Chapter 2
Limitations of Linear Regression Applied on Ecological Data

This chapter revises the basic concepts of linear regression, shows how to apply linear regression in R, discusses model validation, and outlines the limitations of linear regression when applied to ecological data. Later chapters present methods to overcome some of these limitations; but as always before doing any complicated statistical analyses, we begin with a detailed data exploration. The key concepts to consider at this stage are outliers, collinearity, and the type of relationships between the variables. Failure to apply this initial data exploration may result in an inappropriate analysis forcing you to reanalyse your data and rewrite your paper, thesis, or report.

We assume that the reader is 'reasonably' familiar with data exploration and linear regression techniques. This book is a follow-up to *Analysing Ecological Data* by Zuur et al. (2007), which discusses a wide range of exploration and analytical tools (including linear regression and its extensions), together with several related case study chapters. Other useful, non-mathematical textbooks containing regression chapters include Chambers and Hastie (1992), Fox (2002), Maindonald and Braun (2003), Venables and Ripley (2002), Dalgaard (2002), Faraway (2005), Verzani (2005) and Crawley (2002, 2005). At a considerable higher mathematical level, Ruppert et al. (2003) and Wood (2006) are excellent references for linear regression and extensions. All these books discuss linear regression and show how to apply it in R. Other good, but not based on R, textbooks include Montgomery and Peck (1992), Draper and Smith (1998) and Quinn and Keough (2002). Any of the above mentioned texts using R can be also used to learn R, but we highly recommend the book from Dalgaard (2002) or for a slightly different approach, Crawley (2005). However, even if you are completely unfamiliar with R, you should still be able to pick up the essentials from this book and 'learn it as you go along'. It is not that difficult and, once exposed to R, you will never use anything else.

Although various linear regression examples are given in this chapter, a complete example, including all R code and aspects like interaction, model selection and model validation steps, is given in Appendix A.

A.F. Zuur et al., *Mixed Effects Models and Extensions in Ecology with R*, Statistics for Biology and Health, DOI 10.1007/978-0-387-87458-6_2,

2.1 Data Exploration

2.1.1 Cleveland Dotplots

The first step in any data analysis is the data exploration. An important aspect in this step is identifying outliers (we discuss these later) and useful tools for this are boxplots and/or Cleveland dotplots (Cleveland, 1993). As an example of data exploration, we start with data used in Ieno et al. (2006). To identify the effect of species density on nutrient generation in the marine benthos, they applied a two-way ANOVA with nutrient concentration as the response variable with density of the deposit-feeding polychaete *Hediste diversicolor* (*Nereis diversicolor)*, and nutrient type (NH_4-N, PO_4-P, NO_3-N) as nominal explanatory variables. The data matrix consists of three columns labelled concentration, biomass, and nutrient type. The aim is to model Nereis concentration as a function of biomass and nutrient. The following R code reads the data and makes a Cleveland dotplot.

```
> library(AED); data(Nereis)
```

R commands are case sensitive; so make sure you type in commands exactly as illustrated. The data are stored in a data frame called `Nereis`, which is a sort of data matrix. Information in a data frame can be accessed in various ways. First, we need to know what is in there, and this is done by typing the following at the R prompt:

```
> names(Nereis)
```

This command gives the names of all variables in the data frame:

```
[1] "concentration" "biomass"       "nutrient"
```

The following lines of code produce the Cleveland dotplot in Fig. 2.1A.

```
> dotchart(Nereis$concentration,
    ylab = "Order of observations",
    xlab = "Concentration", main = "Cleveland dotplot")
```

The `dotchart` function makes the Cleveland dotplot. Note that the arguments of the `dotchart` function are typed in over multiple rows. When the code runs over more than one line like this, you should ensure that the last symbol on such a line is a slash (\) or a comma (,). So, this works as well:

```
> dotchart(Nereis$concentration, ylab = "Order of \
    observations",
    xlab =" \
    Concentration", main = "Cleveland dotplot")
```

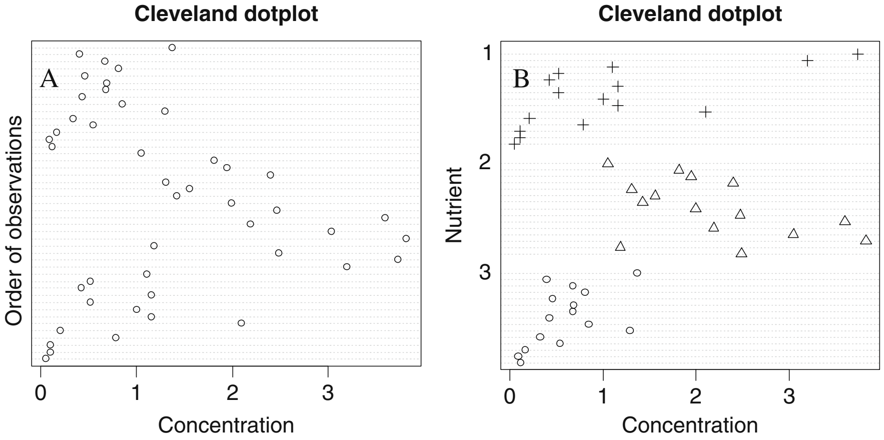

Fig. 2.1 **A**: Cleveland dotplot for Nereis concentration. **B**: Conditional Cleveland dotplot of Nereis concentration conditional on nutrient with values 1, 2 and 3. Different symbols were used, and the graph suggests violation of homogeneity. The *x*-axes show the value at a particular observation, and the *y*-axes show the observations

In a `dotchart`, the first row in the text file is plotted as the lowest value along the *y*-axis in Fig. 2.1A, the second observation as the second lowest, etc. The *x*-axis shows the value of the concentration for each observation. By itself, this graph is not that spectacular, but extending it by making use of the grouping option in `dotchart` (for further details type: `?dotchart` in R) makes it considerably more useful, as can be seen from Fig. 2.1B. This figure was produced using the following command:

```
> dotchart(Nereis$concentration,
    groups = factor(Nereis$nutrient),
    ylab = "Nutrient", xlab = "Concentration",
    main = "Cleveland dotplot", pch = Nereis$nutrient)
```

The `groups = factor(nutrient)` bit ensures that observations from the same nutrient are grouped together, and the `pch` command stands for point character. In this case, the nutrient levels are labelled as 1, 2 and 3. If other characters are required, or nutrient is labelled as alpha-numerical values, then you have to make a new column with the required values. To figure out which number corresponds to a particular symbol is a matter of trial and error, or looking it up in a table, see, for example, Venables and Ripley (2002).

Cleveland dotplots are useful to detect outliers and violation of homogeneity. Homogeneity means that the spread of the data values is the same for all variables, and if this assumption is violated, we call this heterogeneity. Points on the far end along the horizontal axis (extremely large or extremely small values) may be considered outliers. Whether such points are influential in the statistical analysis depends on the technique used and the relationship between the response and explanatory variables. In this case, there are no extremely large of small values for the variable concentration values. The Cleveland dotplot in Fig. 2.1B indicates that we may expect problems with violation of homogeneity in a linear regression model applied on these data, as the spread in the third nutrient is considerable smaller than that in the other two. The mean concentration value of nutrient two seems to be larger, indicating that in a regression model, the covariate nutrient will probably play an important role.

2.1.2 Pairplots

Another essential data exploration tool is the pairplot obtained by the R command

```
> pairs(Nereis)
```

The resulting graph is presented in Fig. 2.2. Each panel is a scatterplot of two variables. The graph does not show any obvious relationships between concentration and biomass, but there seems to be a clear relationship between concentration and

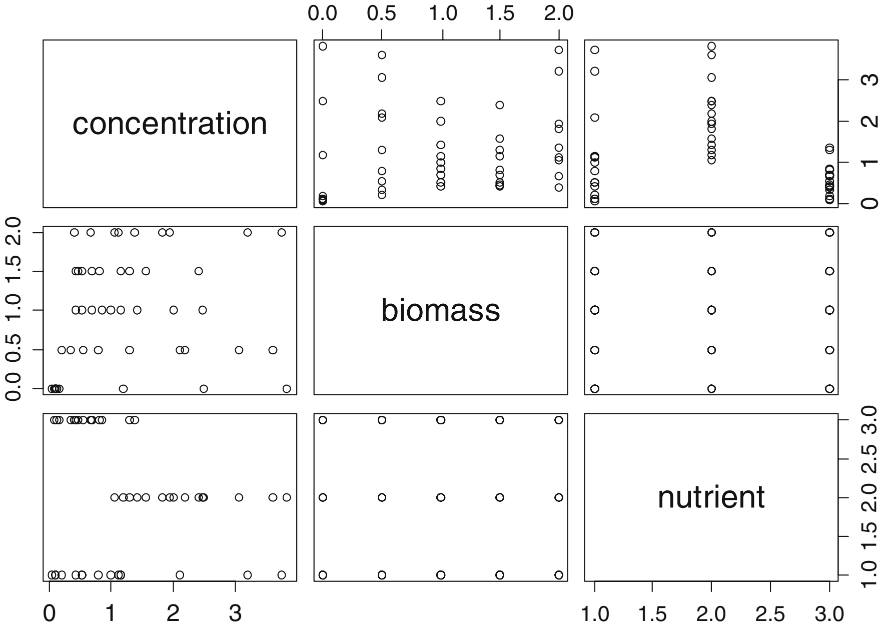

Fig. 2.2 Pairplot for concentration, biomass and nutrient. Each panel is a scatterplot between two variables. It is also possible to add regression or smoothing lines in each panel. In general, it does not make sense to add a nominal variable (nutrient) to a pairplot. In this case, there are only two explanatory variables; hence, it does not do any harm to include nutrient

nutrients, as already suggested by the Cleveland dotplot. More impressive pairplots can be made by using the `panel` option in pairs. The help file for pairs is obtained by typing: `?pairs`. It shows various examples of pairplot code that gives pairplots with histograms along the diagonal, correlations in the lower panels, and scatterplots with smoothers in the upper diagonal panels.

2.1.3 Boxplots

Another useful data exploration tool that should be routinely applied is the boxplot. Just like the Cleveland dotplot, it splits up the data into groups based on a nominal variable (for example nutrient). The boxplot of concentration conditional on nutrient is given in Fig. 2.3. The following code was used to generate the graph:

```
> boxplot(concentration ~ factor(nutrient),
    varwidth = TRUE, xlab = "nutrient",
    main = "Boxplot of concentration conditional on\
    nutrient", ylab = "concentration", data = Nereis)
```

The `varwidth = TRUE` command ensures that the width of each boxplot is proportional to the sample size per level. In this case, the sample size per nutrient (labelled 1, 2, and 3) is about the same.

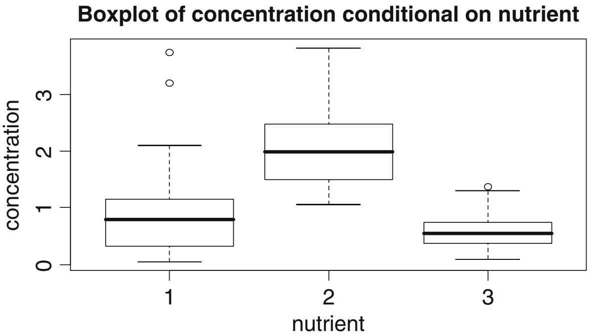

Fig. 2.3 Boxplot of concentration conditional on the nominal variable nutrient. The horizontal line in each box is the median, the boxes define the hinge (25–75% quartile, and the line is 1.5 times the hinge). Points outside this interval are represented as dots. Such points may (or may not) be outliers. One should not label them as outliers purely on the basis of a boxplot! The width of the boxes is proportional to the number of observations per class

2.1.4 xyplot from the Lattice Package

As with the Cleveland dotplot and the pairplot, the boxplot shows that there may be a nutrient effect: higher mean concentration values for nutrient level 2, but also

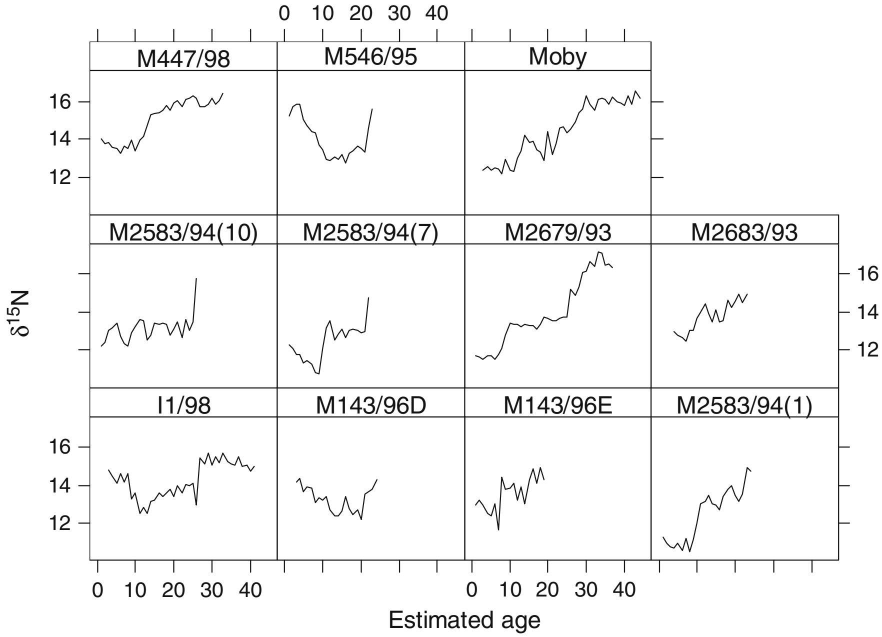

Fig. 2.4 Nitrogen concentration in teeth versus age for each of the 11 whales stranded in Scotland. The graph was made with the `xyplot` from the lattice package

less spread for nutrient level 3, indicating potential heterogeneity problems later on. We now show a more advanced data exploration method. As the Nereis data set has only two explanatory variables, this method is less appropriate for these data, and therefore we use a different data set.

Just like rings in trees, teeth of an animal have rings, and from these it is possible to extract information on how chemical variables have changed during the life of the animal. Mendes et al. (2007) measured the nitrogen isotopic composition in growth layers of teeth from 11 sperm whales stranded in Scotland. The underlying aim of the research was to 'investigate the existence, timing, rate and prevalence of dietary and/or foraging location shifts that might be indicative of ontogenetic benchmarks related to changes in schooling behaviour, movements, environmental conditions, foraging ecology and physiology' (Mendes et al., 2007).

Figure 2.4 shows an `xyplot` from the lattice package. The name lattice is used in R, but in SPLUS it is called a Trellis graph. It consists of a scatterplot of nitrogen isotope ratios versus age for each whale. Working with lattice graphs is difficult, and one of the few books on this topic is Sarkar (2008). One of the underlying questions is whether all whales have similar nitrogen-age relationships, and the graph suggests that some whales indeed have similar patterns. The R code to generate the graph in Fig. 2.4 is

```
> library(AED); data(TeethNitrogen)
> library(lattice)
> xyplot(X15N ~ Age | factor(Tooth), type = "l",
    xlab = "Estimated age", col = 1,
    ylab = expression(paste(delta^{15}, "N")),
    strip = function(bg = 'white', ...)
    strip.default(bg = 'white', ...),
    data = TeethNitrogen)
```

The `xyplot` makes the actual graph, and the rest of the code is merely there to extract the data. The `type = "l"` and `col = 1` means that a line in black colour is drawn. Note that the `l` in `type` stands for lines, not for the 1 from 1, 2, and 3. But the `1` for col is a number! The complicated bit for the y-label is needed for subscripts, and the strip code is used to ensure that the background colour in the strips with whale names is white. It can be difficult to figure out this type of information, but you quickly learn the coding you use regularly. To make some journal editors happy, the following code can be added before the last bracket to ensure that tick marks are pointing inwards: `scales = list(tck = c (-1, 0)`. More data exploration tools will be demonstrated later in this book.

2.2 The Linear Regression Model

In the second step of the data analysis, we have to apply some sort of model, and the 'mother of all models' is without doubt the linear regression model. The *bivariate* linear regression model is defined by

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

The Y_i is the response (or dependent) variable, and X_i is the explanatory (or independent) variable. The unexplained information is captured by the residuals ε_i, and these are assumed to be normally distributed with expectation 0 and variance σ^2. The parameters α and β are the population intercept and slope and are unknown. In practice, we take a sample and use this to come up with estimates a and b and confidence intervals. These confidence intervals tell us that if we repeat the experiment a large number of times, how often the real (fixed and unknown) α and β are in the interval based on the confidence bands (which will differ for each experiment!). A typical choice is the 95% confidence interval. In most cases, β (the slope) is of primary interest as it tells us whether there is a relationship between Y and X.

So, we take a sample of size N and obtain the estimators a and b plus confidence intervals. And then, we make a statement on the population parameters α and β. But this is a big thing to do! You may wonder how it is possible that we can do this. Well, the magic answer is 'assumptions'. The fact that you take sample data and use this to make a statement on population parameters is based on a series of assumptions, namely, normality, homogeneity, fixed X, independence, and correct model specification.

The underlying geometric principle of linear regression is shown in Fig. 2.5 (based on Figs. 5.6 and 5.7 in Zuur et al. (2007), and Fig. 14.4 in Sokal and Rohlf (1995)). The data used in this graph is from a benthic study carried out by RIKZ in The Netherlands. Samples at 45 stations along the coastline were taken and benthic species were counted. To measure diversity, the species richness (the different number of species) per site was calculated. A possible factor explaining species richness is Normal Amsterdams Peil (NAP), which measures the height of a site compared to average sea level, and represents a measure of food for birds, fish, and benthic species. A linear regression model was applied, and the fitted curve is the straight line in Fig. 2.5. The Gaussian density curves on top of the line show the probability of other realisations at the same NAP values. Another 'realisation' can be thought of as going back into the field, taking samples at the same environmental conditions, carry out the species identification, and again determining species richness per site. Obviously, you will not find exactly the same results. The normality assumption means that for each NAP value, we have bell-shaped curves determining the probabilities of the (species richness) values of other realisations or sub-samples. Homogeneity means that the spread of all Gaussian curves is the same at all NAP values.

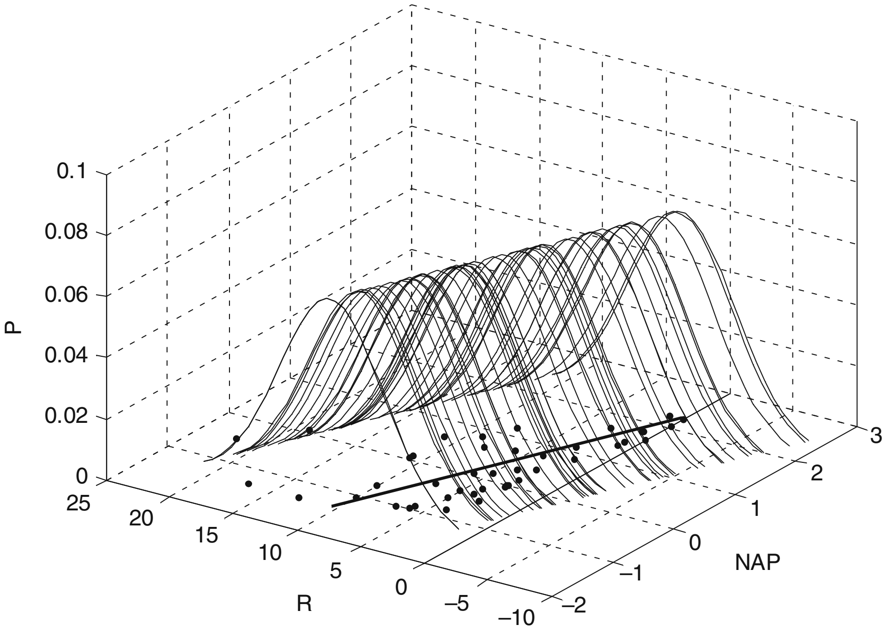

Fig. 2.5 Regression curve for all 45 observations from the RIKZ data discussed in Zuur et al. (2007) showing the underlying theory for linear regression. NAP is the explanatory variable, R (species richness) is the response variable, and the third axis labelled 'P' shows the probability of other realisations

Multiple linear regression is an extension of bivariate linear regression in the sense that multiple explanatory variables are used. The underlying model is given by

$$Y_i = \alpha + \beta_1 \times X_{1i} + \beta_2 \times X_{2i} + \ldots + \beta_M \times X_{Mi} + \varepsilon_i \quad \text{where} \quad \varepsilon_i \sim N(0, \sigma^2)$$

There are now *M* explanatory variables. Visualising the underlying theory as in Fig. 2.5 is not possible, as we cannot draw a high dimensional graph on paper, but the same principle applies. Further information on bivariate and multiple linear regression are discussed in the examples below and in Appendix A.

2.3 Violating the Assumptions; Exception or Rule?

2.3.1 Introduction

One of the questions that the authors of this book are sometimes faced with is: 'Why do we have to do all this GLM, GAM, mixed modelling, GLMM, and GAMM stuff? Can't we just apply linear regression on our data?' The answer is always in a 'Yes you can, but...' format. The 'but...' refers to the following. Always apply the simplest statistical technique on your data, but ensure it is applied correctly! And here is a crucial problem. In ecology, the data are seldom modelled adequately by linear regression models. If they are, you are lucky. If you apply a linear regression model on your data, then you are implicitly assuming a whole series of assumptions, and once the results are obtained, you need to verify all of them. This is called the model validation process. We already mentioned the assumptions, but will do this again; (i) normality, (ii) homogeneity, (iii) fixed *X* (*X* represents explanatory variables), (iv) independence, and (v) a correct model specification. So, how do we verify these assumptions, and what should we do, if we violate some, or all of them? We discuss how to verify these assumptions using five examples later in this section with each example violating at least one assumption. What should we do if we violate all the assumptions? The answer is simple: reject the model. But what do we do if we only violate one of the assumptions? And how much can we violate the assumptions before we are in trouble? We discuss this later.

2.3.2 Normality

Several authors argue that violation of normality is not a serious problem (Sokal and Rohlf, 1995; Zar, 1999) as a consequence of the central limit theory. Some authors even argue that the normality assumption is not needed at all provided the sample size is large enough (Fitzmaurice et al., 2004). Normality at each *X* value should be checked by making a histogram of all observations at *that* particular *X* value. Very often, we don't have multiple observations (sub-samples) at each *X* value. In that case, the best we can do is to pool all residuals and make a histogram of the pooled residuals; normality of the pooled residuals is reassuring, but it does not imply normality of the population data.

We also discuss how not to check for normality as the underlying concept of normality is grossly misunderstood by many researchers. The linear regression model requires normality of the data, and therefore of the residuals at *each* X value. The residuals represent the information that is left over after removing the effect of the explanatory variables. However, the raw data Y (Y represents the response variable) contains the effects of the explanatory variables. To assess normality of the Y data, it is therefore misleading to base your judgement purely on a histogram of all the Y data. The story is different if you have a large number of replicates at each X value. Summarising, unless you have replicated observations for each X value, you should not base your judgment of normality based on a histogram of the raw data. Instead, apply a model, and inspect the residuals.

2.3.3 Heterogeneity

Ok, apparently we can get away with a small amount of non-normality. However, heterogeneity (violation of homogeneity), also called heteroscedasticy, happens if the spread of the data is not the same at each X value, and this can be checked by comparing the spread of the residuals for the different X values. Just as in the previous subsection, we can argue that most of the time, we don't have multiple observations at each X value, at least not in most field studies. The only thing we can do is to pool all the residuals and plot them against fitted values. The spread should be roughly the same across the range of fitted values. Examples of such graphs are provided later. In sexual dimorphism, female species may show more variation than male species (or the other way around depending on species). In certain ecological systems, there may be more spread in the summer than in the winter, or less spread at higher toxicated sites, more spread at certain geographical locations, more variation in time due to accumulation of toxic elements, etc. In fact, we have seldom seen a data set in which there was no heterogeneity of some sort. The easiest option to deal with heterogeneity is a data transformation. And this is where the phrase 'a mean-variance stabilising' transformation comes from.

Many students have criticised us for using graphical techniques to assess homogeneity, which require some level of subjective assessment rather than using one of the many available tests. The problem with the tests reported by most statistical software packages, and we will illustrate some of them later, is that they require normality. For example, Barlett's test for homogeneity is quite sensitive to non-normality (Sokal and Rohlf, 1995). We therefore prefer to assess homogeneity purely based on a graphical inspection of the residuals.

Minor violation of homogeneity is not too serious (Sokal and Rohlf, 1995), but serious heterogeneity is a major problem. It means that the theory underlying the linear regression model is invalid, and although the software may give beautiful *p*-values, *t*-values and *F*-values, you cannot trust them. In this book, we will discuss various ways to deal with heterogeneity.

2.3.4 Fixed X

Fixed *X* is an assumption implying that the explanatory variables are deterministic. You know the values at each sample in advance. This is the case if you a priori select sites with a preset temperature value or if you choose the amount of toxin in a basin. But if you go into the field, take at random a sample, and then measure the temperature or the toxin concentration, then it is random. Chapter 5 in Faraway (2005) gives a very nice overview how serious violation of this assumption results in biased regression parameters. The phrase 'biased' means that the expected value for the estimate parameter does not equal the population value. Fortunately, we can ignore the problem if the error in determining the explanatory variable is small compared to the range of the explanatory variable. So, if you have 20 samples where the temperature varies between 15 and 20 degrees Celsius, and the error of your thermometer is 0.1, then you are ok. But the age determination of the whales in Fig. 2.4 may be a different story as the range of age is from 0 to 40 years, but the error on the age reading may (or may not) be a couple of years. There are some elegant solutions for this (see the references for this in Faraway (2005)), but in Chapter 7 we (shortly) discuss the use of a brute force approach (bootstrapping).

2.3.5 Independence

Violation of independence is the most serious problem as it invalidates important tests such as the *F*-test and the *t*-test. A key question is then how do we identify a lack of independence and how do deal with it. You have violation of independence if the *Y* value at X_i is influenced by other X_i (Quinn and Keough, 2002). In fact, there are two ways that this can happen: either an improper model or dependence structure due to the nature of the data itself. Suppose you fit a straight line on a data set that shows a clear non-linear pattern between *Y* and *X* in a scatterplot. If you plot the residuals versus X, you will see a clear pattern in the residuals: the residuals of samples with similar *X* values are all positive or negative. So, an improper model formulation may cause violation of independence. The solution requires a model improvement, or a transformation to 'linearise the relationship'. Other causes for violation of independence are due to the nature of the data itself. What you eat now depends on what you were eating 1 minute ago. If it rains at 100 m in the air, it will also rain at 200 m in the air. If we have large numbers of birds at time *t*, then it is likely that there were also large numbers of birds at time $t - 1$. The same holds for spatial locations close to each other and sampling pelagic bioluminescence along a depth gradient. This type of violation of independence can be taken care of by incorporating a temporal or spatial dependence structure between the observations (or residuals) in the model.

The case studies later in the book contain various examples of both scenarios, but for now we look at a series of examples where some of these important assumptions have been violated.

2.3.6 Example 1; Wedge Clam Data

Figure 2.6 shows a coplot of biomass (labelled as AFD which stands for ash free dry weight) of 398 wedge clams (*Donax hanleyanus)* plotted against length for six different months (Ieno, unpublished data). The data used in this section were measured on a beach in Argentina in 1997. An initial scatterplot of the data (not shown here) showed a clear non-linear relationship, and therefore, both AFD and length were log-transformed to linearise the relationship. Note this transformation is only necessary if we want to apply linear regression. As an alternative, the untransformed data can be analysed with additive modelling (Chapter 3). The coplot in Fig. 2.6 indicates a clear linear relationship between AFD and length in all months, and it seems sensible to apply linear regression to model this relationship. Due to different stages of the life cycle of wedge clams, the biomass-length relationship may change between months, especially before and after the spawning period in September–October and February–March. This justifies adding a length–month interaction term. This model is also known as an analysis of covariance (ANCOVA). The following R code was used for the coplot (Fig. 2.6) and the linear regression model.

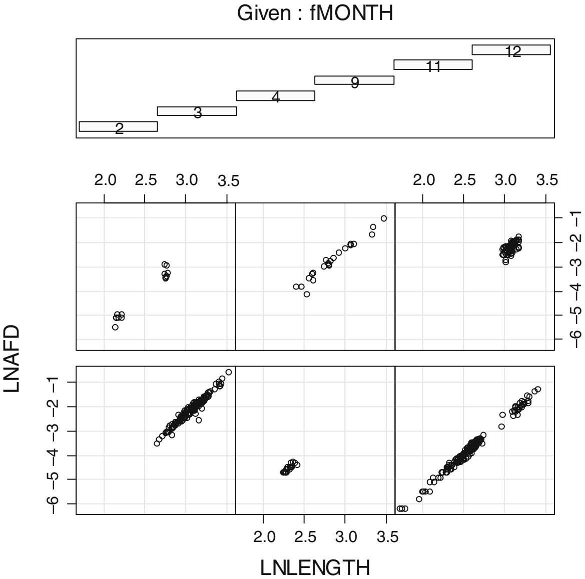

Fig. 2.6 Coplot of the wedge clam data during the spring and summer period. (The data were taken on the southern hemisphere.) The lower left panel contains the data from month 2, the lower right of month 4, the upper left from month 9, and the upper right of month 12

```
> library(AED); data(Clams)
> Clams$LNAFD <- log(Clams$AFD)
> Clams$LNLENGTH <- log(Clams$LENGTH)
> Clams$fMONTH <- factor(Clams$MONTH)
> library(lattice)
> coplot(LNAFD ~ LNLENGTH | fMONTH, data = Clams)
> M1 <- lm(LNAFD ~ LNLENGTH * fMONTH, data = Clams)
> drop1(M1,test = "F")
```

The drop1 command compares the full model with a model in which the interaction is dropped, and an *F*-test is used to compare the residual sum of squares of both the models (Appendix A):

```
Single term deletions
Model: LNAFD ~ LNLENGTH * fMONTH
                Df Sum of Sq     RSS      AIC F value    Pr(F)
<none>                          6.36 -1622.35
LNLENGTH:fMONTH  5      0.23    6.58 -1618.47  2.7385 0.01906
```

On the third line of this output (labelled as none), we have the output of the full model, and the last line shows the output from the model without the interaction. Note that this model is nested within the full model. The *F*-statistic shows that the interaction is significant at the 5% level. However, before trusting the values obtained by the *F*-statistic and use the ‘magic’ 5% as rejection level, we need to be confident that all model assumptions are valid. Hence, we enter the next stage of the analysis, the model validation.

2.3.6.1 Model Validation

Standard model validation graphs are (i) residuals versus fitted values to verify homogeneity, (ii) a QQ-plot or histogram of the residuals for normality, and (iii) residuals versus each explanatory variable to check independence, see Fig. 2.7. We also need to check whether there are any influential observations. The following R code was used to generate Fig. 2.7.

```
> op <- par(mfrow = c(2, 2), mar = c(5, 4, 1, 2))
> plot(M1, add.smooth = FALSE, which = 1)
> E <- resid(M1)
> hist(E, xlab = "Residuals", main = "")
> plot(Clams$LNLENGTH, E, xlab = "Log(Length)",
       ylab = "Residuals")
> plot(Clams$fMONTH, E, xlab = "Month",
       ylab = "Residuals")
> par(op)
```

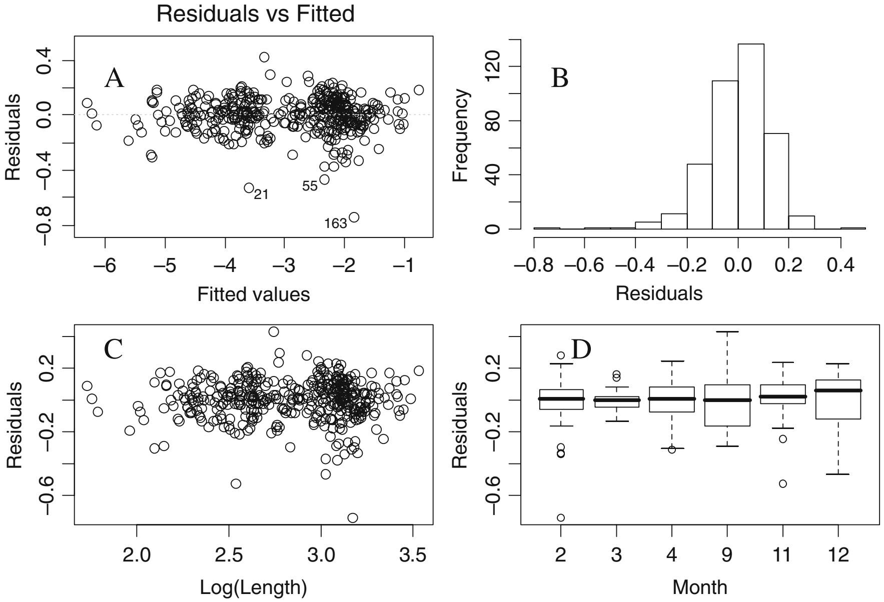

Fig. 2.7 Model validation graphs. **A**: Fitted values versus residuals (homogeneity). **B**: Histogram of the residuals (normality). **C**: Residuals versus length (independence). **D**: Residuals versus month

The first line specifies a graphical window with four panels and a certain amount of white space around each panel. The last command `par(op)` sets the graphical settings back to the default values. There seems to be minor evidence of non-normality (Fig. 2.7B), and more worrying, the spread in the residuals is not the same at all length classes and months (Fig. 2.7A, C, D). In month 3, there is less spread than in other months. A and C of Fig. 2.7 are similar in this case, but if we had a larger number of explanatory variables, these panels would no longer share this similar appearance.

The residuals play an essential part in the model validation process. Residuals are defined as observed values minus fitted values (we call these the ordinary residuals). However, it is also possible to define other types of residuals, namely standardised residuals and Studentised residuals. In Appendix A, we discuss the definition of the standardised residuals. These have certain theoretical advantages over the ordinary residuals, and it better to use these in the code above. Studentised residuals are useful for identifying influential observations. They are obtained by fitting a linear regression model using the full data set, and the same regression model on a data set in which one observation is dropped (in turn), and predicting the value of the dropped observation (Zuur et al., 2007). We do not use Studentised residuals here. However, if you do a good data exploration and deal with outliers at that stage, then ordinary, standardised, and Studentised residuals tend to be very similar (in terms of patterns).

Instead of a visual inspection, it is also possible to apply a test for homogeneity. Sokal and Rohlf (1995) describe three such tests, namely the Barlett's test for homogeneity, Hartley's F_{max} test and the log-anova, or Scheffé-Box test. Faraway (2005) gives an example of the *F*-test. It uses the ratio of variances. Panel 2.7C suggests that the observations for `log(Length)` less than 2.275 have a different spread than those larger than 2.275. The following code applies the *F*-ratio test, and the output is given immediately after the code.

```
> E1 <- E[Clams$LNLENGTH <= 2.75]
> E2 <- E[Clams$LNLENGTH > 2.75]
> var.test(E1, E2)

F test to compare two variances data: E1 and E2
F = 0.73, num df = 161, denom df = 235, p-value = 0.039
alternative hypothesis: true ratio of variances is not
equal to 1
95 percent confidence interval: 0.557 0.985
sample estimates: ratio of variances: 0.738
```

The null hypothesis (H_0) in this test is that the ratio of the two variances is equal to 0, and the test suggests rejecting it at the 5% level. However, $p = 0.04$ is not very convincing. On top of this, the choice for 2.275 is rather arbitrary. We can easily fiddle around with different cut-off levels and come up with a different conclusion. We could also use the F_{max} to test whether residuals in different months have the same spread (see page 397 in Sokal and Rohlf, 1995). We will address the same question with the Bartlett test for homogeneity. The null hypothesis is that variances in all months are the same. The following code and output shows that we can reject the null hypothesis at the 5% level.

```
> bartlett.test(E, Clams$fMONTH)

Bartlett test of homogeneity of variances
data:  E and MONTH
Bartlett's K-squared = 34.28, df = 5, p-value = <0.001
```

The problem with the Bartlett test is that it is rather sensitive to non-normality; hence, one should make histograms of residuals per month. Results are not presented here, but the R command `hist(E[Clams$MONTH = = 12])` gives a bimodal histogram.

The conclusion of the linear regression (or ANCOVA) model is that there is a significant relationship between biomass, length, and month with a weak but significant interaction between the length and the month. However, with a *p*-value of 0.02 for this interaction term, we would have preferred to see no patterns at all in the residuals. Both the tests and graphical output, gave us some reasons to doubt the suitability of this model for these data. In Chapter 4, we discuss extensions of the linear regression model that can be used to test whether we need different variances per month.

2.3.7 Example 2; Moby's Teeth

Figure 2.4 showed nitrogen isotope ratios in teeth of stranded whales. One of which became famous and attracted newspaper headlines when it stranded in Edinburgh, Scotland, and was nicknamed 'Moby the whale'. The graph in Fig. 2.4 indicates that Moby's isotope ratios increased with age, and a linear regression was applied to model this pattern. The following code was used to access the data, rename the object with a very long name (TeethNitrogen) into something much shorter, apply linear regression on Moby's data, and make the validation graphs in Fig. 2.8.

```
> library(AED); data(TeethNitrogen)
> TN <- TeethNitrogen
> M2 <- lm(X15N ~ Age, subset = (TN$Tooth == "Moby"),
           data = TN)
> op <- par(mfrow = c(2, 2))
> plot(M2, add.smooth = FALSE)
> par(op)
```

Figure 2.8 is the typical graphical output produced by the `plot` command in R. Based on the QQ-plot in panel B, the residuals look normally distributed (if the points are in a line, normality can be assumed). Panel D identifies potential and

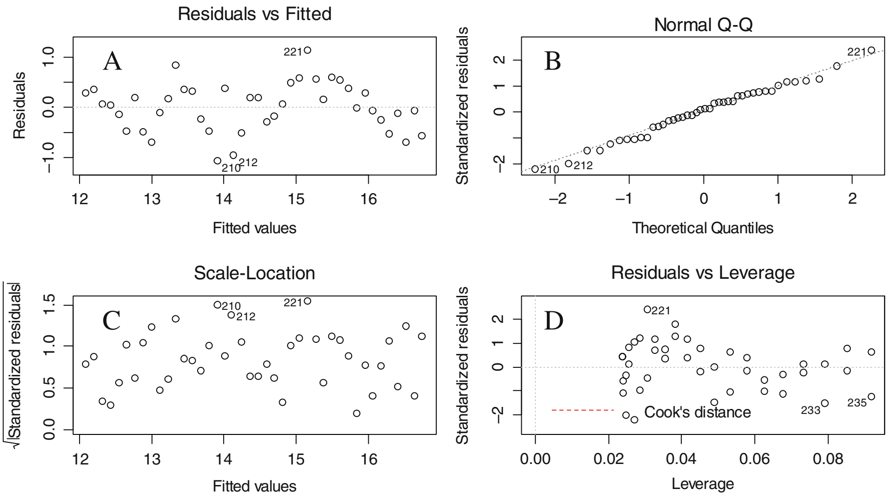

Fig. 2.8 Model validation graphs obtained by applying a linear regression model on the teeth data from Moby. Panel **A** and **C** show residuals versus fitted values; note the clear pattern! Panel **B** is a QQ-plot for normality, and Panel **D** shows the standardised residuals versus leverage and the Cook statistic is superimposed as contour plots. In this case, the Cook values are small and cannot be clearly seen

influential observations. It is a scatterplot of leverage against residuals. Leverage measures whether any observation has extreme values of the explanatory variables. If there is only one explanatory variable, then a Cleveland dotplot or boxplot will identify such points. However, an observation may have a combination of values of explanatory variables that make it unique in terms of 'environmental' conditions. None of the data exploration methods mentioned so far will detect this. If such a point has a 'large' influence on the linear regression model, we may decide to remove it. And this is measured by the Cook distance (a leave-one-out measure of influence), which is superimposed with contour lines in panel D. We will return to the Cook distance later (Appendix A) as the default output of R is not the best way to present the Cook distance. In this case, there are no observations with a Cook distance larger than 1, which is the threshold value upon one should take further action (Fox, 2002). Summarising, leverage indicates how different an individual observation is compared to the other observations in terms of the values of the explanatory variables; the Cook distance tells you how influential an observation is on the estimated parameters.

Figure 2.8A shows residuals versus fitted values. Violation of homogeneity can be detected if this panel shows any pattern in the spread of the residuals. Panel C is based on the same theme. However, in panel C, the residuals are square-root transformed (after taking the absolute values) and weighted by the leverage. Both panels A and C can be used to assess homogeneity. The spread seems to be the same everywhere; however, panel A shows a clear problem: violation of independence. There are in fact two violations to deal with here. The first one can be seen better from Fig. 2.9. It shows the observed values plotted against age with a fitted linear regression curve added. There are groups of sequential residuals that are above and below the regression line.

The graph was obtained by

```
> N.Moby <- TN$X15N[TN$Tooth == "Moby"]
> Age.Moby <- TN$Age[TN$Tooth == "Moby"]
```

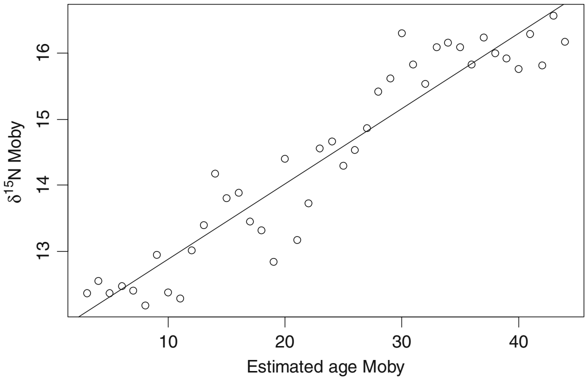

Fig. 2.9 Observed nitrogen isotope ratios plotted versus age for Moby the whale. The line is obtained by linear regression

```
> plot(y = N.Moby, x = Age.Moby,
     xlab = "Estimated age Moby",
     ylab = expression(paste(delta^{15}, "N Moby")))
> abline(M2)
```

To keep the code for the `plot` command simple, we defined the variables `N.Moby` and `Age.Moby`. The `abline` command draws the fitted regression curve. Applying an additive model (Chapter 3) or adding more covariates may solve the misfit. The other form of dependence is due to the nature of these data; high nitrogen isotope ratios at a certain age may be due to high nitrogen values at younger ages. To allow for this type of dependence, some sort of auto-correlation structure on the data is needed, and this is discussed in Chapters 5, 6, and 7.

The relevant numerical output obtained by the `summary(M2)` command is given by

```
            Estimate Std. Error  t-value   p-value
(Intercept) 11.748      0.163     71.83    <0.001
Age.Moby      0.113      0.006     18.40    <0.001

Residual standard error: 0.485 on 40 degrees of freedom
Multiple R-Squared: 0.894, Adjusted R-squared: 0.891
F-statistic: 338.4 on 1 and 40 DF, p-value: < 0.001
```

The output shows the estimated intercept and slope (plus standard errors, t-values and p-values). We also get information on R^2 and the adjusted R^2 (the latter one can be used to select the best model if there are any non-significant terms in the model), the square root of the variance (residual standard error), and the F-statistic (which is testing the null hypothesis whether all slopes, one is this case, are equal to zero). The estimated model is given by

$$y_i = 11.748 + 0.113 \times age_i$$

The estimated slope and intercept are significantly different from 0 at the 5% level. The model explains 89% of the variation; the estimator for σ is equal to $s = 0.486$. But the problem is that we still have to reject this model because there is a clear violation of independence. Solutions will be given in Chapters 6 and 7.

2.3.8 Example 3; Nereis

In the third example, we present the results of a linear regression model applied on the Nereis data, presented earlier in this chapter. The concentration is modelled as a function of nutrient, biomass, and their interaction. This can also be called a 2-way ANOVA with interaction. The following R code accesses the data, defines

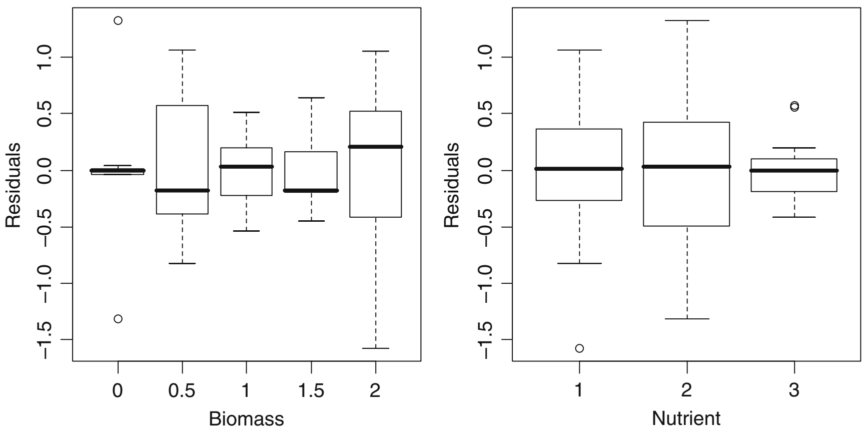

Fig. 2.10 Model validation graphs for the Nereis data showing heterogeneity. Residuals are plotted versus biomass and nutrient

the explanatory variables biomass and nutrient as factors, applies linear regression, and plots the validation graphs in Fig. 2.10. Note that homogeneity is violated!

```
> library(AED); data(Nereis)
> Nereis$fbiomass <- factor(Nereis$biomass)
> Nereis$fnutrient <- factor(Nereis$nutrient)
> M3 <- lm(concentration ~ fbiomass * fnutrient,
           data = Nereis)
> drop1(M3, test = "F")
> op <- par(mfrow = c(1, 2))
> plot(resid(M3) ~ Nereis$fbiomass, xlab = "Biomass",
       ylab = "Residuals")
> plot(resid(M3) ~ Nereis$fnutrient,
       xlab = "Nutrient", ylab = "Residuals")
> par(op)
```

The numerical output obtained by the `drop1` command is printed below and shows that the biomass-nutrient interaction term is significant at the 5% level.

```
Single term deletions
Model: concentration ~ fbiomass * fnutrient
                   Df Sum of Sq     RSS     AIC F value  Pr(F)
<none>                           13.630 -23.746
fbiomass:fnutrient  8    11.553  25.183 -12.121  3.1785 0.0099
```

However, the boxplot of (i) residuals versus nutrient and (ii) residuals versus biomass in Fig. 2.10 shows a clear violation of homogeneity. Applying a transformation on concentration may solve this problem. The disadvantage of a transformation is that we are changing the type of relationship between response and explanatory variables. So, again we need to reject the linear regression model for these data.

2.3.9 Example 4; Pelagic Bioluminescence

In Gillibrand et al. (2007), pelagic bioluminescence along a depth gradient in the northeast Atlantic Ocean is analysed. Figure 2.11 shows an `xyplot` from the lattice package. Each panel represents a station. The underlying questions are (i) how to model the bioluminescent–depth relationship and (ii) how to deal with the data of difference stations. The following code was used to read the data and make the lattice panel.

```
> library(AED); data(ISIT)
> ISIT$fStation <- factor(ISIT$Station)
> library(lattice)
> xyplot(Sources ~ SampleDepth | fStation, data = ISIT,
    xlab = "Sample Depth", ylab = "Sources",
    strip = function(bg = 'white', ...)
    strip.default(bg = 'white', ...),
    panel = function(x, y) {
                panel.grid(h = -1, v = 2)
                I1 <- order(x)
                llines(x[I1], y[I1], col = 1)})
```

You can see this code is slightly more complicated than used for Fig. 2.4. In this code, we used a panel function that automatically splits up the data by station. When R enters this panel function, the *x* and the *y* variables are the data for one particular station. We then have a range of options in the way we can display this *x* and *y* data. First, we add a grid using the `panel.grid` command. If you don't like the grid, just remove this command. The `I1 <- order (x)` determines the order of age as we did not sort the data before importing into R. Finally, we added lines between points with sequential ages. Omitting the `order` command and removing the `[I1]` in the `llines` function produces a spaghetti plot.

There is no point in applying a linear regression model with Sources as the response variable and Depth and Station as explanatory variables (plus an interaction between them) because the relationships are not linear and the variation per station differs. Perhaps it is better to consider station as a random effect (Chapter 5). Another problem is that depth can be seen as a spatial gradient. Hence, there may be spatial correlation along the depth gradient. In Chapter 5, we discuss random effect models, and in Chapter 7 spatial correlation for smoothing models. A full analysis of this data set is presented in Chapter 17.

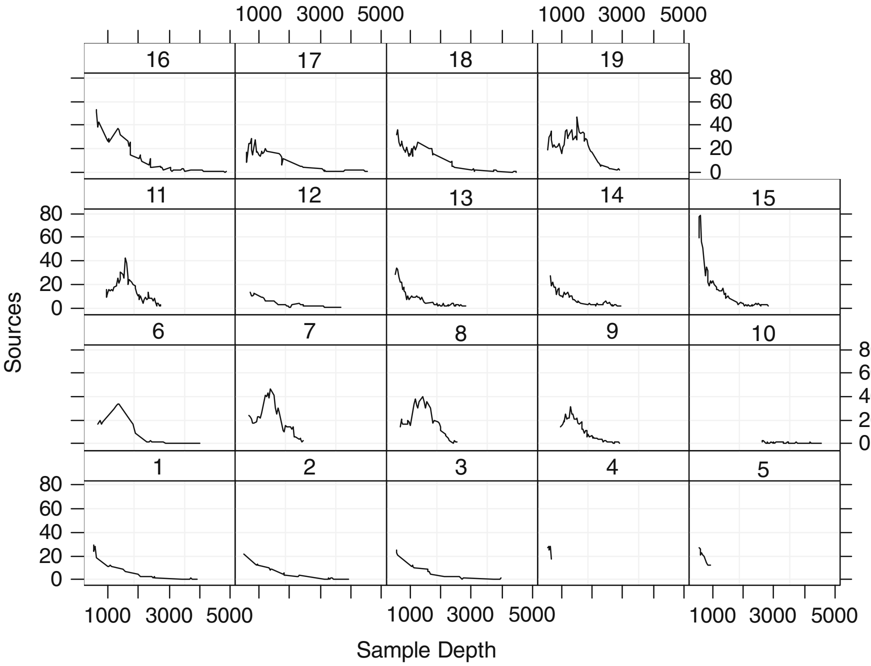

Fig. 2.11 Pelagic bioluminescence (labelled as Sources) along a depth gradient in the northeast Atlantic Ocean. Each panel represents a station

2.4 Where to Go from Here

The data exploration should filter out any typing mistakes (typos), identify possible outliers and the need for a data transformation, and provide some ideas about the follow up analyses. As for typos, these should obviously be corrected before continuing with any analysis, but do not apply a transformation on the response variable yet unless there are strong reasons to do so. Some of the methods discussed in later chapters may be able to deal with (groups) of extreme observations or heterogeneity. Many books will tell you to routinely apply a data transformation to linearise the relationship. Well, if you are particular fond on linear regression then yes, but (generalised) additive (mixed) modelling is especially designed to model non-linear relationships. Even heterogeneity, as for example encountered in Fig. 2.1B can be dealt with (as will be explained in Chapter 4); so you do not need to apply a transformation to stabilise the mean-variance relationship, provided you are willing to read the rest of this book. The only thing we cannot solve with any of the techniques discussed in later chapters is observations with extreme explanatory variables. If this happens for your data, then a transformation on the explanatory variable(s) could well be justified at this stage.

The original aim of this chapter was to simply illustrate the linear regression model for an ecological data set and discuss the numerical and graphical output. However, in preparing this book, we had access to about 15 data sets, and in Zuur et al. (2007), we had access to a further 20 data sets. In none of these real data sets could we find a non-trivial example for a linear regression model for which all assumptions held. This clearly identifies the limitation of linear regression for analysing ecological data. Hence, our choice of the title of this chapter.

So, what can we do? The problem of heterogeneity can be solved by either allowing for different variances in the linear regression model (using generalised least squares estimation) or using a different distribution and model structure (Poisson, negative binomial and Gamma distributions in GLM); the dependence problem requires the use of models that allow for more flexibility than regression (e.g. smoothing methods) and a model for the error structure (e.g. temporal, spatial correlation, or along another gradient like age or depth). We will also need to consider nested data and random effects. Taken together, all these techniques lead to mixed

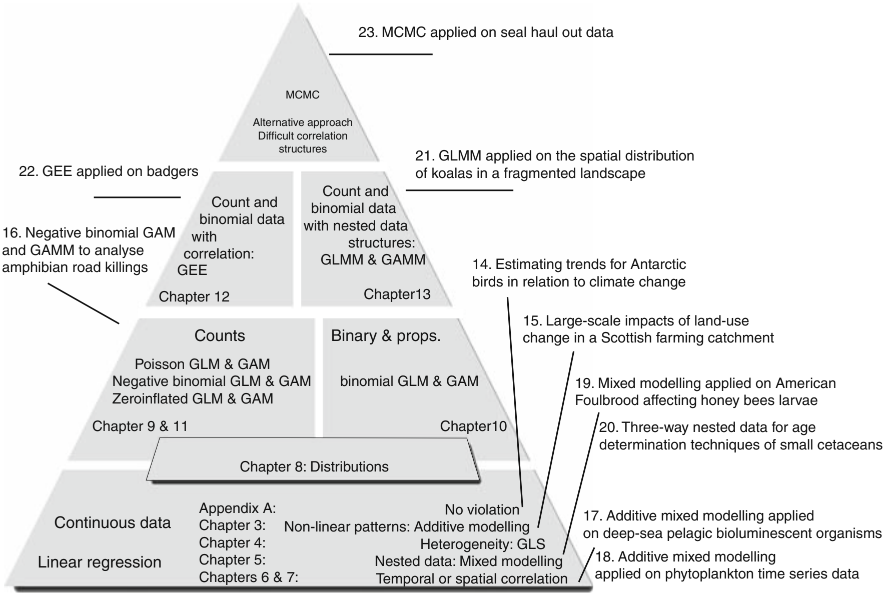
23. MCMC applied on seal haul out data
MCMC
Alternative approach
Difficult correlation structures
22. GEE applied on badgers
21. GLMM applied on the spatial distribution of koalas in a fragmented landscape
Count and binomial data with correlation: GEE
Chapter 12
Count and binomial data with nested data structures: GLMM & GAMM
Chapter13
16. Negative binomial GAM and GAMM to analyse amphibian road killings
14. Estimating trends for Antarctic birds in relation to climate change
15. Large-scale impacts of land-use change in a Scottish farming catchment
Counts
Poisson GLM & GAM
Negative binomial GLM & GAM
Zeroinflated GLM & GAM
Chapter 9 & 11
Binary & props.
binomial GLM & GAM
Chapter10
19. Mixed modelling applied on American Foulbrood affecting honey bees larvae
20. Three-way nested data for age determination techniques of small cetaceans
Chapter 8: Distributions
Continuous data
Linear regression
Appendix A:
Chapter 3:
Chapter 4:
Chapter 5:
Chapters 6 & 7:
No violation
Non-linear patterns: Additive modelling
Heterogeneity: GLS
Nested data: Mixed modelling
Temporal or spatial correlation
17. Additive mixed modelling applied on deep-sea pelagic bioluminescent organisms
18. Additive mixed modelling applied on phytoplankton time series data

Fig. 2.12 Overview of all the chapters in this book. Linear regression is discussed in Appendix A. Additive modeling, generalised least squares (GLS), and mixed modelling techniques are presented in Chapters 4, 5, 6, and 7. Chapter 8 contains an explanation of the Poisson, negative binomial, Bernoulli, binomial, and zero-truncated distributions. GLM and GAM models are discussed in Chapters 9, 10, and 11, and finally, Chapters 12 and 13 contain GEE, GLMM, and GAMM. Associated case studies are printed outside the triangle. Chapter 23 contains an application of Markov Chain Monte Carlo (MCMC), which can be used as an alternative estimation technique or if the correlation structure is more complicated than the R functions for mixed modeling, GLMM and GAMM can cope with

modelling approach, and if combined with GLM and GAM, to generalised linear mixed modelling (GLMM) and generalised additive mixed modelling (GAMM).

Chapter 4 shows how we can deal with heterogeneity in linear regression and smoothing models, random effects for nested data are introduced in Chapter 5, and temporal and spatial correlation structures are discussed in Chapters 6 and 7. In Chapter 8, we introduce different distributions for count data, binary data, proportional data, and zero inflated count data. These are then used in Chapters 9, 10, and 11. Finally, Chapters 12 and 13 discuss how we can incorporate correlation structures and random effects in models for count data, binary data, and proportional data. See Fig. 2.12 for a schematic overview.

Before reading on, we strongly advise to read Appendix A as it provides a more detailed discussion on linear regression. It is essential that you are familiar with all steps discussed in this appendix.