4 Exploration

The first step in analysing data is a graphical data exploration asking the following questions:

1. Where are the data centred? How are they spread? Are they symmetric, skewed, bimodal?
2. Are there outliers?
3. Are the variables normally distributed?
4. Are there any relationships between the variables? Are relationships between the variables linear? Which follow-up analysis should be applied?
5. Do we need a transformation?
6. Was the sampling effort approximately the same for each observation or variable?

We need to address all of these questions because the next step of the analysis needs the data to comply with several assumptions before any conclusions can be considered valid. For example, principal component analysis (PCA) depends on linear relationships between variables, and outlying values may cause non-significant regression parameters and mislead the analysis. Another example is large overdispersion in generalised linear modelling, which can also result in non-significant parameters. We therefore need a range of exploratory tools to address questions 1 to 6 with different tools aimed at answering different questions. For example, a scatterplot might suggest that a particular point is an outlier in the combined *xy*-space, but not identify it as an outlier within in the *x*-space or *y*-space if inspected in isolation. This chapter discusses a range of exploratory tools and suggests how they can be used to ensure the validity of any subsequent analysis. When looking at your data you should use all the techniques discussed and not rely on the results from a single technique to make decisions about outliers, normality or relationships.

Many books have chapters on data exploration techniques, and good sources are Montgomery and Peck (1992), Crawley (2002), Fox (2002a) and Quinn and Keough (2002). We have only presented the methods we find the most useful. Expect to spend at least 20% of your research time exploring your data. This makes the follow-up analysis easier and more efficient.

4.1 The first steps

Boxplots and conditional boxplots

A boxplot, or a box-and-whiskers plot (Figure 4.1), visualises the mean and spread for a univariate variable. Normally, the midpoint of a boxplot is the median, but it can also be the mean. The 25% and 75% quartiles (Q_{25} and Q_{75}) define the hinges (end of the boxes), and the difference between the hinges is called the spread. Lines (or whiskers) are drawn from each hinge to 1.5 times the spread or to the most extreme value of the spread, whichever is the smaller. Any points outside these values are normally identified as outliers. Some computer programmes draw the whiskers to the values covering most data points, such as 10% and 90% of the points, and show minimum and maximum values as separate points.

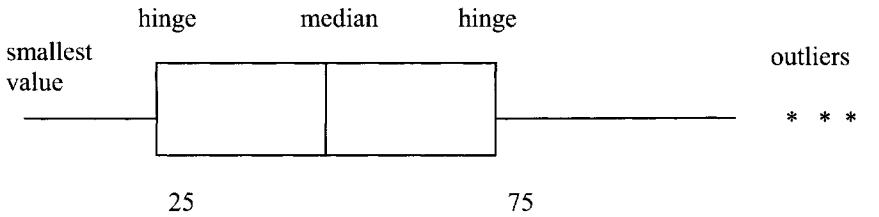

Figure 4.1. Boxplots show the middle of the sampled values, variability, shape of the distribution, outliers and extreme values.

The numbers below give the number of ragworms (*Laeoneris acuta*) recorded in an Argentinean salt marsh, and we use them to explain making a boxplot. The top row identifies the ranked sampling point, and the lower row gives the number of ragworm counted at that point.

					Q_{25}					**M**					Q_{75}					
1	2	3	4	5	**6**	7	8	9	10	**11**	12	13	14	15	**16**	17	18	19	20	21
0	0	0	1	2	**3**	6	7	9	11	**14**	14	14	16	19	**20**	21	24	27	35	121

The median value (denoted by **M**) for *L. acuta* is at observation number 11 (14 ragworms). The end points of the boxes in the boxplot are at Q_{25} and Q_{75}. Therefore, observation numbers 6 and 16 form the hinges. The spread for these data is 20 − 3 = 17, and 1.5 times the spread is 25.5. Adding 25.5 to the upper hinge of 20 (Q_{75}) allows the right line (or whisker) to be drawn up to 45.5. Observation number 21 (121 ragworms) would normally be considered as an extreme value or outlier. The resulting boxplot is given in Figure 4.2.

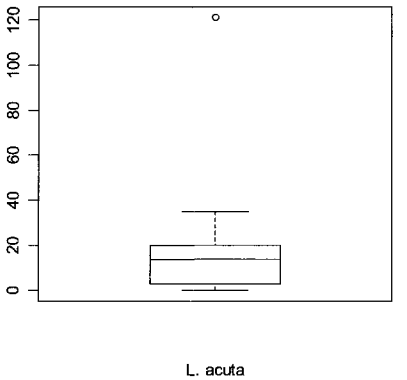

Figure 4.2. Boxplot for the ragworms (*L. acuta*). The upper hinge is calculated as having a value of 45.5, but as the most extreme value within this range is only 35, it is drawn at this latter point.

In Chapter 28, zoobenthic data from a salt marsh in Argentina are analysed. The data consist of measurements on four zoobenthic species in three transects. Each transect contained ten sites, and all sites were measured in Autumn and Spring, resulting in a 60-by-4 data matrix for the species data. Further details can be found in Chapter 28. Several boxplots for the species data are shown in Figure 4.3. Panel A in Figure 4.3 is a boxplot for the four zoobenthic species of the Argentinean zoobenthic dataset introduced in Chapter 2. It shows that some species have potential outliers, which prompted an inspection of the original data to check for errors in data entry. After checking, it was concluded that there were no data errors. However, the presence of outliers (or large observations) is the first sign that you may need to transform the data to reduce or down-weight its influence on the analysis. We decided to apply a square root transformation, and boxplots of the transformed data are shown in Figure 4.3-B. The reasons for choosing a square root transformation is discussed later. Note that the boxplots for the transformed data show that this has removed the outliers. The large number of dots outside the interval defined by 1.5 times the range might indicate a large number of zero observations for *Uca uruguayensis* and *Neanthes succinea*. This is called the double-zero problem, but how big a problem this is depends on the underlying ecological questions. If two variables have many data points with zero abundance, the correlation coefficient will be relatively large as both variables are below average at the same sites. This means that these two variables are identified as similar, only because they are absent at the same sites. It is like saying that butterflies and elephants are similar because they are both absent from the North Pole, the Antarctic and the moon. It sounds trivial, but the first few axes in a principal component analysis could be determined by such variables, and it is a common problem in ecological data of which we need to be aware.

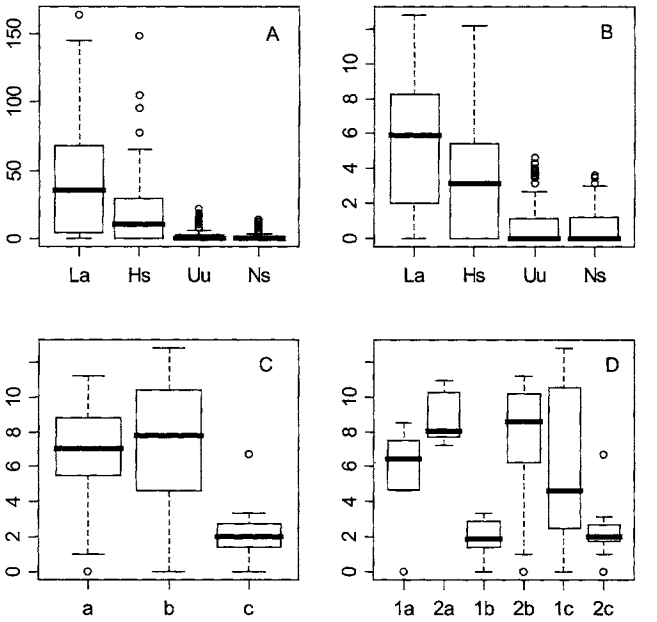

Figure 4.3. Boxplots. A: boxplots of the abundance of four zoobenthic species (La = *Laeonereis acuta*, Hs = *Heteromastus similis*, Uu = *Uca uruguayensis*, Ns=*Neanthes succinea*). B: boxplots of four square root transformed species. C: boxplot of square root transformed *L. acuta* conditional on the nominal variable transect with values a, b and c. D: Boxplot of square root transformed *L. acuta* conditional on season (1 = Autumn, 2 = Spring) and transect.

Boxplots are also useful to find relationships between variables. Panel C in Figure 4.3 shows the boxplot of square root transformed *L. acuta* abundance conditional on the nominal variable transect (a, b and c). It is readily seen that abundances are considerably lower in transect C. Panel D takes this one step further; the same species is now plotted conditional on season and transect. The first two boxplots from the left correspond to *L. acuta* from transect a in Autumn and Spring. Although this shows differences in abundances between the seasons, it also shows there appears to be no seasonal consistency between transects.

Depending on software, boxplots can be modified in various ways. For example, notches can be drawn at each side of the boxes. If the notches of two plots do not overlap, then the medians are significantly different at the 5% (Chambers et al. 1983). It is also possible to have boxplots with widths proportional to the square roots of the number of observations in the groups. Sometimes, it can be useful to plot the boxplot vertically instead of horizontally, where this might better visualise the characteristics of the original data.

Cleveland dotplot

Cleveland dotplots (Cleveland 1985) are useful to identify outliers and homogeneity. Homogeneity means that the variance in the data does not change along the gradients. Violation is called heterogeneity, and as we will see later, homogeneity is a crucial assumption for many statistical methods. Various software programmes use different terminology for dotplots. For example, with S-Plus and R, each observation is presented by a single dot. The value is presented along the horizontal axis, and the order of the dots (as arranged by the programme) is shown along the vertical axis. Cleveland dotplots for the abundance of four zoobenthic species of the Argentinean dataset are given in Figure 4.4. The 60 data points (30 sites in Spring and 30 sites in Autumn) are plotted along the vertical axes and the horizontal axes show the values for each site. Any isolated points on the right- or left-hand side indicate outliers, but in this dataset, none of the points are considered outliers. However, the dotplots also indicate a large number of zero observations, which can be a problem with some of the methods discussed in later chapters. Also note that the boxplots show that *L. acuta* and *H. similis* are considerably more abundant than the other two species. The dotplots were made using different symbols conditional on a nominal explanatory variable, which in this case is Transect. This means that data points from the same transect will have the same symbols. Note that *U. uruguayensis* has zero abundance along transect a in the Autumn (these are the bottom 10 data points along the y-axis); along transect c in the Autumn (these are the data points in the middle with a '+'); along transect a in the Spring (next 10 data points represented by 'o'); and along transect c in Spring (the upper 10 data points represented by '+'). Although we have not done it here, it would also be useful to make Cleveland dotplots for explanatory variables and diversity indeces.

You can also usefully compare boxplots with dotplots, as this can explain why the boxplot identified some points as 'outliers'. The boxplots and dotplots for the Argentinean zoobenthic data tell us that we have many zero observations, two species have larger abundances than the other species, there are no 'real' outliers, and there are differences in species abundances between transects and seasons.

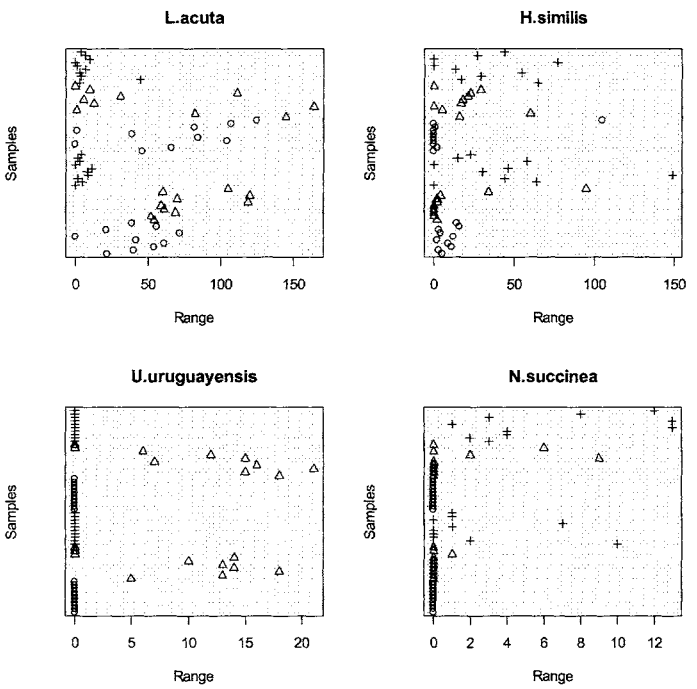

Figure 4.4. Dotplots for species of the Argentinean zoobenthic dataset. The horizontal axes show the value at each data point. The vertical axes represent the identity of the data points. The values at the top of the vertical axes are the data points at the end of the spreadsheet. It is also possible to group data points based on a nominal variable.

Histograms

A histogram shows the centre and distribution of the data and gives an indication of normality. However, applying a data transformation to make the data fit a normal distribution requires care. Panel A in Figure 4.5 shows the histogram for a set of data on the Gonadosomatic index (GSI, i.e., the weight of the gonads relative to total body weight) of squid (Graham Pierce, University of Aberdeen, UK, unpublished data). Measurements were taken from squid caught at various locations, months, and years in Scottish waters. The shape of the histogram shows bimodality, and one might be tempted to apply a transformation. However, a conditional histogram gives a rather different picture. In a conditional histogram the data are split up based on a nominal variable, and histograms of the subsets are plotted above each other. Panels B and C show the conditional histograms for the GSI index conditional on sex. Panel B shows the GSI index for female squid and Panel C for male squid. Notice that there is a clear difference in the shape and centre of the distribution. Hence, part of the first peak in panel A comprises mainly

the male squid. This suggests the need to include a sex effect and interactions rather than transform the full dataset. We also suggest making conditional histograms on year, month and location for these data.

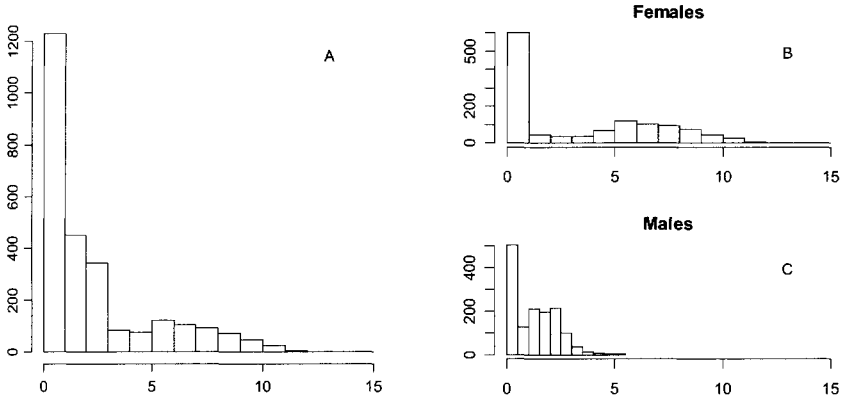

Figure 4.5. Histograms. A: histogram for GSI index of the squid data. B and C: conditional histograms for GSI index for the squid data. Panel B is for the female species and panel C for the male species.

QQ-plots

A Quantile-Quantile plot is a graphical tool used to determine whether the data follow a particular distribution. The QQ-plot for a normal distribution compares the distribution of a given variable to the Gaussian distribution. If the resulting points lie roughly on a straight line, then the distribution of the data is considered to be the same as a normally distributed variable. Let us discuss this in a bit more detail, as it can be quite confusing to understand what exactly it does. First, we need to revise some basic statistics. The p^{th} quantile point q for a random variable y is given by $F(q) = P(y \leq q) = p$. If we want to know which q value belongs to the p, we write $q = F^{-1}(p)$. Suppose we have five observations Y_i with values 1, 2, 3, 4 and 5. We have sorted the observations from the smallest to the highest. By definition the first number is the 0% percentile, the middle is the 50% percentile and 5 is the 100% percentile. The difference between a quantile and percentile point is only a factor 100. QQ-plots are either based on these percentiles, or more typically they use the sample quantile points $(i - 0.5)/n$ where i is from 1 to 5 and $n = 5$ for this example. The sample quantile points for these data are 0.1, 0.3, 0.5, 0.7 and 0.9. These are the sample values for p. In the second step, we compare these sample quantile points with that of a normal distribution. This means that the density function used in $P(y \leq q)$ is now a normal density function and $F()$ is the corresponding normal cumulative distribution function. The QQ-plot is then a plot of the samples values Y_i versus q_i. Some software packages add a straight line to the plot, which is typically obtained by connecting the 25^{th} and 75^{th} quartile points.

It is useful to combine QQ-plots with a power transformation, which is given by

$$\frac{Y^p - 1}{p} \quad \text{if } p \text{ is not equal to 0} \qquad \text{and} \qquad \log(Y) \text{ if } p \text{ is 0} \tag{4.1}$$

Note that this p is not the p that we used for the quantiles. It is also useful to compare several QQ-plots for different values of p, and Figure 4.6 shows an example for the Argentinean data. In this example the square root transformation seems to perform the best, but this could be open to debate.

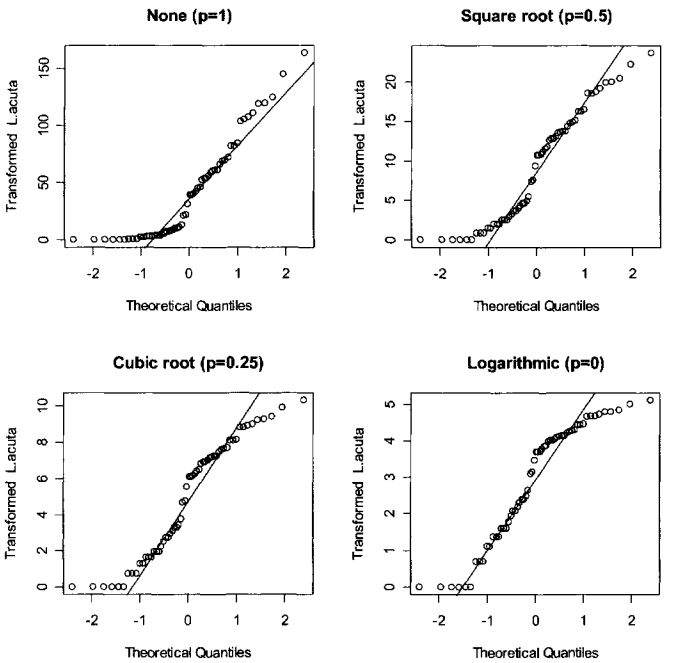

Figure 4.6. QQ-plots for the zoobenthic species *L. acuta* from the Argentinean zoobenthic dataset; for the untransformed data, square root transformed data, the cubic root transformed data, and $\log_{10}$ transformed data. In this example, the square root transformation seems to give the best results.

Scatterplot

So far, the main emphasis has been on detecting outliers, checking for normality, and exploring datasets associated with single nominal explanatory variables. The following techniques look at the relationships *between* variables.

A scatterplot is a tool to find a relationship between two variables. It plots one variable along the horizontal axis and a second variable along the vertical axis. To

help visualise the relationship between the variables, a straight line or smoothing curve is often added to the plot. Figure 4.7 shows the pairplot for the variables biomass and length for the wedge clam *Donax hanleyanus*, measured on a beach in Buenos Aires province, Argentina (Ieno, unpublished data).

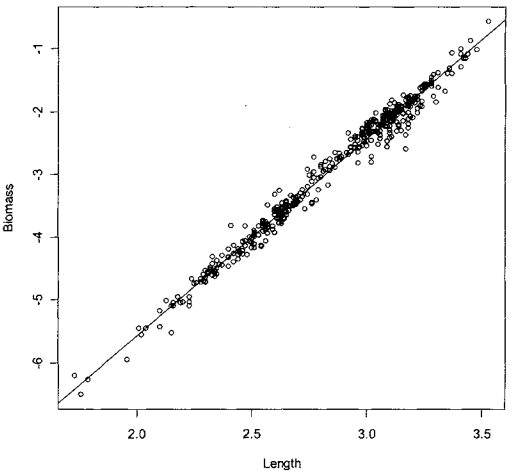

Figure 4.7. Scatterplot for biomass wedge clam dataset, using log transformed biomass, versus log transformed length.

Pairplot

If you have more than two variables, then a series of scatterplots can be produced: one for each pair of variables. However, the number of scatterplots required increases rapidly if you have more than three variables to explore. A better approach, for up to approximately 10 explanatory variables, is the pairplot, or scatterplot matrix (Figure 4.8). These show multiple pair-wise scatterplots in one graph and can be used to detect relationships between variables and to detect collinearity. The example in Figure 4.8 shows a pairplot for the response variable species richness and for four selected environmental variables. Species richness measures the different number of species per observation. The Decapoda zooplankton data form the basis for the case study in Chapter 20. Each panel is a scatterplot between two variables, with the labels for the variables printed in the panels running diagonally through the plot. A smoothing line has been added to help visualise the strength of the relationship. However, you can choose not to add a line, or you can add a regression line, whichever best suits the data. The pairplot in Figure 4.8 suggests a relationship between species richness (R) and temperature (T1m) and between species richness (R) and chlorophyll a (Ch). It also shows some collinearity between salinity at the surface (S1m) and at 35-45 meters (S45_35). Collinearity means that there is a high correlation between explanatory variables.

Figure 4.9 shows another pairplot for the same dataset where all the available explanatory variables have been plotted. The differences between this graph and the previous pairplot is that correlation coefficients between the variables are printed in the lower part of the graph. Note that there is strong collinearity between some of the variables, for example temperature at 1 m and temperature at 45 m.

Pairplots should be made for every analysis. These should include (i) a pairplot of all response variables (assuming that more than one response variable is available); (ii) a pairplot of all explanatory variables; and (iii) a pairplot of all response *and* explanatory variables. The first plot (i) gives information that will help choose the most appropriate multivariate techniques. It is hoped that the response variables will show strong linear relationships (some techniques such as PCA depend on linear relationships). However, if plot (ii) shows a clear linear relationship between the explanatory variables, indicating collinearity, then we know we have a major problem to deal with before further analysis. With plot (iii) we are judging whether the relationships between the response variables and the explanatory variables are linear. If this is not the case, then several options are available. The easiest option is to apply a transformation on response and/or explanatory variables to linearise the relationships. Other options are discussed later in this chapter.

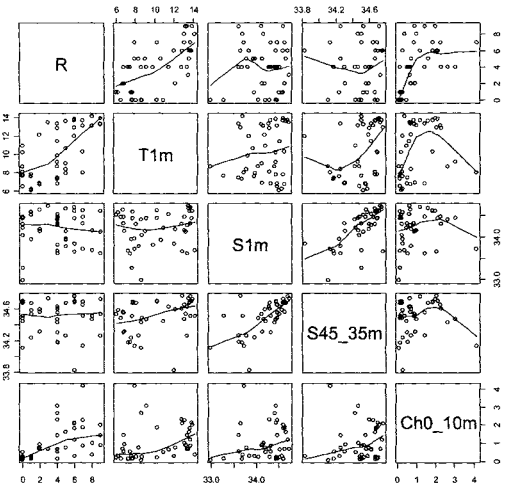

Figure 4.8. Pairplot for the response variable species richness and four selected environmental variables for the Decapoda zooplankton data. The pairplot indicates a linear relationship between richness and temperature. Each smoothing line is obtained by using one variable as the response variable and the other as an explanatory variable in the smoothing procedure. The difference between the smoothing lines in two corresponding graphs above and below the diagonal is due to what is used as the response and explanatory variable in the smoothing method, and therefore, the shape of the two matching smoothers might be different.

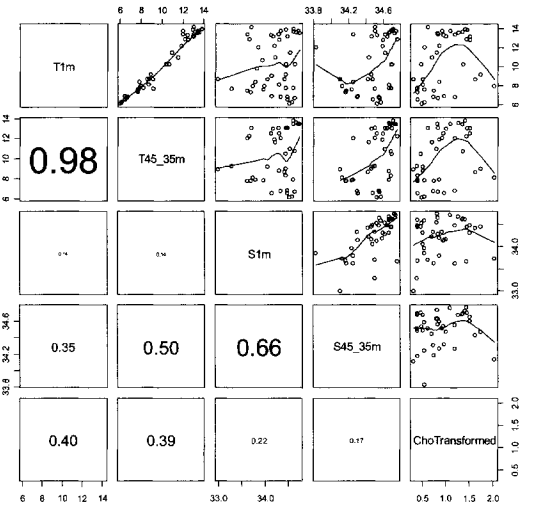

Figure 4.9. Pairplot for all environmental variables in the Decapoda zooplankton data. The lower diagonal part shows the (absolute) correlation coefficient and the upper diagonal part the scatterplots. The font size of the correlation is proportional to its size. There is strong collinearity between some of the variables, e.g., temperature at 1 m and temperature at 45 m.

Coplot

A coplot is a conditional scatterplot showing the relationship between y and x, for different values of a third variable z, or even a fourth variable w. The conditioning variables can be nominal or continuous. Figure 4.10 shows an example for the RIKZ data (Chapter 27). It is a coplot of the species richness versus NAP (which represents the average sea level height at each site), conditional on the nominal variable week. The panels are ordered from the lower left to the upper right. This order corresponds to increasing values of the conditioning explanatory variable. The lower left panel shows the relationship between NAP and the richness index for the samples measured in week 1, the lower right for the week 2 samples, the upper left panel for the week 3 samples, and the upper right panel for the week 4 samples. We did not add a regression line because in the fourth week, only 5 samples were taken. The richness values in week 1 are larger than in weeks 2 and 3, but the NAP range is smaller in week 1.

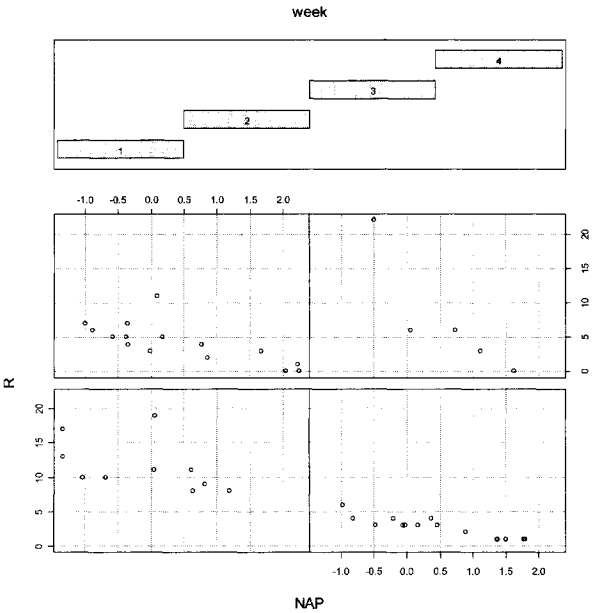

Figure 4.10. Coplot for the species richness index function of the RIKZ data versus NAP, conditional on week. The lower left panel corresponds to week 1, the lower right to week 2, the upper left to week 3, and the upper right for week 4. Note that the number of observations is different for each week.

For nominal variables, as shown in Figure 4.10, there is no overlap in ranges of the conditional variable. For continuous conditioning variables, we can allow for some overlap in the ranges of the conditioning variables, and the number of graphs, as well as the amount of overlap can be modified. This is illustrated in Figure 4.11, which shows another coplot for the RIKZ data. Each panel shows the relationship between the species richness index function and the explanatory variable NAP for a different temperature range. The lower left panel shows sites with temperatures between 15.5 and 17.5 degrees Celsius, and the upper right graph for temperature of 20 degrees and higher. The other panels show a range of different temperature bands between these two extremes. In this instance we have included smoothing curves (Chapter 7) and these highlight a negative relationship between species richness and NAP for all the measured temperature regimes. As this relationship between richness and NAP is common across all the temperature regimes, it suggests that it is not being influenced by temperature. Knowing that a specific variable is unrelated to the response variable is just as important as knowing that it is.

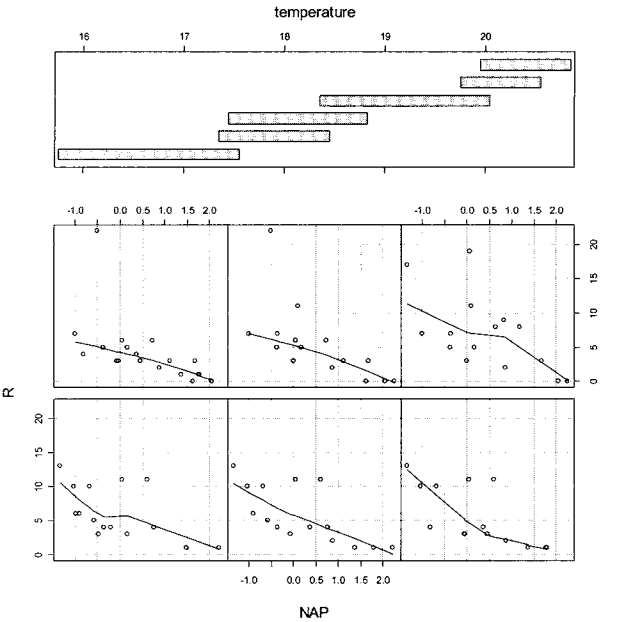

Figure 4.11. Coplot for RIKZ data where NAP is plotted against species richness for different temperature regimes.

Lattice graphs

Another useful tool are lattice graphs (called Trellis graphs in S-Plus). Like coplots these graphs show relationships between two variables, conditional on nominal variables. Lattice graphs have the advantage over coplots because they can work with larger numbers of panels. However, the conditional factor must be nominal. In coplots the conditional factor can be nominal or continuous. We use lattice graphs for time series data exploration and, to a lesser extent, to investigate sampling effort. Unless there are good reasons for deciding otherwise, you should normally use the same sample size and sampling effort across all the explanatory variables. Figure 4.12 shows a lattice graph for the squid data. Each panel shows the relationship between the GSI index and month. The conditional variable is area, and the plots clearly show an unbalance in the sampling effort. In some areas sampling largely took place in one month. Obviously, care is needed if these data were to be analysed in a regression model containing the nominal variables month and area.

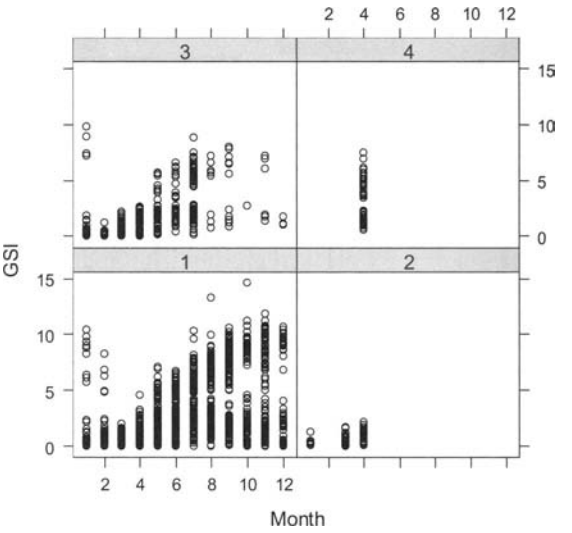

Figure 4.12. GSI index (vertical axis) versus month (horizontal axis) conditional on area (different panels) for the squid data. Note the unbalanced design of the data.

Design and interaction plots

Design and interaction plots are another valuable tool for exploring datasets with nominal variables and are particularly useful to use before applying regression, GLM, mixed modelling or ANOVA. They visualise (i) differences in mean values of the response variable for different levels of the nominal variables and (ii) interactions between explanatory variables. Figure 4.13 shows a design plot for the wedge clam data introduced earlier in this chapter. For these data there are three nominal variables: beach (3 beaches), intertidal or subtidal level on the beach (2 levels) and month (5 months). The design plot allows a direct comparison of the means (or medians) of all the nominal variables in a single graph. The graphs indicate that the mean value of the number of clams for beach 1 is around 0.26, with the mean values at the other two beaches considerably lower. It can also be seen that months 2 and 5 have relatively high mean values. However, the design plot shows little about the interaction *between* explanatory variables, and for this, we use an interaction plot (Figure 4.14). Panel A shows the interaction between month and beach. Mean values at beach 1 are rather different compared with beaches 2 and 3. It also shows that the interaction between season (month) and the mean clam numbers is similar for beaches 1 and 2, but very different for beach 3. Panel B shows the interaction between month and level, with mean values at level 1 in month 5 considerably larger than in the other levels.

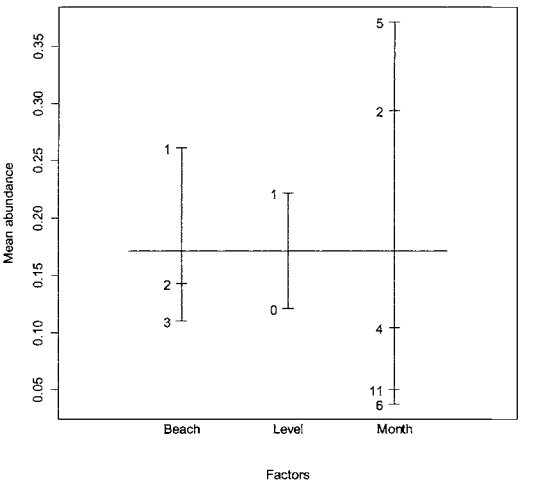

Figure 4.13. Design plot for the wedge clam data. The vertical axis shows the mean value per class for each nominal variable.

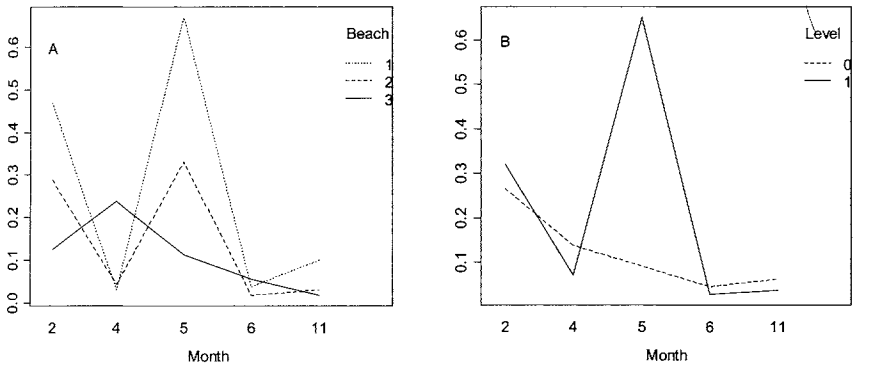

Figure 4.14. Design plot for the wedge clam data. The vertical axis shows the mean value and the horizontal axis the month. A: interaction between month and beach. B: interaction between month and level.

4.2 Outliers, transformations and standardisations

Outliers

An outlier is a data point that, because of its extreme value compared to the rest of the dataset, might incorrectly influence an analysis. So the first question is: 'how can we identify an outlier?' A simple approach might be to quantify everything as an outlier that is beyond a certain distance from the centre of the data. For example, the points outside the hinges of a boxplot could be considered as outliers. However, the dotplots and boxplots for the Argentinean data (Section 4.1) show that this is not always a good decision. Two-dimensional scatterplots can also highlight observations that may be potential outliers. For example, Figure 4.15 is a scatterplot for the variables NAP and species richness for the RIKZ data. These data are analysed in Chapter 27. The data consist of abundance of 75 zoobenthic species measured at 45 sites. NAP represents the height of a site compared with average sea level. The two observations with richness values larger than 19 species are not obviously outliers in the NAP (x) space. Although these sites have large richness values, they are not different enough from the other data points, to consider them extreme or isolated observations. However, as we will see in Chapter 5, these two observations cause serious problems in the linear regression for these data. So, although an observation is not considered an outlier in either the x-space or the y-space, it can still be an outlier in the xy-space. The situation that an observation is an outlier in the x-space, and also in the y-space, but not in the xy-space, is possible as well. A boxplot for the data presented in Figure 4.16 suggests that point A in the left panel would be an outlier in the y-space, but not in the x-space. However, fitting a linear regression line clearly identifies it as a highly influential observation. So, it is also an outlier in the xy-space. Point B is an outlier in the x-space *and* in the y-space, but not in the xy-space as it would not cause any major problems for a linear regression model. Point C is an outlier in the xy-space as it would strongly influence a regression. The right panel in Figure 4.16 shows a more serious problem: including point A in a linear regression or when calculating a correlation coefficient will show a strong positive relationship, whereas leaving out point A will give a strong negative relationship.

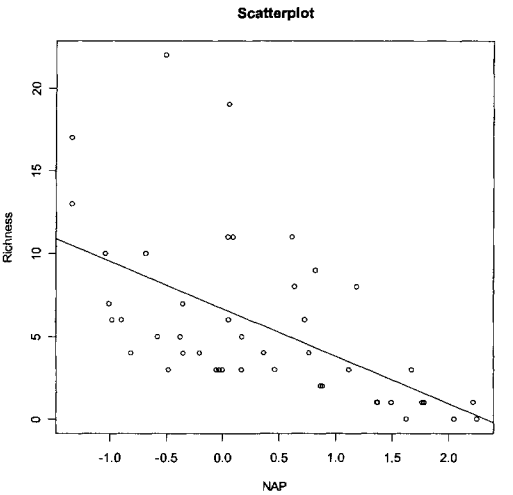

Figure 4.15. Scatterplot of species richness versus NAP for the RIKZ data. The two sites with high species richness are extreme with respect to the overall NAP-richness relationship.

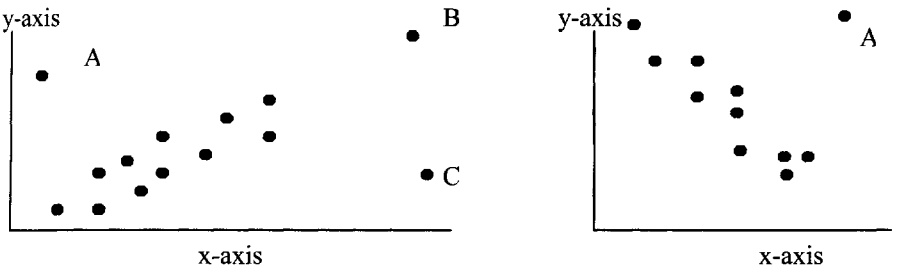

Figure 4.16. Left panel: scatterplot with two outliers. Right panel: scatterplot with 1 outlier.

Transformation

There are many reasons for transforming data, but it is normally because you have data with extreme outliers and non-normal distributions. Data transformation (on the response variables) will also be required when you plan to use discriminant analysis and there is clear evidence (e.g., by using a Cleveland dotplot) of heterogeneity.

Both the response variables and the explanatory variables can be transformed, and different types of transformations can be applied to different variables within the same dataset. Choosing the 'correct' transformation can be difficult and is usually, as least in part, based on experience. Additionally, the choice of transfor-

mation is influenced by the choice of follow-up analysis. For some techniques, such as classification or regression trees, the transformation of the explanatory variables makes no difference to the results. However, most techniques may require some transformation of the raw data before analysis.

The easiest problem to solve is where the extreme observations identified during the data exploration stage turn out to be typing errors. However, we will assume that this easy solution is not available and that we have a dataset with genuine extreme observations. If these extreme observations are in the explanatory variables, then a transformation of the (continuous) explanatory variables is definitely required, especially if regression, analysis of covariance, GLM, GAM or multivariate techniques like redundancy analysis and canonical correspondence analysis are applied. When the extreme observations are in the response variable, there is more than one approach available. You can either transform the data or you can apply a technique that is slightly better in dealing with extreme values, such as a GLM or a GAM with a Poisson distribution. The latter only works if there is an increase in spread of the observed data for larger values. Alternatively, quasi-Poisson models can be used if the data are overdispersed. Note, you should not apply a square root or log transformation on the response variable, and then continue with a Poisson GLM model, as this applies a correction twice. Yet, another option is to use dummy explanatory variables (Harvey 1989) to model the extreme observations. A more drastic solution for extreme observations is to simply omit them from the analysis. However, if you adopt this approach, you should always provide the results of the analysis with, and without, the extreme observations. If the large values are all from one area, or one month, or one sex, then it may be an option to use different variance components within the linear regression model, resulting in generalised least squares (GLS).

As an example, we will assume the aim is to carry out a linear regression. The Cleveland dotplot or boxplot indicate that there are no outliers of any concern, but the scatterplot of a response and explanatory variable shows a clear non-linear relationship. In this case, we should consider transforming one or both variables. But, which transformation should we use? The range of possible transformations for the response and explanatory variables can be selected from

$$\ldots, y^{\frac{1}{4}}, y^{\frac{1}{3}}, y^{\frac{1}{2}}, y, \log(y), y^2, y^3, y^4, \ldots$$

These transformations can be written in one formula, namely the Box–Cox power transformation; see also equation (4.1). It is basically a family of transformations and they can only be applied if the data are non-negative, but a constant can be applied to avoid this problem. Alternative transformations are ranking and converting everything to 0–1 data. For example, if the original data have the values 2, 7, 4, 9, 22, and 40, the rank transformed data will be 1, 3, 2, 4, 5, 6. If the original data are 0, 1, 3, 0, 4, 0, and 100, then converting everything to 0–1 data gives 0 1 1 0 1 0 1. Converting to 0–1 data seems like a last resort, particularly if considerable expense and time has been spent collecting more detailed data. However, our experience shows this is often the only realistic option with difficult ecological datasets where the sample size or sampling quality is less than ideal. This

is a common problem with zoobenthic species community and fisheries data and an example is given in the *Solea solea* case study in Chapter 21.

Several strategies are available for choosing the most appropriate transformation. The first approach is trial and error. Using the graphical data exploration techniques discussed earlier, especially the Cleveland dotplots, boxplots and pairplots, you can apply what appears to be the best transformation and see how well it corrects the identified issues. This is our preferred option, but it does require some existing expertise. It is also important that this trial-and-error approach is fully reported when presenting the results, including details of both the unsuccessful as well as the successful transformation approaches.

When the follow-up analysis is based on linear relationships, then a useful tool is the Mosteller and Tukey's bulging rule (Mosteller and Tukey 1977, Fox 2002a), which is from the Box–Cox family of transformations. This approach relies on identifying non-linear patterns in your data by inspecting a scatterplot. The required transformations for linearising the relationships can be inferred from the bulging rule illustrated in Figure 4.17. For example, if the scatterplot shows a pattern as in the upper right quadrant, then either the y's or the x's need to be increased to transform the data to linear.

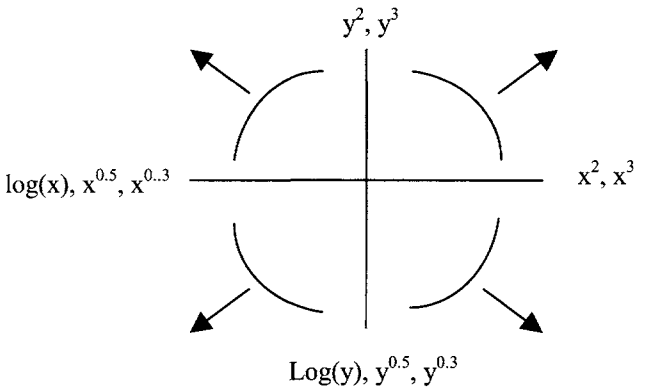

Figure 4.17. Mosteller and Turkey's bulging rule. When the arrow points downwards, y should be made smaller; if it points upwards, it should be increased. If the arrow points towards the left, x should be made larger, etc. See also Fox (2002a).

An example of the bulging rule is presented in Figure 4.18. Panel A shows a scatterplot of length versus biomass for the untransformed wedge clam data. This pattern matches with the lower right quadrant in Figure 4.17, and therefore the bulging rule suggests transforming either length to Length^2 (or higher powers) or taking the log or square root of biomass (panel B and C). Panel D suggests that transforming both length and biomass is the best option.

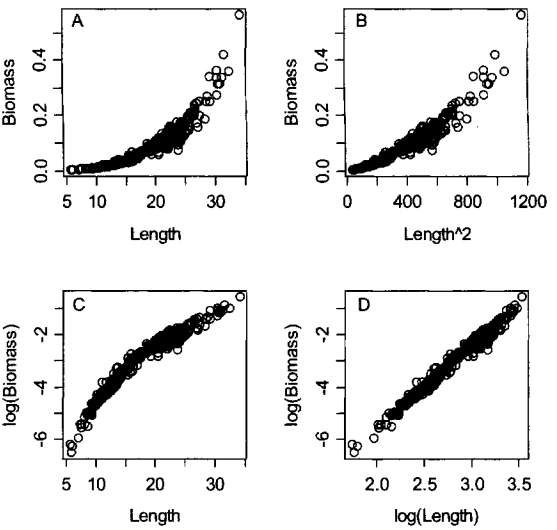

Figure 4.18. Scatterplot of (A) length versus biomass for the wedge clam data, (B) squared length versus biomass, (C) length versus log transformed biomass and (D) log length versus log biomass. The transformations shown in panels B and D follow those suggested by Mosteller and Tukey's bulging rule (Figure 4.17). Length-weight relationships typically require log-log transformations.

Automatic selection of transformations

Automatic transformation selection techniques are described by Montgomery and Peck (1992) and Fox (2002a), among others. Montgomery and Peck apply a series of power transformations, and for each power transformation, they calculate the residual sum of squares. These sums of squares cannot be compared directly as reducing the size of the data with a square root transformation, in most cases, also makes the residuals, and therefore residual sum of squares, smaller. Therefore, this power transformation contains a correction factor using the geometric mean that makes the residual sum of squares directly comparable. So, the challenge is to find the optimal value of p, where p defines the transformation in equation (4.1). Using a grid (range of values), and then increasingly finer grids, if required, the optimal value for p can be found as the one that has the smallest residual sum of squares. It is also possible to calculate a confidence interval for the power transformation parameter p. If this interval contains 1, then no transformation is required. This modified power transformation is

$$Y^p = \begin{cases} \dfrac{Y^p - 1}{p\dot{Y}^{p-1}} & \text{if } p \neq 0 \\ Y^p = \dot{Y}\ln(Y) & \text{if } p \neq 0 \end{cases} \qquad (4.2)$$

$$\text{where } \dot{Y} = \exp(\frac{1}{n}\sum_{i=1}^{n} Y_i)$$

The difference between this formula and equation (4.1) is the correction factor $\dot{Y}$, also called the geometric mean. The confidence interval for p can be found by:

$$SS^* = SS_p(1 + \frac{t^2_{\alpha/2,\nu}}{\nu}) \qquad (4.3)$$

where ν is the residual degrees of freedom (Chapter 5) and SS_p is the lowest residual sum of squares. The confidence interval for p can be visualised by making a plot of various p values against SS_p, and using SS^* to read off the confidence bands. An example of using the biomass and log transformed length variables from the wedge clam dataset is given below. The log transformation for length was used as it contains various observations with rather large values. We are trying to find out which transformation for the biomass data is most optimal for fitting a linear regression model. Initially, the values of p were chosen to be between –3 and 3 with steps of 0.01. However, this was unsatisfactory and a finer grid of p was needed with values between –0.1 and 0.1. Figure 4.19 shows the sum of squares plotted against different values of p. The optimal value is p = 0.025 (lowest point on the curve), and SS^* is represented by the dotted line that allows us to read off the confidence intervals from where it intersects the curve. The 95% confidence band for p is therefore approximately between 0.005 and 0.035. Although 0, which is the log transformation by definition, is just outside this interval, in this instance for ease of interpretation a log transformation would probably be the best option.

Although we have only looked at transforming the response variable Montgomery and Peck (1992) also give a procedure for automatic selection of the transformation on the explanatory variable.

In conclusion, the main reasons for a data transformation are (in order of importance) as follows:

1. Reduce the effect of outliers.
2. Improve linearity between variables.
3. Make the data and error structure closer to the normal distribution.
4. Stabilise the relationship between the mean and the variance (this will be discussed further in Chapter 6).

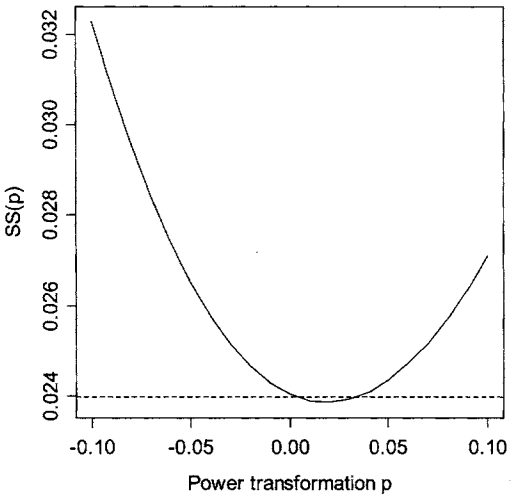

Figure 4.19. Sum of squares for different power transformations. The vertical axis shows the sum of squares for different power transformations p. The dotted line represents the 95% confidence band.

If the follow-up analysis is a generalised additive model with Poisson distribution, points 2 to 4 are irrelevant (Chapter 7). Although there are some rules of thumb for transformations, such as using a cubic or square root transformation for count data, and a log transformation where the relationships are multiplicative, it is still difficult to choose the best option. We suggest the following approach:

- Apply all data exploration techniques on the original data.
- If there are outliers in the explanatory variables, transform them.
- Apply linear regression or related techniques (e.g., GLM, analysis of variance), and judge whether the residuals show any patterns.
- If there is any residual information left, or if there are influential observations, then a data transformation might be an option.
- Choose the best transformation using trial and error, or use an automatic selection routine.

Unless replicates of the response variable are available, we believe it is unwise to apply a transformation purely on the argument that the 'response variable must be normally distributed'. The normality assumption is for the data at each X value (this will be discussed further in Section 5.1)! For example, a transformation on the GSI index for the squid data might remove the differences between male and female species (Figure 4.5). And normality of the explanatory variables is not assumed at all!

Other points to consider are whether to use the same transformation for (i) all response variables, (ii) all explanatory variables, or (iii) all response variables and

all explanatory variables. In general, we recommend applying the same transformation to all response variables, and the same transformation to all explanatory variables. However, you can apply a different transformation to the response variables from the one applied to the explanatory variables. Sometimes, the explanatory variables represent different types of variables; e.g., if some are distance or size related, some are time related and some are nominal. In this case, there is nothing wrong in using a different transformation for each type of variable. Nominal explanatory variables should not be transformed, but distance and size-related variables tend to have a wider spread and might require a transformation. And the same approach should be adopted with response variables. For example, in an EU-funded project (WESTHER) on herring, biological, morphometric, chemical, genetic and parasite variables were measured, and were all considered as response variables. The parasite data were count data and required a square root transformation. The morphometric data were pre-filtered by length and did not require any further transformation, but the chemical data required a log transformation.

A final word on transformation is to be aware that, sometimes, the aim of the analysis is to investigate the outliers, e.g., the relationship between high water levels and the height of sea defences, or the analysis of scarce, and therefore only rarely recorded, species. In these cases, you cannot remove the extreme values, and the choice of analytical approach needs to take this into account. Useful sources on extreme values modelling are Coles (2004) and Thompson (2004), which both discuss sampling species that are rare and elusive.

Standardisations

If the variables being compared are from widely different scales, such as comparing the growth rates of small fish species against large fish species, then standardisation (converting all variables to the same scale) might be an option. However, this depends on which statistical technique is being used. For example, standardising the response variables would be sensible if you intend on using dynamic factor analysis (Chapter 17), but methods like canonical correspondence analysis and redundancy analysis (Chapters 12 and 13) apply their own standardisation before running the analysis. To make it more confusing, applying multidimensional scaling (Chapter 10) with the Euclidean distance function on standardised data is acceptable, but the same standardised data will give a problem if the Bray–Curtis distance function is used. There are several methods for converting data to the same scale, and one option is to centre all variables around zero by

$$Y_i^{\text{new}} = Y_i - \overline{Y}$$

where $\overline{Y}$ is the sample mean and Y_i the value of the i^{th} observation. However, the most common used standardisation is given by:

$$Y_i^{\text{new}} = (Y_i - \overline{Y}) / s_y$$

where s_y is the sample standard deviation. The transformed values Y_i^{new} are now centred around zero, have a variance of one, and are unit-less. This transformation is also called normalisation. Other, less-used transformations are

$$Y_i^{new} = Y_i / Y_{max} \quad \text{and} \quad Y_i^{new} = (Y_i - Y_{min}) / (Y_{max} - Y_{min})$$

They rescale the data between zero and one. Centering or standardisation can be applied on response and/or explanatory variables. To illustrate the difference between no transformation, centring and normalisation, we use a North-American sea surface temperature (SST) time series. These data come from the COADS datasets (Slutz et al. 1985, Woodruff et al. 1987), and details on obtaining the mean monthly values used here can be found in Mendelssohn and Schwing (2002). The upper left panel in Figure 4.20 shows lattice graphs for four time series from this dataset. In the upper right panel, all series are centred around zero and are in their original units. This transformation takes away the differences in absolute value. However, there is still a difference in variation; the fluctuation in the upper left and lower right panels is considerably smaller. The standardisation (or: normalisation) removes these differences (lower left panel in Figure 4.20), and the amount of fluctuation becomes similar. This transformation is the most commonly used approach and rescales all time series around zero. The time series are now without units, and the normalisation makes all time series equally scaled, even if some time series had large fluctuations and others only small fluctuations. Centring only removes the differences in absolute values between series.

As with other transformations the decision to standardise your data depends on the statistical technique you plan to use. For example, if you want to compare regression parameters, you might consider it useful to standardise the explanatory variables before the analysis, especially if they are in different units or have different ranges. Some techniques such as principal component analysis automatically normalise or centre the variables.

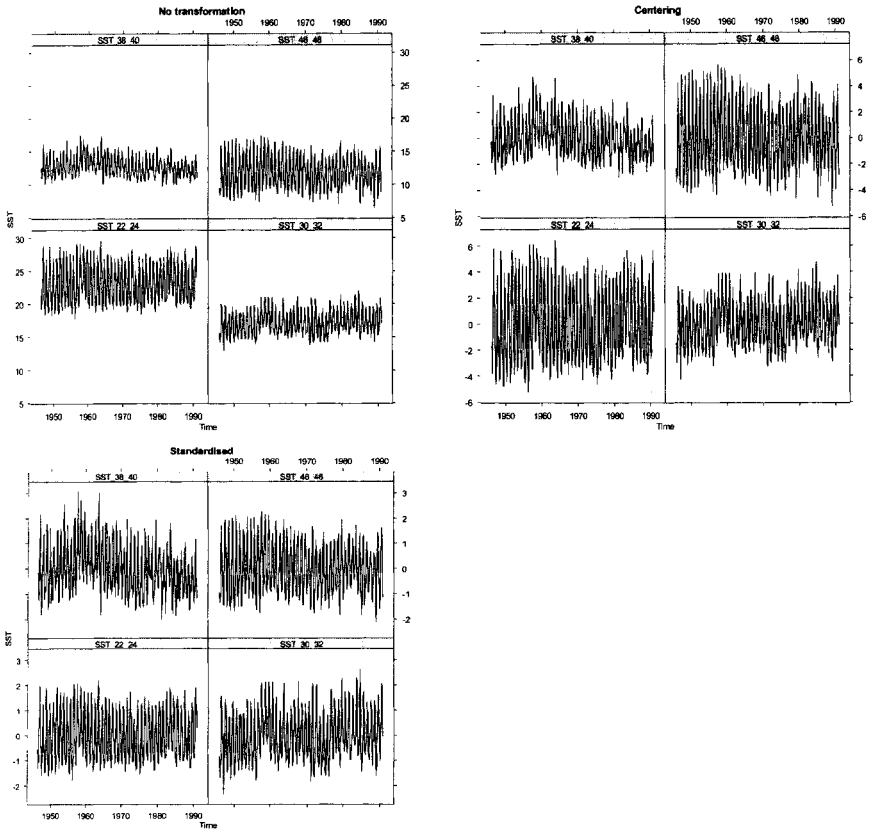

Figure 4.20. The upper left panel shows four time series from the North American SST dataset with no transformation applied. The upper right panel shows the same time series centred, and the lower left shows the same time series normalised.

4.3 A final thought on data exploration

Even if you don't see it, it might still be there

Even if the scatterplots suggest the absence of a relationship between Y and X, this does not necessarily mean one does not exist. A scatterplot only shows the relationship between two variables, and including a third, fourth or even fifth variable might force a different conclusion. To illustrate this, we have used the GSI index of the squid data again (Section 4.1). The left panel in Figure 4.21 shows the scatterplot of month against the GSI index. The most likely conclusion based on this graph is that that there is no strong seasonal effect in the GSI index. However,

using a four-dimensional scatterplot, or coplot (right panel in Figure 4.21), a strong seasonal pattern is apparent for female squid in areas 1 and 3, and a weak seasonal pattern for males in area 3.

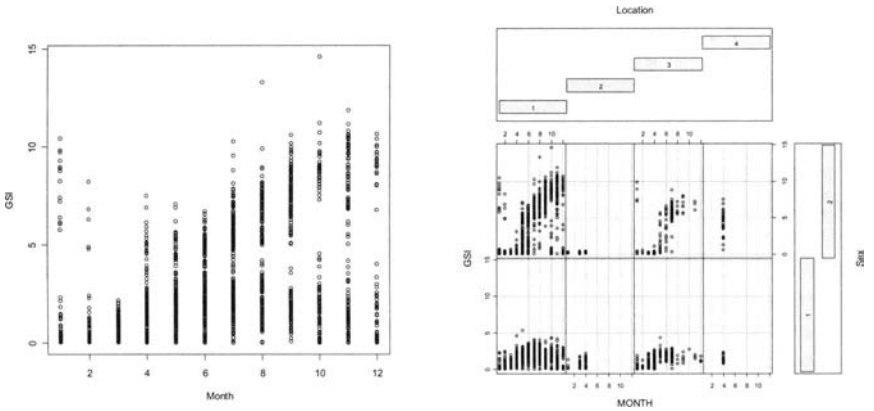

Figure 4.21. Left: scatterplot of GSI data. Right: coplot of GSI index for the squid data. The conditioning variables are Location (four areas) and Sex (1 is male, 2 is female).

This final thought here is to re-enforce the argument that a thorough data exploration stage is essential before moving on to the analysis stage of the data investigation.

What Next?

After completing the data exploration, the next step is to verify and investigate the patterns and relationships this step identified. Assuming the scatterplot indicates a linear relationship between the variables, then linear regression is the obvious next step. However, if the scatterplot suggests a clear non-linear pattern, then a different approach needs to be taken, which might include (i) using interactions and/or quadratic terms in the linear regression model, (ii) transforming the data, (iii) continuing with a non-linear regression model, (iii) using generalised linear modelling, (iv) applying generalised additive modelling techniques, or (v) applying (additive) mixed modelling techniques. All these approaches are investigated in later chapters. The first option means that you proceed with the linear regression model, but you need to ensure that all assumptions are met (e.g. no residual patterns).

To choose which approach is the most appropriate requires knowledge of the assumptions of the selected methods, and tools to detect violations (using residuals). These are all discussed in later chapters, but basically it all comes down to something very basic: learn from your errors.