5 Linear regression

In Chapter 4, we used various graphical tools (Cleveland dotplots, boxplots, histograms) to explore the shape of our data (normality), look for the presence of outliers, and assess the need for data transformations. We also discussed more complex methods (coplot, lattice graphs, scatterplot, pairplots, conditional boxplots, conditional histograms) that helped to see the relationships between a single response variable and more than one explanatory variable. This is the essential first step in any analysis that allows the researcher to get a feel for the data before moving on to formal statistical tools such as linear regression.

Not all datasets are suitable for linear regression. For data with counts or presence–absence data, generalised linear modelling (GLM) is more suitable. And where the parametric models used by linear regression and GLM give a poor fit, non-parametric techniques like additive modelling and generalised additive modelling (GAM) are likely to give better results. In this book we look at a range of tools suitable for analysing the univariate data commonly found in ecological or environmental studies, including linear regression, partial linear regression, GLM, additive modelling, GAM, regression and classification trees, generalised least squares (GLS) and mixed modelling. Techniques like GLM, GAM, GLS and mixed modelling are difficult to understand and even more difficult to explain. So we start by briefly summarising the underlying principles of linear regression, as this underpins most of the univariate techniques we are going to discuss. However, if you feel the need for more than a brief refresher then have a look at one of the many standard texts. Useful starting points are Fowler et al. (1998) and Quinn and Keough (2002), with more detailed discussions found in Montgommery and Peck (1992) and Draper and Smith (1998).

5.1 Bivariate linear regression

In Chapter 27 a detailed analysis of the RIKZ data is presented. Abundances of around 75 invertebrate species from 45 sites were measured on various beaches along the Dutch coast. In this study, the variable "NAP" measured the height of the sample site compared with average sea level, and indicated the time a site is under water. A site with a low NAP value will spend more time under water than a site with a high NAP value, and sites with high NAP values normally lie further up the beach. The tidal environment creates a harsh environment for the animals living there, and it is reasonable to assume that different species and species abun-

dances will be found in beaches with different NAP values. A simple starting point is therefore to compare species diversity (species richness) with the NAP values from different areas of the beach. Although ecological knowledge suggests the relationship between species richness and NAP values is unlikely to be linear, we start with a bivariate linear regression model using only one explanatory variable. It is always better to start with a simple model first, only moving on to more advanced models when this approach proves inadequate.

The first step in a regression analysis is a scatterplot. Panel A in Figure 5.1 shows the scatterplot of species richness versus NAP. The scatter of points suggests that a straight line might provide a reasonable fit for these data. Panel B in Figure 5.1 shows the same scatterplot, but with a fitted regression line added. The slope and direction of the line suggests there is a negative linear relationship between richness and NAP. The slope is significantly different from 0 at the 5% level, and this suggests that the relationship between species richness and NAP values is significant. A histogram of the residuals suggests the residuals are approximately normally distributed, suggesting the data have a Gaussian or normal distribution, and the Cook distance function (which we explain later) shows there are no influential observations. Later in this chapter, we discuss the difference between influential and extreme observations. For some researchers, this evidence would be enough to decide there is a significant negative relationship between richness and NAP and allow them to consider the analysis complete. However, it is more complicated than this, and to understand why, we need to look at some linear regression basics.

Back to the basics

Figure 5.2-A shows the same scatterplot of richness versus NAP we used earlier, but this time, to keep the figure simple, we have only used 7 of the 45 available samples. The species richness values are essentially the result of a random process. It is random because we would expect to find a different range of values every time the field sampling was repeated, provided the environmental conditions were the same. To illustrate this randomness, Figure 5.2-B shows 30 simulated values of the same data. The large dots represent the observed values, and the small dots represent the simulated values (other realisations). However, in practise we do not have these extra observations, and therefore we need to make a series of assumptions before using a linear regression model.

At this point, we need to introduce some mathematical notation. Let Y_i be the value of the response variable (richness) at the i^{th} site, and X_i the value of the explanatory variable (NAP) for the same site.

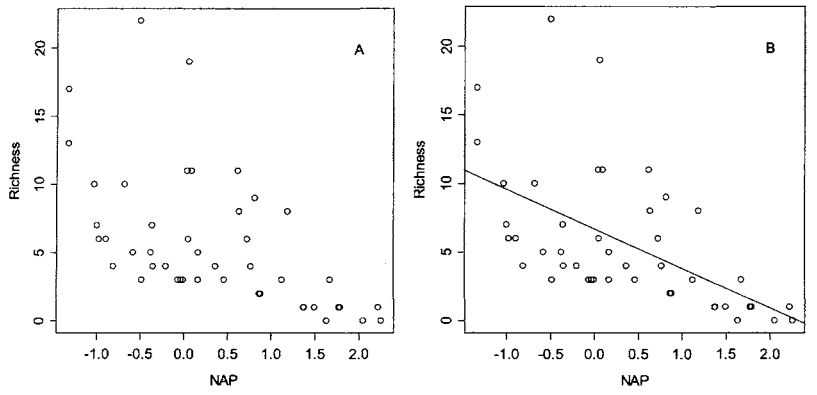

Figure 5.1. A: scatterplot of species richness versus NAP for the RIKZ data. B: scatterplot and regression line for RIKZ data.

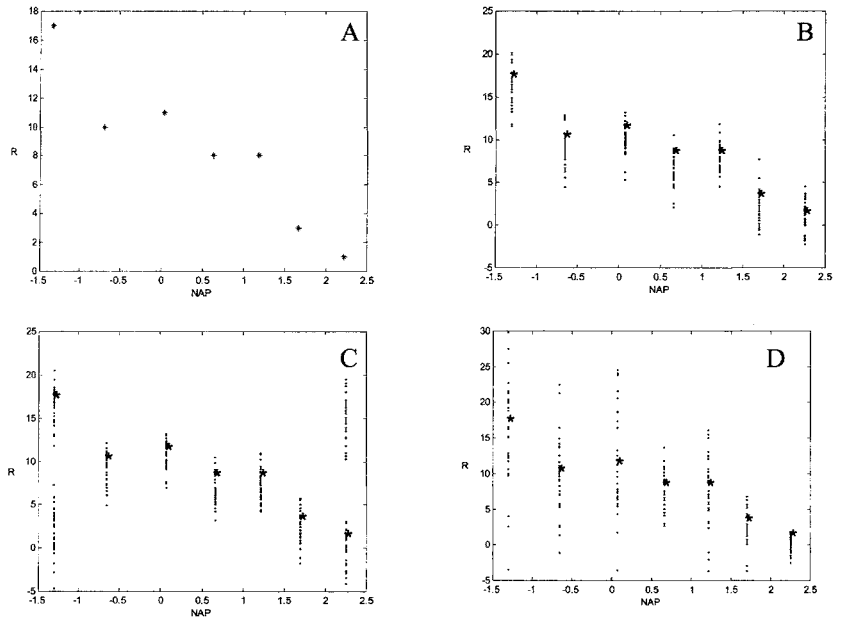

Figure 5.2. A: Scatter plot for NAP and species richness, B: Possible values of the same process. C: Violation of the normality and homogeneity assumptions. D: Violation of homogeneity assumption.

The bivariate (meaning two variables) linear regression model is given by

$$Y_i = \alpha + X_i\beta + \varepsilon_i \tag{5.1}$$

where α is the population intercept, β is the population slope, and ε_i is the residual, or the information that is not explained by the model. This model is based on the entire population, but as explained above, we only have a sample of the population, and somehow we need to use this sample data to estimate the values of α and β for the whole population. To do this, we need to make four assumptions about our data that will allow a mathematical procedure to produce estimated values for α and β. These estimators, called a and b, based on the sample data then act as estimators for their equivalent population parameters, α and β, respectively. The four assumptions that allow the sample data to be used to estimate the population data are (i) normality, (ii) homogeneity, (iii) independence and (iv) fixed X.

Normality

The normality assumption means that if we repeat the sampling many times under the same environmental conditions, the observations will be normally distributed for each value of X. We have illustrated this in the upper right panel of Figure 5.2 where observations at the same value of X are centred around a specific Y value. In a three-dimensional space you can imagine them as bell-shaped histograms. The observed value (indicated by a *) does not need to be exactly in the middle of the realisations at each X value. Figure 5.2-D shows another example, where the observations at each value of X are fairly evenly distributed around a middle value suggesting a normal distribution. Figure 5.2-C is a clear violation of the normality assumption as there are two distinct clouds of observations for some of the X values.

Up to now, we have discussed the normality assumption in terms of Y. However, because we have multiple Y observations for every X value, we also have multiple residuals for every value of X. As X is assumed to be fixed (see below), the assumption of normality implies that the errors ε are normally distributed (at each X value) as well. Using the normality assumption, the bivariate linear regression model is written as

$$Y_i = \alpha + X_i\beta + \varepsilon_i \quad \text{where } \varepsilon_i \sim N(0, \sigma_i^2) \tag{5.2}$$

The notation $N(0, \sigma_i^2)$ stands for a Normal distribution with expectation 0 and variance σ_i^2. If σ_i^2 is relatively large, then there is a large amount of unexplained variation in the response variable. Note that σ_i^2 has an index i.

Homogeneity

To avoid estimating a variance component for every X value, we assume the variance for all X values is the same. This is called the homogeneity assumption. This is important and implies that the spread of all possible values of the population is the same for every value of X. As a result, we have

$$\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \ldots = \sigma^2$$

And instead of having to estimate many variances, we only have to estimate one. Figure 5.2-B shows an example of this assumption where the spread of values for Y is the same at each value of X. Figure 5.2-C, however, shows a much wider variation at two values of X. In Figure 5.2-D, we simulated an example in which sites with low richness have a much smaller variation.

In ecological field studies, violation of the homogeneity assumption is common, as organisms often have a clumped distribution.

Independence

This assumption means the Y values for one observation (X_i) should not influence the Y values for other observations (X_j), for $i \neq j$. Expressed differently, the position of a realisation (or residual) at a particular X value should be independent of realisations at other X values. Clear violations of this assumption are time series data and spatial data. For example, when sampling vegetation over time, the vegetation present at the initial sampling period is likely to strongly influence the vegetation present at subsequent sampling periods. We will revisit independence later.

Fixed X

The term 'fixed' means that X is not random, and 'not random' implies that we know the exact values of X and there is no noise involved. In ecological field studies, X is hardly ever fixed, and it is normally assumed that any measurement error in X is small compared with the noise in Y. For example, using modern GPS, determining NAP can be reasonably accurate. An example of where X can be considered as fixed, is measurements of toxic concentrations in mesocosm studies in eco-toxicology. Other examples are nominal variables such as time and transects. Chapter 5 in Faraway (2004) contains a nice discussion on this topic.

The regression model

The four assumptions above give the linear regression model:

$$Y_i = \alpha + X_i \beta + \varepsilon_i \quad \text{where } \varepsilon_i \sim \text{N}(0, \sigma^2) \tag{5.3}$$

Note the variance term no longer contains an index i. The population parameters α, β and σ^2 (population variance) are estimated by a, b and s^2 (sample variance). The underlying mathematical tool for this is either ordinary least squares (OLS) or maximum likelihood estimation. OLS finds parameters a and b, based on minimising the residual sum of squares (*RSS*), where *RSS* is defined by $RSS = \Sigma_i e_i^2$ and $e_i = Y_i - a - X_i b$. We used an e instead of an ε to indicate that it is sample data, and not population data. Figure 5.3 shows the residuals, together with the observed and fitted values. The residual error (e_i) is the difference between the observed point Y_i and the fitted value. Note that β represents the change in the Y for a

1-unit change in the X, and α is the expected Y value if $X = 0$. Of main interest is the β as it measures the strength of the relationship between Y and X.

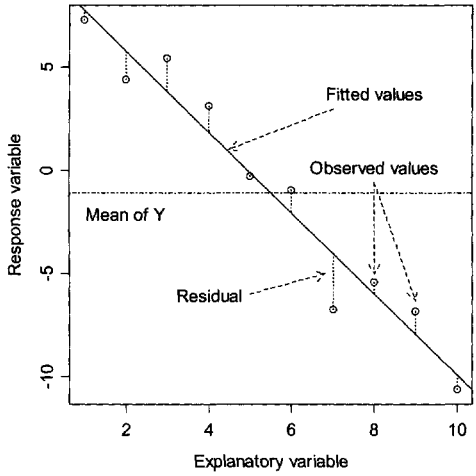

Figure 5.3. Illustration showing the fitted line and residuals. Observed values are the dots, and the straight line is the fitted regression curve. Residuals are obtained by the difference between these two sets of values. The mean of Y is the horizontal dotted line and is needed for the total sum of squares.

Now that we know what the underlying model and assumptions are, we show the results for the seven sampling points in Figure 5.2. The estimated regression parameters were $a = 10.37$ and $b = -3.90$, and the fitted values are obtained by

$$\hat{Y}_i = 10.37 - 3.90 NAP_i$$

The ^ on the Y is used to indicate that it is a fitted value. The equation produces a straight line shown in Figure 5.4 where the observed values are indicated by the squares. The implications of the assumptions are shown in Figure 5.5. The observed values Y_i are plotted as dots in the space defined by the R (richness) and NAP axes (which we will call the R-NAP space). The straight line in the same space defines the fitted values, and the fitted values at a particular value of X are represented by an 'x'. The Gaussian curves on top of the fitted values have their centre at the fitted values. The widths of these curves are all the same and are determined by the estimator of the standard deviation s, a function of the residuals and the number of observations calculated by

$$s = \sqrt{\sum_{i=1}^{n} \frac{e_i^2}{n-2}}$$

Each Gaussian curve shows the probability of other values for Y at any particular value of X. If these curves are wide, then other samples could have a range of very different Y values, compared with the observed value. From this figure it is easily seen why an extreme outlier can cause a large s.

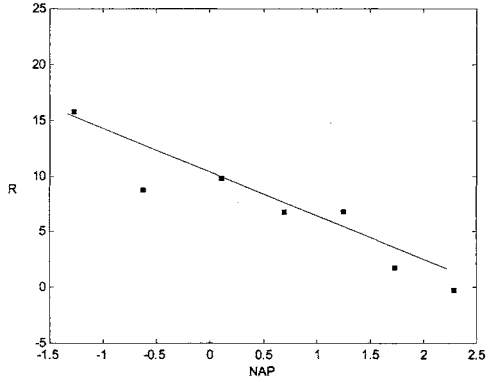

Figure 5.4. Fitted regression curve for seven points in the upper left panel in Figure 5.2.

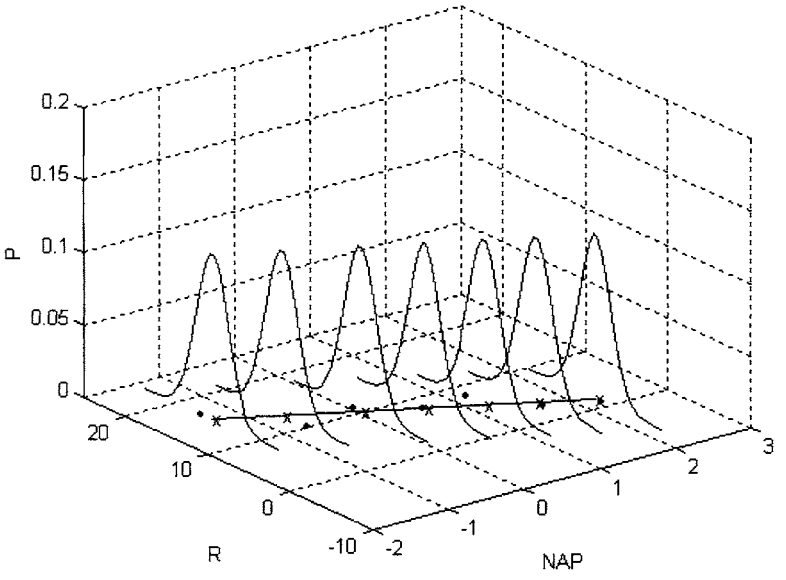

Figure 5.5. Implications of assumptions in linear regression. A dot represents the observed value, and the 'x' is the fitted value at a particular X value. The straight line contains all fitted values. Each Gaussian density curve is centred at a fitted value. The vertical axis shows the probability of finding a particular Y value.

In Figures 5.2, 5.4 and 5.5, we only used seven arbitrarily selected sample points from the 45 available. Using only seven points avoids cluttering the graphs with too much information. However, having gained some familiarity with the graphical interpretation of the Gaussian model, it is now valuable to look at the linear regression model using all 45 observations. Figure 5.6 shows the same three-dimensional graph as in Figure 5.5, but using all 45 observations. The regression coefficients are slightly different in this example compared with the earlier one because they are based on a different sample size. Looking at these new figures highlights two concerns. The results show that the regression model now predicts negative values for species richness from sites with a NAP value of slightly larger than two metres. This is obviously impossible as the species richness can only ever be zero or greater than zero. Not only is the fitted value close to zero at two metres, but the Gaussian probability density function suggests there is a large probability of finding negative values for the response variable at these sites. Another concern becomes clear when viewing Figure 5.6 from a slightly different angle (Figure 5.7). At least three points in the tail of the distribution have high values for species richness. If there had been only one such observation, then it could probably be dismissed as an outlier, but with three points, we need to look at this in more detail. These three points are causing the wide shape of the Gaussian probability density functions (they have a high contribution to the standard deviation s). Imagine how much smaller the width of the curves would be if the three points were much closer to the fitted regression curve. They also cause large confidence intervals for the regression parameters. Figure 5.7 also shows a clear violation of the homogeneity assumption, and later in this section, we discuss some graphical techniques to detect this violation.

The final important assumption of regression analysis is independence. The independence assumption means that if an observed value is larger than the fitted value (positive residual) at a particular X value, then this should be independent of the Y value for neighbouring X values. We therefore do not want to see lots of positive (or negative) residuals next to one another, as this might indicate a violation of the independence assumption.

Explaining linear regression in this way simplifies introducing GLM and GAM, which can both be explained using a graph similar to Figure 5.6, except that the fitted curve has a slightly different form and the Gaussian density curve is replaced by a more appropriate distribution. GLM and GAM are discussed in a later chapter, but before moving on to these techniques we need to discuss a few linear regression concepts (ANOVA-tables, t-values, coefficient of determination and the AIC) that have equivalents in GLM and GAM. We also discuss the tools used to detect violation of the underlying linear regression assumptions described earlier.

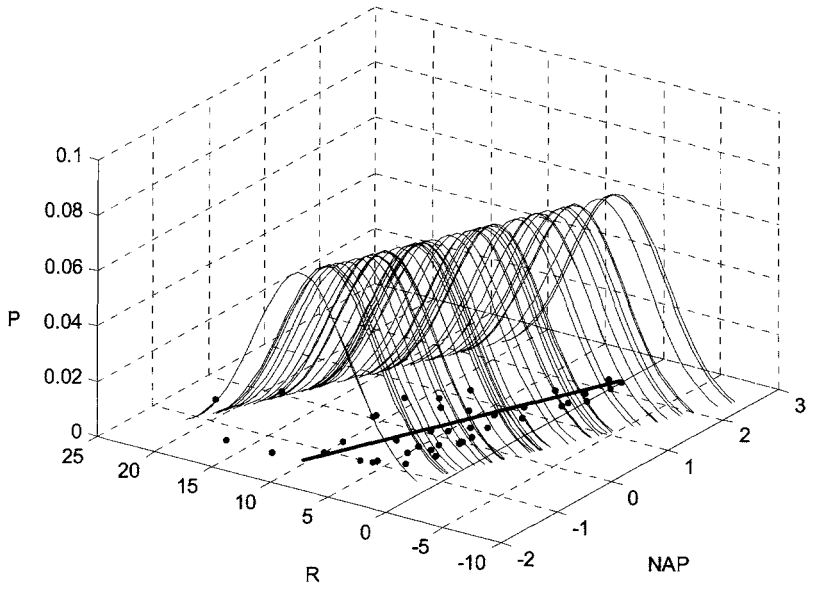

Figure 5.6. Regression curve for all 45 observations from the RIKZ data discussed in the text showing the underlying theory for linear regression.

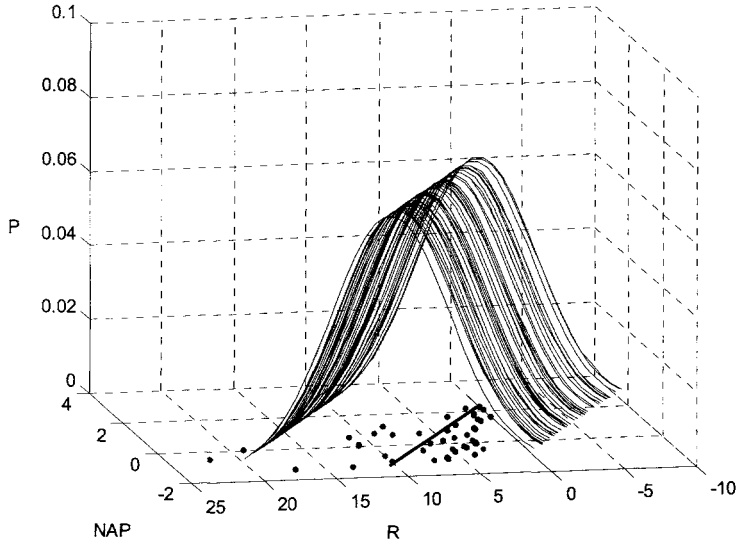

Figure 5.7. Regression curve for all 45 observations, viewed from a different angle.

Assessing the significance of regression parameters

An important aspect in linear regression is the partitioning of the total variability. The total variance in Y, denoted by SS_{total}, can be split up into the part explained by X ($SS_{regression}$) and the part not explained by X ($SS_{residual}$). $SS_{regression}$ measures how well the regression part (X) explains Y, and $SS_{residual}$ shows the amount of variability in the Y values that cannot be explained by the regression model. The values for these components are calculated using the formulae in Table 5.1, and Figure 5.3 shows a geometric interpretation. Most statistics programmes produce ANOVA tables in the style of Table 5.2. The sum of squares depends on the sample size n. If more observations are used, the sums of squares get larger. Therefore, the sums of squares are transformed to variance components by dividing them by the degrees of freedom[1]. The degrees of freedom for the regression sum of squares is the number of regression parameters minus 1. In this case: $2 - 1 = 1$. The degrees of freedom for the total sum of squares is $n - 1$. If there were no explanatory variables, the variance would be estimated from the ratio of the total sum of squares and $n - 1$. The degrees of freedom for the residual sum of squares is $n - 2$; two regression parameters were estimated to calculate this component: the intercept and the slope. The ratio of the two variance components is called the mean square (MS). The MSs are sample variances and, therefore, estimate parameters. $MS_{residual}$ estimates σ_ε^2 and $MS_{regression}$ estimates σ_ε^2 plus an extra term dependent on β and X. The fifth column in Table 5.2 shows what the MS components are estimating.

In bivariate regression, the ANOVA table is used to test the null-hypothesis that the slope of the regression line is equal to zero (H_0: $\beta = 0$). Under this null hypothesis, the expected MS for the regression component is equal to one. So, the ratio of the two variance components $MS_{regression}$ and $MS_{residual}$ is also 1. If for our sample data, the ratio is larger than one, then there is evidence that the null hypothesis is false. Assuming the four regression assumptions hold, then the ratio of $MS_{regression}$ and $MS_{residual}$ will follow an F-distribution with $df_{regression}$ and $df_{residual}$ degrees of freedom. The results for the RIKZ data are shown in Table 5.3.

---

[1] The degrees of freedom for a statistic is the number of observations that are free to vary. Suppose we want to calculate the standard deviation of the observations 1, 2, 3, 4 and 5. The formula for the standard deviation is given by

$$s = \sqrt{\sum_{i=1}^{n} \frac{(y_i - \overline{y})^2}{n-1}}$$

We first have to calculate the mean, which is three. In the next step, we have to calculate the sum of the squared deviations from the mean. Although there are five such squared deviations, only four of them are free to assume any value. The last one must contain the value of Y such that the mean is equal to three. For example, suppose we have the squared components 4 (=2(1 −3)), 1, 0 and 4. We now know that the last square component is calculated using $Y_i = 5$. For this reason, the standard deviation is said to have n-1 degrees of freedom.

Table 5.1. Three variance components.

Notation	Variance in	Sum of squared deviations of	Formula
SS_{total}	Y	Observed data from the mean	$\sum_{i=1}^{n}(Y_i - \overline{Y})^2$
$SS_{regression}$	Y explained by X	Fitted values from the mean value	$\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$
$SS_{residual}$	Y not explained by X	Observed values from fitted values	$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$

Table 5.2. ANOVA table for simple regression model. df stands for degrees of freedom.

Source of variation	SS	df	MS	Expected MS
Regression	$\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$	1	$\sum_{i=1}^{n}\frac{(\hat{Y}_i - \overline{Y})^2}{1}$	$\sigma_\varepsilon^2 + \beta^2\sum_{i=1}^{n}(X_i - \overline{X})^2$
Residual	$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	$n-2$	$\sum_{i=1}^{n}\frac{(Y_i - \hat{Y}_i)^2}{n-2}$	σ_ε^2
Total	$\sum_{i=1}^{n}(Y_i - \overline{Y})^2$	$n-1$		

Table 5.3. ANOVA table for the RIKZ data.

	df	SS	MS	F-value	P(>F)
NAP	1	357.53	357.53	20.66	<0.001
residuals	43	744.12	17.31		

The ANOVA table (Table 5.3) shows that the ratio $MS_{regression}/MS_{residual}$ is 20.66. Under the null hypothesis that all slopes are equal to 0, the ratio of 20.66 is unlikely ($p < 0.001$), and H_0 can be rejected. The ANOVA table therefore provides a test to identify whether there is a linear relationship between Y and X.

For the bivariate regression model, a mathematically identical test to an ANOVA is the single parameter t-test. The null hypothesis is H_0: $\beta = 0$, and the test statistic is

$$t = \frac{b_1}{s_{b_1}}$$

The t-value can be compared with a t-distribution with $n - 2$ degrees of freedom. The estimated regression parameters, standard errors, t-values and p-values for the RIKZ data are shown in Table 5.4. Note that these values are slightly different then the earlier model where only seven observations were used, as we are now using all 45 observations. The t-statistic for the regression parameter β is

$$t = \frac{-2.87}{0.63} = -4.55$$

This statistic follows a t-distribution. The critical value for t is 2.02 (significance level is 0.05, two-sided, df = 43). So the null hypothesis can be rejected. Alternatively, the p-value can be used. So, assuming the four assumptions underlying the linear regression model are valid, we can conclude there is a significant negative relationship between species richness and NAP.

Table 5.4. Estimated regression parameters, standard errors, t-values and p-values for the RIKZ data using all 45 observations.

	Estimated Value	Std. Error	t-value	p-value
Intercept	6.69	0.66	10.16	<0.001
NAP	−2.87	0.63	−4.55	<0.001

Model validation in bivariate linear regression

Coefficient of determination

The proportion of total variance in Y explained by X can be measured by R^2, also called the coefficient of determination. It is defined by

$$R^2 = \frac{SS_{regression}}{SS_{total}} = 1 - \frac{SS_{residual}}{SS_{total}}$$

The higher this value, the more the model explains. For the RIKZ data, R^2 = 0.32, which means that NAP explains 32% of the variation in the species richness data. However, the value of R^2 should not be used to compare models with different data transformations. Nor should it be used for model selection, as a model with more explanatory variables will always have a higher R^2. R^2 can also often have high values for some non-linear models, even when the regression provides a poor fit with the data. This is shown in Figure 5.8 using data from Anscombe (1973). All four panels show data that share the same intercept, slope and confidence bands. Both the F-statistics and the t-values indicate that the regression

parameter is significantly different from zero, and more worrying, all four R^2 values are equal to 0.67! Provided all assumptions hold and there are no patterns in the residuals, there is nothing wrong with an R^2 of 0.32 for the RIKZ data.

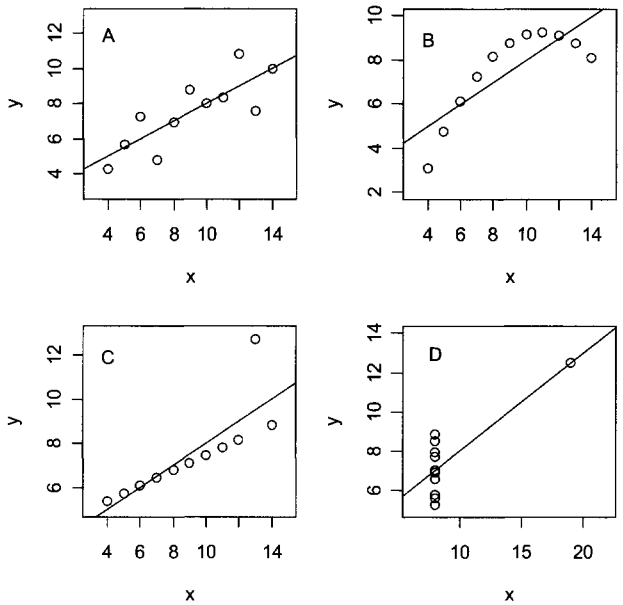

Figure 5.8. Famous Anscombe data. All regression parameters, t-values, F-values and R^2 values are the same.

Assessing the four assumptions

Linear regression is based on the four assumptions described above, and these need to be verified before placing any confidence in your regression model. Normality was the first assumption, and unless multiple observations at the same X value are available, normality cannot be confirmed. So, in practise, the normality assumption is checked using a histogram of the residuals. Quinn and Keough (2002) provide several references that consider normality to be an important but not a crucial assumption. More important is homogeneity, but without replicate observations, this assumption cannot be tested, either. However, homogeneity can be assessed by plotting the residuals against X to check for any increases (or decreases) in the spread of residuals along the x-axis. Alternatively, plotting fitted values against residuals can show increases in the spread for larger fitted values: a strong indicator of heterogeneity. So, we can pool residuals from different X values and use them to assess normality and homogeneity. As to independence, if the data are a time series, then residuals can be checked by the auto-correlation function found in most statistics programmes. This is further discussed in Chapter 16 and in various case study chapters. Spatial correlation is discussed in Chapters 18, 19 and 37.

As well as normality, homogeneity and independence, you should also check for residual patterns in the data. This assesses model misspecification and model fit. A useful tool is to plot residuals against each explanatory variable to check that no clear patterns are shown by the plotted residuals. Ideally the residuals should be scattered equally across the whole graph. To help with visual interpretation, a smoothing curve (Chapter 7) plus confidence bands can be added to this graph.

If the plotted residuals show an obvious non-random structure, several options are available:

1. Apply a transformation.
2. Add other explanatory variables.
3. Add interactions.
4. Add non-linear terms of the explanatory variables (e.g. quadratic terms).
5. Use smoothing techniques like additive modelling.
6. Allow for different spread using generalised least squares (GLS); see Chapters 8 and 23.
7. Apply mixed modelling (Pinheiro and Bates, 2000).

Some common problems and solutions are as follows:

1. There is a violation of homogeneity indicated by residuals versus fitted values. However, the residuals plotted against the explanatory variables do not show a clear pattern. Possible solutions are a transformation on the response variable, adding interactions or using generalised linear modelling with a Poisson distribution (if the data are counts).
2. There is a violation of homogeneity (as above), and the residuals plotted against the explanatory variables show a clear pattern. Possible solutions are as follows. Add interactions or non-linear terms of the explanatory variable (e.g., quadratic terms). Alternatively, consider generalised additive modelling.
3. There is no violation of homogeneity, but there are clear patterns in the residuals plotted against the explanatory variables. Possible solutions: Consider a transformation on the explanatory variables or apply additive modelling.

Instead of GLM with a Poisson distribution it is also possible to model the heterogeneity explicitly using generalised least squares (or mixed modelling). For example, instead of assuming that $\varepsilon_i \sim \mathrm{N}(0,\sigma^2)$ in the linear regression model we can allow for heterogeneity using variance structures like (Pinheiro and Bates 2000):

- $\varepsilon_i \sim N(0, NAP_i \times \sigma^2)$
- $\varepsilon_i \sim N(0, |NAP_i|^{2\delta} \times \sigma^2)$
- $\varepsilon_i \sim N(0, \sigma^2 \times \exp(2\delta NAP_i))$
- $\varepsilon_i \sim N(0, \sigma_j^2)$

where δ is an unknown parameter. Further variance structures are described in Chapter 5 in Pinheiro and Bates (2000). The first three options allow for an in-

crease (or decrease) in residual variance depending on the values of the *variance-covariate* NAP. The fourth option allows for different spread per level of a nominal variable. Only the fourth option is used in the case study chapters (23, 26, 36).

The Decapod case study chapter shows some of these approaches, but for this chapter we return to the RIKZ data. The linear regression for the RIKZ data is illustrated in Figure 5.9, where the assumptions discussed above can be checked. The upper left graph shows a scatterplot of the residuals versus the fitted values. This graph shows an increase in the spread of the residuals for the larger values of the fitted values, indicating a violation of the homogeneity assumption. The lower left graph shows a Scale-Location plot, which is a plot of square root transformed absolute standardised residuals versus fitted values. Standardised residuals are explained later in this section. Taking the square root of the absolute values reduces the skewness and makes non-constant variance more noticeable. This graph should show no pattern, but in this graph the values for Y increase as X increases, suggesting that variance is not constant. The upper right panel shows a QQ-plot of the residuals, which checks how closely the data follow a normal distribution. Normally distributed data points should lie approximately on a straight line, which is not the case here. Normality can also be checked using a histogram of the residuals (not shown) and, as with the QQ plot, suggests the data are not approximately normally distributed. The lower right panel in Figure 5.9 is discussed in the next section.

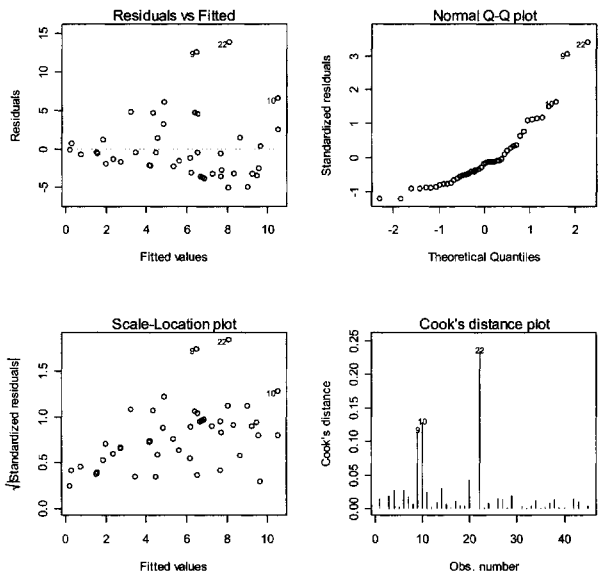

Figure 5.9. Graphical output for the linear regression model for the RIKZ data, allowing the assumptions of normality, homogeneity and independence to be checked. The Cook distance plot is used to check for influential observations.

Influential points

If a particular explanatory variable has one or more values that are much larger than the other observations, these observations could strongly influence the regression results. Leverage is a tool that identifies observations that have rather extreme values for the explanatory variables and may potentially bias the regression results. In Chapter 4, the left panel in Figure 4.16 showed three extreme observations A, B and C. Along the x-axis, A does not have any extremely large or small values, but B and C are both rather large and have a high leverage. However, applying a regression analysis with and without point B gives similar results, showing that B might have a high leverage, but it is not influential on the regression parameters. In contrast, C has high leverage and is also influential. A better measure for influential points is the Cook's distance statistic. This statistic identifies single observations that are influential on all regression parameters. It does this by calculating a test statistic D that measures the change in all regression parameters when omitting an observation. So, point B in the left panel in Figure 4.16 has a low Cook distance statistic, but A and C have high Cook distance statistics. The lower right panel in Figure 5.9 shows the Cook distance function for the RIKZ data. Each bar shows the value of the Cook distance statistic (which can be read from the vertical axis) for the corresponding observation along the x-axis. Relatively large bars indicate influential observations, and for the RIKZ data, observation 22 (as well as observations 9 and 10) looks influential. Fox (2002a) reports that the value $4/(n - k - 1)$ can be used as a rough cut off for noteworthy values of D_i where n is the number of observations, k is the number of regression slopes in the model, and D_i is the Cook value of observation i. Montgomery and Peck (1992) compare the Cook value with an F-value of approximately 1, and all D_i values larger than 1 could be influential. In this case, the Cook statistic at observation 22 is smaller than one and we can assume that this observation is not influential. We suggest using these graphs to inspect for points noticeably different from the majority. A mathematical formula for both the Cook distance statistic and the leverage can be found in Montgomery and Peck (1992) or Fox (2002a). A slightly modified Cook statistic is given in Garthwaite et al. (1995).

Summarising, leverage identifies observations with extreme explanatory variables and the Cook statistic detects points that are influential. It is easier to justify omitting influential points if they have extreme explanatory variables (these are points with a large Cook and a large leverage).

A related method is the Jackknife in which each observation i is omitted in turn, and regression parameters are estimated for the remaining $n - 1$ observations. Large changes in the regression parameters between iterations indicate influential points. Figure 5.10 shows the changes in the intercept and slope if the i^{th} observation is omitted. Note the slope changes considerably if observation 22 is omitted, and leaving out this observation also reduces the slope (it becomes more negative). Similar changes are noted when observations 9 and 10 are omitted.

Note that these measures of influence (leverage, Cook distance, and change in parameters) only assess the effect of one observation at a time. If two observations

have similar explanatory variables, and both are influential, these methods might not detect them.

We have already discussed using residuals to assess normality and homogeneity. As well as these ordinary residuals (calculated as observed values minus fitted values), alternative residuals can be defined as standardised residuals and Studentised residuals. These give a more detailed assessment on the likely influence of outlying values. The standardised residuals are defined as:

$$\frac{e_i}{\sqrt{MS_{residual}(1-h_i)}}$$

where e_i is the difference between the observed and fitted value and h_i is the leverage for observation i. Standardised residuals are assumed to be normally distributed with expectation 0 and variance 1; N(0,1). Consequently, large residual values (>2) indicate a poor fit of the regression model. Studentised residuals are a leave-one-out measure of influence. To obtain the i^{th} Studentised residual, the regression model is applied on all data except for observation i, and the $MS_{residual}$ is based on the $n - 1$ points (but not the residual e_i or hat value h_i: These are based on the full dataset). If the i^{th} Studentised residual is much larger than a standardised residual, then this is an influential observation because the variance without this point is smaller. Both types of residuals are shown in Figure 5.11. The Standardised residuals indicate that two observations have values larger than 2 and could be potential outliers.

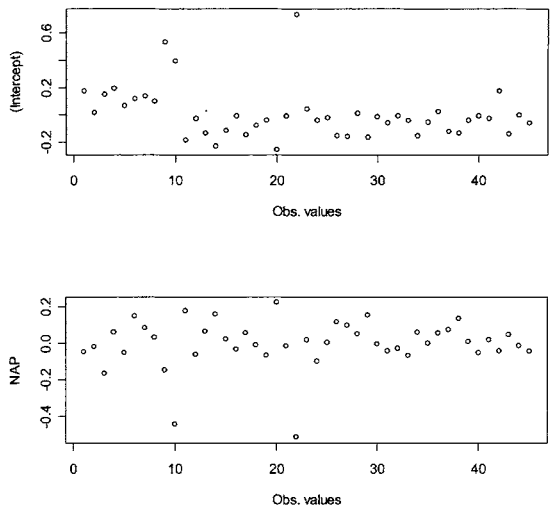

Figure 5.10. Changes in the intercept (upper panel) and slope for NAP (lower panel) if the i^{th} observation is omitted. It can be seen that observations 9, 10 and 22 have a considerable influence on the value of the intercept regression parameter.

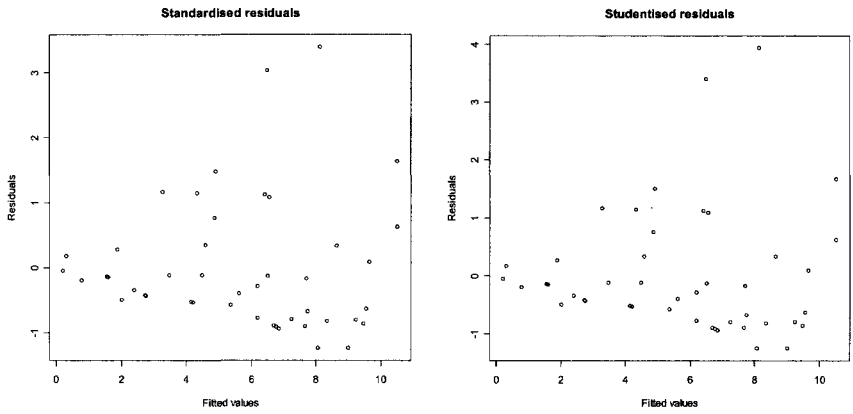

Figure 5.11. Standardised (left panel) and Studentised (right panel) residuals for the RIKZ data. The vertical axes show the values of the residuals and the horizontal axes the fitted values.

It is important to realise what the standardised residuals represent. Recall that Figure 5.5 showed three observations in the tails of the distributions. Suppose the NAP values of these three observations were very different in value from the other observations and located, say, on the far left-hand side of the NAP axis. In that case, you might conclude that the environmental conditions (in terms of NAP) are rather different from the other observations, and they could justifiably be excluded from the regression analysis. And this is exactly what leverage (or hat value) is measuring, the influence of a particular observation in the X space. An observation that has a relatively high leverage (compared with the other observations) is likely to have a relatively high standardised residual (because we are dividing by a small number). Figure 5.12 shows the leverage values, and there are no observations with much higher values than any other observations. As the observations with extreme NAP values do not have large standardised residuals, they cannot be left out of the analysis. This shows that comparing the leverage with the Cook distance can be useful.

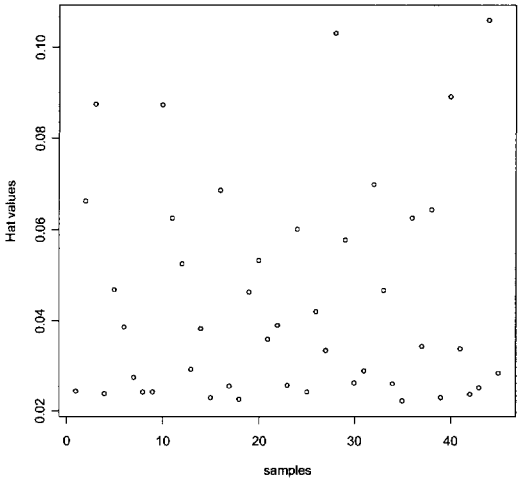

Figure 5.12. Leverage (or hat) values for the RIKZ data. The vertical axis shows the value of the leverage, and the horizontal axis indicates the identity of the observation (corresponding to the order of the data in the spreadsheet).

5.2 Multiple linear regression

In the previous section, we arbitrarily chose NAP as the explanatory variable in the bivariate regression model. However, this was only one of several explanatory variables available (e.g., grain size, humus, angle of the beach, exposure, and week, etc. were also measured; see Chapter 27). In this section, we discuss linear regression techniques that allow modelling a response variable (e.g., species richness) as a linear function of multiple explanatory variables, hence, the name multiple linear regression. This section expands our investigation into the RIKZ data using multiple regression; however, it does not seek to find the most optimal model as this is done later in the book. The general mathematical formula for a multiple regression model is

$$Y_i = \alpha + \beta_1 X_{1i} + \ldots + \beta_p X_{pi} + \varepsilon_i$$

For the species richness R in the RIKZ data this becomes:

$$R_i = \text{constant} + \beta_1 \text{NAP}_i + \beta_2 \text{Grainsize}_i + \beta_3 \text{Humus}_i + \text{Week}_i + \beta_4 \text{Angle}_i + \text{noise}_i$$

See Chapter 27 for an explanation of these explanatory variables. Week is fitted as a nominal variable. To reduce the numerical output, we only concentrate on these five explanatory variables. Selecting and assessing the best explanatory variables are discussed later in this section. On interpreting the regression parameters, β_1 shows the change in species richness for a one-unit change in NAP, while keep-

ing all other variables constant. And β_2 represents the richness change for a one-unit change in grainsize, while keeping all other variables constant. These parameters are called partial regression slopes as they measure the change in Y for a particular value while keeping the remaining $p - 1$ values constant. The ANOVA table (Table 5.5) for a multivariate regression model is similar to the table produced for a bivariate regression model (Table 5.2).

Table 5.5. ANOVA table for multiple linear regression model. $\overline{Y}$ is the mean value, and $\hat{Y}$ is the fitted value.

Source of variation	SS	df	MS
Regression	$\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$	p	$\frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}{p}$
Residual	$\sum_{i=1}^{n}(Y - \hat{Y}_i)^2$	n − p − 1	$\frac{\sum_{i=1}^{n}(Y - \hat{Y}_i)^2}{n-p-1}$
Total	$\sum_{i=1}^{n}(Y_i - \overline{Y})^2$	n − 1	

The null hypothesis tested by the ANOVA table is that all slope parameters are equal to 0. In formula: H_0: $\beta_1 = \beta_2 = .. = \beta_p = 0$. Just as in bivariate linear regression, the ratio of $MS_{regression}$ and $MS_{residual}$ follows an F-distribution and can be used to test the null hypothesis. If there is no evidence to reject the null hypothesis, then it can be concluded that none of the explanatory variables is related to the response variable. For the RIKZ model, the F-statistic is 11.18, which is highly significant ($p < 0.001$). This means the null hypothesis that all slope parameters are equal to 0 can be rejected. And consequently this means that at least one of the explanatory variables is significantly related to species richness. However, the F-statistic does not say which explanatory variables are significant. To identify the significant explanatory variables, the t-statistic introduced for the bivariate regression model can be used. The t-values of the regression parameters for NAP and Week (Table 5.6) indicate that both variables are significantly differently from zero at the 5% level. Occasionally a t-statistics will indicate non-significance even when you have a significant F-statistic. An explanation for this can be found in Montgommery and Peck (1992).

Table 5.6. Multiple linear regression results for the RIKZ data.

	Estimate	Std. Error	*t*-value	*p*-value
Intercept	9.30	7.97	1.17	0.25
angle2	0.02	0.04	0.39	0.70
NAP	−2.27	0.53	−4.30	<0.001
grainsize	0.00	0.02	0.11	0.92
humus	0.52	8.70	0.06	0.95
factor(week)2	−7.07	1.76	−4.01	<0.001
factor(week)3	−5.72	1.83	−3.13	<0.001
factor(week)4	−1.48	2.72	−0.55	0.58

Note that there are no entries for week 1. Most software will create dummy variables with values 0 and 1, indicating in which week an observation was taken. For week (with four levels), we get four of those dummy variables, which we could call W_1 to W_4. Using all of them as explanatory variables will result in an error or warning message as there is 100% collinearity; if an observation was not taken in week 1, 2 or 3, then it must be from week 4. A possible solution is to omit one of them, say W_1, and apply the GLM using W_2, W_3, W_4, and the other explanatory variables. The notation 'factor(week)2' just means W_2. Its estimated regression parameter is −7.07, which means that the species richness in week 2 is 7.07 lower than for week 1. The *p*-value suggests that it is significantly different from the baseline week. It is also possible to re-level the week and use a different baseline, for example week 2. Dalgaard (2002) showed how each level can be used as baseline in turn, and how to correct the *p*-values for multi-comparisons.

Instead of the *t*-statistic, you can compare two nested models. Two models are called nested if the explanatory variables in one model are a subset of those in the other model. For example, suppose we want to compare the following two models.

Model 1: $Y_i = \alpha + \varepsilon_i$
Model 2: $Y_i = \alpha + \beta\,\text{Angle}_i + \varepsilon_i$

We have already introduced one form of *F*-statistic, but a more general form is

$$F = \frac{(RSS_1 - RSS_2)/(p-q)}{RSS_2/(n-p)}$$

RSS_1 and RSS_2 are the residual sum of squares of model 1 (nested model) and model 2 (the full model), respectively, and n is the number of observations. The number of parameters in models 2 and 1 are $p + 1$ and $q + 1$ respectively ($p > q$). The '+1' is because the intercept, and p and q, are the number of slopes in each model. Because model 1 contains fewer parameters ($q + 1$), the fit will always be equal or worse than model 2 and the same holds for the total sum of residuals. If both models give a similar fit, the *F*-statistic will be small. So, large values of the *F*-statistic indicate that the slope is not equal to zero (there is a relationship). Indeed, the null-hypothesis in this test is that the regression parameter for the extra term in model 2 is equal to zero (H_0: $\beta = 0$). Most statistics programmes produce analysis of variance tables in the following form

	df	Sum Sq	Mean Sq	*F*	Pr(>*F*)
angle	1	124.86	124.86	13.06	0.001
NAP	1	319.32	319.32	33.41	<0.001
grainsize	1	106.76	106.76	11.17	0.002
humus	1	19.53	19.53	2.04	0.161
factor(week)	3	177.51	59.17	6.19	0.003
Residuals	37	353.66	9.56		

The first line compares model 1 with model 2, and the F-value of 13.06 (p = 0.001) indicates that angle is significantly related to richness. The second line compares the following two models:

Model 2 $Y_i = \alpha + \beta_1 \text{Angle}_i + \varepsilon_i$
Model 3 $Y_i = \alpha + \beta_1 \text{Angle}_i + \beta_2 \text{NAP}_i + \varepsilon_i$

The F-value is equal to 33.41 ($p < 0.001$) and indicates that adding NAP to a model that already contains angle gives a better model. Adding humus to a model that already contains angle, NAP and grain size does not result in a model improvement (the F-statistic is 2.04, which is not significant). Adding week to the model that contains all four explanatory variables still improves the model. The disadvantage of this table is that it depends on the order of the explanatory variables. Problems of collinearity may result in terms added at the end being not significant that would have been significant if added at the beginning.

This general F-statistic can also compare nested models in the following form:

Model 2 $Y_i = \alpha + \beta_1 \text{Angle}_i + \varepsilon_i$
Model 4 $Y_i = \alpha + \beta_1 \text{Angle}_i + \beta_2 \text{NAP}_i + \beta_3 \text{Grainsize}_i + \beta_4 \text{Humus}_i + \varepsilon_i$

The underlying null hypothesis in the F-test is that the regression parameters for NAP, grain size and humus are equal to zero (H_0: $\beta_2 = \beta_3 = \beta_4 = 0$). In this case, both models give the same fit. The advantage of this general F-test is that it would give one p-value for a nominal variable that contains more than two classes.

Model selection — finding the best set of explanatory variables

One aim of regression modelling is to find the optimal model that identifies the parameters that best explain the collected data. In this instance these are the parameters that best explain why any specific species richness value occurs at any particular site on the beach. This means we are looking for the best subset of explanatory variables. The reasons for this are (i) a subset is easier to interpret, and (ii) precision of predicted intervals and confidence bands will be smaller (using fewer parameters). Defining 'optimal' is subjective, and to remove part of this subjectivity, statistical criteria are available, for example the AIC, the adjusted R^2 value, and the BIC. Here, we discuss the AIC (Akaike Information Criteria) and the adjusted R^2. They are defined by:

$$\text{AIC} = n \log(\text{SS}_{\text{residual}}) + 2(p + 1) - n\log(n)$$

$$\text{Adjusted } R^2 = 1 - \frac{SS_{residual}/(n-(p+1))}{SS_{total}/(n-1)}$$

The first part of the AIC definition is a measure of goodness of fit. The second part is a penalty for the number of parameters in the model. The AIC can be calculated for each possible combination of explanatory variables, and the model with the smallest AIC is chosen as the most optimal model. The disadvantage of R^2 is that the more explanatory variables are used, the higher the R^2. The adjusted R^2 accounts for different degrees of freedom and, hence, the extra regression parameters.

These criteria can be obtained for every possible combination of explanatory variables, and the model with the optimal values (lowest for AIC, highest for the adjusted R^2) can be selected as 'the' optimal model. Selecting the explanatory variables can be done manually, but if there are more than five explanatory variables, this can become tedious. Instead, automatic selection procedures can be used to apply a forward selection, backward selection, or a combination of forward and backward selection. The forward selection method first applies bivariate linear regression using each explanatory variable in turn. The variable with the lowest AIC is selected, and a multiple linear regression model is applied using the selected variable and each of the remaining variables in turn. The variable that gives the lowest AIC in combination with the first selected variable is then selected, and the process is repeated until the AIC starts to increase. Alternatively, a backward selection (starting with all variables and dropping one at a time) can be used, or even a combination of both. The output below shows the results for the RIKZ data. As most software is capable of applying a backward selection, we show the results for this approach.

$$R_i = \text{constant} + \beta_1 NAP_i + \beta_2 \text{Grain size}_i + \beta_3 \text{Humus}_i + \text{Week}_i + \beta_4 \text{Angle}_i + \text{noise}_i$$

Note that week is fitted as a nominal variable. The first iteration is presented:

Start: AIC= 108.78
R ~ angle + NAP + grainsize + humus + factor(week)

	df	AIC
- humus	1	106.78
- grain size	1	106.79
- angle	1	106.96
<none>		108.78
- factor(week)	3	121.08
- NAP	1	124.98

This step shows that leaving out humus results in a drop in the AIC from 108.78 to 106.78, which means that humus is not important. Leaving out NAP or week gives a large increase in AIC (meaning these are important variables). Hence, the selection algorithm will continue without humus and produce the following output:

Step: AIC= 106.78. R ~ angle + NAP + grain size + as.factor(week)

	df	AIC
- grain size	1	104.80
- angle	1	104.98
<none>		106.78
- factor(week)	3	120.70
- NAP	1	123.32

This steps shows that grain size is the next variable that should be removed. We will omit the output from the next step and present the results of the final step:

Step: AIC= 103.2. R ~ NAP + as.factor(week)

	df	AIC
<none>		103.20
- NAP	1	122.04
- factor(week)	3	130.25

Further model simplifications lead to larger AIC values indicating a reduced fit. Hence, the optimal model contains both NAP and week. A selection method using both forward and backward selection gave the same results. Instead of a full selection procedure, you can drop one explanatory variable at a time and apply the general F-test. This was discussed earlier in this section, and the output is printed below. Note that one variable is dropped in turn.

	df	AIC	F-value	p-value
angle	1	106.96	0.15	0.70
NAP	1	124.98	18.45	0.00
Grain size	1	106.79	0.01	0.92
humus	1	106.78	0.00	0.95
factor(week)	3	121.08	6.19	0.00

Each row compares the full model versus the model with one variable dropped. For example, leaving out NAP resulted in an F-statistic of 18.45. Results indicate that angle, grain size and humus can be dropped from the model as their removal did not significantly affect the F-value. Note that they should not be dropped all at once!

The AIC should only be used as a general guide. Sometimes the AIC comes up with an 'optimal' model that has one or two non-significant regression parameters. In such cases, further model selection steps are required using, for example the F-test. On the other hand, things like model fit and residual patterns can also influence the choice of the final model. One can even argue that a model with a non-significant regression parameter, but with no clear residual pattern is better than a model where all parameters are significant but with clear residual patterns.

A problem with selection procedures is (i) multiple comparisons and (ii) collinearity. With multiple comparisons, every time we apply a regression model, there is a 5% chance of deciding a regression parameter is significantly differently from

0, even when it is not. This risk increases every time the regression is applied, and running a large number of forward and backward selections increases this chance. In ecological studies, there are three ways people deal with this problem; ignore it, avoid using selection methods, or apply a correction method such as the Bonferroni method where p-values (or significance levels) are adjusted for the number of tests carried out. When explanatory variables are highly correlated with each other, a forward selection and a backward selection might give different results due to collinearity. Assuming model simplification is the aim, it is better to avoid using explanatory variables that are likely to vary together (collinearity) because both variables may be reacting in the same way to changes in some other variable.

5.3 Partial linear regression

There are three reasons to discuss partial linear regression. The first reason is to answer why a particular explanatory variable is in the model. Is it significant because of a few outliers, because of collinearity, or is there a genuine relationship? The second reason is because it is used in some multivariate techniques (redundancy analysis and canonical correspondence analysis). The third reason is that variance partitioning (Chapter 12) is easier to explain now, using partial linear regression rather than trying to explain it when we get to the multivariate analysis chapters.

Based on the results in Chapter 4 for the Argentinean marine benthic data, we carried out a multiple regression analysis (results not shown). A backwards selection suggested the variables mud and transect were important in explaining biodiversity (Shannon–Weaver biodiversity index). The next question is how big a contribution does the mud variable make in explaining the different biodiversity indices recorded. Perhaps mud is only significant because it is collinear with transect; transects b and c might be muddier. Partial linear regression identifies the relationship between the biodiversity index and mud, while filtering out the effects of transect (or indeed any set of explanatory variables). In this section we are going to look at two slightly different approaches to partial linear regression. The first approach is discussed in Quinn and Keough (2002) and consists of three steps.

Step 1.

Assume there are one response variable Y and three explanatory variables X, W and Z. The basic linear regression model for these variables is given by

$$Y_i = \text{constant} + \beta_1 X_i + \beta_2 W_i + \beta_3 Z_i + \varepsilon_i$$

The residuals ε_i are estimated from the observed values minus the fitted values. Formulated differently, the residuals represent the information in Y that cannot be explained with X, W and Z. Suppose we fit the model

$$Y_i = \text{constant} + \beta_4 W_i + \beta_5 Z_i + \varepsilon_{1i}$$