RMA regression, it is assumed that σ_ε^2 and σ_δ^2 are proportional to σ_Y^2 and σ_X^2 respectively so λ equals σ_Y^2/σ_X^2. The RMA estimate of β_1 is simply the ratio of standard deviation of Y to the standard deviation of X:

$$b_1 = \frac{s_Y}{s_X} \tag{5.22}$$

This is also the average of the OLS estimate of the slope of Y on X and the reciprocal of the OLS estimate of the slope of X on Y. The standard error for the RMA estimate can be determined by bootstrapping but it turns out that the standard error of β_1 is, conveniently, the same as the standard error of the OLS estimate. Confidence intervals for β_1 can then be determined in the usual manner (Section 5.3.3). The H_0 that β_1 equals some specified value (except zero) can also be tested with a T-statistic (McArdle 1988, modified from Clarke 1980):

$$T = \frac{|\log b_1 - \log \beta_1^*|}{\sqrt{(1-r^2)(n-2)}} \tag{5.23}$$

where b_1 is the RMA estimate of β_1, β_1^* is the value of β_1 specified in the H_0 and the denominator is

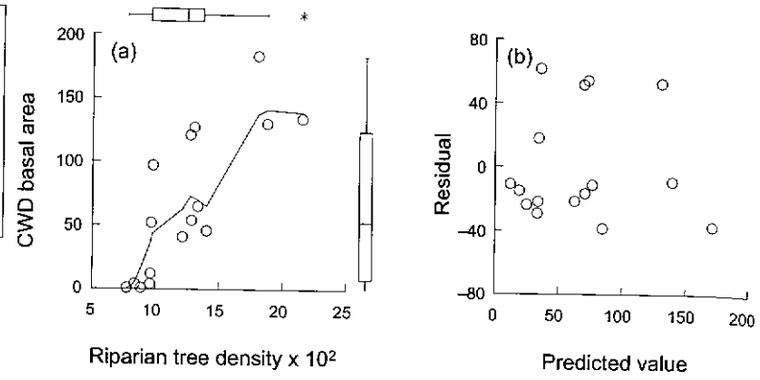

Figure 5.13 (a) Scatterplot (with Loess smoother, smoothing parameter = 0.5) of CWD basal area against riparian tree density. (b) Scatterplot of residuals against predicted CWD basal area from linear regression of CWD basal area against riparian tree density.

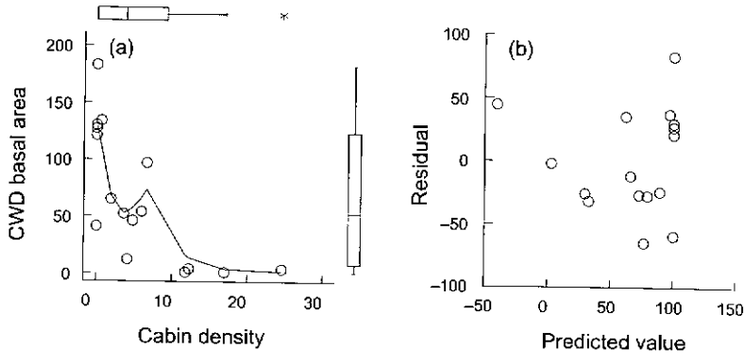

Figure 5.14 (a) Scatterplot (with Loess smoother, smoothing parameter = 0.5) of CWD basal area against cabin density. (b) Scatterplot of residuals against predicted CWD basal area from linear regression of CWD basal area against cabin density.

the standard error of the correlation coefficient (r). Note again the close relationship between RMA regression and the correlation coefficient. Testing β_1 against a specific non-zero value is applicable in many aspects of biology, such as the scaling of biological processes with body size of organisms (LaBarbera 1989). The H_0 that β_1 equals zero cannot be tested because log zero is undefined; the RMA regression slope is related to λ and cannot be strictly zero unless σ_Y^2 is also zero, an unlikely occurrence in practice (Legendre & Legendre 1998, McArdle 1988, Sokal & Rohlf 1995). The inability to formally test the H_0 that β_1 equals zero is actually a trivial problem because the H_0 that the population correlation coefficient (ρ) equals zero is essentially the same.

Prairie *et al.* (1995) proposed the slope-range method, which estimates β_1 when X is random from the relationship between the OLS estimate and $(1/s_X^2)$ for subsets of the data covering different ranges of X. This is a modification of methods based on instrumental variables (a third variable which may separate the data into groups). The main limitation of the method is that it needs a reasonably large sample size – at least ten potential groups in the data set with $n > 20$ in each group.

The intercepts are straightforward to calculate for any of these estimates of the slope because each regression line passes through the point $(\bar{y}, \bar{x})$ – see Section 5.3.3. The MA and RMA regression lines can be related to principal components analysis (see Chapter 17); the former is the first principal component of the covariance matrix between Y and X and the latter is the first principal component of the correlation matrix between Y and X. The RMA regression line is also the long axis of the bivariate confidence ellipse (Figure 5.4), indicating a close relationship between the correlation coefficient and the RMA regression line that we will elaborate on below.

Note that fitting a regression model of Y on X will produce a different OLS regression line than a regression model of X on Y for the same data because the first is minimizing deviations from the fitted line in Y and the latter is minimizing deviations from the fitted line in X. Interestingly,

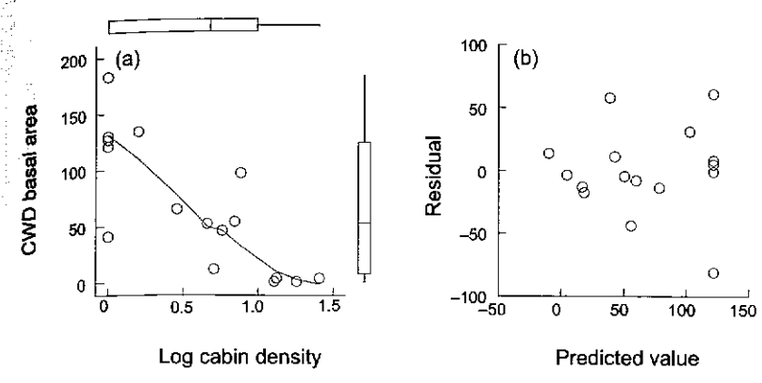

Figure 5.15 (a) Scatterplot (with Loess smoother, smoothing parameter = 0.5) of CWD basal area against $\log_{10}$ cabin density. (b) Scatterplot of residuals against predicted CWD basal area from linear regression of CWD basal area against $\log_{10}$ cabin density.

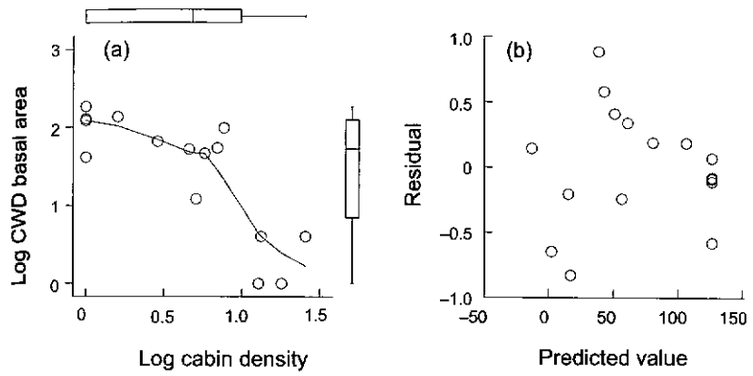

Figure 5.16 (a) Scatterplot (with Loess smoother, smoothing parameter = 0.5) of $\log_{10}$ CWD basal area against $\log_{10}$ cabin density. (b) Scatterplot of residuals against predicted $\log_{10}$ CWD basal area from linear regression of $\log_{10}$ CWD basal area against $\log_{10}$ cabin density.

as pointed out by Jackson (1991), the RMA line seems to most observers a more intuitive and better "line-of-best-fit" than the OLS line since it lies half way between the OLS line for Y on X and the OLS line for X on Y.

Simulations by McArdle (1988) comparing OLS, MA and RMA regression analyses when X is random showed two important results. First, the RMA estimate of β_1 is less biased than the MA estimate and is preferred, although he did not consider the ranged MA method. Second, if the error variability in X is more than about a third of the error variability in Y, then RMA is the preferred method; otherwise OLS is acceptable. As the correlation coefficient between Y and X approaches one (positive or negative), the difference between the OLS and RMA estimates of β_1, and therefore the difference between the fitted regression lines, gets smaller. Legendre & Legendre (1998) preferred the ranged MA over RMA, partly because the former permits a direct test of the H_0 that β_1 equals zero. We don't regard this as a crucial issue because the test of the H_0 that the correlation coefficient equals zero is the same test. A more sophisticated decision tree for choosing between methods for Model II regression is provided by Legendre & Legendre (1998), in addition to a detailed but very readable discussion of the issues.

Examples of the application of Model II regression analyses are most common in studies of scaling of aspects of biology with body size of organisms. Herrera (1992) calculated the OLS, MA and RMA estimates of the slope of the linear regression of log fruit width on log fruit length for over 90 species of plants from the Iberian Peninsula. He showed that, averaging across the species, the RMA estimate of the regression slope was greater than MA, which in turn was greater than OLS. He argued that MA regression was appropriate because the error variabilities for log width and log length were similar. Trussell (1997) used RMA regression for describing relationships between morphological characteristics (e.g. shell height, shell length, foot size, etc.) of an intertidal snail. However, he used OLS regressions to compare between shores as part of an analysis of

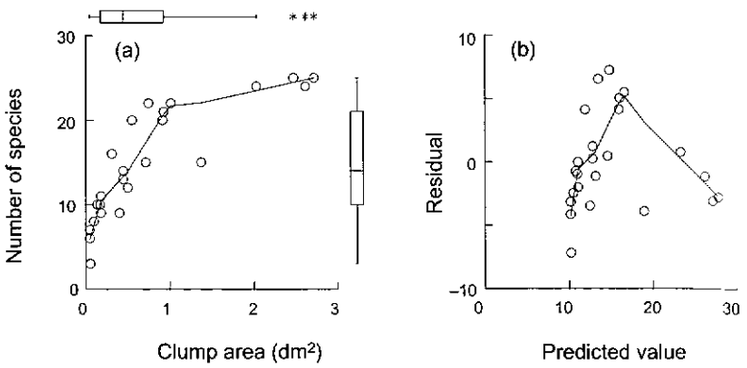

Figure 5.17 (a) Scatterplot (with Loess smoother, smoothing parameter = 0.5) of number of species against clump area. (b) Scatterplot (with Loess smoother, smoothing parameter = 0.5) of residuals against predicted number of species from linear regression of number of species against clump area.

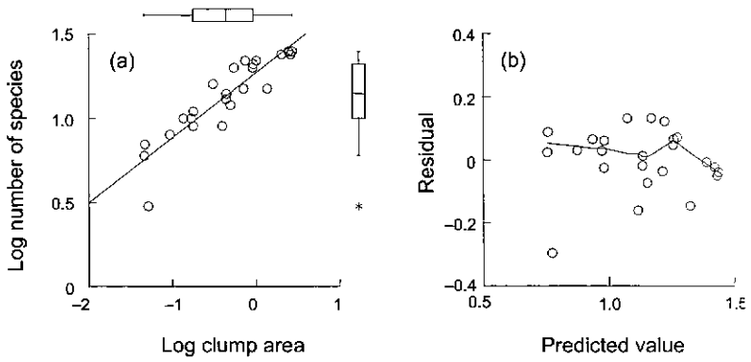

Figure 5.18 (a) Scatterplot (with linear regression line fitted) of $\log_{10}$ number of species against $\log_{10}$ clump area. (b) Scatterplot (with Loess smoother, smoothing parameter = 0.5) of residuals against predicted number of species from linear regression of $\log_{10}$ number of species against $\log_{10}$ clump area.

covariance (see Chapter 12). Both Herrera (1992) and Trussell (1997) tested whether their regression slopes were significantly different from unity, the value predicted if the relationships were simply allometric.

It is surprising that there are not more uses of Model II regression, or acknowledgment of the potential biases of using OLS estimates when both Y and X are random, in biological research literature, particularly given the extensive discussion in the influential biostatistics text by Sokal & Rohlf (1995). This may be partly because many excellent linear models textbooks are based on examples in industry or business and marketing where the assumption of fixed X is commonly met, so the issue X being random is not discussed in detail. Also, biologists seem primarily interested in the test of the H_0 that β_1 equals zero. Since the test is identical for OLS regression of Y on X and X on Y, and both are identical to the test that the correlation coefficient (ρ) equals zero, then it essentially does not matter whether OLS or RMA regression is used for this purpose. Biologists less commonly compare their estimates of β_1 with other values, so underestimating the true slope may not be costly.

5.3.15 Robust regression

One of the limitations of OLS is that the estimates of model parameters, and therefore subsequent hypothesis tests, can be sensitive to distributional assumptions and affected by outlying observations, i.e. ones with large residuals. Even generalized linear model analyses (GLMs; see Chapter 13) that allow other distributions for error terms besides normal, and are based on ML estimation, are sensitive to extreme observations. Robust regression techniques are procedures for fitting linear regression models that are less sensitive to deviations of the underlying distribution of error terms from that specified, and also less sensitive to extreme observations (Birkes & Dodge 1993).

Least absolute deviations (LAD)

LAD, sometimes termed least absolute residuals (LAR; see Berk 1990), is where the estimates of β_0 and β_1 are those that minimize the sum of absolute values of the residuals:

$$\sum_{i=1}^{n} |e_i| = \sum_{i=1}^{n} |(y_i - \hat{y}_i) \quad (5.24)$$

rather than the sum of squared residuals ($\Sigma_{i=1}^{n} e_i^2$) as in OLS. By not squaring the residuals, extreme observations have less influence on the fitted model. The difficulty is that the computations of the LAD estimates for β_0 and β_1 are more complex than OLS estimates, although algorithms are available (Birkes & Dodge 1993) and robust regression techniques are now common in statistical software (often as part of nonlinear modeling routines).

M-estimators

These were introduced in Chapter 2 for estimating the mean of a population. In a regression context, M-estimators involve minimizing the sum of some function of e_i, with OLS (minimizing $\Sigma_{i=1}^{n} e_i^2$) and LAD (minimizing $\Sigma_{i=1}^{n} |e_i|$) simply being special cases (Birkes & Dodge 1993). Huber M-estimators, described in Chapter 2, weight the observations differently depending how far they are from the center of the distribution. In robust regression analyses, Huber M-estimators weight the residuals (e_i) differently depending on how far they are from zero (Berk 1990) and use these new residuals to calculate adjusted Y-values. The estimates for β_0 and β_1 are those that minimize both $\Sigma_{i=1}^{n} e_i^2$ (i.e. OLS) when the residuals are near zero and $\Sigma |e_i|$ (i.e. LAD) when the residuals are far from zero. We need to choose the size of the residual at which the method switches from OLS to LAD; this decision is somewhat subjective, although recommendations are available (Huber 1981, Wilcox 1997). You should ensure that the default value used by your statistical software for robust regression seems reasonable. Wilcox (1997) described more sophisticated robust regression procedures, including an M-estimator based on iteratively reweighting the residuals. One problem with M-estimators is that the sampling distributions of the estimated coefficients are unlikely to be normal, unless sample sizes are

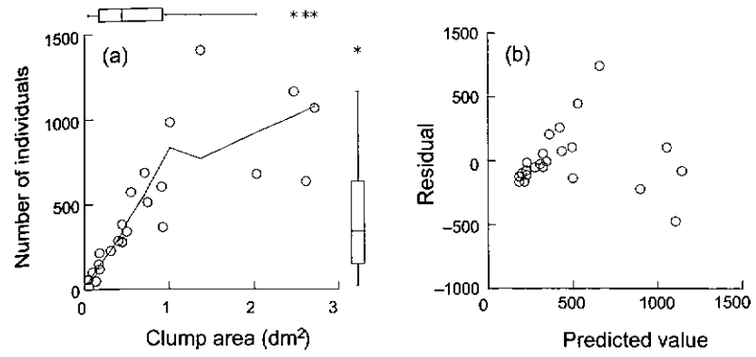

Figure 5.19 (a) Scatterplot (with Loess smoother, smoothing parameter = 0.5) of number of individuals against clump area. (b) Scatterplot of residuals against predicted number of individuals from linear regression of number of individuals against clump area.

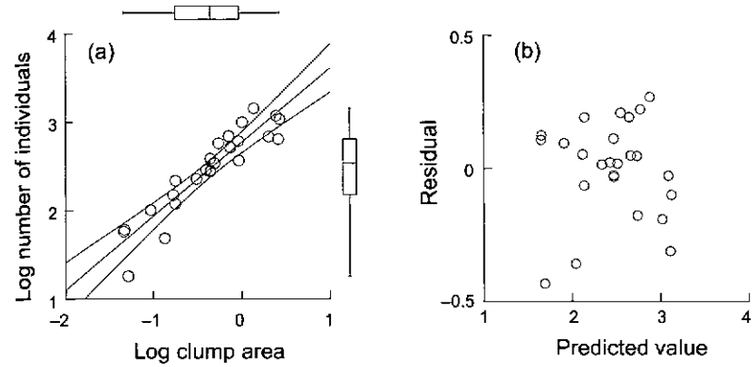

Figure 5.20 (a) Scatterplot (with linear regression line and 95% confidence band fitted) of $\log_{10}$ number of individuals against $\log_{10}$ clump area. (b) Scatterplot of residuals against predicted number of individuals from linear regression of $\log_{10}$ number of individuals against $\log_{10}$ clump area.

large, and the usual calculations for standard errors, confidence intervals and hypothesis testing may not be valid (Berk 1990). Resampling methods such as bootstrap (Chapter 2) are probably the most reliable approach (Wilcox 1997).

Rank-based ("non-parametric") regression

This approach does not assume any specific distribution of the error terms but still fits the usual linear regression model. This approach might be particularly useful if either of the two variables is not normally distributed and nonlinearity is evident but transformations are either ineffective or misrepresent the underlying biological process. The simplest non-parametric regression analysis is based on the $[n(n-1)]/2$ OLS slopes of the regression lines for each pair of X values (the slope for y_1x_1 and y_2x_2, the slope for y_2x_2 and y_3x_3, the slope for y_1x_1 and y_3x_3, etc.). The non-parametric estimator of β_1 (b_1) is the median of these slopes and the non-parametric estimator of β_0 (b_0) is the median of all the $y_i - b_1x_i$ differences (Birkes & Dodge 1993, Sokal & Rohlf 1995, Sprent 1993). A t test for β_1 based on the ranks of the Y-values is described in Birkes & Dodge (1993); an alternative is to simply use Kendall's rank correlation coefficient (Sokal & Rohlf 1995).

Randomization test

A randomization test of the H_0 that β_1 equals zero can also be constructed by comparing the observed value of b_1 to the distribution of b_1 found by pairing the y_i and x_i values at random a large number of times and calculating b_1 each time (Manly 1997). The P value then is the % of values of b_1 from this distribution equal to or larger than the observed value of b_1.

5.4 Relationship between regression and correlation

The discussion on linear regression models when both Y and X are random variables in Section 5.3.14 indicated the close mathematical and conceptual similarities between linear regression and correlation analysis. We will formalize those similarities here, summarizing points we have made throughout this chapter. The population slope of the linear regression of Y on X (β_{YX}) is related to the correlation between Y and X (ρ_{YX}) by the ratio of the standard deviations of Y and X:

$$\beta_{YX} = \rho_{YX}\frac{\sigma_Y}{\sigma_X} \tag{5.25}$$

Therefore, the OLS estimate of β_1 from the linear regression model for Y on X is:

$$b_{YX} = r_{YX}\frac{s_Y}{s_X} \tag{5.26}$$

The equivalent relationship also holds for the population slope of the linear regression of X on Y with the ratio of standard deviations reversed. Therefore the sample correlation coefficient between Y and X can be calculated from the standardized slope of the OLS regression of Y on X (Rodgers & Nicewander 1988).

These relationships between regression slopes and correlation coefficients result in some interesting equivalencies in hypothesis tests. The test of the H_0 that β_{YX} equals zero is also identical to the test of the H_0 that β_{XY} equals zero, although the estimated values of the regression slopes will clearly be different. These tests that β_{YX} or β_{XY} equal zero are also identical to the test of the H_0 that ρ_{YX} equals zero, i.e. the test of the OLS regression slope of Y on X is identical to the test of the OLS regression slope of X on Y and both are identical to the test of the Pearson correlation coefficient between Y and X, although neither estimated value of the slope will be the same as the estimated value of the correlation coefficient. The sample correlation coefficient is simply the geometric mean of these two regression slopes (Rodgers & Nicewander 1988):

$$r = \pm\sqrt{b_{YX}b_{XY}} \tag{5.27}$$

Simple correlation analysis is appropriate when we have bivariate data and we simply wish to measure the strength of the linear relationship (the correlation coefficient) between the two variables and test an H_0 about that correlation coefficient. Regression analysis is called for when we can biologically distinguish a response (Y) and a predictor variable (X) and we wish to describe the form of the model relating Y to X and use our estimates of the parameters of the model to predict Y from X.

5.5 Smoothing

The standard OLS regression analysis, and the robust regression techniques, we have described in this chapter specify a particular model that we fit to our data. Sometimes we know that a linear model is an inappropriate description of the relationship between Y and X because a scatterplot shows obvious nonlinearity or because we know theoretically that some other model should apply. Other times we simply have no preconceived model, linear or nonlinear, to fit to the data and we simply want to investigate the nature of the relationship between Y and X. In both situations, we require a method for fitting a curve to the relationship between Y and X that is not restricted to a specific model structure (such as linear). Smoothers are a broad class of techniques that describe the relationship between Y and X, etc., with few constraints on the form the relationship might take (Goodall 1990, Hastie & Tibshirani 1990). The aim of the usual linear model analysis is to separate the data into two components:

$$\text{model} + \text{residual (error)} \tag{5.28}$$

Smoothing also separates data into two components:

$$\text{smooth} + \text{rough} \tag{5.29}$$

where the rough component should have as little information or structure as possible (Goodall 1990). The logic of smoothing is relatively simple.

- Each observation is replaced by the mean or the median of surrounding observations or the predicted value from a regression model through these local observations.
- The surrounding observations are those within a window (sometimes termed a band or a neighbourhood) that covers a range of observations along the X-axis and the X-value on which the window is centered is termed the target. The size of the window, i.e. the number of observations it includes, is determined by a smoothing parameter for most smoothers (Hastie & Tibshirani 1990).
- Successive windows overlap so that the resulting line is smooth.
- The mean or median in one window are not affected by observations in other windows so smoothers are robust to extreme observations.
- Windows at the extremes of the X-axis often extend beyond the smallest or largest X-value and must be handled differently (see Section 5.5.5).

Smoothing functions are sometimes termed non-parametric regressions; here, non-parametric refers to the absence of a specified form of the relationship between Y and X rather than the distribution of the error terms from the fit of a model. Smoothing functions don't set any specific conditions for Y or X. For example, the observations may come from a joint distribution of Y and X (both Y and X random) or X may be considered fixed (Hastie & Tibshirani 1990). There are numerous varieties of smoothers and our descriptions are based on Goodall (1990) and Hastie & Tibshirani (1990).

5.5.1 Running means

A running (moving) means (averages) smoother is determined from the means of all the observations in a window. Each window is centered on the target X-value and the remaining X-values included in the window can be determined in two ways: (i) including a fixed number of observations both sides of the target X-value, or (ii) including a fixed number of nearest observations to the target x_i irrespective of which side of the target they occur (Hastie & Tibshirani 1990, Neter *et al.* 1996). The latter tend to perform better (Hastie & Tibshirani 1990), especially for locally weighted smoothers (see Cleveland's Loess below). Note that any observation might be included in a number of neighbouring windows. Using running medians instead of means makes the smoothing more resistant to extreme observations, i.e. more robust (Figure 5.21(a,b)). Running means or medians have been used commonly for analyzing data from simple time series (Diggle 1990), although the resulting line is rarely smooth (Hastie & Tibshirani 1990).

5.5.2 LO(W)ESS

A simple modification of running means or medians is to calculate the OLS regression line

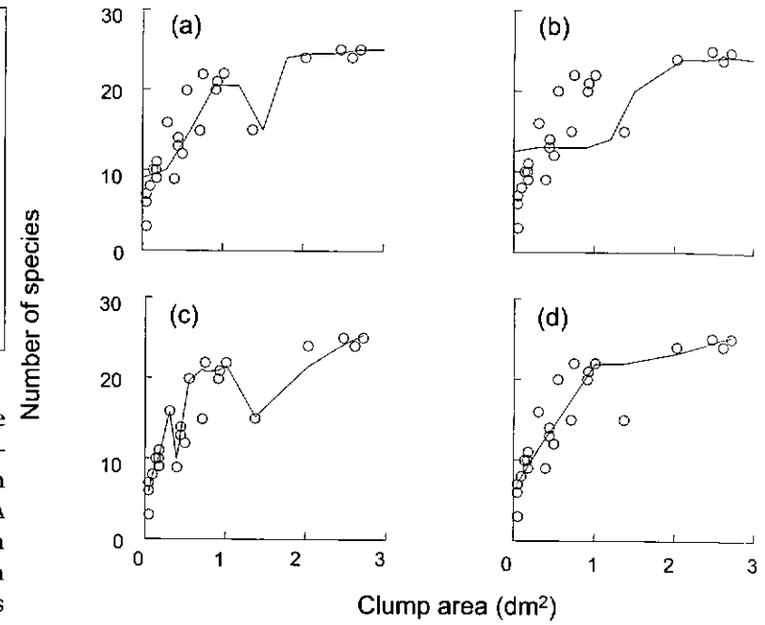

Figure 5.21 Smoothing functions through species–area data from Peake & Quinn (1993). (a) Running median smoother with smoothing parameter of 0.25, (b) running median smoother with smoothing parameter of 0.75, (c) Loess smoother with smoothing parameter of 0.25, and (d) Loess smoother with smoothing parameter of 0.75. Plotted in SYSTAT.

within a window and replace the observed y_i with that predicted by the local regression line for the target X-value. A modification of this approach is locally weighted regression scatterplot smoothing (Loess or Lowess; Cleveland 1979, 1994; see Figure 5.21). Here, the observations in a window are weighted differently depending on how far they are from the target X-value using a tri-cube weight function (see Hastie & Tibshirani 1990 and Trexler & Travis 1993 for details). In essence, observations further from the target X-value are downweighted compared with values close to the target X-value (Goodall 1990). Further refinement can be achieved by repeating the smoothing process a number of times during which observations with large residuals (difference between observed y_i and those predicted by the smooth) are downweighted. The final Loess smooth is often an excellent representation of the relationship between Y and X, although the choice of smoothing parameter (window size) can be important for interpretation (see Section 5.5.5). A related smoother is distance weighted least squares (DWLS) that also weights observations differently within each window. DWLS is slightly less sensitive to extreme observations than Loess for a given smoothing parameter.

5.5.3 Splines

Splines approach the smoothing problem by fitting polynomial regressions (see Chapter 6), usually cubic polynomials, in each window. The final smoother is termed a piecewise polynomial. The windows are separated at user-defined breakpoints termed knots and the polynomials within each window are forced to be continuous between windows, i.e. two adjacent polynomials join smoothly at a knot (Hastie & Tibshirani 1990). The computations are complex and a rationale for the choice of the number of knots, that will influence the shape of the smooth, is not obvious. Our experience is that regression splines are less useful than Loess smoothers as an exploratory tool for bivariate relationships.

5.5.4 Kernels

We have already discussed kernel functions as non-parametric estimators of univariate (Chapter 2) and bivariate (Section 5.1.3) probability density functions. Hastie & Tibshirabni (1990) also described a kernel smoother for Y versus X relationships. Within a window, observations are weighted based on a known function (e.g. normal distribution), termed the kernel, so that the weights decrease the further the observation is from the target X-value (just like in Loess smoothing). The estimated smoother results from the means of the Y-values within each window. Again, a smoothing parameter sets the size of the window and this, along with the kernel (the function that sets the weights of the observations within each window), are defined by the user. Kernels are not often used as smoothers for estimating the relationship between Y and X but are useful as more general univariate or bivariate density estimators.

5.5.5 Other issues

All the above smoothers describe the relationship between Y and X, and a predicted Y-value ($\hat{y}_i$) can be determined for each x_i. Therefore, residuals ($y_i - \hat{y}_i$) can also be calculated for each observation, and are produced from some statistical software. These residuals can be used in a diagnostic fashion to assess the fit of the smooth to the data, similar to the methods described in Section 5.3.10 for OLS linear regression. In particular, large residuals might indicate influential observations, although most smoothing techniques are considered robust to outliers because the components of the smoother are fitted to local observations within windows. Also, standard errors for $\hat{y}_i$ can be determined using bootstrap techniques (Efron & Tibshirani 1991; Chapter 2) and hypotheses about $\hat{y}_i$ tested with randomization procedures (Chapter 3).

There are several important issues related to the practical application of all the smoothers described here. First, whichever smoothing method is used, an important decision for the user is the value for the smoothing parameter, i.e. how many observations to include in each window. Hastie & Tibshirani (1990) have discussed this in some detail. Increasing the number of observations in each window (larger smoothing parameter) produces a flatter and "smoother" smooth that has less variability (Figure 5.21(a,c)) but is less likely to represent the real relationship between Y and X well (the smooth is probably biased). In contrast, fewer observations in each window (smaller smoothing parameter) produces a "jerkier", more variable, smooth (Figure 5.21(b,d)) but which may better match the pattern in the data (less biased). Hastie & Tibshirani (1990) have described complex, data-based methods for choosing the smoothing parameter (window-size) and producing a smooth that best minimizes both variance and bias. These methods might be useful if low variance is important because the smooth is being used as part of a modeling process, e.g. generalized additive modeling (GAM; see Chapter 13). Lower variance will result in predictions from such models being more precise. Trexler & Travis (1993) recommended the approach of Cleveland (1994) for Loess smoothing whereby the smoothing parameter (window-size) is as large as possible without resulting in any relationship between the residuals and X. In our experience, such a relationship is not common irrespective of the value of the smoothing parameter so this recommendation does not always work. Since smoothers are most commonly used as an exploratory tool rather than for model-fitting, we recommend trying different values of smoothing functions as part of the phase of exploring patterns in data before formal analyses.

A second issue is what we do when the endpoints (the smallest and largest X-values) are the targets, because their windows will usually exceed the range of the data. Goodall (1990) suggested a step-down rule so that the window size decreases as the largest and smallest X-values are approached, although he emphasized that definitive recommendations are not possible.

In summary, smoothing functions have a number of applications. First, they are very useful for graphically describing a relationship between two variables when we have no specific model in mind. Second, they can be used as a diagnostic check of the suitability of a linear model or help us decide which form of nonlinear model might be appropriate. Third, they can be used for modeling and prediction, particularly as part of generalized additive models (Chapter 13).

5.6 Power of tests in correlation and regression

Since H_0s about individual correlation and regression coefficients are tested with t tests, power calculations are relatively straightforward based on non-central t distributions (Neter *et al.* 1996; see also Chapters 3 and 7). In an *a priori* context, the question of interest is "How many observations do we need to be confident (at a specified level, i.e. power) that we will detect a regression slope of a certain size if it exists, given a preliminary estimate of σ_ε^2?" Equivalent questions can be phrased

for correlation coefficients. As always with power analyses, the difficult part is determining what effect size, e.g. size of regression slope, is important (see Chapter 7).

5.7 General issues and hints for analysis

5.7.1 General issues

- Estimating and testing correlations are straightforward for linear (straight-line) relationships. Use robust methods (e.g. non-parametric) if relationships are nonlinear but monotonic.
- Classical linear regression models fitted by OLS assume that X is a fixed variable (Model I). In biology, both Y and X are usually random (Model II) and alternative methods are available for estimating the slope. Even with X random, predictions and tests of hypotheses about the regression slope can be based on Model I analyses.
- The null hypothesis that the slope of the Model I regression equals zero can be tested with either a t test or an ANOVA F-ratio test. The conclusions will be identical and both are standard output from statistical software. These are also identical to the tests of the null hypotheses that the correlation coefficient equals zero and the slope of the RMA (Model II) regression equals zero.
- The standardized regression slope provides a measure of the slope of the linear relationship between the response and the predictor variable that is independent of their units.
- The assumptions of linear regression analysis (normality, homogeneity of variance, independence) apply to the error terms from the model and also to the response variable. Violations of these assumptions, especially homogeneity of variances and independence, can have important consequences for estimation and testing of the linear regression model.
- If transformations are ineffective or inapplicable, robust regression based on M-estimation or on ranks should be considered to deal with outliers and influential observations.
- Smoothing functions are very useful exploratory tools, suggesting the type of model that may be most appropriate for the data, and also for presentation, describing the relationship between two variables without being constrained by a specific model.

5.7.2 Hints for analysis

- Tests of null hypotheses for non-zero values of the correlation coefficient are tricky because of complex distribution of r; use procedures based on Fishers's z transformation.
- A scatterplot should always be the first step in any correlation or simple regression analysis. When used in conjunction with a smoothing function (e.g. Loess), scatterplots can reveal nonlinearity, unequal variances and outliers.
- As always when fitting linear models, use diagnostic plots to check assumptions and adequacy of model fit. For linear regression, plots of residuals against predicted values are valuable checks for homogeneity of residual variances. Checks for autocorrelation, especially if the predictor variable represents a time sequence, should also precede any formal analysis. Cook's D_i statistic (or $DFITS_i$) is a valuable measure of the influence each observation has on the fitted model.
- Transformations of either or both variables can greatly improve the fit of linear regression models to the data and reduce the influence of outliers. Try transforming the response variable to correct for non-normality and unequal variances and the predictor if variances are already roughly constant.
- Think carefully before using a no-intercept model. Forcing the model through the origin is rarely appropriate and renders measures of fit (e.g. r^2) difficult to interpret.

Chapter 6

Multiple and complex regression

In Chapter 5, we examined linear models with a single continuous predictor variable. In this chapter, we will discuss more complex models, including linear models with multiple predictor variables and models where one predictor interacts with itself in a polynomial term, and also nonlinear models. Note that this chapter will assume that you have read the previous chapter on bivariate relationships because many aspects of multiple regression are simply extensions from bivariate (simple) regression.

6.1 Multiple linear regression analysis

A common extension of simple linear regression is the case where we have recorded more than one predictor variable. When all the predictor variables are continuous, the models are referred to as multiple regression models. When all the predictor variables are categorical (grouping variables), then we are dealing with analysis of variance (ANOVA) models (Chapters 8–11). The distinction between regression and ANOVA models is not always helpful as general linear models can include both continuous and categorical predictors (Chapter 12). Nonetheless, the terminology is entrenched in the applied statistics, and the biological, literature. We will demonstrate multiple regression with two published examples.

Relative abundance of C_3 and C_4 plants

Paruelo & Lauenroth (1996) analyzed the geographic distribution and the effects of climate variables on the relative abundance of a number of plant functional types (PFTs) including shrubs, forbs, succulents (e.g. cacti), C_3 grasses and C_4 grasses. The latter PFTs represent grasses that utilize the C from the atmosphere differently in photosynthesis and are expected to have different responses to CO_2 and climate change. They used data from 73 sites across temperate central North America and calculated the relative abundance of each PFT, based on cover, biomass and primary production, at each site. These relative abundance measures for each PFT were the response variables. The predictor variables recorded for each site included longitude and latitude (centesimal degrees), mean annual temperature (°C), mean annual precipitation (mm), the proportion of precipitation falling in winter between December and February, the proportion of precipitation falling in summer between June and August, and a categorical variable representing biome (one for grassland, two for shrubland). The analyses of these data are in Box 6.1.

Abundance of birds in forest patches

Understanding which aspects of habitat and human activity affect the biodiversity and abundance of organisms within remnant patches of forest is an important aim of modern conservation biology. Loyn (1987) was interested in what characteristics of habitat were related to the abundance and diversity of forest birds. He selected 56 forest patches in southeastern Victoria, Australia, and recorded the number of species and abundance of forest birds in each patch as two response variables. The predictor variables recorded for

Box 6.1 Worked example of multiple linear regression: relative abundance of plant functional types

Paruelo & Lauenroth (1996) analyzed the geographic distribution and the effects of climate variables on the relative abundance of a number of plant functional types (PFTs) including shrubs, forbs, succulents (e.g. cacti), C_3 grasses and C_4 grasses. There were 73 sites across North America. The variables of interest are the relative abundance of C_3 plants, the latitude in centesimal degrees (LAT), the longitude in centesimal degrees (LONG), the mean annual precipitation in mm (MAP), the mean annual temperature in °C (MAT), the proportion of MAP that fell in June, July and August (JJAMAP) and the proportion of MAP that fell in December, January and February (DJFMAP). The relative abundance of C_3 plants was positively skewed and transformed to $\log_{10}+0.1$ ($\log_{10}C_3$).

A correlation matrix between the predictor variables indicated that some predictors are strongly correlated.

	LAT	LONG	MAP	MAT	JJAMAP	DJFMAP
LAT	1.00					
LONG	0.097	1.000				
MAP	−0.247	−0.734	1.000			
MAT	−0.839	−0.213	0.355	1.000		
JJAMAP	0.074	−0.492	0.112	−0.081	1.000	
DJFMAP	−0.065	0.771	−0.405	0.001	−0.792	1.00

Note the high correlations between LAT and MAT, LONG and MAP, and JJAMAP and DJFMAP, suggesting that collinearity may be a problem with this analysis.

With six predictor variables, a linear model with all possible interactions would have 64 model terms (plus an intercept) including four-, five- and six-way interactions that are extremely difficult to interpret. As a first pass, we fitted an additive model:

$$(\log_{10}C_3)_i = \beta_0 + \beta_1(\text{LAT})_i + \beta_2(\text{LONG})_i + \beta_3(\text{MAP})_i + \beta_4(\text{MAT})_i + \beta_5(\text{JJAMAP})_i + \beta_6(\text{DJFMAP})_i + \varepsilon_i$$

Coefficient	Estimate	Standard error	Standardized coefficient	Tolerance	t	P
Intercept	−2.689	1.239	0		−2.170	0.034
LAT	0.043	0.010	0.703	0.285	4.375	<0.001
LONG	0.007	0.010	0.136	0.190	0.690	0.942
MAP	<0.001	<0.001	0.181	0.357	1.261	0.212
MAT	−0.001	0.012	−0.012	0.267	−0.073	0.942
JJAMAP	−0.834	0.475	−0.268	0.316	−1.755	0.084
DJFMAP	−0.962	0.716	−0.275	0.175	−1.343	0.184

It is clear that collinearity is a problem with tolerances for two of the predictors (LONG & DJFMAP) approaching 0.1.

Paruelo & Lauenroth (1996) separated the predictors into two groups for their analyses. One group included LAT and LONG and the other included MAP, MAT, JJAMAP and DJFMAP. We will focus on the relationship between log-transformed relative abundance of C_3 plants and latitude and longitude. We fitted a multiplicative model including an interaction term that measured how the relationship between C_3 plants and latitude could vary with longitude and vice versa:

$$(\log_{10}C_3)_i = \beta_0 + \beta_1(\text{LAT})_i + \beta_2(\text{LONG})_i + \beta_3(\text{LAT}\times\text{LONG})_i + \varepsilon_i$$

Coefficient	Estimate	Standard error	Tolerance	t	P
Intercept	7.391	3.625		2.039	0.045
LAT	−0.191	0.091	0.003	−2.102	0.039
LONG	−0.093	0.035	0.015	−2.659	0.010
LAT × LONG	0.002	0.001	0.002	2.572	0.012

Note the very low tolerances indicating high correlations between the predictor variables and their interactions. An indication of the effect of collinearity is that if we omit the interaction and refit the model, the partial regression slope for latitude changes sign. We refitted the multiplicative model after centring both LAT and LONG.

Coefficient	Estimate	Standard error	Tolerance	t	P
Intercept	−0.553	0.027		20.130	<0.001
LAT	0.048	0.006	0.829	8.483	<0.001
LONG	−0.003	0.004	0.980	−0.597	0.552
LAT × LONG	0.002	0.001	0.820	2.572	0.012

Now the collinearity problem has disappeared. Diagnostic checks of the model did not reveal any outliers nor influential values. The boxplot of residuals was reasonably symmetrical and although there was some heterogeneity in spread of residuals when plotted against predicted values, and a 45° line representing sites with zero abundance of C_3 plants, this was not of a form that could be simply corrected (Figure 6.2).

The estimated partial regression slope for the interaction hasn't changed and we would reject the H_0 that there is no interactive effect of latitude and longitude on log-transformed relative abundance of C_3 plants. This interaction is evident in the DWLS smoother fitted to the scatterplot of relative abundance of C_3 plants against latitude and longitude (Figure 6.11). If further interpretation of this interaction is required, we would then calculate simple slopes for relative abundance of C_3 plants against latitude for specific values of longitude or vice versa. We will illustrate the simple slopes analysis with Loyn's (1987) data in Box 6.2.

Out of interest, we also ran the full model with all six predictors through both a forward and backward selection routine for stepwise multiple regression. For both methods, the significance level for entering and removing terms based on partial F statistics was set at 0.15.

The backward selection is as follows.

Coefficient	Estimate	Standard error	t	P
JJAMAP	−1.002	0.433	−2.314	0.024
DJFMAP	−1.005	0.486	−2.070	0.042
LAT	0.042	0.005	8.033	<0.001

The forward selection is as follows.

Coefficient	Estimate	Standard error	t	P
MAP	<0.001	<0.001	1.840	0.070
LAT	0.044	0.005	66.319	<0.001

Note the marked difference in the final model chosen by the two methods, with only latitude (LAT) in common.

each patch included area (ha), the number of years since the patch was isolated by clearing (years), the distance to the nearest patch (km), the distance to the nearest larger patch (km), an index of stock grazing history from 1 (light) to 5 (heavy), and mean altitude (m). The analyses of these data are in Box 6.2.

6.1.1 Multiple linear regression model

Consider a set of $i = 1$ to n observations where each observation was selected because of its specific X-values, i.e. the values of the p ($j = 2$ to p) predictor variables X_1, X_2,...X_j...X_p were fixed by the investigator, whereas the Y-value for each observation was sampled from a population of possible Y-values. Note that the predictor variables are usually random in most biological research and we will discuss the implications of this in Section 6.1.17. The multiple linear regression model that we usually fit to the data is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_j x_{ij} + \ldots + \beta_p x_{ip} + \varepsilon_i \quad (6.1)$$

The details of the linear regression model, including estimation of its parameters, are provided in Box 6.3.

For Loyn's (1987) data, p equals six and a linear model with all predictors would be:

$$(\text{bird abundance})_i = \beta_0 + \beta_1(\text{patch area})_i + \beta_2(\text{years isolated})_i + \beta_3(\text{nearest patch distance})_i + \beta_4(\text{nearest large patch distance})_i + \beta_5(\text{stock grazing})_i + \beta_6(\text{altitude})_i + \varepsilon_i \quad (6.2)$$

Using the data from Paruelo & Lauenroth (1996), we might fit a model where p equals two to represent geographic pattern of C_3 grasses:

$$(\text{relative abundance of } C_3 \text{ grasses})_i = \beta_0 + \beta_1(\text{latitude})_i + \beta_2(\text{longitude})_i + \varepsilon_i \quad (6.3)$$

A multiple regression model cannot be represented by a two-dimensional line as in simple regression and a multidimensional plane is needed (Figure 6.1). We can only graphically present such a model with two predictor variables although such graphs are rarely included in research publications.

Note that this is an additive model where all the explained variation in Y is due to the additive effects of the response variables. This model does not allow for interactions (multiplicative effects) between the predictor variables, although such interactions are possible (even likely) and will be discussed in Section 6.1.12.

We have the following in models 6.1 and 6.3.

y_i is the value of Y for the ith observation when the predictor variable X_1 equals x_{i1}, X_2 equals x_{i2}, X_j equals x_{ij}, etc.

β_0, β_1, β_2, β_j etc. are population parameters, also termed regression coefficients, where

β_0 is the population intercept, e.g. the true mean value of the relative abundance of C_3 grasses when latitude and longitude equal zero.

β_1 is the population slope for Y on X_1 holding X_2, X_3, etc., constant. It measures the change in relative abundance of C_3 grasses for a one

Box 6.2 Worked example of multiple linear regression: abundance of birds in forest patches

Loyn (1987) selected 56 forest patches in southeastern Victoria, Australia, and related the abundance of forest birds in each patch to six predictor variables: patch area (ha), distance to nearest patch (km), distance to nearest larger patch (km), grazing stock (1 to 5 indicating light to heavy), altitude (m) and years since isolation (years). Three of the predictor variables (patch area, distance to nearest patch or dist, distance to nearest larger patch or ldist) were highly skewed, producing observations with high leverage, so these variables were transformed to $\log_{10}$. A correlation matrix indicated some moderate correlations between predictors, especially between $\log_{10}$ dist and $\log_{10}$ ldist, $\log_{10}$ area and graze, and graze and years.

	$\log_{10}$ dist	$\log_{10}$ ldist	$\log_{10}$ area	Grazing	Altitude	Years
$\log_{10}$ dist	1.000					
$\log_{10}$ ldist	0.604	1.000				
$\log_{10}$ area	0.302	0.382	1.000			
Grazing	−0.143	−0.034	−0.559	1.000		
Altitude	−0.219	−0.274	0.275	−0.407	1.000	
Years	−0.020	0.161	−0.278	0.636	−0.233	1.000

As for the data set from Paruelo & Lauenroth (1996), a multiple linear regression model relating abundance of forest birds to all six predictor variables and their interactions would have 64 terms plus an intercept, and would be unwieldy to interpret. So an additive model was fitted:

$$(\text{bird abundance})_i = \beta_0 + \beta_1(\log_{10}\text{ area})_i + \beta_2(\log_{10}\text{ dist})_i + \beta_3(\log_{10}\text{ ldist})_i + \beta_4(\text{grazing})_i + \beta_5(\text{altitude})_i + \beta_6(\text{years})_i + \varepsilon_i$$

	Estimate	Standard error	Standardized coefficient	Tolerance	t	P
Intercept	20.789	8.285	0		2.509	0.015
$\log_{10}$ area	7.470	1.465	0.565	0.523	5.099	<0.001
$\log_{10}$ dist	−0.907	2.676	−0.035	0.604	−0.339	0.736
$\log_{10}$ ldist	−0.648	2.123	−0.035	0.498	−0.305	0.761
Grazing	−1.668	0.930	−0.229	0.396	−1.793	0.079
Altitude	0.020	0.024	0.079	0.681	0.814	0.419
Years	−0.074	0.045	−0.176	0.554	−1.634	0.109

Diagnostic checks of the model did not reveal any outliers or influential values. The response variable (bird abundance) was not skewed, the boxplot of residuals was reasonably symmetrical and although there was some heterogeneity of spread of residuals when plotted against predicted values, this was not of a form that could be simply corrected (Figure 6.3). The r^2 was 0.685, indicating that about 69% of the variation in bird abundance can be explained by this combination of predictors. Note that none of the tolerances were very low suggesting that despite some correlations among the predictors, collinearity may not be a serious issue for this data set. There was a significant positive partial regression slope for bird abundance against $\log_{10}$ area. No other partial regression slopes were significant.

Source	df	MS	F	P
Regression	6	723.513	17.754	<0.001
Residual	49	40.752		

The H_0 that all partial regression slopes equal zero was also rejected.

Now we will fit a second model to investigate possible interactions between predictor variables. A model with six predictors plus interactions is unwieldy so we will simplify the model first by omitting those predictors that contributed little to the original model ($\log_{10}$ dist, $\log_{10}$ ldist, altitude). The first two were correlated with each other and with $\log_{10}$ area anyway. Refitting the additive model with these three predictors omitted changed the estimated regression slopes of the remaining terms only slightly, suggesting that any bias in the estimates of the remaining predictors from omitting other predictors is small. This leaves us with a model with three predictors and their interactions:

$$(\text{bird abundance})_i = \beta_0 + \beta_1(\log_{10}\text{area})_i + \beta_2(\text{grazing})_i + \beta_3(\text{years})_i + \beta_4(\log_{10}\text{area} \times \text{grazing})_i + \beta_5(\log_{10}\text{area} \times \text{years})_i + \beta_6(\text{grazing} \times \text{years})_i + \beta_7(\log_{10}\text{area} \times \text{grazing} \times \text{years})_i + \varepsilon_i$$

Tolerance values were unacceptably low (all <0.10) unless the predictor variables were centered so the model was based on centered predictors.

	Estimate	Standard error	Standardized coefficient	Tolerance	t	P
Intercept	22.750	1.152	0		19.755	<0.001
Log_{10} area	8.128	1.540	0.615	0.373	5.277	<0.001
Grazing	−2.979	0.837	−0.408	0.386	−3.560	0.001
Years	0.032	0.057	0.076	0.280	0.565	0.574
Log_{10} area × Grazing	2.926	0.932	0.333	0.450	3.141	0.003
Log_{10} area × Years	−0.173	0.063	−0.305	0.411	−2.748	0.008
Grazing × Years	−0.101	0.035	−0.343	0.362	−2.901	0.006
Log_{10} area × Grazing × Years	−0.011	0.034	−0.037	0.397	−0.329	0.743

The three-way interaction was not significant so we will focus on the two-way interactions. The $\log_{10}$ area × grazing term indicates how much the effect of grazing on bird density depends on $\log_{10}$ area. This interaction is significant, so we might want to look at simple effects of grazing on bird density for different values of $\log_{10}$ area. We chose mean $\log_{10}$ area (0.932) ± one standard deviation (0.120, 1.744). Because the three-way interaction was not significant, we simply set years since isolation to its mean value (33.25). We could also just have ignored years since isolation and calculated simple slopes as for a two predictor model and got similar patterns. The simple slopes of bird abundance against grazing for different $\log_{10}$ area values and mean of years since isolation were as follows.

Log_{10} area	Simple slopes	Standard error	Standardized slope	t	P
0.120	−5.355	1.223	−0.734	−4.377	<0.001
0.932	−2.979	0.837	−0.408	−3.560	0.001
1.744	−0.603	1.024	−0.083	−0.589	0.558

As we predicted, the negative effect of grazing on bird abundance is stronger in small fragments and there is no relationship between bird abundance and grazing in the largest fragments.

Box 6.3 The multiple linear regression model and its parameters

Consider a set of $i = 1$ to n observations where each observation was selected because of its specific X-values, i.e. the values of the p ($j = 2$ to p) predictor variables $X_1, X_2, \ldots X_j \ldots X_p$ were fixed by the investigator, whereas the Y-value for each observation was sampled from a population of possible Y-values. The multiple linear regression model that we usually fit to the data is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_j x_{ij} + \ldots + \beta_p x_{ip} + \varepsilon_i \quad (6.1)$$

In model 6.1 we have the following.

y_i is the value of Y for the ith observation when the predictor variable X_1 equals x_{i1}, X_2 equals x_{i2}, X_j equals x_{ij}, etc.

β_0 is the population intercept, the true mean value of Y when X_1 equals zero, X_2 equals zero, X_j equals zero, etc.

β_1 is the partial population regression slope for Y on X_1 holding X_2, X_3, etc., constant. It measures the change in Y per unit change in X_1 holding the value of all other X-variables constant.

β_2 is the partial population regression slope for Y on X_2 holding X_1, X_3, etc., constant. It measures the change in Y per unit change in X_2 holding the value of all other X-variables constant.

β_j is the partial population regression slope for Y on X_j holding X_1, X_2, etc., constant; it measures the change in Y per unit change in X_j holding the value of the other $p - 1$ X-variables constant.

ε_i is random or unexplained error associated with the ith observation. Each ε_i measures the difference between each observed y_i and the mean of y_i; the latter is the value of y_i predicted by the population regression model, which we never know. We assume that when the predictor variable X_1 equals x_{i1}, X_2 equals x_{i2}, X_j equals x_{ij}, etc., these error terms are normally distributed, their mean is zero ($E(\varepsilon_i)$ equals zero) and their variance is the same and is designated σ^2_ε. This is the assumption of homogeneity of variances. We also assume that these ε_i terms are independent of, and therefore uncorrelated with, each other. These assumptions (normality, homogeneity of variances and independence) also apply to the response variable Y when the predictor variable X_1 equals x_{i1}, X_2 equals x_{i2}, X_j equals x_{ij}, etc.

Fitting the multiple regression model to our data and obtaining estimates of the model parameters is an extension of the methods used for simple linear regression, although the computations are complex. We need to estimate the parameters (β_0, $\beta_1, \beta_2, \ldots, \beta_p$ and σ_ε^2) of the multiple linear regression model based on our random sample of n ($x_{i1}, x_{i2}, \ldots, x_{ip}, y_i$) observations. Once we have estimates of the parameters, we can determine the sample regression line:

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_j x_{ij} + \ldots + b_p x_{ip}$$

where:

$\hat{y}_i$ is the value of y_i for $x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{ip}$ predicted by the fitted regression line,
b_0 is the sample estimate of β_0, the *Y*-intercept,
$b_1, b_2, \ldots, b_j, \ldots b_p$ are the sample estimates of $\beta_1, \beta_2, \ldots, \beta_j, \ldots, \beta_p$, the partial regression slopes.

We can estimate these parameters using either (ordinary) least squares (OLS) or maximum likelihood (ML). If we assume normality, the OLS estimates of β_0, β_1, etc., are the same as the ML estimates. As with simple regression, we will focus on OLS estimation. The actual calculations for the OLS estimates of the model parameters involve solving a set of simultaneous normal equations, one for each parameter in the model, and are best represented with matrix algebra (Box 6.4).

The OLS estimates of $\beta_0, \beta_1, \beta_2$, etc., are the values that produce a sample regression line ($\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_j x_{ij} + \ldots + b_p x_{ip}$) that minimizes $\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2$. These are the sum of the squared deviations (SS) between each observed y_i and the value of y_i predicted by the sample regression line for each x_{ij}. Each $(y_i - \hat{y}_i)$ is a residual from the fitted regression plane and represents the vertical distance between the regression plane and the *Y*-value for each observation (Figure 6.1). The OLS estimate of σ_ε^2 (the variance of the model error terms) is the sample variance of these residuals and is the Residual (or Error) Mean Square from the analysis of variance (Section 6.1.3).

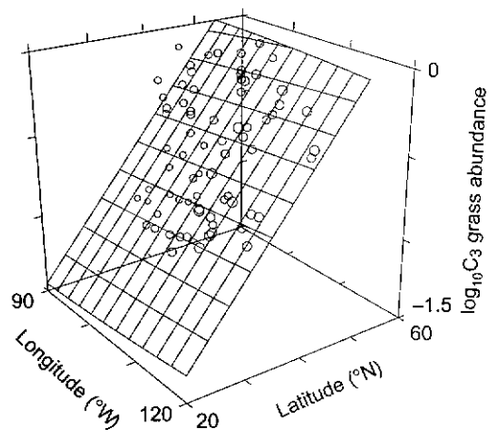

Figure 6.1 Scatterplot of the log-transformed relative abundance of C_3 plants against longitude and latitude for 73 sites from Paruelo & Lauenroth (1996) showing OLS fitted multiple regression linear response surface.

centesimal degree change in latitude, holding longitude constant.

β_2 is the population slope for *Y* on X_2 holding X_1, X_3, etc., constant. It measures the change in relative abundance of C_3 grasses for a one centesimal degree change in longitude, holding latitude constant.

β_j is the population slope for *Y* on X_j holding X_1, X_2, etc., constant; it measures the change in *Y* per unit change in X_j holding the value of the other $p-1$ *X*-variables constant.

ε_i is random or unexplained error associated with the *i*th observation of relative abundance of C_3 grasses not explained by the model.

The slope parameters (β_1, β_2,..., β_j,..., β_p) are termed partial regression slopes (coefficients) because they measure the change in *Y* per unit

change in a particular *X* holding the other $p-1$ *X*-variables constant. It is important to distinguish these partial regression slopes in multiple linear regression from the regression slope in simple linear regression. If we fit a simple regression model between *Y* and just one of the *X*-variables, then that slope is the change in *Y* per unit change in *X*, ignoring the other $p-1$ predictor variables we might have recorded plus any predictor variables we didn't measure. Again using the data from Paruelo & Lauenroth (1996), the partial regression slope of the relative abundance of C_3 grasses against longitude measures the change in relative abundance for a one unit (one centesimal degree) change in longitude, holding latitude constant. If we fitted a simple linear regression model for relative abundance of C_3 grasses against longitude, we completely ignore latitude and any other predictors we didn't record in the interpretation of the slope. Multiple regression models enable us to assess the relationship between the response variable and each of the predictors, adjusting for the remaining predictors.

6.1.2 Estimating model parameters

We estimate the parameters (β_0, β_1, β_2,..., β_p and σ_ε^2) of the multiple linear regression model, based on our random sample of n ($x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{ip}, y_i$) observations, using OLS methods (Box 6.3). The fitted regression line is:

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_j x_{ij} + \ldots + b_p x_{ip} \qquad (6.4)$$

where:

$\hat{y}_i$ is the value of relative abundance of C_3 grasses for $x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{ip}$ (e.g. a given combination of latitude and longitude) predicted by the fitted regression model,
b_0 is the sample estimate of β_0, the *Y*-intercept,
$b_1, b_2, \ldots, b_j, \ldots b_p$ are the sample estimates of $\beta_1, \beta_2, \ldots, \beta_j, \ldots, \beta_p$, the partial regression slopes. We can also determine standardized partial regression slopes that are independent of the units in which the variables are measured (Section 6.1.6).

The OLS estimates of these parameters are the values that minimize the sum of squared deviations (SS) between each observed value of relative abundance of C_3 grasses and the relative abundance of C_3 grasses predicted by the fitted regression model. This difference between each observed y_i and each predicted $\hat{y}_i$ is called a residual (e_i). We will use the residuals for checking the fit of the model to our data in Section 6.1.8.

The actual calculations for the OLS estimates of the model parameters involve solving a set of simultaneous normal equations (see Section 5.2.3), one for each parameter in the model, and are best represented with matrix algebra (Box 6.4). The computations are tedious but the estimates, and their standard errors, should be standard output from multiple linear regression routines in your statistical software. Confidence intervals for the parameters can also be calculated using the *t* distribution with $n-p$ df. New *Y*-values can be predicted from new values of any or all of the p *X*-variables by substituting the new *X*-values into the regression equation and calculating the predicted *Y*-value. As with simple regression, be careful about predicting from values of any of the *X*-variables outside the range of your data. Standard errors and prediction intervals for new *Y*-values can be determined (see Neter *et al.* 1996). Note that the confidence intervals for model parameters (slopes and intercept) and prediction intervals for new *Y*-values from new *X*-values depend on the number of observations and the number of predictors. This is because the divisor for the $MS_{Residual}$, and the df for the *t* distribution used for confidence intervals, is $n-(p+1)$. Therefore, for a given standard error, our confidence in predicted *Y*-values from our fitted model is reduced when we include more predictors.

6.1.3 Analysis of variance

Similar to simple linear regression models described in Chapter 5, we can partition the total variation in *Y* (SS_{Total}) into two additive components (Table 6.1). The first is the variation in *Y* explained by its linear relationship with $X_1, X_2, \ldots, X_p$, termed $SS_{Regression}$. The second is the variation in *Y* not explained by the linear relationship with $X_1, X_2, \ldots, X_p$, termed $SS_{Residual}$ and which is measured as the difference between each observed y_i and the *Y*-value predicted by the regression model ($\hat{y}_i$). These SS in Table 6.1 are identical to those in Table 5.1 for simple regression models. In fact, the

Box 6.4 Matrix algebra approach to OLS estimation of multiple linear regression models and determination of leverage values

Consider an additive linear model with one response variable (Y) and p predictor variables ($X_1, X_2 \ldots X_p$) and a sample of n observations. The linear model will have $p+1$ parameters, a slope term for each X-variable and an intercept. Let $\mathbf{Y}$ be a vector of observed Y-values with n rows, $\hat{\mathbf{Y}}$ be a vector of predicted Y-values with n rows and $\mathbf{X}$ be an $n \times (p+1)$ matrix of the values of the X-variables (one X-variable per column) plus a column for the intercept. The linear model can be written as:

$$\mathbf{Y} = \beta\mathbf{X} + \varepsilon$$

where β is a vector of model parameters ($\beta_0, \beta_1, \ldots, \beta_p$) with $p+1$ rows and ε is a vector of error terms with n rows. The OLS estimate of β can be found by solving the normal equations:

$$\mathbf{X'Xb} = \mathbf{X'Y}$$

The OLS estimate of β then is:

$$\mathbf{b} = (\mathbf{X'X})^{-1}(\mathbf{X'Y})$$

where $\mathbf{b}$ is a vector of sample partial regression coefficients ($b_0, b_1, \ldots, b_p$) with $p+1$ rows. Note that $(\mathbf{X'X})^{-1}$ is the inverse of $(\mathbf{X'X})$ and is critical to the solution of the normal equations and hence the OLS estimates of the parameters. The calculation of this inverse is very sensitive to rounding errors, especially when there are many parameters, and also to correlations (linear dependencies – see Rawlings *et al.* 1998) among the X-variables, i.e. collinearity. Such correlations exaggerate the rounding errors problem and make estimates of the parameters unstable and their variances large (see Box 6.5).

The matrix containing the variances of, and the covariances between, the sample partial regression coefficients ($b_0, b_1, \ldots, b_p$) is:

$$\mathbf{s_b}^2 = \mathrm{MS}_{\mathrm{Residual}}(\mathbf{X'X})^{-1}$$

From the variances of the sample partial regression coefficients, we can calculate standard errors for each partial regression coefficient.

We can also create a matrix $\mathbf{H}$ whereby:

$$\mathbf{H} = \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}$$

$\mathbf{H}$ is an $n \times n$ matrix, usually termed the hat matrix, whose n diagonal elements are leverage values (h_{ii}) for each observation (Neter *et al.* 1996). These leverage values measure how far an observation is from the means of the X-variables. We can then relate $\mathbf{Y}$ to $\hat{\mathbf{Y}}$ by:

$$\hat{\mathbf{Y}} = \mathbf{HY}$$

So the hat matrix transforms observed Y into predicted Y (Bollen & Jackman 1990).

Table 6.1 Analysis of variance table for a multiple linear regression model with an intercept, p predictor variables and n observations

Source of variation	SS	df	MS
Regression	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	p	$\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{p}$
Residual	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$n-p-1$	$\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-p-1}$
Total	$\sum_{i=1}^{n}(y_i - \bar{y})^2$	$n-1$	

partitioning of the $\mathrm{SS}_{\mathrm{Total}}$ for the simple linear regression model is just a special case of the multiple regression model where p equals one, although the calculation of the SS for multiple regression models is more complex. These SS can be converted into variances (mean squares) by dividing by the appropriate degrees of freedom. For example, using the data from Paruelo & Lauenroth (1996) and the regression model 6.3, the $\mathrm{SS}_{\mathrm{Total}}$ in relative abundance of C_3 grasses across the 73 sites is partitioned into the SS explained by the linear regression on latitude and longitude and that unexplained by this regression.

The expected values of these two mean squares are again just an extension of those we described for simple regression (Table 6.2). The expected value for $\mathrm{MS}_{\mathrm{Residual}}$ is σ_ε^2, the variance of the error terms (ε_i), and of y_i, which are assumed to be constant across each combination of $x_{i1}, x_{i2}, \ldots, x_{ij}$, etc. The expected value for $\mathrm{MS}_{\mathrm{Regression}}$ is more complex (Neter *et al.* 1996) but importantly it includes the square of each regression slope plus σ_ε^2.

6.1.4 Null hypotheses and model comparisons

The basic null hypothesis we can test when we fit a multiple linear regression model is that all the partial regression slopes equal zero, i.e. H_0: $\beta_1 = \beta_2 = \ldots = \beta_j = \ldots = 0$. For example, Paruelo & Lauenroth (1996) might have tested the H_0 that the partial regression slopes for abundance of C_3 plants on latitude and longitude both equal zero. We test this H_0 with the ANOVA partitioning of the total variation in Y into its two components, that explained by the linear regression with X_1, X_2, etc., and the residual variation. If the H_0 is true, then $\mathrm{MS}_{\mathrm{Regression}}$ and $\mathrm{MS}_{\mathrm{Residual}}$ both estimate σ_ε^2 and their F-ratio should be one. If the H_0 is false, then at least one of the partial regression slopes does not equal zero and $\mathrm{MS}_{\mathrm{Regression}}$ estimates σ_ε^2 plus a positive term representing the partial regression slopes, so the F-ratio of $\mathrm{MS}_{\mathrm{Regression}}$ to $\mathrm{MS}_{\mathrm{Residual}}$ should be greater than one. So we can test this H_0 by comparing the F-ratio statistic to the appropriate F distribution, just as we did with simple linear regression in Chapter 5.

Irrespective of the outcome of this test, we would also be interested in testing null hypotheses about each partial regression coefficient, i.e. the H_0 that any β_j equals zero. We can use the process of comparing the fit of full and reduced models that we introduced in Chapter 5 to test these null hypotheses. Imagine we have a model with three predictor variables (X_1, X_2, X_3). The full model is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i \qquad (6.5)$$

Using the data from Loyn (1987), we might model the abundance of forest birds against patch area, years since isolation and grazing intensity:

$$(\text{bird abundance})_i = \beta_0 + \beta_1(\text{patch area})_i + \beta_2(\text{years isolated})_i + \beta_3(\text{stock grazing})_i + \varepsilon_i \qquad (6.6)$$

To test the H_0 that the partial regression slope for bird abundance against patch area holding years since isolation and grazing intensity constant (i.e. β_1) equals zero, we compare the fit of models 6.5 and 6.6 to the reduced models:

$$y_i = \beta_0 + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i \qquad (6.7)$$

$$(\text{bird abundance})_i = \beta_0 + \beta_2(\text{years isolated})_i + \beta_3(\text{stock grazing})_i + \varepsilon_i \qquad (6.8)$$

Models 6.7 and 6.8 assume the H_0 (β_1 equals zero) is true. If the explained variance ($\mathrm{SS}_{\mathrm{Regression}}$) of models 6.6 and 6.8 is not different, then there is no evidence to reject H_0; if there is an increase in explained variation for the full model compared

to the reduced model, we have evidence suggesting the H_0 is false. We calculate the extra SS explained by including β_1 in the model:

$$SS_{Extra} = \text{Full } SS_{Regression} - \text{Reduced } SS_{Regression} \quad (6.9)$$

This SS_{Extra} is sometimes expressed as $SS_{Regression}(X_1|X_2,X_3)$, the increase in $SS_{Regression}$ when X_1 is added to a model already including X_2 and X_3, e.g. $SS_{Regression}$(patch area | years isolated, grazing stock). This is identical to measuring the drop in unexplained variation by omitting β_1 from the model:

$$SS_{Drop} = \text{Reduced } SS_{Residual} - \text{Full } SS_{Residual} \quad (6.10)$$

also expressed as $SS_{Residual}(X_1|X_2,X_3)$, the decrease in $SS_{Residual}$ when X_1 is added to a model already including X_2 and X_3. We convert the SS_{Extra} or SS_{Drop} into a MS by dividing by the df. There is one df in this case because we are testing a single regression parameter. In general, the df is the number of predictor variables in the full model minus the number of predictor variables in the reduced model. We can then use an F test, now termed a partial F test, to test the H_0 that a single partial regression slope equals zero:

$$F_{1,n-p} = \frac{MS_{Extra}}{\text{Full } MS_{Residual}} \quad (6.11)$$

For any predictor variable X_j, we can also test the H_0 that β_j equals zero with a t statistic with $(n-(p+1))$ df:

$$t = \frac{b_j}{s_{b_j}} \quad (6.12)$$

where s_{b_j} is the standard error of b_j (see Box 6.4). These t tests are standard multiple regression output from statistical software. Note that the F and t tests for a given H_0 are equivalent and F equals t^2. We prefer the F tests, however, because the model fitting procedure (comparing full and reduced models) can be used to test any subset of regression coefficients, not just a single coefficient. For example, we could calculate the $SS_{Regression}(X_2,X_3|X_1)$ to test the H_0 that β_2 equals β_3 equals zero. We just need to fit a full and a reduced (H_0 is true) model. In general, the full model will contain all the predictor variables and the reduced model omits those predictors that are specified in H_0 to be zero. In Section 6.1.15, we will see that it is also possible to test partial regression coefficients in a sequential fashion, omitting those terms found to be not significantly different from zero from the model.

The H_0 that β_0 (population intercept) equals zero can also be tested, either with a t test or with an F test by comparing a full model with an intercept to a reduced model without. The test of zero intercept is usually of much less interest because it is testing a parameter using an estimate that is usually outside the range of our data (see Chapter 5).

6.1.5 Variance explained

The multiple r^2 is the proportion of the total variation in Y explained by the regression model:

$$r^2 = \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{SS_{Residual}}{SS_{Total}} = 1 - \frac{\text{Full } SS_{Residual}}{\text{Reduced } SS_{Residual}} \quad (6.13)$$

Here the reduced model is one with just an intercept and no predictor variables (i.e. $\beta_1 = \beta_2 = \ldots = \beta_j = \ldots = 0$). Interpretation of r^2 in multiple linear regression must be done carefully. Just like in simple regression, r^2 is not directly comparable between models based on different transformations (Anderson-Sprecher 1994; Chapter 5). Additionally, r^2 is not a useful measure of fit when comparing models with different numbers of, or combinations of, predictor variables (e.g. interaction terms, see Section 6.1.12). As more predictors are added to a model, r^2 cannot decrease so that models with more predictors will always appear to fit the data better. Comparing the fit of models with different numbers of predictors should use alternative measures (see Section 6.1.15).

6.1.6 Which predictors are important?

Once we have fitted our multiple linear regression model, we usually want to determine the relative importance of each predictor variable to the response variable. There are a number of related approaches for measuring relative importance of each predictor variable in multiple linear regression models.

Tests on partial regression slopes

The simplest way of assessing the relative importance of the predictors in a linear regression model is to use the F or t statistics, and their associated P values, from the tests of the null hypotheses that each β_j equals zero. These tests are straightforward to interpret but only tell us the probability of observing our sample observations or ones more extreme for these variables if the H_0 for a given predictor is true. Also, some statisticians (Neter *et al.* 1996, Rawlings *et al.* 1998) have argued that we are testing null hypotheses about a number of regression coefficients simultaneously from a single data set, so we should adjust the significance level for each test to limit the overall probability of at least one Type I error among all our tests to α. Such an adjustment will reduce the power of individual tests, and as we discussed in Chapter 3, seems unnecessarily harsh. If you deem such an adjustment necessary, however, one of the sequential Bonferroni procedures is appropriate.

Change in explained variation

The change in variation explained by the model with all predictors and the model with a specific predictor omitted is also a measure of importance of that predictor. This is basically comparing the fit of two models to the data; because the number of predictors differs between the two models, the choice of measure of fit is critical and will be discussed further when we consider model selection in Section 6.1.15. To measure the proportional reduction in the variation in Y when a predictor variable X_j is added to a model already including the other predictors (X_1 to X_p except X_j) is simply:

$$r_{X_j}^2 = \frac{SS_{Extra}}{\text{Reduced } SS_{Residual}} \quad (6.14)$$

where SS_{Extra} is the increase in $SS_{Regression}$, or the decrease in $SS_{Residual}$, when X_j is added to the model and Reduced $SS_{Residual}$ is unexplained SS from the model including all predictor variables except X_j. This $r_{X_j}^2$ is termed the coefficient of partial determination for X_j and its square root is the partial correlation coefficient between Y and X_j holding the other predictor variables constant (i.e. already including them in the model).

A related approach is hierarchical partitioning (Chevan & Sutherland 1991, Mac Nally 1996), which quantifies the independent correlation of each predictor variable with the response variable. It works by partitioning any measure of explained variance (e.g. r^2) into components measuring the independent contribution of each predictor. It is an important tool for multivariate inference, especially in multiple regression models, and we will describe it in more detail in Section 6.1.16.

Standardized partial regression slopes

The sizes of the individual regression slopes are difficult to compare if the predictor variables are measured in different units (see Chapter 5). We can calculate standardized regression slopes by regressing the standardized response variable against the standardized predictor variables, or alternatively, calculate for predictor X_j:

$$b_j^* = b_j \frac{s_{X_j}}{s_Y} \quad (6.15)$$

These standardized regression slopes are comparable independently of the scales on which the predictors are measured. Note that the regression model based on standardized variables doesn't include an intercept, because its OLS (and ML) estimate will always be zero. Note also that if the predictor variables are not correlated with each other, then the standardized regression slopes relating Y to each X_j are the same as the correlation coefficients relating Y to X_j.

For model 6.3, standardized regression slopes would not assist interpretation because both predictors (latitude and longitude) are in the same units (centesimal degrees). However, if we included mean annual temperature (°C) and mean annual precipitation (mm) in the model, then the magnitudes of the unstandardized regression slopes would not be comparable because of the different units, so standardization would help.

Bring (1994) suggested that the size of each standardized slope should relate to the reduction in explained variation when each predictor is omitted from the full model (see Equation 6.14). He argued that standardization should be based on partial standard deviations rather than ordinary standard deviations, so that the size of the b_j^* relates to the reduction in r^2 when that X_j is omitted from the model. The partial standard deviation of predictor variable j (X_j) is:

$$s_{X_j}^* = \frac{s_{X_j}}{\sqrt{VIF_j}} \sqrt{\frac{n-1}{n-p}} \quad (6.16)$$

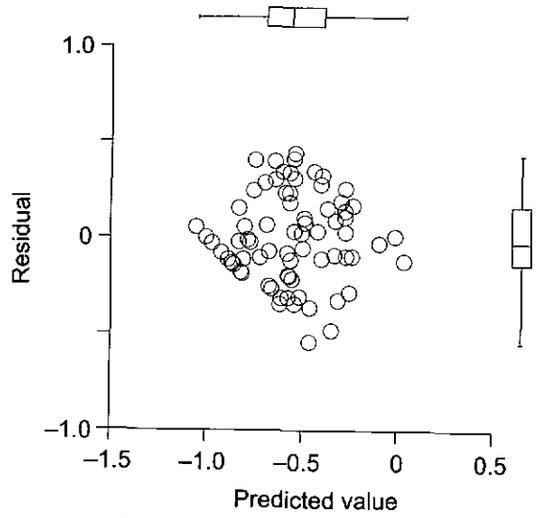

Figure 6.2 Plot of residuals against predicted values (with boxplots) from fitting the multiplicative model $(\log_{10}C_3)_i = \beta_0 + \beta_1(\text{LAT})_i + \beta_2(\text{LONG})_i + \beta_3(\text{LAT} \times \text{LONG})_i + \varepsilon_i$ to data with centered predictors from Paruelo & Lauenroth (1996).

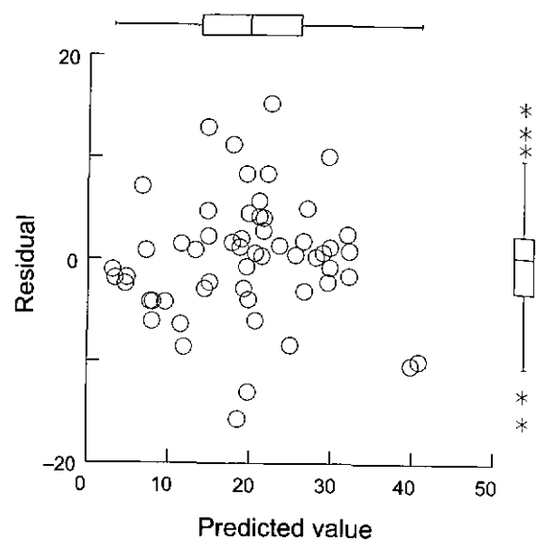

Figure 6.3 Plot of residuals against predicted values (with boxplots) from multiple linear regression of bird abundance in forest patches against patch area, distance to nearest patch, distance to nearest larger patch (these three variables $\log_{10}$ transformed), grazing intensity, altitude, and years since isolation for the 56 patches surveyed by Loyn (1987).

VIF is the variance inflation factor and will be defined in Section 6.1.11 when we examine the problem of multicollinearity. This partial standard deviation can then be incorporated in the formula for the standardized regression slope (Equation 6.15).

Regressions on standardized variables will produce coefficients (except for the intercept) that are the same as the standardized coefficients described above. The hypothesis tests on individual standardized coefficients will be identical to those on unstandardized coefficients. Standardization might be useful if the variables are on very different scales and the magnitude of coefficients for variables with small values may not indicate their relative importance in influencing the response variable. However, it is the predictor variables that are important here and standardizing the response variable may not be necessary and will make predicted values from the model more difficult to interpret. Regression models using standardized (or simply centered) predictors are very important for detecting and treating multicollinearity and interpreting interactions between predictors (Sections 6.1.11 and 6.1.12).

6.1.7 Assumptions of multiple regression

As with simple linear regression (Chapter 5), interval estimation and hypothesis tests of the parameters of the multiple linear regression model rely on a number of assumptions about the model error terms at each combination of $x_{i1}, x_{i2}, \ldots, x_{ip}$. We assume that the error terms, and therefore the Y-values, are normally distributed, they have constant variance and they are independent of each other. Checks of these assumptions are carried out as for simple linear regression (Chapter 5). Boxplots and probability plots of the residuals can be used to check for normality, plots of residuals against $\hat{y}_i$ can detect heterogeneity of variance (Section 6.1.9; Figure 6.2, Figure 6.3) and plots of residuals against each X_j can detect autocorrelation if X_j is a time sequence.

We also assume that each X is a fixed variable with the values x_{i1}, x_{i2}, etc., being constants that would not vary from sample to sample. This is unlikely in biological research with some or all of the predictors likely to be random variables and our observations actually coming from a multivariate distribution that we assume is normal.

Both of our examples illustrate this point: Paruelo & Lauenroth (1996) did not choose specific latitudes and longitudes for their sampling sites and Loyn (1987) did not choose forest patches with specifically chosen values of area, number of years since the patch was isolated by clearing, distance to the nearest patch, distance to the nearest larger patch, stock grazing history, or altitude. Our inferences are then conditional on the particular values of x_{i1}, x_{i2}, etc., that we have in our sample. Model II multiple regression when the predictor variables are random will be discussed in Section 6.1.17.

An additional assumption that affects multiple linear regression is that the predictor variables must be uncorrelated with each other. Violation of this assumption is called (multi)collinearity and is such an important issue for multiple regression that we will discuss it separately in Section 6.1.11.

Finally, the number of observations must exceed the number of predictor variables or else the matrix calculations (Box 6.4) will fail. Green (1991) proposed specific minimum ratios of observations to predictors, such as $p + 104$ observations for testing individual predictor variables, and these guidelines have become recommendations in some texts (*e.g.* Tabachnick & Fidell 1996). These numbers of observations are probably unrealistic for many biological and ecological research programs. Neter *et al.* (1996) are more lenient, recommending six to ten times the number of predictors for the number of observations. We can only suggest that researchers try to maximize the numbers of observations and if trade-offs in terms of time and cost are possible, reducing the numbers of variables to allow more observations is nearly always preferable to reducing the number of observations.

6.1.8 Regression diagnostics

Diagnostic checks of the assumptions underlying the fitting of linear models and estimating their parameters, and to warn of potential outliers and influential observations, are particularly important when there are multiple predictor variables. We are usually dealing with large data sets and scanning the raw data or simple bivariate scatterplots (see Section 6.1.9) that might have worked for simple regression models will rarely be adequate for checking the appropriateness of a multiple regression model. Fortunately, the same diagnostic checks we used for simple regression in Chapter 5 apply equally well for multiple regression. All are standard output from regression or linear model routines in good statistical software.

Leverage

Leverage measures how extreme each observation is from the means of all the X_j (the centroid of the p X-variables), so in contrast to simple regression, leverage in multiple regression takes into account all the predictors used in the model. Leverage values greater than $2(p/n)$ should be cause for concern, although such values would also be detected as influential by Cook's D_i.

Residuals

Residuals in multiple regression are interpreted in the same way as for simple regression, the difference between the observed and predicted Y-values for each observation $(y_i - \hat{y}_i)$. These residuals can be standardized and studentized (see Chapter 5) and large residuals indicate outliers from the fitted model that could be influential.

Influence

Measures of how influential each observation is on the fitted model include Cook's D_i and DFITS_i and these are as relevant for multiple regression as they were for simple regression (Chapter 5). Observations with a D_i greater than one are usually considered influential and such observations should be checked carefully.

6.1.9 Diagnostic graphics

As we emphasized for simple regression models, graphical techniques are often the most informative checks of assumptions and for the presence of outliers and influential values.

Scatterplots

Bivariate scatterplots between the X_js are important for detecting multicollinearity (see Section 6.1.11) and scatterplots between Y and each X_j, particularly in conjunction with smoothing functions, provide an indication of the nature of relationships being modeled. Scatterplot matrices

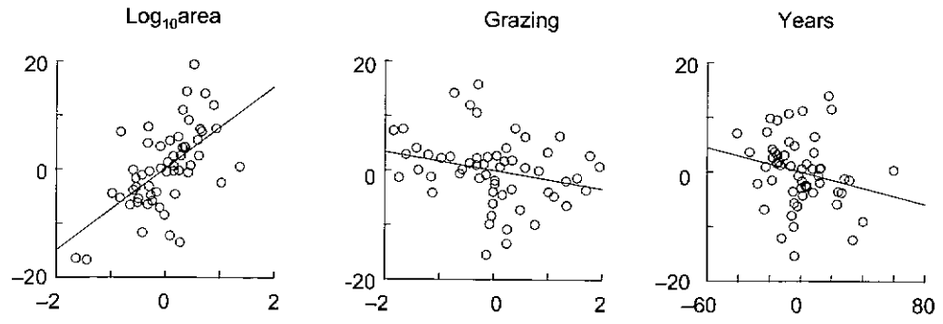

Figure 6.4 Partial regression plots for three of the predictors from a linear model relating bird abundance in forest patches to patch area, distance to nearest patch, distance to nearest larger patch (these three variables $\log_{10}$ transformed), grazing intensity, altitude, and years since isolation for the 56 patches surveyed by Loyn (1987). Vertical axis is residuals from OLS regression of bird abundance against all predictors except the one labelled, horizontal axis is residuals from OLS regression of labelled predictor against remaining predictors. See Section 6.1.9 for full analysis.

(SPLOMs; see Chapter 4) are the easiest way of displaying these bivariate relationships. However, scatterplots between Y and X_1, Y and X_2, etc., ignore the other predictor variables in the model and therefore do not represent the relationship we are modeling, i.e. the relationship between Y and X_j holding all other Xs constant.

A scatterplot that does show this relationship for each predictor variable is the added variable, or partial regression, plot, which is a plot between two sets of residuals. Let's say we are fitting a model of Y against p predictor variables and we want a scatterplot to show the relationship between Y and X_j, holding the other $p-1$ X-variables constant. The residuals for the vertical axis of the plot (e_{i1}) come from the OLS regression of Y against all p predictors except X_j. The residuals for the horizontal axis of the plot (e_{i2}) come from the OLS regression of X_j against all p predictors except X_j. This scatterplot of e_{i1} against e_{i2} shows the relationship between Y and X_j holding the other X-variables constant and will also show outliers that might influence the regression slope for X_j. If we fit an OLS regression of e_{i1} against e_{i2}, the fitted slope of this line is the partial regression slope of Y on X_j from the full regression model of Y on all p predictors.

Three partial regression plots are illustrated in Figure 6.4 from a model relating bird abundance in forest patches to patch area, distance to nearest patch, distance to nearest larger patch (these three variables $\log_{10}$ transformed), stock grazing, altitude, and years since isolation for the 56 patches surveyed by Loyn (1987). The partial regression plot for patch area (Figure 6.4, left) has the residuals from a model relating bird abundance to all predictors except patch area on the vertical axis and the residuals from a model relating patch area to the other predictors on the horizontal axis. Note the strong positive relationship for $\log_{10}$ area and the weak negative relationships for grazing and years since isolation. There was little pattern in the plots for the other three predictors. The slopes of the OLS regression lines fitted to these residual plots are the partial regression slopes from the multiple regression model relating bird abundance to these predictors.

Residual plots

There are numerous ways residuals from the fit of a multiple linear regression model can be plotted. A plot of residuals against $\hat{y}_i$, as we recommended for simple regression (Chapter 5), can detect heterogeneity of variance (wedge-shaped pattern) and outliers (Figure 6.2 and Figure 6.3). Plots of residuals against each X_j can detect outliers specific to that X_j, nonlinearity between Y and that X_j and can also detect autocorrelation if X_j is a time sequence. Finally, residuals can be plotted against predictors, or interactions between predictors, not included in the model to assess whether these predictors or their interactions might be important, even if they were deleted from the model based on other criteria (Neter *et al.* 1996).

6.1.10 Transformations

Our general comments on transformations from Chapter 4, and specifically for bivariate regression in Chapter 5, are just as relevant for multiple regression. Transformations of the response variable can remedy non-normality and heterogeneity of variance of error terms and transformations of one or more of the predictor variables might be necessary to deal with nonlinearity and influential observations due to high leverage. For example, the abundance of C_3 plants in the study by Paruelo & Lauenroth (1996) was transformed to logs to reduce strong skewness and three of the predictor variables in the study by Loyn (1987) were also log transformed to deal with observations with high leverage (Box 6.2). Transformations can also reduce the influence of interactions between predictors on the response variable, i.e. make an additive model a more appropriate fit than a multiplicative model (see Section 6.1.12).

6.1.11 Collinearity

One important issue in multiple linear regression analysis, and one that seems to be ignored by many biologists who fit multiple regression models to their data, is the impact of correlated predictor variables on the estimates of parameters and hypothesis tests. If the predictors are correlated, then the data are said to be affected by (multi)collinearity. Severe collinearity can have important, and detrimental, effects on the estimated regression parameters. Lack of collinearity is also very difficult to meet with real biological data, where predictor variables that might be incorporated into a multiple regression model are likely to be correlated with each other to some extent. In the data set from Loyn (1987), we might expect heavier grazing history the longer the forest patch has been isolated and lighter grazing history for bigger patches since domestic stock cannot easily access larger forest fragments (Box 6.2).

The calculations for multiple linear regression analysis involve matrix inversion (Box 6.4). Collinearity among the X-variables causes computational problems because it makes the determinant of the matrix of X-variables close to zero and matrix inversion basically involves dividing by the determinant. Dividing by a determinant that is close to zero results in values in the inverted matrix being very sensitive to small differences in the numbers in the original data matrix (Tabachnick & Fidell 1996), i.e. the inverted matrix is unstable. This means that estimates of parameters (particularly the partial regression slopes) are also unstable (see Philippi 1993). Small changes in the data or adding or deleting one of the predictor variables can change the estimated regression coefficients considerably, even changing their sign (Bowerman & O'Connell 1990).

A second effect of collinearity is that standard errors of the estimated regression slopes, and therefore confidence intervals for the model parameters, are inflated when some of the predictors are correlated (Box 6.5). Therefore, the overall regression equation might be significant, i.e. the test of the H_0 that all partial regression slopes equal zero is rejected, but none of the individual regression slopes are significantly different from zero. This reflects lack of power for individual tests on partial regression slopes because of the inflated standard errors for these slopes.

Note that as long as we are not extrapolating beyond the range of our predictor variables and we are making predictions from data with a similar pattern of collinearity as the data to which we fitted our model, collinearity doesn't necessarily prevent us from estimating a regression model that fits the data well and has good predictive power (Rawlings *et al.* 1998). It does, however, mean that we are not confident in our estimates of the model parameters. A different sample from the same population of observations, even using the same values of the predictor variables, might produce very different parameter estimates.

Detecting collinearity

Collinearity can be detected in a number of ways (e.g. Chaterjee & Price 1991, Neter *et al.* 1996, Philippi 1993) and we illustrate some of these in Box 6.1 and Box 6.2 with our example data sets. First, we should examine a matrix of correlation coefficients (and associated scatterplots) between the predictor variables and look for large correlations. A scatterplot matrix (SPLOM) is a very useful graphical method (Chapter 4) and, if the response variable is included, also indicates nonlinear relationships between the response variable and any of the predictor variables.

Box 6.5 | Collinearity

Here is a simple illustration of the effects of collinearity in a multiple regression model with one response variable (Y) and two predictor variables (X_1, X_2). Two artificial data sets were generated for the three variables from normal distributions. In the first data set, X_1 and X_2 are relatively uncorrelated ($r = 0.21$). A multiple linear regression model, including an intercept, was fitted to these data.

	Coefficient	Standard error	Tolerance	t	P
Intercept	−1.045	1.341		−0.779	0.447
Slope X_1	0.893	0.120	0.954	7.444	<0.001
Slope X_2	−0.002	0.112	0.954	−0.017	0.987

Note that tolerance is 0.95 indicating no collinearity problems and standard errors are small. The partial regression slope for Y on X_1 holding X_2 constant is significant.

For the second data set, the values of X_2 were re-arranged between observations (but the values, their mean and standard deviation were the same) so that they are highly correlated with X_1 ($r = 0.99$), which along with Y is unchanged. Again a multiple linear regression model, including an intercept, was fitted.

	Coefficient	Standard error	Tolerance	t	P
Intercept	0.678	1.371		0.495	0.627
Slope X_1	−0.461	0.681	0.024	−0.678	0.507
Slope X_2	1.277	0.634	0.024	2.013	0.060

Note that tolerance is now very low indicating severe collinearity. The standard error for the partial regression slope of Y against X_1 is much bigger than for the first data set and the test of the H_0 that this slope equals zero is now not significant, despite the values of Y and X_1 being identical to the first data set.

Now let's add a third predictor (X_3) that is correlated with both X_1 and X_2.

	Coefficient	Standard error	Tolerance	t	P
Intercept	−0.306	1.410		−0.217	0.831
Slope X_1	−0.267	0.652	0.023	−0.410	0.687
Slope X_2	0.495	0.746	0.015	0.664	0.516
Slope X_3	0.657	0.374	0.068	1.758	0.098

Note that the estimated regression coefficients for X_1 and X_2 have changed markedly upon the addition of X_3 to the model.

Second, we should check the tolerance value for each predictor variable. Tolerance for X_j is simply $1 - r^2$ from the OLS regression of X_j against the remaining $p - 1$ predictor variables. A low tolerance indicates that the predictor variable is correlated with one or more of the other predictors. An approximate guide is to worry about tolerance values less than 0.1. Tolerance is sometimes expressed as the variance inflation factor (VIF), which is simply the inverse of tolerance (and can also be calculated from the eigenvectors and eigenvalues derived from a PCA on the predictor variables – see Chapter 17); VIF values greater than ten suggest strong collinearity.

Third, we can extract the principal components from the correlation matrix among the predictor variables (see Chapter 17). Principal components with eigenvalues (i.e. explained variances) near zero indicate collinearity among the original predictor variables, because those components have little variability that is independent of the other components. Three statistics are commonly used to assess collinearity in this context. First, the condition index is the square root of the largest eigenvalue divided by each eigenvalue ($\sqrt{\lambda_{max}/\lambda}$). There will be a condition index for each principal component and values greater than 30 indicate collinearities that require attention (Belsley *et al.* 1980, Chaterjee & Price 1991). The second is the condition number, which is simply the largest condition index ($\sqrt{\lambda_{max}/\lambda_{min}}$). Third, Hocking (1996) proposed an indicator of collinearity that is simply λ_{min} and suggested values less than 0.5 indicated collinearity problems.

It is worth noting that examining eigenvalues from the correlation matrix of the predictor variables implicitly standardizes the predictors to zero mean and unit variance so they are on the same scale. In fact, most collinearity diagnostics give different results for unstandardized and standardized predictors and two of the solutions to collinearity described below are based on standardized predictor variables.

Dealing with collinearity

Numerous solutions to collinearity have been proposed. All result in estimated partial regression slopes that are likely to be more precise (smaller standard errors) but are no longer unbiased. The first approach is the simplest: omit predictor variables if they are highly correlated with other predictor variables that remain in the model. Multiple predictor variables that are really measuring similar biological entities (e.g. a set of morphological measurements that are highly correlated) clearly represent redundant information and little can be gained by including all such variables in a model. Unfortunately, omitting variables may bias estimates of parameters for those variables that are correlated with the omitted variable(s) but remain in the model. Estimated partial regression slopes can change considerably when some predictor variables are omitted or added. Nonetheless, retaining only one of a number of highly correlated predictor variables that contain biologically and statistically redundant information is a sensible first step to dealing with collinearity.

The second approach is based on a principal components analysis (PCA) of the X-variables (see Chapter 17) and is termed principal components regression. The p principal components are extracted from the correlation matrix of the predictor variables and Y is regressed against these principal components, which are uncorrelated, rather than the individual predictor variables. Usually, components that contribute little to the total variance among the X-variables or that are not related to Y are deleted and the regression model of Y against the remaining components refitted. The regression coefficients for Y on the principal components are not that useful, however, because the components are often difficult to interpret as each is a linear combination of all p predictor variables. Therefore, we back-calculate the partial regression slopes on the original standardized variables from the partial regression slopes on the reduced number of principal components. The back-calculated regression slopes are standardized because the PCA is usually based on a correlation matrix of X-variables, so we don't have to worry about an intercept term. Because principal components regression requires an understanding of PCA, we will describe it in more detail in Chapter 17; see also Jackson (1991), Lafi & Kaneene (1992) and Rawlings *et al.* (1998).

Note that deciding which components to omit is critical for principal components regression. Simply deleting those with small eigenvalues (little relative contribution to the total variation in the X-variables) can be very misleading (Jackson 1991, Hadi & Ling 1998). The strength of the relationship of each component with Y must also be considered.

The third approach is ridge regression, another biased regression estimation technique that is somewhat controversial. A small biasing constant is added to the normal equations that are solved to estimate the standardized regression coefficients (Chaterjee & Price 1991, Neter *et al.* 1996). Adding this constant biases the estimated regression coefficients but also reduces their

Table 6.2 Expected values of mean squares from analysis of variance for a multiple linear regression model with two predictor variables

Mean square	Expected value
$MS_{Regression}$	$\sigma_\varepsilon^2 + \dfrac{\beta_1^2\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2 + \beta_2^2\sum_{i=1}^{n}(x_{i2}-\bar{x}_2) + 2\beta_1\beta_2\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2)}{2}$
$MS_{Residual}$	σ_ε^2

variability and hence their standard errors. The choice of the constant is critical. The smaller its value, the less bias in the estimated regression slopes (when the constant is zero, we have an OLS regression); the larger its value, the less collinearity (increasing the constant reduces the VIF). Usually a range of values is tried (say, increasing from 0.001) and a diagnostic graphic (the ridge trace) used to determine the smallest value of the constant that is the best compromise between reducing the variation in the estimated regression slopes and reducing their VIFs. Neter *et al.* (1996) provided a clear worked example.

Careful thought about the predictor variables to be included in a multiple linear regression model can reduce collinearity problems before any analysis. Do not include clearly redundant variables that are basically measuring similar biological entities. If the remaining predictor variables are correlated to an extent that might affect the estimates of the regression slopes, then we prefer principal components regression over ridge regression for two reasons. First, it is relatively straightforward to do with most statistical software that can handle multiple regression and PCA, although some hand calculation might be required (e.g. for standard errors). Second, PCA is also a useful check for collinearity so is often done anyway. The calculations required for ridge regression, in contrast, are complex and not straightforward in most statistical software.

6.1.12 Interactions in multiple regression

The multiple regression model we have been using so far is an additive one, i.e. the effects of the predictor variables on Y are additive. In many biological situations, however, we would anticipate interactions between the predictors (Aiken & West 1991, Jaccard *et al.* 1990) so that their effects on Y are multiplicative. Let's just consider the case with two predictors, X_1 and X_2. The additive multiple linear regression model is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i \quad (6.17)$$

This assumes that the partial regression slope of Y on X_1 is independent of X_2 and vice-versa. The multiplicative model including an interaction is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \varepsilon_i \quad (6.18)$$

The new term $(\beta_3 x_{i1} x_{i2})$ in model 6.18 represents the interactive effect of X_1 and X_2 on Y. It measures the dependence of the partial regression slope of Y against X_1 on the value of X_2 and the dependence of the partial regression slope of Y against X_2 on the value of X_1. The partial slope of the regression of Y against X_1 is no longer independent of X_2 and vice versa. Equivalently, the partial regression slope of Y against X_1 is different for each value of X_2.

Using the data from Paruelo & Lauenroth (1996), model 6.2 indicates that we expect no interaction between latitude and longitude in their effect on the relative abundance of C_3 plants. But what if we allow the relationship between C_3 plants and latitude to vary for different longitudes? Then we are dealing with an interaction between latitude and longitude and our model becomes:

$$(\text{relative abundance of } C_3 \text{ grasses})_i = \beta_0 + \beta_1(\text{latitude})_i + \beta_2(\text{longitude})_i + \beta_3(\text{latitude})_i \times (\text{longitude})_i + \varepsilon_i \quad (6.19)$$

One of the difficulties with including interaction terms in multiple regression models is that lower-order terms will usually be highly correlated with their interactions, *e.g.* X_1 and X_2 will be highly correlated with their interaction X_1X_2. This results in

all the computational problems and inflated variances of estimated coefficients associated with collinearity (Section 6.1.11). One solution to this problem is to rescale the predictor variables by centering, i.e. subtracting their mean from each observation, so the interaction is then the product of the centered values (Aiken & West 1991, Neter *et al.* 1996; see Box 6.1 and Box 6.2). If X_1 and X_2 are centered then neither will be strongly correlated with their interaction. Predictors can also be standardized (subtract the mean from each observation and divide by the standard deviation) which has an identical affect in reducing collinearity.

When interaction terms are not included in the model, centering the predictor variables does not change the estimates of the regression slopes nor hypothesis tests that individual slopes equal zero. Standardizing the predictor variables does change the value of the regression slopes, but not their hypothesis tests because the standardization affects the coefficients and their standard errors equally. When interaction terms are included, centering does not affect the regression slope for the highest-order interaction term, nor the hypothesis test that the interaction equals zero. Standardization changes the value of the regression slope for the interaction but not the hypothesis test. Centering and standardization change all lower-order regression slopes and hypothesis tests that individual slopes equal zero but make them more interpretable in the presence of an interaction (see below). The method we will describe for further examining interaction terms using simple slopes is also unaffected by centering but is affected by standardizing predictor variables.

We support the recommendation of Aiken & West (1991) and others that multiple regression models with interaction terms should be fitted to data with centered predictor variables. Standardization might also be used if the variables have very different variances but note that calculation and tests of simple slopes must then be based on analyzing standardized variables but using the unstandardized regression coefficients (Aiken & West 1991).

Probing interactions

Even in the presence of an interaction, we can still interpret the partial regression slopes for other terms in model 6.18. The estimate of β_1 determined by the OLS fit of this regression model is actually the regression slope of Y on X_1 when X_2 is zero. If there is an interaction (β_3 does not equal zero), this slope will obviously change for other values of X_2; if there is not an interaction (β_3 equals zero), then this slope will be constant for all levels of X_2. In the presence of an interaction, the estimated slope for Y on X_1 when X_2 is zero is not very informative because zero is not usually within the range of our observations for any of the predictor variables. If the predictors are centered, however, then the estimate of β_1 is now the regression slope of Y on X_1 for the mean of X_2, a more useful piece of information. This is another reason why variables should be centered before fitting a multiple linear regression model with interaction terms.

However, if the fit of our model indicates that interactions between two or more predictors are important, we usually want to probe these interactions further to see how they are structured. Let's express our multiple regression model as relating the predicted y_i to two predictor variables and their interaction using sample estimates:

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i1} x_{i2} \quad (6.20)$$

This can be algebraically re-arranged to:

$$\hat{y}_i = (b_1 + b_3 x_{i2}) x_{i1} + (b_2 x_{i2} + b_0) \quad (6.21)$$

We now have $(b_1 + b_3 x_{i2})$, the simple slope of the regression of Y on X_1 for any particular value of X_2 (indicated as x_{i2}). We can then choose values of X_2 and calculate the estimated simple slope, for either plotting or significance testing. Cohen & Cohen (1983) and Aiken & West (1991) suggested using three different values of X_2: $\bar{x}_2$, $\bar{x}_2 + s$, $\bar{x}_2 - s$, where s is the sample standard deviation of X_2. We can calculate simple regression slopes by substituting these values of X_2 into the equation for the simple slope of Y on X_1.

The H_0 that the simple regression slope of Y on X_1 for a particular value of X_2 equals zero can also be tested. The standard error for the simple regression slope is:

$$\sqrt{s_{11}^2 + 2x_2 s_{13}^2 + x_2^2 s_{33}^2} \quad (6.22)$$

where s_{11}^2 and s_{33}^2 are the variances of b_1 and b_3 respectively, s_{13}^2 is the covariance between b_1 and b_3

and x_2 is the value of X_2 chosen. The variance and covariances are obtained from a covariance matrix of the regression coefficients, usually standard output for regression analyses with most software. Then the usual t test is applied (simple slope divided by standard error of simple slope). Fortunately, simple slope tests can be done easily with most statistical software (Aiken & West 1990, Darlington 1990). For example, we use the following steps to calculate the simple slope of Y on X_1 for a specific value of X_2, such as $\bar{x}_2 + s$.

1. Create a new variable (called the conditional value of X_2, say CVX_2), which is x_{i2} minus the specific value chosen.
2. Fit a multiple linear regression model for Y on X_1, CVX_2, X_1 by CVX_2.
3. The partial slope of Y on X_1 from this model is the simple slope of Y on X_1 for the specific value of X_2 chosen.
4. The statistical program then provides a standard error and t test.

This procedure can be followed for any conditional value. Note that we have calculated simple slopes for Y on X_1 at different values of X_2. Conversely, we could have easily calculated simple slopes for Y on X_2 at different values of X_1.

If we have three predictor variables, we can have three two-way interactions and one three-way interaction:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i1} x_{i2} + \beta_5 x_{i1} x_{i3} + \beta_6 x_{i2} x_{i3} + \beta_7 x_{i1} x_{i2} x_{i3} + \varepsilon_i \quad (6.23)$$

In this model, β_7 is the regression slope for the three-way interaction between X_1, X_2 and X_3 and measures the dependence of the regression slope of Y on X_1 on the values of different combinations of both X_2 and X_3. Equivalently, the interaction is the dependence of the regression slope of Y on X_2 on values of different combinations of X_1 and X_3 and the dependence of the regression slope of Y on X_3 on values of different combinations of X_1 and X_2. If we focus on the first interpretation, we can determine simple regression equations for Y on X_1 at different combinations of X_2 and X_3 using sample estimates:

$$\hat{y}_i = (b_1 + b_4 x_{i2} + b_5 x_{i3} + b_7 x_{i2} x_{i3}) x_{i1} + (b_2 x_{i2} + b_3 x_{i3} + b_6 x_{i2} x_{i3} + b_0) \quad (6.24)$$

Now we have $(b_1 + b_4 x_{i2} + b_5 x_{i3} + b_7 x_{i2} x_{i3})$ as the simple slope for Y on X_1 for specific values of X_2 and X_3 together. Following the logic we used for models with two predictors, we can substitute values for X_2 and X_3 into this equation for the simple slope. Aiken & West (1991) suggested using $\bar{x}_2$ and $\bar{x}_3$ and the four combinations of $\bar{x}_2 \pm s_{x_2}$ and $\bar{x}_3 \pm s_{x_3}$. Simple slopes for Y on X_2 or X_3 can be calculated by just reordering the predictor variables in the model. Using the linear regression routine in statistical software, simple slopes, their standard errors and t tests for Y on X_1 at specific values of X_2 and X_3 can be calculated.

1. Create two new variables (called the conditional values of X_2 and X_3, say CVX_2 and CVX_3), which are x_{i2} and x_{i3} minus the specific values chosen.
2. For each combination of specific values of X_2 and X_3, fit a multiple linear regression model for Y on X_1, CVX_2, CVX_3, X_1 by CVX_2, X_1 by CVX_3, CVX_2 by CVX_3, and X_1 by CVX_2 by CVX_3.
3. The partial slope of Y on X_1 from this model is the simple slope of Y on X_1 for the chosen specific values of X_2 and X_3.

With three or more predictor variables, the number of interactions becomes large and they become more complex (three-way interactions and higher). Incorporating all possible interactions in models with numerous predictors becomes unwieldy and we would need a very large sample size because of the number of terms in the model. There are two ways we might decide which interactions to include in a linear regression model, especially if our sample size does not allow us to include them all. First, we can use our biological knowledge to predict likely interactions and only incorporate this subset. For the data from Loyn (1987), we might expect the relationship between bird density and grazing to vary with area (grazing effects more important in small fragments?) and years since isolation (grazing more important in new fragments?), but not with distance to any forest or larger fragments. Second, we can plot the residuals from an additive model against the possible interaction terms (new variables formed by simply multiplying the predictors) to see if any of these interactions are related to variation in the response variable.

There are two take-home messages from this section. First, we should consider interactions between continuous predictors in multiple linear regression model because such interactions may be common in biological data. Second, these interactions can be further explored and interpreted using relatively straightforward statistical techniques with most linear regression software.

6.1.13 Polynomial regression

Generally, curvilinear models fall into the class of nonlinear regression modeling (Section 6.4) because they are best fitted by models that are nonlinear in the parameters (e.g. power functions). There is one type of curvilinear model that can be fitted by OLS (i.e. it is still a linear model) and is widely used in biology, the polynomial regression.

Let's consider a model with one predictor variable (X_1). A second-order polynomial model is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2 + \varepsilon_i \quad (6.25)$$

where β_1 is the linear coefficient and β_2 is the quadratic coefficient. Such models can be fitted by simply adding the x_{i1}^2 term to the right-hand side of the model, and they have a parabolic shape. Note that x_{i1}^2 is just an interaction term (i.e. x_{i1} by x_{i1}). There are two questions we might wish to ask with such a model (Kleinbaum *et al.* 1988). First, is the overall regression model significant? This is a test of the H_0 that β_1 equals β_2 equals zero and is done with the usual F test from the regression ANOVA. Second, is a second-order polynomial a better fit than a first-order model? We answer this with a partial F statistic, which tests whether the full model including X^2 is a better fit than the reduced model excluding X^2 using the principle of extra SS we described in Section 6.1.4:

$$F(X^2 | X) = \frac{(SS_{Extra} \text{ due to added } X^2)/1}{\text{Full } MS_{Residual}} \quad (6.26)$$

where the SS_{Extra} is the difference between the $SS_{Regression}$ for the full model with the second-order polynomial term and the $SS_{Regression}$ for the reduced model with just the first-order term.

For example, Caley & Schluter (1997) examined the relationship between local and regional species diversity for a number of taxa and geographic regions at two spatial scales of sampling (1% of region and 10% of region). Regional species diversity was the predictor variable and local species diversity was the response variable and Caley & Schluter (1997) showed that adding a quadratic term to the model explained significantly more of the variance in local species diversity compared with a simple linear model (Box 6.6; Figure 6.5).

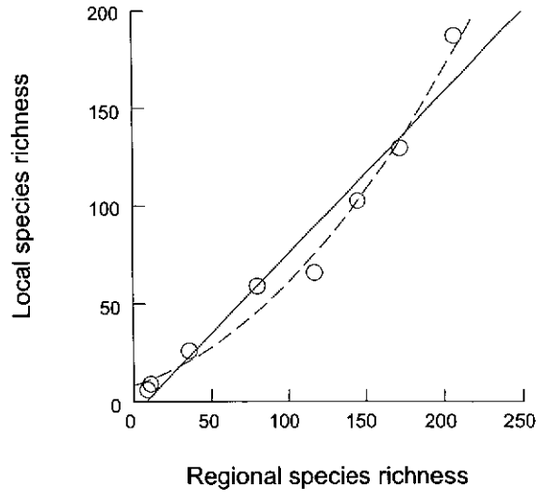

Figure 6.5 Scatterplot of local species richness against regional species richness for 10% of regions sampled in North America for a range of taxa (Caley & Schluter 1997) showing linear (solid line) and second-order polynomial (quadratic; dashed line) regression functions.

Polynomial regressions can be extended to third-order (cubic) models, which have a sigmoid shape:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2 + \beta_3 x_{i1}^3 + \varepsilon_i \quad (6.27)$$

Polynomial models can also contain higher orders (quartic, quintic, etc.) and more predictors. We have to be very careful about extrapolation beyond the range of our data with polynomial regression models. For example, a quadratic model will have a parabolic shape although our observations may only cover part of that function. Imagine fitting a quadratic model to the species area data in Figure 5.17. Predicting species number for larger clumps using this quadratic model would be misleading as theory suggests that species number would not then decline with increasing clump area.

Box 6.6 | Worked example of polynomial regression

We will use the data set from Caley & Schluter (1997), examining the regression of local species richness against regional species richness just for North America and at a sampling scale of 10% of the region. Although there was some evidence that both local and regional species richness were skewed, we will, like the original authors, analyze untransformed variables. Caley & Schluter (1997) forced their models through the origin, but because that makes interpretation difficult, we will include an intercept in the models. First, we will fit a second-order polynomial to the data:

$$(\text{local species richness})_i = \beta_0 + \beta_1(\text{regional species richness})_i + \beta_2(\text{regional species richness})_i^2 + \varepsilon_i$$

	Coefficient	Standard error	Tolerance	t	P
β_0	8.124	6.749		1.204	0.283
β_1	0.249	0.170	0.066	1.463	0.203
β_2	0.003	0.001	0.066	3.500	0.017

We would reject the H_0 that β_2 equals zero. Note that the tolerances are very low, indicating collinearity between regional species richness and (regional species richness)^2 as we would expect. This collinearity might affect the estimate and test of β_1 but won't affect the partitioning of the variance and the calculation of SS_{Extra} [(regional species richness)^2 | regional species richness], so we, like Caley & Schluter (1997) will continue the analysis with uncentered data.

The partitioning of the variation resulted in the following ANOVA.

Source	SS	df	MS	F	P
Regression	2.781×10^4	2	1.390×10^4	184.582	<0.001
Residual	376.620	5	75.324		

Note the $SS_{Regression}$ has two df because there are three parameters in the model. We would reject the H_0 that β_1 equals β_2 equals zero.

Now we fit a reduced model without the quadratic term:

$$(\text{local species richness})_i = \beta_0 + \beta_1(\text{regional species richness})_i + \varepsilon_i$$

Source	SS	df	MS	F	P
Regression	2.688×10^4	1	2.688×10^4	124.152	<0.001
Residual	1299.257	6	216.543		

The $SS_{Regression}$ from the full model is 2.781×10^4 and the $SS_{Regression}$ from the reduced model is 2.688×10^4. Therefore SS_{Extra} is 922.7 with one df and F [(regional species richness)^2 | regional species richness] equals 12.249 with $P < 0.018$. We would conclude that adding the second-order polynomial term to this model contributes significantly to explained variation in local species richness. It is apparent from Figure 6.5, despite the small sample size, that the second-order polynomial model provides a better visual fit than a simple linear model. Note that quadratic models were not better fits than linear for any of the other combinations of region (worldwide, Australia, North America) and spatial scale (1% and 10% of region).

Table 6.3 | Dummy variable coding for grazing effect from Loyn (1987)

Grazing intensity	$Grazing_1$	$Grazing_2$	$Grazing_3$	$Grazing_4$
Zero (reference category)	0	0	0	0
Low	1	0	0	0
Medium	0	1	0	0
High	0	0	1	0
Intense	0	0	0	1

Polynomial terms in these models will always be correlated with lower-order terms, so collinearity can be a problem, causing unstable estimates of the coefficients for the lower order terms and increasing their standard errors. Since the polynomial term is just an interaction, centring the predictors will reduce the degree of collinearity, without affecting the estimate and test of the slope for the highest-order term in the model nor the partitioning of the SS. However, the estimate of the slope for the lower-order terms will be different but also more reliable with smaller standard errors once collinearity has been reduced.

6.1.14 Indicator (dummy) variables

There are often situations when we would like to incorporate a categorical variable into our multiple regression modeling. For example, Loyn (1987) included a predictor variable indicating the historical intensity of grazing in each of his forest patches. This variable took values of 1, 2, 3, 4 or 5 and was treated as a continuous variable for the analysis. We could also treat this as a categorical variable, with five categories of grazing. While the values of this variable actually represent a quantitative scale (from low grazing intensity to high grazing intensity), many categorical variables will be qualitative. For example, Paruelo & Lauenroth (1996) included a categorical variable that separated sites into shrubland and grassland. To include categorical variables in a regression model, we must convert them to continuous variables called indicator or dummy variables. Commonly, dummy variables take only two values, zero or one, although other types of coding are possible.

In the example from Paruelo & Lauenroth (1996) where there are only two categories, we could code grasslands as zero and shrublands as one, although the authors used coding of one and two. As long as the interval is the same, the coding doesn't matter in this case. For Loyn's (1987) grazing history variable, there are five categories that we will call zero, low, medium, high, and intense grazing. The dummy variables would be as follows.

X_1 1 if low
0 if not

X_2 1 if medium
0 if not

X_3 1 if high
0 if not

X_4 1 if intense
0 if not

This defines all our categories (Table 6.3) and we would fit a linear model including each of these dummy variables as predictors. For a predictor variable with c categories, we only need $c-1$ dummy variables. Interpreting the regression coefficients is a little tricky. The coefficients for X_1, X_2, X_3 and X_4 indicate how different the effects of low, medium, high and intense grazing respectively are compared to zero grazing, i.e. the coefficients for dummy variables measure the differential effects of each category compared to a reference category (in which all dummy variables equal zero). The choice of the reference category should be made prior to analysis. In this example, we used the zero grazing category ("control") as the reference category. An alternative method of coding dummy variables is using the deviation of each category mean from the overall mean, which is commonly used in analysis of variance models (see Chapter 8 onwards) and is termed effects coding.

Box 6.7 | Worked example of indicator (dummy) variables

We will consider a subset of the data from Loyn (1987) where abundance of forest birds is the response variable and grazing intensity (1 to 5 from least to greatest) and $\log_{10}$ patch area are the predictor variables. First, we treat grazing as a continuous variable and fit model 6.28.

Coefficient	Estimate	Standard error	t	P
Intercept	21.603	3.092	6.987	<0.001
Grazing	−2.854	0.713	−4.005	<0.001
$\log_{10}$ area	6.890	1.290	5.341	<0.001

Note that both the effects of grazing and $\log_{10}$ area are significant and the partial regression slope for grazing is negative, indicating that, holding patch area constant, there are fewer birds in patches with more intense grazing.

Now we will convert grazing into four dummy variables with no grazing (level 1) as the reference category (Table 6.3) and fit model 6.29.

	Estimate	Standard error	t	P
Intercept	15.716	2.767	5.679	<0.001
Grazing_1	0.383	2.912	0.131	0.896
Grazing_2	−0.189	2.549	−0.074	0.941
Grazing_3	−1.592	2.976	−0.535	0.595
Grazing_4	−11.894	2.931	−4.058	<0.001
$\log_{10}$ area	7.247	1.255	5.774	<0.001

The partial regression slopes for these dummy variables measure the difference in bird abundance between the grazing category represented by the dummy variable and the reference category for any specific level of $\log_{10}$ area. Note that only the effect of intense grazing (category: 5; dummy variable: grazing_4) is different from the no grazing category.

If our linear model only has categorical predictor variables ("factors"), then they are usually considered as classical analyses of variance models. Commonly, we have linear models with a mixture of categorical and continuous variables. The simplest case is one categorical predictor (converted to dummy variables) and one continuous predictor. For example, consider a subset of the data from Loyn (1987) where we will model the abundance of forest birds against grazing intensity (1 to 5 indicating no grazing to intense grazing) and patch area (transformed to $\log_{10}$) – see Box 6.7. Because the levels of grazing categories are quantitative, grazing intensity can be treated as a continuous variable with the following typical multiple regression model:

$$(\text{bird abundance})_i = \beta_0 + \beta_1(\text{grazing})_i + \beta_2(\log_{10}\text{area})_i + \varepsilon_i \quad (6.28)$$

Alternatively, we could consider grazing intensity as a categorical variable and we would create four dummy variables (Table 6.3) and include these in our model:

$$(\text{bird abundance})_i = \beta_0 + \beta_1(\text{grazing}_1)_i + \beta_2(\text{grazing}_2)_i + \beta_3(\text{grazing}_3)_i + \beta_4(\text{grazing}_4)_i + \beta_5(\log_{10}\text{area})_i + \varepsilon_i \quad (6.29)$$

This model can be envisaged as separate linear regression models between Y and $\log_{10}$ area for each level of the categorical predictor (grazing). The partial regression slope for each dummy variable measures the difference in the predicted value of Y between that category of grazing and the reference category (zero grazing) for any specific value of $\log_{10}$ area. Using analysis of covariance terminology (Chapter 12), each regression slope measures the difference in the adjusted mean of Y between that category and the reference category (Box 6.7). Interaction terms between the dummy variables and the continuous variable could also be included. These interactions measure how much the slopes of the regressions between Y and the $\log_{10}$ area differ between the levels of grazing. Most statistical software now automates the coding of categorical variables in regression analyses, although you should check what form of coding your software uses. Models that incorporate continuous and categorical predictors will also be considered as part of analysis of covariance in Chapter 12.

6.1.15 Finding the "best" regression model

In many uses of multiple regression, biologists want to find the smallest subset of predictors that provides the "best fit" to the observed data. There are two apparent reasons for this (Mac Nally 2000), related to the two main purposes of regression analysis – explanation and prediction. First, the "best" subset of predictors should include those that are most important in explaining the variation in the response variable. Second, other things being equal, the precision of predictions from our fitted model will be greater with fewer predictor variables in the model. Note that, as we said in the introduction to Chapter 5, biologists, especially ecologists, seem to rarely use their regression models for prediction and we agree with Mac Nally (2000) that biologists are usually searching for the "best" regression model to explain the response variable.

It is important to remember that there will rarely be, for any real data set, a single "best" subset of predictors, particularly if there are many predictors and they are in any way correlated with each other. There will usually be a few models, with different numbers of predictors, which provide similar fits to the observed data. The choice between these competing models will still need to be based on how well the models meet the assumptions, diagnostic considerations of outliers and other influential observations and biological knowledge of the variables retained.

Criteria for "best" model

Irrespective of which method is used for selecting which variables are included in the model (see below), some criterion must be used for deciding which is the "best" model. One characteristic of such a criterion is that it must protect against "overfitting", where the addition of extra predictor variables may suggest a better fit even when these variables actually add very little to the explanatory power. For example, r^2 cannot decrease as more predictor variables are added to the model even if those predictors contribute nothing to the ability of the model to predict or explain the response variable (Box 6.8). So r^2 is not suitable for comparing models with different numbers of predictors.

We are usually dealing with a range of models, with different numbers of predictors, but all are subsets of the full model with all predictors. We will use P to indicate all possible predictors, p is the number of predictors included in a specific model, n is the number of observations and we will assume that an intercept is always fitted. If the models are all additive, i.e. no interactions, the number of parameters is $p+1$ (the number of predictors plus the intercept). When interactions are included, then p in the equations below should be the number of parameters (except the intercept) in the model, including both predictors and their interactions. We will describe four criteria for determining the fit of a model to the data (Table 6.4).

The first is the adjusted r^2 which takes into account the number of predictors in the model and, in contrast to the usual r^2, basically uses mean squares instead of sum of squares and can increase or decrease as new variables are added to the model. A larger value indicates a better fit. Using the $\text{MS}_{\text{Residual}}$ from the fit of the model is equivalent where a lower value indicates a better fit.

The second is Mallow's C_p, which works by comparing a specific reduced model to the full model with all P predictors included. For the full model with all P predictors, C_p will equal $P+1$ (the number of parameters including the intercept). The choice of the best model using C_p has two components: C_p should be as small as possible and as close to p as possible.

Box 6.8 | Hierarchical partitioning and model selection.

The data from Loyn (1987) were used to compare model selection criteria. Only the best two models (based on the BIC) for each number of predictors are presented as well as the full model. The model with the lowest BIC is in bold.

No. predictors	Model	r^2	Adj r^2	C_p	AIC	Schwarz (BIC)
1	$\log_{10}$ area	0.548	0.539	18.4	224.39	228.45
1	grazing	0.466	0.456	31.1	223.71	237.76
2	**$\log_{10}$ area + grazing**	**0.653**	**0.640**	**4.0**	**211.59**	**217.67**
2	$\log_{10}$ area + years	0.643	0.630	5.4	213.06	219.14
3	$\log_{10}$ area + grazing + years	0.673	0.654	2.8	210.19	218.29
3	$\log_{10}$ area + grazing + $\log_{10}$ ldist	0.664	0.644	4.3	211.77	219.88
4	$\log_{10}$ area + grazing + years + altitude	0.682	0.657	3.4	210.60	220.73
4	$\log_{10}$ area + grazing + years + $\log_{10}$ ldist	0.679	0.654	3.9	211.15	221.28
5	$\log_{10}$ area + grazing + years + $\log_{10}$ ldist + $\log_{10}$ dist	0.681	0.649	5.1	212.89	225.05
5	$\log_{10}$ area + grazing + altitude + $\log_{10}$ ldist + $\log_{10}$ dist	0.668	0.635	5.1	215.11	227.27
6	$\log_{10}$ area + grazing + years + altitude + $\log_{10}$ ldist + $\log_{10}$ dist	0.685	0.646	7.0	214.14	228.32

The Schwarz criterion (BIC) selects a model with just two predictors ($\log_{10}$ area and grazing). In contrast, the AIC and Mallow's C_p selected a model that included these two predictors and years since isolation, and the adjusted r^2 selected a four-predictor model that added altitude to the previous three predictors. Note that the unadjusted r^2 is highest for the model with all predictors.

For these data, automated forward and backward selection procedures (the significance level for entering and removing terms based on partial F-ratio statistics was set at 0.15) produced the same final model including $\log_{10}$ area, grazing and years since isolation. The results from a hierarchical partitioning of r^2 from the model relating abundance of forest birds to all six predictor variables from Loyn (1987) are shown below.

	Independent	Joint	Total
Log_{10} area	0.315	0.232	0.548
Log_{10} dist	0.007	0.009	0.016
Log_{10} ldist	0.014	<0.001	0.014
Altitude	0.057	0.092	0.149
Grazing	0.190	0.275	0.466
Years	0.101	0.152	0.253

Clearly, $\log_{10}$ area and grazing contribute the most to the explained variance in abundance of forest birds, both as independent effects and joint effects with other predictors, with some contribution also by years since isolation.

Table 6.4 | Criteria for selecting "best" fitting model in multiple linear regression. Formulae are for a specific model with p predictors included. Note that p excludes the intercept

Criterion	Formula
Adjusted r^2	$1 - \frac{SS_{Residual}/[n-(p+1)]}{SS_{Total}/(n-1)}$
Mallow's C_p	$\frac{\text{Reduced } SS_{Residual}}{\text{Full } MS_{Residual}} - [n-2(p+1)]$
Akaike Information Criterion (AIC)	$n[\ln(SS_{Residual})] + 2(p+1) - n\ln(n)$
Schwarz Bayesian Information Criterion (BIC)	$n[\ln(SS_{Residual})] + (p+1)\ln(n) - n\ln(n)$

The remaining two measures are in the category of information criteria, introduced by Akaike (1978) and Schwarz (1978) to summarize the information in a model, accounting for both sample size and number of predictors (Table 6.4). Although these information criteria are usually based on likelihoods, they can be adapted for use with OLS since the estimates of parameters will be the same when assumptions hold. The first of these criteria is the Akaike information criterion (AIC), which tends to select the same models as Mallow's C_p as n increases and the $MS_{Residual}$ becomes a better estimate of σ_ε^2 (Christensen 1997; see Box 6.8). The Bayesian (or Schwarz) information criterion (BIC) is similar but adjusts for sample size and number of predictors differently. It more harshly penalizes models with a greater number of predictors than the AIC (Rawlings *et al.* 1998).

For both AIC and BIC, smaller values indicate better, more parsimonious, models (Box 6.8). We recommend the Schwarz criterion for determining the model that best fits the data with the fewest number of parameters (see also Mac Nally 2000). It is simple to calculate and can be applied to linear and generalized linear models (see Chapter 13).

Selection procedures

The most sensible approach to selecting a subset of important variables in a complex linear model is to compare all possible subsets. This procedure simply fits all the possible regression models (i.e. all possible combinations of predictors) and chooses the best one (or more than one) based on one of the criteria described above. Until relatively recently, automated fitting of all subsets was beyond the capabilities of most statistical software because of the large number of possible models. For example, with six predictors, there are 64 possible models! Consequently, stepwise procedures were developed that avoided fitting all possible models but selected predictor variables based on some specific criteria. There are three types of stepwise procedures, forward selection, backward selection and stepwise selection.

Forward selection starts off with a model with no predictors and then adds the one (we'll call X_a) with greatest F statistic (or t statistic or correlation coefficient) for the simple regression of Y against that predictor. If the H_0 that this slope equals zero is rejected, then a model with that variable is fitted. The next predictor (X_b) to be added is the one with the highest partial F statistic for X_b given that X_a is already in the model $[F(X_b|X_a)]$. If the H_0 that this partial slope equals zero is rejected, then the model with two predictors is refitted and a third predictor added based on $F(X_c|X_a,X_b)$. The process continues until a predictor with a non-significant partial regression slope is reached or all predictors are included.

Backward selection (elimination) is the opposite of forward selection, whereby all predictors are initially included and the one with the smallest and non-significant partial F statistic is dropped. The model is refitted and the next predictor with the smallest and non-significant

partial *F* statistic is dropped. The process continues until there are no more predictors with non-significant partial *F* statistics or there are no predictors left.

Stepwise selection is basically a forward selection procedure where, at each stage of refitting the model, predictors can also be dropped using backward selection. Predictors added early in the process can be omitted later and vice versa.

For all three types of variable selection, the decision to add, drop or retain variables in the model is based on either a specified size of partial *F* statistics or significance levels. These are sometimes termed *F*-to-enter and *F*-to-remove and obviously, the values chosen will greatly influence which variables are added or removed from the model, especially in stepwise selection. Significance levels greater than 0.05, or small *F* statistics, are often recommended (and are default settings in stepwise selection routines of most regression software) because this will result in more predictors staying in the model and reduce the risk of omitting important variables (Bowerman & O'Connell 1990). However, as always, the choice of significance levels is arbitrary. Note that so many *P* values for tests of partial regression slopes are generated in variable selection procedures that these *P* values are difficult to interpret, due to the multiple testing problem (see Chapter 3) and lack of independence. Variable selection is not suited to the hypothesis testing framework.

It is difficult to recommend any variable selection procedure except all subsets. The logical and statistical problems with the forward, backward and stepwise procedures have been pointed out elsewhere (e.g. James & McCulloch 1990, Chaterjee & Price 1991, Neter *et al.* 1996). They all use somewhat arbitrary statistical rules (significance levels or the size of *F* statistics) for deciding which variables enter or leave the model and these rules do not consider the increased probability of Type I errors due to multiple testing. These approaches seem to be an abuse of the logic of testing *a priori* statistical hypotheses; statistical hypothesis testing and significance levels are ill-suited for exploratory data-snooping. Also, the forward, backward and stepwise approaches for including and excluding variables can produce very different final models even from the same set of data (James & McCulloch 1990, Mac Nally 2000), particularly if there are many predictors. Additionally, simulation studies have shown that these stepwise procedures can produce a final model with a high r^2, even if there is really no relationship between the response and the predictor variables (Flack & Chang 1987, Rencher & Pun 1980). Finally, variable selection techniques are sensitive to collinearity between the predictors (Chaterjee & Price 1991). This is because collinearity will often result in large variances for some regression slopes that may result in those predictor variables being excluded from the model irrespective of their importance.

The all-subsets procedure is limited by the large number of models to be compared when there are many predictor variables, although most statistical software can now compare all subsets for reasonable numbers of predictors. It is difficult to envisage a data set in biology with too many variables for all subsets comparisons that is also not plagued by serious collinearity problems, which would invalidate any variable selection procedure.

If the number of observations is large enough, then we recommend using cross-validation techniques to check the validity of the final model. The simplest form of cross-validation is randomly to split the data set in two and fit the model with half the data set and then see how well the model predicts values of the response variable in the other half of the data set. Unfortunately, splitting the data for cross-validation is not always possible because of small sample sizes often encountered in biology.

In the end, however, the best argument against stepwise variable selection methods is that they do not necessarily answer sensible questions in the current age of powerful computers and sophisticated statistical software. If a regression model is required for explanation, then we wish to know which variables are important, and the criteria we described above, combined with hierarchical partitioning (Section 6.1.16), are the best approaches. If a model is required for prediction, with as few predictor variables as possible, then comparing all-subsets is feasible and probably the most sensible, although more complex procedures are possible (Mac Nally 2000). We conclude with a quote from James & McCulloch (1990, pp. 136–137): "Many authors have documented the folly of using stepwise procedures with any multivariate method. Clearly, stepwise regression is not able to select from a set of variables those that are most influential."

6.1.16 Hierarchical partitioning

Hierarchical partitioning is a method that has been around for some time but its utility for interpreting the importance of variables in linear models has only recently been appreciated in the statistical (Chevan & Sutherland 1991) and biological literature (Mac Nally 1996). Its purpose is to quantify the "independent" correlation of each predictor variable with the response variable. It works by measuring the improvement in the fit of all models with a particular predictor compared to the equivalent model without that predictor and the improvement in fit is averaged across all possible models with that predictor. We can use any of a number of measures of fit, but for linear models, it is convenient to use r^2.

Consider a model with a response variable (*Y*) and three predictor variables (X_1, X_2, X_3). There are 2^p possible models when there are p "independent" predictor variables, so here, there are 2^3 equals eight models. We can calculate r^2 for the eight possible models listed in Table 6.5. Note that there are four hierarchical levels of model complexity, representing the number of predictors in the model. Hierarchical partitioning splits the total r^2 for each predictor, i.e. the r^2 for the linear relationship between *Y* and each predictor by itself (as in Models 2, 3 and 4), into two additive components.

- The "independent" contributions of each predictor variable, which is a partitioning of the r^2 for the full model with all predictors (Model 8).
- The "joint" contributions of each predictor in conjunction with other predictors.

For the independent contributions, we calculate for each predictor variable the improvement in fit by adding that predictor to reduced models without that predictor at each hierarchical level.

Table 6.5 Eight possible models with one response variable and three predictor variables

Label	Model	Level of hierarchy
1	No predictors, r^2 equals zero	0
2	X_1	1
3	X_2	1
4	X_3	1
5	$X_1 + X_2$	2
6	$X_1 + X_3$	2
7	$X_2 + X_3$	2
8	$X_1 + X_2 + X_3$	3

For example, for X_1, we would compare the following r^2 values:

$r^2(X_1)$ vs r^2(Null)

$r^2(X_1,X_2)$ vs $r^2(X_2)$

$r^2(X_1,X_3)$ vs $r^2(X_3)$

$r^2(X_1,X_2,X_3)$ vs $r^2(X_2,X_3)$

The differences in r^2 values are averaged within each hierarchical level (first order, second order, third order) and then averaged across the levels to produce the independent contribution of X_1 to the explained variance in *Y*. The same procedure is followed for the other predictor variables. These independent contributions of all the predictor variables represent a partitioning of the r^2 from the full model with all predictors included. For example, the sum of the independent contributions of $\log_{10}$ area, $\log_{10}$ dist, $\log_{10}$ ldist, altitude, grazing and years to forest bird abundance for the data from Loyn (1987) equals the r^2 from the fit of the full model with all these predictors (Box 6.8).

If the predictor variables are completely independent of (i.e. uncorrelated with) each other, then there will be no joint contributions and the sum of the r^2 for Models 2, 3 and 4 (Table 6.5) will equal the total r^2 from the full model. This latter r^2 can be unambiguously partitioned into the independent contributions of each predictor and the analysis would be complete. We know, however, that correlations between predictors nearly always occur within real data sets so the

sum of the r^2 for Models 2, 3 and 4 will exceed the total r^2 from the full model because of the joint effects of predictors. These joint effects represent the variation in Y that is shared between two or more predictors. The joint effects for each predictor are calculated from the difference between the squared partial correlation for the model relating Y to that predictor and the average r^2 representing the independent contribution of that predictor already determined. This simply uses the additive nature of the independent and joint contributions of each predictor to the total r^2 for each predictor, as described above.

The sum of the average independent and average joint contribution to r^2 is the total contribution of each predictor variable to the variation in the response variable, measured by the r^2 for the model relating Y to each predictor. We might like to test the H_0 that this total contribution equals zero for each predictor. Unfortunately, hypothesis tests for r^2 are not straightforward, although Mac Nally (1996) suggested an expedient solution of using the appropriate critical value of the correlation coefficient ($\sqrt{r^2}$).

As Mac Nally (1996) has pointed out, hierarchical partitioning uses all possible models and averages the improvement in fit for each predictor variable, both independently and jointly, across all these models. Note that hierarchical partitioning does not produce a predictive model nor does it provide estimates of, and tests of null hypotheses about, parameters of the regression model. With anything more than a few predictors, hierarchical partitioning cannot be done manually and the algorithm of Chevan & Sutherland (1991) needs to be programmed.

Mac Nally (1996) illustrated the utility of hierarchical partitioning for a data set relating breeding passerine bird species richness to seven habitat variables. The two predictor variables retained by hierarchical partitioning were the same as those with significant bivariate correlations with the response variable but were quite different from those chosen by a full model multiple regression and variable selection (backwards and forwards) procedures (Box 6.8).

6.1.17 Other issues in multiple linear regression

Regression through the origin

We argued in Chapter 5 that forcing a regression model through the origin by omitting an intercept was rarely a sensible strategy. This is even more true for multiple regression because we would need to be sure that Y equals zero when all X_j equal zero. Even if this was the case, forcing our model through the origin will nearly always involve extrapolating beyond the range of our observed values for the predictor variables and measures of fit for no-intercept models are difficult to interpret.

Weighted least squares

Weighting each observation by a value related to the variance in y_i is one way of dealing with heterogeneity of variance although determining the appropriate weights is not straightforward (Chapter 5). As with simple linear regression, our preference is to transform Y and/or the X-variables if the heterogeneity of variance is due to skewed distributions of the variables, particularly if our understanding of the biology suggests a different scale of measurement is more appropriate for one or more of the variables. Alternatively, generalized linear models with an appropriate non-normal distribution of the error terms should be used (Chapter 13).

X random (Model II regression)

The extension of Model II bivariate regression techniques (Chapter 5) to the situation with multiple predictor variables was reviewed by McArdle (1988). To calculate the RMA equivalent estimates for each β_j, first produce a correlation matrix among all the variables (Y and all p X-variables). Then run a principal components analysis (see Chapter 17) on this correlation matrix and extract the eigenvector for the last component with the smallest eigenvalue (explained variance). The estimate of the regression slope for each predictor variable (X_j) is:

$$b_j = \frac{\alpha_j}{\alpha_Y} \qquad (6.30)$$

where b_j is the regression slope for X_j, α_j is the coefficient for X_j and α_Y is the coefficient for Y from the eigenvector for the principal component with the smallest eigenvalue. McArdle (1988) refers to this method as the standard minor axis (SMA) and simply becomes the RMA method when p equals one. Note that these are standardized regression slopes, because they are based on a correlation matrix, so the regression model does not include an intercept.

The choice between OLS and SMA is not as straightforward as that between OLS and RMA for simple bivariate regression. McArdle's (1988) simulations suggested that if the error variance in X_j is greater than about half the error variance in Y, then SMA is better. However, the relative performance of OLS and SMA depended on the correlation between Y and X_j so definitive guidelines cannot be given.

Robust regression

When the underlying distribution of error terms may not be normal, especially if extreme observations (outliers) occur in the data that we cannot deal with via deletion or transformation, then the usual OLS procedure may not be reliable. One approach is to use robust fitting methods that are less sensitive to outliers. The methods described in Chapter 5, least absolute deviations, Huber M-estimation and non-parametric (rank-based) regression, all extend straightforwardly to multiple predictor variables. The major difficulty is that the computations and associated algorithms are complex (Birkes & Dodge 1993). Fortunately, robust regression procedures are now common components of good statistical software.

The randomization test of the H_0 that β_1 equals zero in simple linear regression can also be extended to multiple regression. We compare the observed partial regression slopes to a distribution of partial regression slopes determined by randomly allocating the y_i to observations but not altering the x_{i1}, x_{i2}, etc., for each observation (Manly 1997). Other randomization methods can be used, including using the residuals, although the different methods appear to give similar results (Manly 1997).

Missing data

It is common for biological data comprising two or more variables to have missing data. In data sets suited to multiple regression modeling, we may be missing values for some of the predictor variables or the response variable for some sampling units. It is important to distinguish missing values (no data) from zero values (data recorded but the value was zero) – see Chapter 4. If missing values for the response variable reflect a biological process, e.g. some organisms died during an experiment and therefore growth rate could not be measured, then analyzing the pattern of missing values in relation to the predictor variables may be informative. More commonly, we have missing values for our predictor variables, often due to random events such as equipment failure, incorrect data entry or data being subsequently lost. In these circumstances, most linear models software will omit the entire sampling unit from analysis, even if data are only missing for one of the variables. Alternatives to deletion when missing data occur, including imputing replacement values, will be discussed in Chapter 15.

Power of tests

The tests of whether individual partial regression coefficients equal zero are based on t statistics and therefore the determination of power of these tests is the same as for any simple t test that a single population parameter equals zero (Chapters 3 and 7). Our comments on power calculations for simple regression analyses (Chapter 5) apply similarly for multiple regression.

6.2 Regression trees

An alternative to multiple linear regression analysis for developing descriptive and predictive models between a response variable and one or more predictor variables is regression tree analysis (Brieman *et al.* 1984, De'ath & Fabricius 2000). A "upside-down" tree is created where the root at the top contains all observations, which are divided into two branches at a node, then each branch is further split into two at subsequent nodes and so on. A branch that terminates without further branching is called a leaf.

Consider the data from Loyn (1987), where we have a continuous response variable (abundance of forest birds) and six predictor variables describing 56 forest patches, in this case all continuous.

All possible binary splits of the observations are assessed for each predictor variable. The first split is based on the predictor that results in two groups with the smallest within-group (residual) sums-of-squares for the response variable. Other measures of (lack of) fit can be used, including absolute deviations around the mean or median for a more robust measure of fit (see Chapter 5). These splitting criteria are different indices of impurity, a measure of heterogeneity of the groups at a split (De'ath & Fabricius 2000). This "recursive binary-partitioning" process is repeated within each of the two groups for all the predictors, again choosing the next split based on the predictor that results in the minimum residual SS within groups. Groups further along in the splitting process are more homogeneous than those higher up. The regression tree looks like a dendrogram from cluster analysis (Chapter 18), but is really a tree with the root (the undivided complete data set) at the top, branches with nodes for each division and leaves where branches terminate (terminal nodes).

Regression trees produce a predictive model. For any observation, a predicted value is the mean of the observations at a leaf, i.e. in a terminal group. Obviously, predicted values for observations in the one group (leaf) will be the same. This is in contrast to the usual linear model, which will have different predicted values for all observations unless they have identical values for all predictors. Because we have observed and predicted values, we can also calculate residuals for each observation and use these residuals as a diagnostic check for the appropriateness of the model and whether assumptions have been met. Normality of predictor variables is not a concern because only the rank order of a variable governs each split, although transformation of the response variable to alleviate variance heterogeneity may be important (De'ath & Fabricius 2000).

The splitting process (tree building) could continue until each leaf contains a single observation and for the Loyn (1987) data, we would have 56 terminal nodes. In this situation, the tree would predict the observed values of the response variable perfectly and explain all the variance in the response variable, the equivalent of fitting a saturated linear regression model (Section 6.1.4). Usually, we want the best compromise between tree simplicity (few nodes) and explained variance in the response variable. In practice, therefore, *a priori* stopping criteria are used, such as a maximum number of nodes allowed, a minimum number of objects in each group or a minimum reduction in explained variance from adding more nodes. Different software for building trees will use different measures of fit and different default stopping rules so don't expect trees based on the same data built using different programs to be the same unless these criteria are set to be the same. Once the tree is built, using the stopping criteria, we can also "prune" or "shrink" trees to produce simpler models that achieve a better compromise between fit and simplicity, often using criteria similar to those used for model selection in standard multiple regression (Section 6.1.15). Alternatively, we can assess the predictive capabilities of different sized trees and choose the "best" tree as the one with the smallest prediction error, i.e. the model that provides the most accurate predictions.

De'ath & Fabricius (2000) argue strongly that the best approach for determining prediction error and thus appropriate tree size is using cross-validation (Section 6.1.15; De'ath & Fabricius 2000). One method for cross-validation is where the observations are divided randomly into two groups of a specified size, e.g. 10% and 90% of the observations, and the regression tree model is fitted to the larger group ("training group") to predict values in the smaller group ("validation group"). The difference between the observed and predicted values of the response variable in the smaller group is a measure of prediction error. Of interest is how much of the total variation in the observed values of the response variable is explained by the predicted values. Cross-validation is usually repeated many times, each with a new random allocation of observations to the groups of pre-defined size, i.e. in a randomization testing framework. Randomization testing can also be used to test whether the derived regression tree explains more of the variation in the response variable than we would expect by chance. Brieman *et al.* (1984) and De'ath & Fabricius (2000) provide more detail on cross-validation for regression trees.

Regression trees are often included in statistical software under the acronym CART (classification and regression tree analyses). The main distinction between classification and regression trees is that the former is based on categorical response variables and the latter on continuous response variables. Two common algorithms are AID (Automatic Interaction Detection) for regression trees and CHAID (Chi-squared Automatic Interaction Detection) for classification trees.

We will use two biological examples of regression tree analysis. The first comes from Rejwan *et al.* (1999), who used both standard multiple regression and regression trees to analyze the relationship between the density of nests of smallmouth bass (continuous response variable) and four predictor variables (wind/wave exposure, water temperature, shoreline reticulation and littoral-floor rugosity) for 36 sites in Lake Opeongo, Canada. There were nonlinear relationships between both exposure and littoral-floor rugosity and nest density. The standard multiple regression analysis showed that shoreline reticulation, temperature and (temperature)^2, and exposure were significant predictors, the final model explaining 47% of the variation in nest density between sites. However, cross-validation analysis showed that the model had little predictive power, with almost none of the variation in nest density in random samples of 10% of the sites predictable from the model fitted to the other 90% of the sites.

Their regression tree analysis split the sites based on a temperature cut-off of 17.05 °C into two initial groups of 28 and 8 sites, and then split the latter group into two groups of four sites each based on shoreline reticulation below and above 100m. This tree explained 58% of the variation in nest density and cross-validation analysis showed that the tree model had more predictive power and could explain about 20% of the variation nest density in random samples of 10% of sites.

The second example, illustrated in Box 6.9, uses the data set from Loyn (1987), who recorded the abundance of forest birds in 56 forest fragments and related this response variable to six predictors that described aspects of each patch (area, distance to nearest patch and nearest larger patch, stock grazing, altitude and years since isolation) – see Box 6.2. We built a regression tree model for these data, after transforming area and the two distances to logs. The first split was between patches with grazing indices from one to four and those with a grazing index of five. This former group was further split into two groups with $\log_{10}$ area $\pm$ 1.176 (approx. 15 ha). The final tree is presented in Figure 6.6. This tree is a little different from the results of the multiple linear regression analysis of these data in Box 6.2. There, $\log_{10}$ area was a significant predictor, with grazing not significant ($P = 0.079$), although model selection and hierarchical partitioning both resulted in a model with $\log_{10}$ area and grazing as the two predictors (Box 6.8). The fit of the regression tree model was 0.699. The equivalent multiple linear regression model including just grazing and $\log_{10}$ area as predictors resulted in an r^2 of 0.653 so the regression tree model produced a slightly better fit.

This brief introduction might encourage you to explore these methods further. The standard reference is Brieman *et al.* (1984), and De'ath & Fabricius (2000) provide an excellent and up-dated overview with ecological applications.

6.3 Path analysis and structural equation modeling

The linear model we fit for a multiple regression represents our best guess at causal relationships. The model is postulating that the predictor variables we have incorporated may have biological effects on our response variable. The multiple regression model is, however, a conveniently simple representation of potential causal pathways among our variables as it only considers direct effects of each predictor, adjusting for the others, on the response variable. We may hypothesize much more complex causal links between variables. For example, we may include indirect effects where one predictor affects a second predictor, which in turn affects the response variable, and we may have two or more response variables that can affect each other. The statistical technique we use to analyze models of potential causal relationships was first developed over 50 years ago by Wright (1920, 1934) and is called path analysis

Box 6.9 Worked example of regression trees: abundance of birds in forest patches

A regression tree for the data from Loyn (1987) related the abundance of forest birds in 56 forest fragments to log area, log distance to nearest patch and nearest larger patch, grazing intensity, altitude and years since isolation. We used OLS as our measure of fit and set stopping criteria so that no split would result in less than five observations in a group, the maximum number of nodes was less than 20 (although this latter criterion turned out to be irrelevant) and the minimum proportional reduction in residual variance was 5%. The first node in the tree was between 43 habitat patches with grazing indices from one to four and the 13 patches with a grazing index of five (Figure 6.6). This former group was further split into two groups, 24 patches with $\log_{10}$ area less than 1.176 (approx. 15 ha) and 19 patches with $\log_{10}$ area greater than 1.176.

The fit of this tree model was 0.699. The plot of residuals from the tree model is shown in Figure 6.8(a) with four observations in the group of small patches with low grazing (less than five) standing out from the others and warranting checking and possibly re-running the analysis after their omission to evaluate their influence.

Out of interest, we refitted the tree with looser stopping criteria (smaller allowable reduction in residual variance) to see what subsequent splits in the data would have occurred (Figure 6.7). On one side of the tree, the 13 patches with a grazing index of five were further split by $\log_{10}$ dist. On the other side, the 24 small patches were further split by age (and then by $\log_{10}$ area and $\log_{10}$ dist) and the 19 larger patches were further split by $\log_{10}$ area again. The fit of the model was improved to 0.84 but the model is much more complex with additional variables, some repeated throughout the tree (e.g. $\log_{10}$ area) so the improvement in fit is at least partly a consequence of the increased number of predictors in the tree model. The residuals show a more even pattern, with no obvious outliers (Figure 6.8(b)).

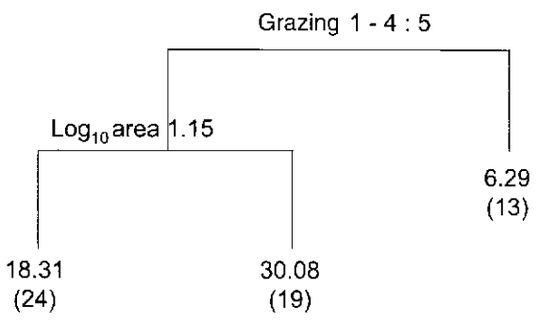

Figure 6.6 Regression tree modeling bird abundance in forest patches against patch area, distance to nearest patch, distance to nearest larger patch (these three variables $\log_{10}$ transformed), grazing intensity, altitude, and years since isolation for the 56 patches surveyed by Loyn (1987). The criteria for each node are included, with left-hand branches indicating observations with values for that predictor below the cut-off and right-hand branches indicating observations with values for that predictor above the cut-off. The predicted value (mean) and number of observations for each leaf (terminal group) are also provided.

(see also Mitchell 1993 for a review). Path analysis was originally designed for simple multiple regression models and is now considered a subset of a more sophisticated collection of analytical tools called structural equation modeling (SEM), also called analysis of covariance (correlation) structure (Tabachnick & Fidell 1996). It is very important to remember that causality can only really be demonstrated by carefully designed and analyzed manipulative experiments, not by any specific statistical procedure. SEM and path analysis are basically analyses of correlations, although

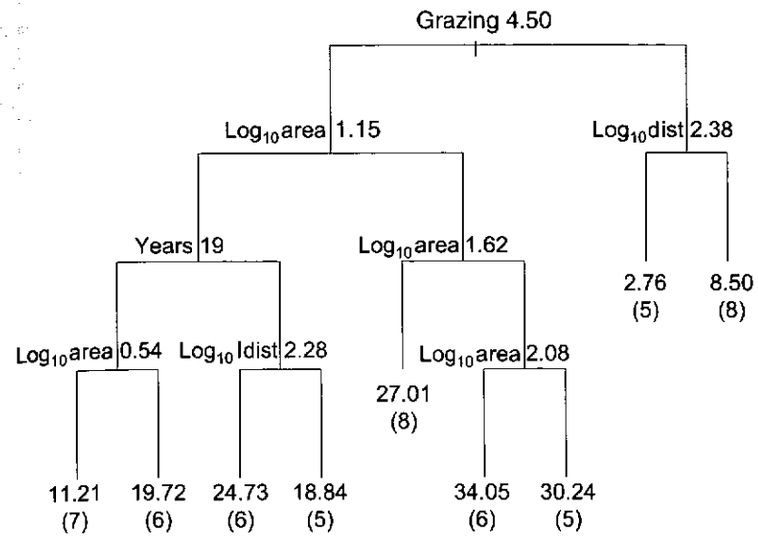

Figure 6.7 Regression tree based on the same data as in Figure 6.6, except that branching was continued to a lower level.

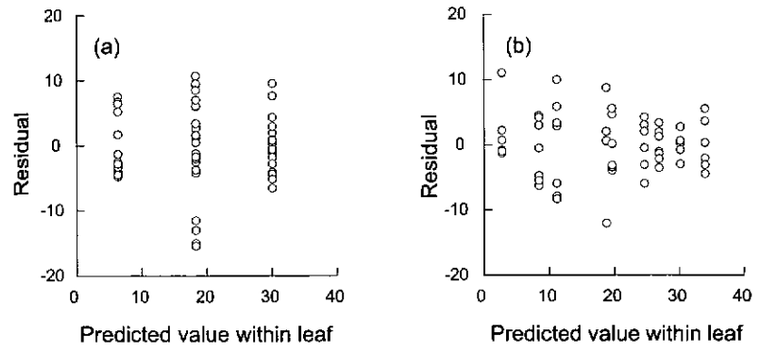

Figure 6.8 Plots of residuals against predicted values from (a) the regression tree model illustrated in Figure 6.6 and (b) the model illustrated in Figure 6.7.

they can be used to analyze experimental data (Smith *et al.* 1997), and simply test how well postulated causal pathways fit the observed data in a modeling context.

The fundamental component of SEM or path analysis is the *a priori* specification of one or more causal models, although most published applications of path analysis in biology do not seem to compare competing models. Let's consider a simple path diagram, based on the data from Loyn (1987), that relates the abundance of forest birds in isolated patches of remnant forest to a number of predictor variables (Figure 6.9). We will include three of these predictors ($\log_{10}$ patch area, years since isolation, grazing) and include all correlations among the predictors and all supposed causal links between each predictor and

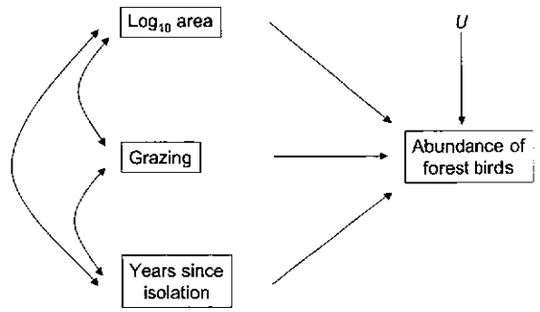

Figure 6.9 Path diagram for simple multiple regression model relating three predictor variables ($\log_{10}$ patch area, grazing, years since isolation) to one response variable (abundance of forest birds) using the data from Loyn (1987).

the response variable in our path diagram. Single-headed arrows represent supposed causal links between variables and double-headed arrows represent correlations between variables with no directional causality postulated. U represents unexplained causes (variables we have not measured) that might affect a response variable.

The process starts by specifying the model for each response variable. In our simple example, there is only one response variable and the model is a standardized multiple regression model without an intercept:

$$(\text{bird abundance})_i = \beta_1(\log_{10}\text{area})_i + \beta_2(\text{years})_i + \beta_3(\text{grazing})_i + \varepsilon_i \quad (6.31)$$

Path analyses basically represent a restructuring of the correlations (or covariances) between all the variables under consideration (Mitchell 1993). The correlation (r_{jY}) between any predictor variable X_j and the response variable Y can be partitioned into two components: the direct and the indirect effects (Mitchell 1993). This partitioning simply represents the normal equations that we used for fitting the regression model using OLS (Box 6.3). The direct effect is measured by the standardized partial regression coefficient between Y and X_j, holding all other predictor variables constant. This direct effect is now the path coefficient relating Y to X_j. Path coefficients are identical to standardized regression coefficients if all correlations between predictor variables are included in our path diagram. The indirect effect is due to the correlations between X_j and the other predictors, which may in turn have direct effects on Y.

Mathematically, this decomposition of the correlations can be derived from the set of normal equations used for estimating the parameters of the multiple regression model (Petraitis *et al.* 1996). For example, for predictor variable one ($\log_{10}$ area):

$$r_{1Y} = b_1 + r_{12}b_2 + r_{13}b_3 \quad (6.32)$$

where r represents simple correlations and b represents standardized partial regression coefficients.

For the Loyn (1987) data:

$$r_{\log_{10}\text{area.abundance}} = b_{\log_{10}\text{area.abundance}} + r_{\log_{10}\text{area.years}}b_{\text{years.abundance}} + r_{\log_{10}\text{area.grazing}}b_{\text{grazing.abundance}} \quad (6.33)$$

The direct effect of $\log_{10}$ area on bird abundance is represented by the standardized regression slope. The indirect effect of $\log_{10}$ area on bird abundance via the former's correlation with years since isolation and with grazing is calculated from the sum of the last two terms in the right hand side of Equation 6.33 above. The correlations between years since isolation and bird abundance and between grazing and bird abundance can be similarly decomposed into direct and indirect effects. The path identified by U (unexplained effects) can be determined from $\sqrt{(1-r^2)}$ from the fit of the model for a given response variable (Mitchell 1993). The results are summarized in Box 6.10 and Figure 6.9.

Complex path models, with multiple response variables, are not as easily handled by the multiple regression approach to path analysis we have just described (Mitchell 1992). More sophisticated forms of structural equation modelling, such as those implemented in software based on CALIS (Covariance Analysis of Linear Structural equations; in SAS) and LISREL (Linear Structural Relations; in SPSS) algorithms, offer some advantages, especially in terms of model testing and comparison. These procedures estimate the path coefficients and the variances and covariances of the predictor variables simultaneously from the data using maximum likelihood, although other estimation methods (including OLS) are available (Tabachnick & Fidell 1996). A covariance matrix is then determined by combining these parameter estimates and this covariance matrix is compared to the actual covariance matrix based on the data to assess the fit of the model. Most software produces numerous measures of model fit, the AIC (see Section 6.1.15) being one of the preferred measures. As pointed out by Mitchell (1992) and Smith *et al.* (1997), such goodness-of-fit statistics can only be determined when there are more correlations between variables than there are coefficients being estimated, i.e. the model is over-identified. For example, we cannot test the fit of the path model in Figure 6.9 because we have estimated all the direct and indirect effects possible, i.e. there are no unestimated correlations. The number of unestimated correlations contributes to the df of the goodness-of-fit statistic (Mitchell 1993).

Box 6.10 Worked example of path analysis: abundance of birds in forest patches

We will use the data from Loyn (1987) to relate the abundance of forest birds in isolated patches of remnant forest to three predictor variables: $\log_{10}$ patch area, years since isolation, grazing (Figure 6.9). Our path model includes all correlations among the predictors and all supposed causal links between each predictor and the response variable. The path model outlined in Figure 6.9 was evaluated by calculating both direct and indirect effects of predictors on the response variable. The full correlation matrix was as follows.

	Abundance	Log_{10} area	Years	Grazing
Abundance	1.000			
Log_{10} area	0.740	1.000		
Years	−0.503	−0.278	1.000	
Grazing	−0.683	−0.559	0.636	1.000

The direct and indirect effects for $\log_{10}$ area were calculated from:

$$r_{\log_{10}\text{area.abundance}} = b_{\log_{10}\text{area.abundance}} + r_{\log_{10}\text{area.years}}b_{\text{years.abundance}} + r_{\log_{10}\text{area.grazing}}b_{\text{grazing.abundance}}$$

where $b_{\log_{10}\text{area.abundance}}$ is the direct effect of $\log_{10}$ area on abundance (the partial regression coefficient), $r_{\log_{10}\text{area.years}}b_{\text{years.abundance}}$ is the indirect effect of $\log_{10}$ area on abundance via years and $r_{\log_{10}\text{area.grazing}}b_{\text{grazing.abundance}}$ is the indirect effect of $\log_{10}$ area on abundance via grazing. Equivalent equations were used for the other predictors. Correlations between predictor variables were also calculated. The final results were as follows.

Predictor	Direct effects	Indirect effects	Total effects
Log_{10} area	0.542	0.198	0.740
via years		0.542	
via grazing		0.146	
Years since isolation	−0.187	−0.317	−0.503
via $\log_{10}$ area		−0.151	
via grazing		−0.166	
Grazing	−0.261	−0.422	−0.683
via $\log_{10}$ area		−0.303	
via years		−0.119	

It is clear that the "effect" of $\log_{10}$ area on bird abundance is primarily a direct effect whereas the "effects" of grazing and years since isolation are primarily indirect through the other predictors. Our use of quotation marks around "effect" here emphasizes that this is simply a correlation analysis; attributing causality to any of these predictor variables can only be achieved by using manipulative experiments. The r^2 for this model is 0.673 so the coefficient of the path from U to bird abundance is 0.572.

These programs also allow for latent (unmeasured) variables, which are unfortunately termed factors in the SEM literature. Latent variables are not commonly included in path models in the biological literature, although Kingsolver & Schemske (1991) discussed the inclusion of unmeasured phenotypic factors in path analyses of selection in evolutionary studies. The difficulty with these sophisticated SEM programs is they are more complex to run. For example, LISREL requires that a number of matrices be specified, representing the variances, covariances and relationships between variables. Detailed comparisons of these different programs, including required input and interpretation of the output, are available in Tabachnick & Fidell (1996)

The limitations and assumptions of classical path analysis are the same as those for multiple regression. The error terms from the model are assumed to be normally distributed and independent and the variances should be similar for different combinations of the predictor variables. Path analysis will also be sensitive to outliers and influential observations, and missing observations will have to be addressed, either by replacement or deletion of an entire observation (see Chapters 4 and 15). Collinearity among the predictor variables can seriously distort both the accuracy and precision of the estimates of the path coefficients, as these are simply partial regression coefficients (Petraitis *et al.* 1996; Section 6.1.11). There is still debate over whether more sophisticated SEM techniques, such as those based on LISREL, are more robust to these issues (Petraitis *et al.* 1996, Pugusek & Grace 1998). Diagnostics, such as residual plots, should be an essential component of any path analysis. Irrespective of which method is used, all estimates of path coefficients are sensitive to which variables are included or which coefficients (correlation or path) are set to zero (Mitchell 1992, Petraitis *et al.* 1996). This is no different to multiple regression, where estimates of partial regression slopes are sensitive to which predictors are included or not.

Finally, we repeat our earlier caution that, although structural equation modeling analyzes postulated causal relationships, it cannot "confirm or disprove the existence of causal links" (Petraitis *et al.* 1996 p. 429). Such causal links can only be demonstrated by manipulative experiments. SEM and path analyses do allow complex linear models to be evaluated and path diagrams provide a useful graphical representation of the strengths of these relationships.

6.4 | Nonlinear models

When the relationship between Y and X is clearly curvilinear, there are a number of options. We have already discussed using a polynomial model (Section 6.1.13) or linearizing transformations of the variables (Section 6.1.10), but these are not always applicable. For example, the relationship between Y and X might be complex and cannot be approximated by a polynomial nor can it be linearized by transformations of the variables. The third option is to fit a model that is nonlinear in the parameters. For example, the relationship between number of species (S) and island area (A) can be represented by the power function:

$$S = \alpha A^{\beta} \tag{6.34}$$

where α and β are the parameters to be estimated (Loehle 1990) – see Box 6.11. This is a two parameter nonlinear model. A three parameter nonlinear model which is very useful for relating a binary variable (e.g. presence/absence, alive/dead) to an independent variable is the logistic model:

$$Y = \frac{\alpha}{1 + e^{(\beta - \delta X)}} \tag{6.35}$$

where α, β and δ are the parameters to be estimated. Ratkowsky (1990) has described a large range of multiparameter nonlinear models, both graphically and statistically, and some of their practical applications.

OLS or ML methods can be used for estimation in nonlinear regression modeling, as we have described for linear models. The OLS estimates of the parameters are the ones that minimize the sum of squared differences between the observed and fitted values and are determined by solving a set of simultaneous normal equations. Solving these equations is much trickier than in linear models and some sort of iterative search procedure is required, whereby different estimates are tried in a sequential fashion. Obviously, with two or more parameters, the number of possible combinations of values for the parameters is essentially infinite so these searching procedures are sophisticated in that they only try values that improve the fit of the model (i.e. reduce the $SS_{Residual}$).

The most common method is the Gauss–Newton algorithm or some modification of it (Myers 1990). Starting values of the parameters must be provided and these are our best guess of what the values of the parameters might be. The more complex the model, the more important it is for the starting values to be reasonably close to the real parameter values. Starting values may come from fits of the equivalent model to other, similar, data (e.g. from the published literature),

Box 6.11 | Worked example of nonlinear regression: species richness of macroinvertebrates in mussel clumps

As described in Chapter 5, Peake & Quinn (1993) collected 25 clumps of an intertidal mussel from a rocky shore at Phillip Island in Victoria. The relationship between the number of species (Y) per clump and clump area in m^2 (X) was examined. The scatterplot suggested a nonlinear relationship between number of species and clump area (Figure 5.17) and theory suggests that a power function might be appropriate:

$$\text{species} = \alpha(\text{area})^{\beta}$$

This power function was fitted using a modified Gauss–Newton method (quasi-Newton). No starting values were provided. The algorithm took six iterations to converge on the following estimates, with their approximate standard errors.

Parameter	Estimate	Standard error	t	P
α	18.540	0.630	29.449	<0.001
β	0.334	0.035	9.532	<0.001

The $MS_{Residual}$ was 7.469. The fitted model was, therefore:

$$\text{species} = 18.540(\text{area})^{0.334}$$

Note that the $MS_{Residual}$ for the nonlinear power function (7.469) is about half that for a linear model (14.133), indicating the former is a better fit to the data. The fitted model is shown in Figure 6.10(a) and the residual plot (Figure 6.10(b)) suggested no strong skewness in the response variable and there were no unusual outliers.

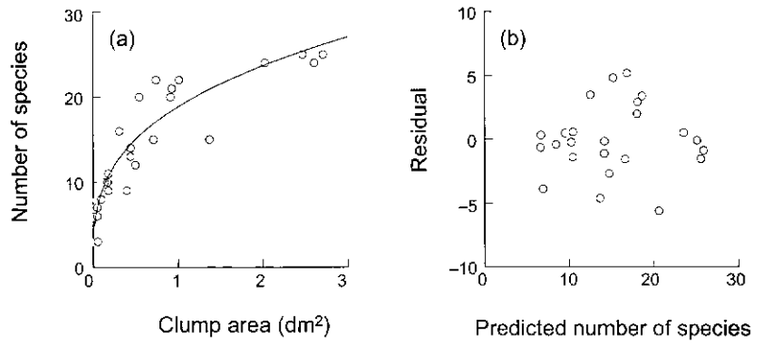

Figure 6.10 (a) Plot of number of species against mussel clump area from Peake & Quinn (1993) showing fitted nonlinear model: number of species = $18.540 \times (\text{area})^{0.334}$. (b) Plot of residuals against predicted values (with boxplots) from fitted nonlinear model in (a) fitted to number of species against mussel clump area from Peake & Quinn (1993).

theoretical considerations or, for relationships that can be linearized by transformation, back-transformed values from a linear model fitted to transformed data. The Gauss–Newton search method is complex, using partial derivatives from the starting values and *X*-values to fit an iterative series of essentially linear models and using OLS to estimate the parameters. The best estimates are reached when the sequential iterations converge, i.e. don't change the estimates by very much. Variances and standard errors for the parameter estimates can be determined; the calculations are tedious but most statistical software provides this information. Confidence intervals, and *t* tests for null hypotheses, about parameters can also be determined (Box 6.11).

There are a number of difficulties with nonlinear modeling. First, sometimes the iterative Gauss–Newton procedure won't converge or converges to estimates that are not the best possible ("local minimum"). Most statistical software use modified Gauss–Newton procedures, which help convergence, and choosing realistic starting values is very important. It is usually worth refitting nonlinear models with different starting values just to be sure the final model can be achieved consistently. Second, OLS works fine for linear models if the errors (residuals) are independent, normally distributed with constant variance; however, for nonlinear models, even when these assumptions are met, OLS estimators and their standard errors, and confidence intervals and hypothesis tests for the parameters, are only approximate (Myers 1990; Rawlings *et al.* 1998). We can be more certain of our estimates and confidence intervals if different combinations of search algorithms and starting values produce similar results. Finally, measuring the fit of nonlinear models to the data is tricky; r^2 cannot be easily interpreted because the usual SS_{Total} for the response variable cannot always be partitioned into two additive components ($SS_{Regression}$ and $SS_{Residual}$). Comparing different models, some of which might be nonlinear, can only be done with variables measured on the same scale (i.e. untransformed; see Chapter 5) and the $MS_{Residual}$ is probably the best criterion of fit.

Once a nonlinear model has been estimated, diagnostic evaluation of its appropriateness is essential. Residuals can be calculated in the usual manner and large values indicate outliers. Because OLS estimation is commonly used for nonlinear models, assumptions of normality, homogeneity of variance and independence of the error terms from the model are applicable. Boxplots of residuals and scatterplots of residuals against predicted values (Figure 6.10) can detect problems with these assumptions as described for linear models. Other estimation methods, such as maximum likelihood, might be more robust than OLS.

For simple nonlinear structures, transforming the variables to achieve linearity is usually recommended, particularly if the transformed variables can be easily interpreted because the transformed scale is a natural alternative scale of measurement for that variable. Note that the transformed model is not the same as the untransformed nonlinear model, in the same way that a *t* test on untransformed data is not testing the same H_0 as a *t* test on the same data transformed. Our parameter estimates from the transformed model cannot easily be interpreted in terms of the original nonlinear model, which may have the stronger theoretical basis.

6.5 Smoothing and response surfaces

The linear plane representing the linear regression model of *Y* against X_1 and X_2 illustrated in Figure 6.1 is sometimes referred to as a response surface, a graphical representation of the relationship between a response variable and two predictors. Response surfaces obviously also exist when there are more than two predictors but we cannot display them graphically. Response surfaces, in this graphical context, are often used to display the model chosen as the best fit based on the model-fitting techniques we have already described. Additionally, exploring a range of response surfaces may help decide what sort of model is best to use and detect patterns we might have missed by being restricted to a specific model.

Model-based surfaces that are linear in parameters include linear and curvilinear relationships.

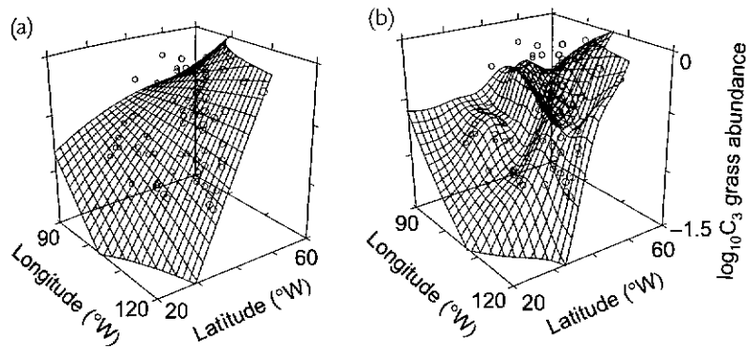

Figure 6.11 Response surfaces relating log-transformed relative abundance of C_3 plants to latitude and longitude for 73 sites in North America (Paruelo & Lauenroth 1996). (a) Quadratic model fitted, and (b) distance-weighted least squares (DWLS) fitted.

For example, polynomial models (quadratic, cubic, etc.) are often good approximations to more complex relationships and provide a more realistic representation of the relationship between *Y* and X_1 and X_2 than a simple linear model. Figure 6.11(a) shows a quadratic response surface, representing a model including linear and quadratic terms for both predictors as well as their interaction, fitted to the data from Paruelo & Lauenroth (1996). Note that compared with the first-order linear model in Figure 6.1, the quadratic model allows a hump-shaped response of log-transformed C_3 plant abundance to longitude for a given latitude. The choice of whether to use this response surface would depend on the results of fitting this model compared with a simpler first-order model.

Smoothing functions, like we discussed in Chapter 5, can sometimes also be applied to three-dimensional surfaces. While the Loess smoother cannot easily be extended to three dimensions, DWLS can and allows a flexible exploration of the nature of the relationship between *Y* and X_1 and X_2 unconstrained by a specific model. For the data from Paruelo & Lauenroth (1996), the DWLS surface (Figure 6.11(b)) suggests a potentially complex relationship between log transformed C_3 plant abundance and longitude in the northern, high latitude, sites, a pattern not revealed by the linear or polynomial models. Note that, like the bivariate case, parameters for these smoothing functions cannot be estimated because they are not model-based; they are exploratory only.

Response surfaces also have other uses. For example, comparing the fitted response surfaces for linear models with and without an interaction between two predictors can help interpret the nature of such an interaction. Again for the data from Paruelo & Lauenroth (1996), the DWLS smoothing function suggests that the relationship between log-transformed abundance of C_3 plants and latitude depends on longitude and vice versa (Figure 6.11(b)). Most statistical software can plot a range of model-based and smoothing response surfaces on three-dimensional scatterplots.

6.6 General issues and hints for analysis

6.6.1 General issues

- Multiple regression models are fitted in a similar fashion to simple regression models, with parameters estimated using OLS methods.
- The partial regression slopes in a multiple regression model measure the slope of the relationship between *Y* and each predictor, holding the other predictors constant. These relationships can be represented with partial regression plots.
- Comparisons of fit between full and reduced models, the latter representing the model when a particular H_0 is true, are an important method for testing null hypotheses about model parameters, or combinations of parameters, in complex models.
- Standardized partial regression slopes should be used if the predictors and the response variable are measured in different units.
- Collinearity, correlations between the predictor variables, can cause estimates of parameters to be unstable and have artificially large

variances. This reduces the power of tests on individual parameters.
- Interactions between predictors should be considered in multiple regression models and multiplicative models, based on centered predictors to avoid collinearity, should be fitted when appropriate.
- Hierarchical partitioning is strongly recommended for determining the relative independent and joint contribution of each predictor to the variation in the response variable.
- Regression trees provide an alternative to multiple linear models for exploring the relationships between response and predictor variables through successive binary splits of the data, although cross-validation is necessary for evaluation of predictive power and hypothesis testing.
- Path analysis can be a useful technique for graphically representing possible causal links between response and predictor variables, and also between predictor variables themselves.
- Nonlinear models can be fitted using OLS, although the estimation procedure is more complex. The trick is deciding *a priori* what the most appropriate theoretical model is.

6.6.2 Hints for analysis

- Multiple regression analyses are sensitive to outliers and influential values. Plots of residuals and Cook's D_i statistic are useful diagnostic checks.
- Information criteria, such as Akaike's (AIC) or Schwarz's (BIC) are the best criteria for distinguishing the fit of different models, although $MS_{Residual}$ is also applicable for regression models fitted using OLS.
- Avoid automated selection procedures (forward, backward, etc.) in model fitting. Their results are inconsistent and hard to interpret because of the large number of significance tests. For moderate numbers of predictors, compare the fit of all possible models.
- Use simple slopes for further interpretation of interactions between predictor variables in multiple regression models.
- Causality can only be demonstrated by careful research and experimentation, not by a particular statistical analysis. For example, path analysis is a method for summarizing correlation structures among variables and cannot show causality.
- Always examine scatterplots and correlations among your variables, to detect nonlinear relationships but also to detect collinearity among predictors. Tolerance (or the variance inflation factor) will also indicate collinearity. Choose which predictor variables to include in the final model carefully, avoiding variables that are highly correlated and measuring a similar quantity.

Chapter 7

Design and power analysis

7.1 Sampling

Fundamental to any statistical analysis, including the regression models we described in the previous two chapters, is the design of the sampling regime. We are assuming that we can clearly define a population of interest, including its spatial and temporal boundaries, and that we have chosen an appropriate type and size of sampling unit. These units may be natural units (e.g. stones, organisms, and lakes) or artificially delineated units of space (e.g. plots or quadrats). Our aim is to design a sampling program that provides the most efficient (in terms of costs) and precise estimates of parameters of the population. It is important to remember that we are talking about a statistical population, all the possible sampling or experimental units about which we wish to make some inference. The term population has another meaning in biology, a group of organisms of the same species (Chapter 2), although this might also represent a statistical population of interest.

We will only provide a brief overview of some sampling designs. We recommend Levy & Lemeshow (1991), Manly (2001) and Thompson (1992), the latter two having more of a biological emphasis, as excellent references for more detail on the design of sampling programs and using them to estimate population parameters.

7.1.1 Sampling designs

Simple random sampling was introduced in Chapter 2 and is where all the possible sampling units in our population have an equal chance of being selected in a sample. Technically, random sampling should be done by giving all possible sampling units a number and then choosing which units are included in the sample using a random selection of numbers (e.g. from a random number generator). In practice, especially in field biology, this method is often difficult, because the sampling units do not represent natural distinct habitat units (e.g. they are quadrats or plots) and cannot be numbered in advance or because the sampling units are large (e.g. 20 m^2 plots) and the population covers a large area. In these circumstances, biologists often resort to "haphazard" sampling, where sampling units are chosen in a less formal manner. We are assuming that a haphazard sample has the same characteristics as a random sample.

The formulae provided in Chapter 2 for estimating population means and variances, standard errors of the estimates and confidence intervals for parameters assume simple random sampling. If the size of the total population of sampling units is finite, then there are correction factors that can be applied to the formulae for variances and standard errors, although many populations in biological research are essentially infinite.

You can't really go wrong with simple random sampling. Estimates of the parameters of the population, especially the mean, will be ML estimators and generally unbiased. The downside of simple random sampling is that it may be less efficient than other sampling designs, especially when there is identified heterogeneity in the population or we wish to estimate parameters at a range of spatial or temporal scales.