Appendix A
Required Pre-knowledge: A Linear Regression and Additive Modelling Example

This appendix introduces the essential background needed to understand the core material of this book. Using linear regression and additive modelling in R we discuss key issues such as interactions, model selection, model validation, and model interpretation. We do not deal with the underlying theory of linear regression and additive modelling in this appendix as this is discussed in Chapters 2 and 3.

To illustrate the linear regression model, we use bird data originally analysed in Loyn (1987), and again in Quinn and Keough (2002). This data set is especially good for introducing linear regression and extensions like additive modelling.

A.1 The Data

Forest bird densities were measured in 56 forest patches in south-eastern Victoria, Australia. The aim of the study was to relate bird densities to six habitat variables; size of the forest patch, distance to the nearest patch, distance to the nearest larger patch, mean altitude of the patch, year of isolation by clearing, and an index of stock grazing history (1 = light, 5 = intensive). The variables are given in Table A.1.

Table A.1 Description of variables for the Loyn bird data

Variable name	Description	Type
ABUND	Density of birds in a forest patch	Continuous response variable
AREA	Size of the forest patch	Continuous explanatory variable
DIST	Distance to the nearest patch	Continuous explanatory variable
LDIST	Distance to the nearest larger patch	Continuous explanatory variable
ALTITUDE	Mean altitude of the patch	Continuous explanatory variable
YEAR.ISOL	Year of isolation by clearance	Continuous explanatory variable
GRAZE	Index of stocking grazing intensity	Nominal (ordinal) explanatory variable with levels 1 (light) to 5 (intensive)

A.2 Data Exploration

As with any analysis, before starting the linear regression, we apply a data exploration focussing on the following points:

1. Outliers in the response and explanatory variables.
2. Collinearity of the explanatory variables.
3. Relationships between the response variable and the explanatory variables.

The results of these three steps should guide us how to proceed with the follow-up analysis, e.g. a linear regression analysis or an additive model. It also indicates whether a data transformation is needed.

A.2.1 Step 1: Outliers

First, we look at the outliers in the response variable and the outliers in the explanatory variables. Useful tools for this are boxplots and Cleveland dotplots. Figure A.1 shows dotplots for all variables. The R function `dotchart` was used to make these graphs. Values of a variable can be read from the *x*-axis, and by default the *y*-axis

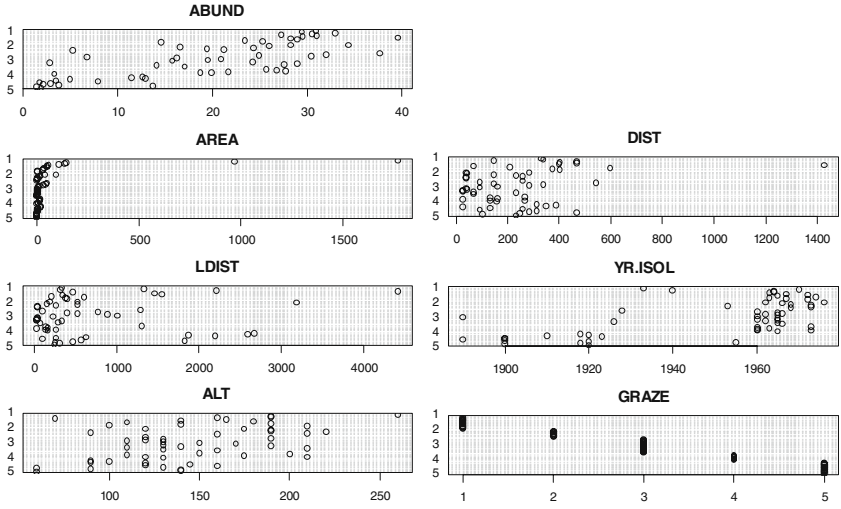

Fig. A.1 Cleveland dotplots for all variables. Each panel corresponds to a variable. The *x*-axes show the values of a variable and the *y*-axes the observations, which are grouped by the values of GRAZE

shows the order of the observations in the variable (from bottom to top). However, with the `group` option, you can group observations, e.g. by the levels of GRAZE. The `group` argument must be a nominal variable.

Isolated points at the far ends, and on either side in a dotplot, suggest potential outliers. This is the case for two observations with high AREA values, one observation with a high DIST value, and a couple of observations with high LDIST values (note that these are all different forest patches). If two or three observations (the same) have larger values for *all* variables, then the decision what to do is easy, just drop them from the analysis. However, if we do this here, we lose too many observations. The alternative is to apply a transformation on AREA, DIST and LDIST. Based on the values of these three variables, a strong transformation is needed, for example, a logarithmic (base 10) transformation or a natural logarithmic transformation. Note that all three variables are related as they measure size and distance. Variables like size, distance, and volume often need a transformation. It is easier to justify a transformation on only a subset of variables if they are somehow ecologically 'related'.

The R code to produce Fig. A.1 is given below.

```
> library(AED); data(Loyn)
> Loyn$fGRAZE <- factor(Loyn$GRAZE)
> op <- par(mfrow = c(4, 2), mar = c(3, 3, 3, 1))
> dotchart(Loyn$ABUND, main = "ABUND", group = Loyn$fGRAZE)
> plot(0, 0, type = "n", axes = FALSE)
> dotchart(Loyn$AREA, main = "AREA", group = Loyn$fGRAZE)
> dotchart(Loyn$DIST, main = "DIST", group = Loyn$fGRAZE)
> dotchart(Loyn$LDIST, main = "LDIST", group = Loyn$fGRAZE)
> dotchart(Loyn$YR.ISOL, main = "YR.ISOL", group = Loyn$fGRAZE)
> dotchart(Loyn$ALT, main = "ALT", group = Loyn$fGRAZE)
> dotchart(Loyn$GRAZE, main = "GRAZE", group = Loyn$fGRAZE)
> par(op)
```

The first four commands are used to access the data, and set up the graphical window with eight panels and the amount of white space between the panels. The rest of the code produces the seven Cleveland dotplots. We used the ordinary `plot` command to ensure that the ABUND dotplot is the only panel in the top row.

A.2.2 Step 2: Collinearity

We continue by checking for collinearity (i.e. high correlation between the explanatory variables). The initial question is whether we should transform variables before or after looking at collinearity. In this particular case, it is quite obvious that we

should apply the transformation on AREA, DIST, and LDIST before continuing the analysis as the large values will dominate any correlation coefficient between variables that involve them. But sometimes a Cleveland dotplot suggests that there are no outliers, and a pairplot (a typical tool used for finding relationships and detecting collinearity) shows that there are outliers. We decided to apply the transformation on AREA, DIST, and LDIST before checking for collinearity:

```
> Loyn$L.AREA <- log10(Loyn$AREA)
> Loyn$L.DIST <- log10(Loyn$DIST)
> Loyn$L.LDIST <- log10(Loyn$LDIST)
```

To assess collinearity, we will use three tools: Pairwise scatterplots, correlation coefficients, and variance inflation factors (VIF). The first two can be combined in one graph with some clever R code; see Fig. A.2.

We included GRAZE in the pairplot, but there is an argument that it should not be included as it is a nominal variable. However, it is ordinal as 2 is larger than 1, 3 is larger than 2, etc.; so, it does make some sense to include it. However, you should interpret the correlation coefficients involving GRAZE with care as the difference between 4 and 5 may not be the same as the difference between 1 and 2 (the correlation coefficient assumes it is).

We also included the response variable ABUND in the pairplot as it avoids repeating the same graph when we look at relationships between response and

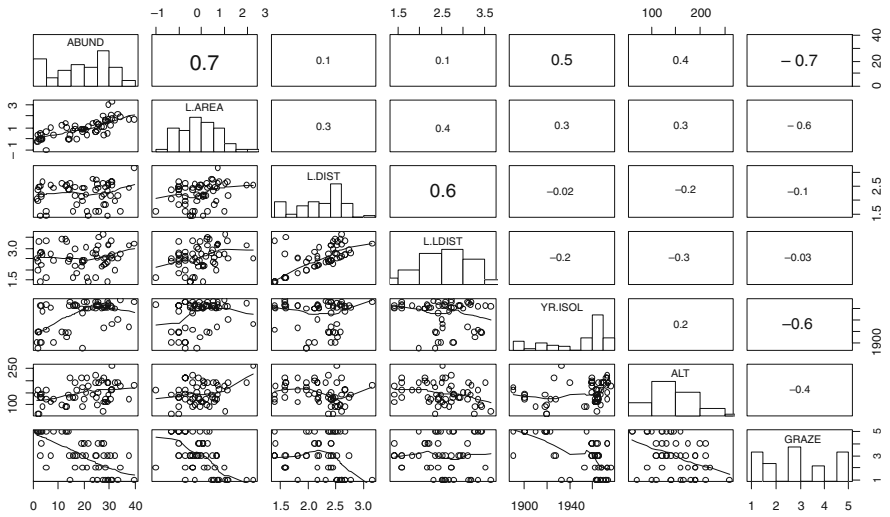

Fig. A.2 Pairplot of all variables. The upper panel contains estimated pair-wise correlations, and the font size is proportional to the absolute value of the estimated correlation coefficient. The diagonal panel contains histograms and the lower panel scatterplots with a LOESS smoother added to aid visual interpretation. The R code to generate this graph was taken from the pairs help file, and the modified code can be found in our `AED` package

explanatory variables in the next paragraph. However, if you have a large number of explanatory variables, you would not normally use the response variable at this stage.

The histograms along the diagonal also require some explanation. Some people are obsessed with normality, which sometimes results in making histograms of every variable, including nominal variables like sex, location, month, or in this case GRAZE. Not all these histograms make sense! A histogram of GRAZE only shows how many observations we have per level. Histograms of continuous explanatory variables also need to be interpreted with care. If, for example, altitude is normally distributed, then this implies that the majority of the observations have similar altitude values. However, we would like to have observations that cover a wide range of altitude values, not just around the average. Hence, it would be nice if the histograms of continuous explanatory variables show the shape of a uniform distribution (flat line).

Focussing only on the explanatory variables in Fig. A.2, there seems to be some correlation between L.DIST and L.LDIST; GRAZE and L.AREA; and GRAZE and YR.ISOL. However, the value of 0.6 (and –0.6) is not large enough to worry us.

The R code to generate Fig. A.2 is given below.

```
> Z <- cbind(Loyn$ABUND, Loyn$L.AREA, Loyn$L.DIST,
             Loyn$L.LDIST, Loyn$YR.ISOL, Loyn$ALT,
             Loyn$GRAZE)
> colnames(Z) <- c("ABUND", "L.AREA", "L.DIST",
                   "L.LDIST", "YR.ISOL", "ALT", "GRAZE")
> pairs(Z, lower.panel = panel.smooth2,
       upper.panel = panel.cor, diag.panel = panel.hist)
```

The functions `panel.smooth2` and `panel.cor` are external functions that we took (and modified) from the `pairs` help file and are stored in our `AED` package.

The last tool we use for detecting collinearity is VIF values. These can be obtained by typing in

```
> corvif(Z[, c(-1,-7)])

Correlations of the variables
           L.AREA      L.DIST    L.LDIST     YR.ISOL         ALT
L.AREA  1.0000000  0.30216662  0.3824795  0.27841452  0.2751428
L.DIST  0.3021666  1.00000000  0.6038664 -0.01957223 -0.2190070
L.LDIST 0.3824795  0.60386637  1.0000000 -0.16111611 -0.2740438
YR.ISOL 0.2784145 -0.01957223 -0.1611161  1.00000000  0.2327154
ALT     0.2751428 -0.21900701 -0.2740438  0.23271541  1.0000000

Variance inflation factors
            GVIF
L.AREA  1.622200
L.DIST  1.622396
```

```
L.LDIST 2.008157
YR.ISOL 1.201719
ALT     1.347805
```

Again, this function uses our `AED` package, but you can also use the VIF values calculated by functions in the `car` package from John Fox. All VIF values are below 3 (see Chapter 26 in Zuur et al. (2007)), indicating there is no collinearity in these variables (at least not without GRAZE). We decided to keep all variables in the analysis.

A.2.3 Relationships

We now look at relationships between the response variable and the explanatory variables. The most obvious tool for this task is a pairplot that contains the response variable and a set of explanatory variables (Fig. A.2). If you have a large number of explanatory variables (10−15), then multiple pairplots may be needed. If you have more than 10−15 explanatory variables, then pairplots are less useful. However, if you have more than 10 explanatory variables, then it is very likely that there will be high collinearity.

Based on the pairplot in Fig. A.2, we expect that the variables L.AREA and GRAZE will play an important role in the analyses. Other graphical tools that can help find relationships between a response variable and multiple explanatory variables are a `coplot` and `xyplot` (from the lattice package, which is part of the R base installation) and design and interaction plots (see `plot.design` from the `design` package, which you need to download). These are all described in Chapter 4 of Zuur et al. (2007) and are also useful to explore potential interactions between the explanatory variables.

A.3 Linear Regression

Before enthusiastically typing in the R code for linear regression and running it, we should first think about what we want to do. The aim of the analysis is to find a relationship between bird densities (ABUND) and the six explanatory variables. But it could well be that birds perceive the AREA effect differently if GRAZE values are low compared to when GRAZE values are high. If that is the case, we have to include an interaction term between GRAZE and L.AREA. The problem is that there are a large number of potential two-way interactions. And there could also be three-way interactions; birds may respond in a different way to AREA if GRAZE values are low in combination with low values for altitude (ALT). The smaller the data set (56 observations is small), the more difficult it is to include multiple inter-

action terms, especially if there are multiple nominal explanatory variables with more than two levels involved. Sometimes, you may not have enough observations per combination. In such cases, individual observations may become particularly influential.

Many statistical newsgroups have long threads on the subject of interaction, and the most common opinions seem to fall into the following categories:

1. Start with a model with no interactions. Apply the model, model selection, and model validation. If the validation shows that there are patterns in the residuals, investigate why. Adding interactions may be an option to improve the model.
2. Use biological knowledge to decide which, if any, interactions are sensible to add.
3. Apply a good data exploration to see which interactions may be important.
4. Identify the prime explanatory variable(s) of interest. In the bird example, this would be GRAZE and AREA as we can control them using management decisions. Include interactions between these variables and any of the other variables.
5. Only include the main terms and two-way interaction terms.
6. Only include higher interactions terms if you have good reasons (i.e. biological justification) to do so.
7. Include all interactions by default.

An important aspect to keep in mind is that if interactions are included, then you must include the corresponding main terms; see Underwood (1997) for a discussion on the interpretation of *p*-values of main terms if interactions are included. Which option you choose from the list above is your own choice. We prefer options one, two and three.

On research projects, where we were unable to convince the biologists to exclude 4-way interactions, we only ended up with confusing models and other misery (e.g. combinations of nominal variables with only three points and therefore large Cook distance values, non convergence, etc.).

We will start the bird data analysis with no interactions. The following R code applies a linear regression in R.

```
> M1 <- lm(ABUND ~ L.AREA + L.DIST + L.LDIST +
           YR.ISOL + ALT + fGRAZE, data = Loyn)
```

The question now is: Should we look at the numerical output first or the graphical output? There is no point in applying a detailed model validation if nothing is significant. On the other hand, why look at the numerical output if all the assumptions are violated? Perhaps starting with the numerical output is better as it takes less time and is easier. There are multiple ways of getting numerical output for our linear regression model:

```
> summary(M1)
> drop1(M1, test="F")
> anova(M1)
```

Each of these commands presents the output in a slightly different way, and they are all useful in different ways. The `summary` command gives the following output:

```
Call: lm(formula = ABUND ~ L.AREA + L.DIST + L.LDIST + YR.ISOL + ALT+
    fGRAZE, data = Loyn)
Residuals:
     Min       1Q   Median       3Q      Max
-15.8992  -2.7245  -0.2772   2.7052  11.2811

Coefficients:
             Estimate Std. Error t value  Pr(>|t|)
(Intercept)  36.68025  115.16348   0.319    0.7515
L.AREA        6.83303    1.50330   4.545 3.97e-05
L.DIST        0.33286    2.74778   0.121    0.9041
L.LDIST       0.79765    2.13759   0.373    0.7107
YR.ISOL      -0.01277    0.05803  -0.220    0.8267
ALT           0.01070    0.02390   0.448    0.6565
fGRAZE2       0.52851    3.25221   0.163    0.8716
fGRAZE3       0.06601    2.95871   0.022    0.9823
fGRAZE4      -1.24877    3.19838  -0.390    0.6980
fGRAZE5     -12.47309    4.77827  -2.610    0.0122

Residual standard error: 6.105 on 46 degrees of freedom
Multiple R-squared: 0.7295,     Adjusted R-squared: 0.6766
F-statistic: 13.78 on 9 and 46 DF,  p-value: 2.115e-10
```

The first part of the output tells you which model was applied and some basic information on the residuals. The part below 'Coefficients' gives the estimated regression parameters, standard errors, t-values, and p-values. The only confusing part of this output is perhaps the absence of GRAZE level 1. It is used as a baseline. Hence, a patch that has GRAZE level 2 has 0.52 birds (density) more than a patch with level 1, and a patch with GRAZE level 5 has 12.4 birds less than a patch with level 1. The corresponding p-values tell you whether a patch is significantly different from level 1. Dalgaard (2002) shows how to change the baseline and adjust for multiple comparisons. Note that you should not assess the significance of a factor by the individual p-values. We will give a better method for this in a moment. You should not drop individual levels of a nominal variable. They all go in or you drop the whole variable. The last bit of the code gives the R^2 and adjusted R^2 (for model selection). The rest of the output you should, hopefully, be familiar with.

The function `drop1` does exactly what you think it does: it drops one variable, each one in turn. Its output is given as follows:

```
Single term deletions
Model:
ABUND ~ L.AREA + L.DIST + L.LDIST + YR.ISOL + ALT + fGRAZE
```

```
         Df Sum of Sq     RSS     AIC F value    Pr(F)
                      1714.43  211.60
L.AREA    1    770.01 2484.44  230.38 20.6603 3.97e-05
L.DIST    1      0.55 1714.98  209.62  0.0147  0.90411
L.LDIST   1      5.19 1719.62  209.77  0.1392  0.71075
YR.ISOL   1      1.81 1716.24  209.66  0.0485  0.82675
ALT       1      7.47 1721.90  209.85  0.2004  0.65650
fGRAZE    4    413.50 2127.92  215.70  2.7736  0.03799
```

The full model has a sum of squares of 1714.43. Each time, *one* term is dropped in turn, and each time, the residual sum of squares is calculated. These are then used to calculate an F-statistic and a corresponding p-value. For example, to get the output on the first line, R fits two models. The first model contains all explanatory variables and the second model all, but L.AREA. It then uses the residual sums of squares of each model in the following F-statistic.

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

The terms RSS_1 and RSS_2 are the residual sum of squares of model M_1 and model M_2, respectively, and n is the number of observations. The number of parameters in models 2 and 1 are p and q, respectively ($p > q$). The models are nested in the sense that one model is obtained from the other by setting certain parameters equal to 0. The null hypothesis underlying this statistic is that omitted parameters are equal to 0: H_0: $\beta = 0$. The larger the value of the F-statistic, the more evidence there is to reject this hypothesis. In fact, the F-statistic follows an F-distribution, assuming homogeneity, normality, independence and no residual patterns. In this case, we can reject the null hypothesis.

In linear regression, the p-values from the `drop1` function are the same as those obtained by the t-statistic from the `summary` command, but for non-Gaussian GLMs, this is not necessarily the case. The null-hypothesis underlying the F-statistic is that the regression parameter from the term that was dropped is equal to 0. Basically, we are comparing a full and (repeatedly) a nested model.

If the model has multiple nominal variables, the `drop1` function gives one p-value for each variable, which is handy.

The `anova` command gives the following output.

```
Analysis of Variance Table
Response: ABUND
           Df Sum Sq Mean Sq F value    Pr(>F)
L.AREA      1 3471.0  3471.0 93.1303 1.247e-12
L.DIST      1   65.5    65.5  1.7568  0.191565
L.LDIST     1  136.5   136.5  3.6630  0.061868
YR.ISOL     1  458.8   458.8 12.3109  0.001019
ALT         1   78.2    78.2  2.0979  0.154281
fGRAZE      4  413.5   103.4  2.7736  0.037992
Residuals  46 1714.4    37.3
```

R uses the mean square of the full model (37.3) and the mean square on each row in a similar F-test as above. So, 93.13 is obtained by dividing 3471.0 by 37.3, and 1.75 is equal to 65.5/37.3. The mean squares are calculated from the sum of squares divided by the degrees of freedom. The sum of squares on the first row, 3471.0, is the regression sum of squares from the model $ABUND_i = \alpha + \beta \times L.AREA_i + \varepsilon_i$. The 65.5 on the second line is the decrease in residual sum of squares if L.DIST is added to this model (to see this, fit a model with only the intercept and L.AREA, and a model with intercept, L.AREA, and L.DIST and compare the two residual sum of squares obtained from the `anova` commands; the difference will be 65.5). A theoretical justification for this table can be found in Section 1.3 in Wood (2006).

The nice thing about this approach is that the last line gives us one p-value for the nominal variable GRAZE (as it is the last variable that is added), and we need this to assess whether GRAZE is significant. The disadvantage of this way of testing is that the p-values will depend on the order the variables: Change the order and you get a different conclusion.

Note that the last line of the `anova` command and the `drop1` are identical. That is because the same nested models are being compared.

The `anova` function can also be used to compare models that are nested. Suppose we fit a linear model with all explanatory variables, and a model with all explanatory variables, except GRAZE. These models are nested as the second model is a special case of the first, assuming all four regression parameters for the GRAZE levels are equal to zero (see below). The R code and its output are

```
> M2 <- lm(ABUND ~ L.AREA + L.DIST + L.LDIST +
           YR.ISOL + ALT, data = Loyn)
> anova(M1, M2)

Analysis of Variance Table
Model 1: ABUND ~ L.AREA + L.DIST + L.LDIST + YR.ISOL + ALT + fGRAZE
Model 2: ABUND ~ L.AREA + L.DIST + L.LDIST + YR.ISOL + ALT
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     46 1714.4
2     50 2127.9 -4    -413.5 2.7736 0.03799
```

The null-hypothesis underlying the F-statistic is that the four regression parameters for GRAZE (levels 2–5) are equal to 0, which is rejected at the 5% level. Note that the p-value is identical to the p-value obtained for the same test with the `anova(M1)` command. Then why do the comparison? The advantage of the `anova(M1, M2)` command is that we can control which terms are dropped. This is especially useful with multiple interaction terms.

A.3.1 Model Selection

Not all explanatory variables are significantly different from 0 as can be seen from the p-values of the t-statistics (`summary` command) or the p-values of the F-statistic (`drop1` command) presented above. If the aim of the analysis is

to understand which explanatory variables are driving bird abundances, then we could decide to drop explanatory variables that are not significant. Note this is again a subject that statisticians disagree about. There are basically three main approaches:

1. Drop individual explanatory variables one by one based on hypothesis testing procedures.
2. Drop individual explanatory variables one by one (and each time refit the model) and use a model selection criteria like the AIC or BIC to decide on the optimal model.
3. Specify a priori chosen models, and compare these models with each other. This approach is further discussed in Appendix A.6.

An example of approach three is given in the Koala case study chapter. Approach one means that you drop the least significant term, either based on the t and p-values obtained by the `summary` command or the `anova` command for comparing nested models if there are nominal variables and/or interactions. In the second approach, we use a selection criterion like the Akaike information criteria (AIC). It measures goodness of fit and model complexity. The advantage of the AIC is that R has tools to apply an automatic backwards or forwards selection based on the AIC, which makes life easy! The disadvantage is that the AIC can be conservative, and you may need to apply some fine tuning (using hypothesis testing procures from approach one) once the AIC has selected an optimal model. A backwards selection is applied by the command `step(M1)`, and its output is given as

```
Start:  AIC=211.6
ABUND ~ L.AREA + L.DIST + L.LDIST + YR.ISOL + ALT + fGRAZE
          Df Sum of Sq     RSS     AIC
- L.DIST   1      0.55 1714.98  209.62
- YR.ISOL  1      1.81 1716.24  209.66
- L.LDIST  1      5.19 1719.62  209.77
- ALT      1      7.47 1721.90  209.85
<none>                1714.43  211.60
- fGRAZE   4    413.50 2127.92  215.70
- L.AREA   1    770.01 2484.44  230.38

Step:  AIC=209.62
ABUND ~ L.AREA + L.LDIST + YR.ISOL + ALT + fGRAZE
          Df Sum of Sq     RSS     AIC
- YR.ISOL  1      1.73 1716.71  207.68
- ALT      1      7.07 1722.05  207.85
- L.LDIST  1      8.57 1723.55  207.90
<none>                1714.98  209.62
- fGRAZE   4    413.28 2128.25  213.71
- L.AREA   1    769.64 2484.62  228.38
```

```
Step:  AIC=207.68
ABUND ~ L.AREA + L.LDIST + ALT + fGRAZE
          Df Sum of Sq     RSS     AIC
- L.LDIST  1      8.32 1725.03  205.95
- ALT      1      9.71 1726.42  205.99
<none>                 1716.71  207.68
- fGRAZE   4    848.77 2565.47  222.17
- L.AREA   1    790.20 2506.90  226.88

Step:  AIC=205.95
ABUND ~ L.AREA + ALT + fGRAZE
         Df Sum of Sq     RSS     AIC
- ALT     1      5.37 1730.40  204.12
<none>                1725.03  205.95
- fGRAZE  4    914.23 2639.26  221.76
- L.AREA  1   1130.78 2855.81  232.18

Step:  AIC=204.12
ABUND ~ L.AREA + fGRAZE
         Df Sum of Sq     RSS     AIC
>none>                1730.40  204.12
- fGRAZE  4   1136.54 2866.94  224.40
- L.AREA  1   1153.85 2884.24  230.73
```

The first part of the code shows that the model containing all explanatory variables has an AIC of 211.6 before each term is dropped. The lower the AIC, the better is the model, *as judged* by the AIC. Hence, we should drop L.DIST. The procedure then goes on by dropping YR.ISOL, L.LDIST, and ALT. At this stage, no further terms are dropped; the model with both L.AREA and fGRAZE has an AIC of 204.12, and dropping any of these terms gives an higher AIC. This means that the optimal model based on the AIC contains fGRAZE and L.AREA. You should reapply this model and see whether both terms are significant. Note that both the `summary` command and the `anova` command are needed for this. Both terms are significant at the 5% level.

You could also try to see whether adding interaction between L.AREA and fGRAZE improves the model. You should be able to get one p-value for this interaction term.

A.3.2 Model Validation

Once the optimal model has been found, it is time to apply a model validation. This process consists (as a minimum) of the following steps:

- Plot (standardised) residuals against fitted values to assess homogeneity.
- Make a histogram of the residuals to verify normality. You can also use a QQ-plot.

- Plot the residuals against each explanatory variable that was used in the model. If you see a pattern, you are violating the independence assumption.
- Plot the residuals against each explanatory variable not used in the model. If you see a pattern, include the omitted explanatory variable and refit the model. If the residuals patterns disappear, include the term, even if it is not significant.
- Asses the model for influential observations. A useful tool is the Cook distance function.

Note that most bullet points use graphical tools to assess the underlying assumptions of homogeneity, normality, and independence. Statisticians tend to use graphs, but non-statisticians seem to prefer tests. Care is needed with tests as some tests that are used to assess homogeneity heavily depend on normality. In Chapter 3, we show that without a large number of replicate samples, you cannot test for normality or homogeneity.

The following R code provides just a small selection of possible graphs.

```
> M3 <- lm(ABUND ~ L.AREA + fGRAZE, data = Loyn)
> op <- par(mfrow = c(2, 2))
> plot(M3) #standard graphical output
> win.graph(); op <- par(mfrow = c(2, 2))
> #Check for normality
> E <- rstandard(M3)
> hist(E)
> qqnorm(E)
> #Check for independence and homogeneity: residuals
> #versus individual explanatory variables
> plot(y = E, x = Loyn$L.AREA, xlab = "AREA",
       ylab = "Residuals")
> abline(0,0)
> plot(E ~ Loyn$fGRAZE, xlab = "GRAZE",
       ylab = "Residuals")
> abline(0, 0)
> par(op)
```

We have not presented the graphs here, but there is some evidence of heterogeneity (as can be seen from the graph with residuals against fitted values) and non-normality. It seems that there is less spread at sites with GRAZE level 5. Based on the graph that shows residuals versus L.AREA, we seem to have some violation of independence.

A.3.3 Model Interpretation

Sometimes it is useful to include a graphical presentation of your model; one graph can tell more than many lines of text. Hence, we have given a graph that shows what

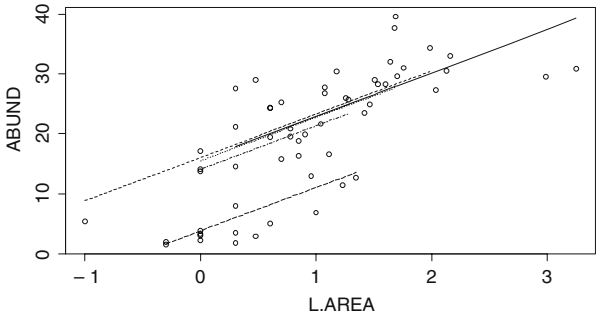

Fig. A.3 Observed bird abundance versus L.AREA with fitted values per GRAZE level. The lower line corresponds to GRAZE level 5. See the help file of `predict` how to obtain confidence intervals (either for the mean or the population) around the predicted values

the model is doing, see Fig. A.3. Observed abundances are plotted against L.AREA. Each line represents a grazing level and the lowest line is level 5. It is also possible to use different colours for the lines or different symbols for the points. The R code for this graph looks complicated, but it is really fairly simple. The first block of code is given by

```
> plot(Loyn$L.AREA,Loyn$ABUND)
> D1 <- data.frame(L.AREA = Loyn$L.AREA[Loyn$GRAZE==1],
                   fGRAZE = "1")
> D2 <- data.frame(L.AREA = Loyn$L.AREA[Loyn$GRAZE==2],
                   fGRAZE = "2")
> D3 <- data.frame(L.AREA = Loyn$L.AREA[Loyn$GRAZE==3],
                   fGRAZE = "3")
> D4 <- data.frame(L.AREA = Loyn$L.AREA[Loyn$GRAZE==4],
                   fGRAZE = "4")
> D5 <- data.frame(L.AREA = Loyn$L.AREA[Loyn$GRAZE==5],
                   fGRAZE = "5")
```

The first line makes a scatterplot of bird abundance and L.AREA. `D1` is a data frame that only contains L.AREA for which grazing equals 1. This is due to the `L.AREA = L.AREA[GRAZE == 1]`argument. Furthermore, in the data frame `D1`, GRAZE is set to 1. We added the symbols `" "` around 1 because fGRAZE is defined as a categorical variable. The values of D1 are as follows.

```
> D1
     L.AREA fGRAZE
1  0.301030      1
2  0.698970      1
3  1.414973      1
4  1.505150      1
```

```
5  1.531479      1
6  1.690196      1
7  1.698970      1
8  1.755875      1
9  2.033424      1
10 2.127105      1
11 2.158362      1
12 2.988113      1
13 3.248219      1
```

D2 to D5 are defined in a similar way. The second block of R code is

```
> P1 <- predict(M3, newdata = D1)
> P2 <- predict(M3, newdata = D2)
> P3 <- predict(M3, newdata = D3)
> P4 <- predict(M3, newdata = D4)
> P5 <- predict(M3, newdata = D5)
```

The `predict` function takes as input the object from the linear regression function `lm`, (see above where the object `M3` is created) and `newdata` specifies the values for the explanatory variables for which abundance predictions should be made. For the first line, this means that a prediction is made for GRAZE = 1, and those L.AREA values for which GRAZE equals one. This means that the fitted lines only cover that part of the gradient for which there are observed measurements for that particular grazing level. It is also possible to use

```
> D1 <- data.frame(L.AREA = Loyn$L.AREA,fGRAZE = "1")
> P1 <- predict(M3, newdata = D1)
```

But now it will predict abundances for GRAZE = 1 and *all* L.AREA values; including those L.AREA values that are outside the range of observed values, which would result in lines that cover the entire L.AREA gradient. The last bit of the R code draws the lines:

```
> lines(D1$L.AREA, P1, lty = 1)
> lines(D2$L.AREA, P2, lty = 2)
> lines(D3$L.AREA, P3, lty = 3)
> lines(D4$L.AREA, P4, lty = 4)
> lines(D5$L.AREA, P5, lty = 5)
```

In this particular case, the code produces straight lines because the L.AREA was sorted from small to large in a spreadsheet before importing into R. If this is not the case, you may end up with a spaghetti plot as the lines command connects consecutive points. If this happens, you should determine the order of the continuous variable, sort it from small to large, order the nominal variable accordingly, and then run the code for the data frame. Something along the lines of

```
> I1 <- order(L.AREA)
> SGRAZE <- GRAZE[I1]      #Use this in remaining code
> SL.AREA <- sort(L.AREA)  #Use this in remaining code
```

The lines in Fig. A.3 are parallel because there is no interaction term between L.AREA and fGRAZE in the model. If an interaction term is significant, then the lines would have different slopes. We advise you try and plot the results of a linear regression model whenever possible as it makes the interpretation much easier.

A.4 Additive Modelling

The scatterplot of ABUND against L.AREA in Fig. A.2, the residuals against L.AREA (not shown here), and the fit of the lines in Fig. A.3 all suggest that imposing a linear L.AREA effect may be incorrect. From a biological point of view, it also makes more sense to assume that the larger the forest patches, the higher the number of birds, but only up to a certain level. A generalised additive model (GAM) is a method that can be used to verify the type of model required. If the GAM indicates that the smoother is a straight line, then we know that the linear regression model is correct.

We will use a GAM with a Gaussian distribution and apply the following model:

$$\begin{aligned} ABUND_i = \alpha &+ f_1(L.AREA_i) + f_2(L.DIST_i) + f_3(L.LDIST_i) \\ &+ f_4(YR.ISOL_i) + f_5(ALT_i) + factor(GRAZE_i) + \varepsilon_i \end{aligned}$$

By default, the smoothing functions f_j are estimated by a thin plate regression spline (Chapter 3), but various alternatives like cubic regression splines exist; see the help file `?s`. It is not essential to know the difference between all these smoothers, but it becomes an issue for very large data sets. The R code to run the GAM is

```
> library(mgcv)
> AM1 <- gam(ABUND ~ s(L.AREA) + s(L.DIST) +
             s(L.LDIST) + s(YR.ISOL) + s(ALT) + fGRAZE,
             data = Loyn)
> anova(AM1)
```

We deliberately started with a model that contains all explanatory variables and not with the subset of explanatory variables (L.AREA and GRAZE) that were selected in the optimal linear regression model. The reason for this is that some variables may have a non-linear effect, which may cause them not to be significant in a linear regression model. However, if our question is: 'Is the L.AREA effect in the optimal linear regression model really linear?' as compared to 'What is the optimal model?', we could compare the optimal linear regression model containing only L.AREA and GRAZE with a GAM model that only contains a smoothing function of L.AREA and GRAZE (as a nominal variable).

The anova command does not apply a sequential F-test as it did for the linear regression model. Instead, it gives the Wald test (approximate!) that shows the significance of each term in the model. Its output is given as

```
y: gaussian
Link function: identity

Formula:
ABUND ~ s(L.AREA) + s(L.DIST) + s(L.LDIST) +
        s(YR.ISOL)+ s(ALT) + fGRAZE

Parametric Terms:
             df     F p-value
fGRAZE        4 3.348  0.0184

Approximate significance of smooth terms:
            edf Est.rank     F  p-value
s(L.AREA)  2.749    6.000 5.703 0.000221
s(L.DIST)  2.531    6.000 1.151 0.350929
s(L.LDIST) 1.000    1.000 0.107 0.745803
s(YR.ISOL) 2.650    6.000 1.425 0.228513
s(ALT)     1.000    1.000 0.510 0.479027
```

The summary command will give the estimated values for the regression parameters for each level. Note that various smoothers are not significant at the 5% level. This means that we are back to the data selection process. Again, there are various approaches, see also the linear regression section above. We can either compare a priori selected models (not discussed here), use hypothesis testing procedures or a model selection tool like the AIC. And, in this case, there is a further option, which we mention at the end of this section.

The hypothesis testing approach is the easiest; just drop the least significant term from the model, refit the model, and repeat this process until all terms are significant. This is a bit a quick and dirty approach, but is useful if computing time is long.

You can also use the AIC obtained by the AIC(AM1) command, but in gam there is no function step that will do the work for you; you have to drop each term in turn, write down the AIC, and choose the variable to drop from the model, and repeat this process a couple of times. This can be a time consuming process.

There is one other option. The optimal amount of smoothing is estimated with a method called cross-validation (Wood, 2006), where one degree of freedom produces a straight line and 10 degrees of freedom is a highly non-linear curve. In linear regression, a non-significant term is still consuming one degree of freedom. The gam function is able to produce smoothers with 0 degrees of freedom, which basically removes the need to refit the model without the terms. It only works with thin plate regression splines and cubic regression spline. The code is

```
 > AM2 <- gam(ABUND ~ s(L.AREA, bs = "cs") +
         s(L.DIST, bs = "cs") + s(L.LDIST,bs = "cs") +
         s(YR.ISOL, bs = "cs") + s(ALT, bs = "cs") +
         fGRAZE, data = Loyn)
> anova(AM2)
```

The new bit is the `bs = "cs"` part. It tells R to use the cubic regression spline with shrinkage. Again, it is not that important for you to fully understand the differences between these different types of smoothers. In practise, they look similar. Thin plate smoothers tend to be slightly more linear.

The relevant output of the `anova` command is

```
Family: gaussian
Link function: identity

Formula:
ABUND ~ s(L.AREA, bs = "cs") + s(L.DIST, bs = "cs") + s(L.LDIST,
    bs = "cs") + s(YR.ISOL, bs = "cs") + s(ALT, bs = "cs") +
    fGRAZE

Parametric Terms:
       df     F p-value
fGRAZE  4 13.06 3.9e-07

Approximate significance of smooth terms:
                edf  Est.rank      F  p-value
s(L.AREA)  2.160e+00 5.000e+00 16.336 3.66e-09
s(L.DIST)  2.499e+00 5.000e+00  1.790    0.134
s(L.LDIST) 4.077e-11 1.000e+00  0.161    0.690
s(YR.ISOL) 2.218e+00 5.000e+00  1.442    0.228
s(ALT)     7.437e-11 1.000e+00  0.673    0.416
```

Note that the smoothers for L.LDIST and ALT have 0 degrees of freedom. However, there is still some work to be done as the L.DIST and YR.ISOL smoothers are not significant at the 5% level. If you drop these two variables (one by one), you will see that the optimal model only contains an L.AREA effect and a GRAZE effect. The smoother for L.AREA of this model is presented in Fig. A.4, and the following R code is used:

```
> AM3 <- gam(ABUND ~ s(L.AREA, bs = "cs") + fGRAZE,
             data = Loyn)
> plot(AM3)
```

The model validation process should follow nearly the same steps as in linear regression. The only differences are that the residuals are obtained by the command

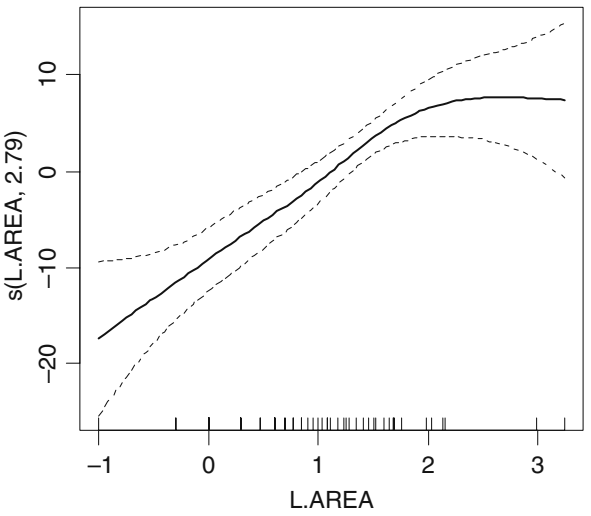

Fig. A.4 Smoothing function of L.Area in the optimal GAM. The estimated degrees of freedom is 2.79

`resid (AM3)` and there is no function that plots residuals against fitted values. You have to do this manually using the following code:

```
> E.AM3 <- resid(AM3)
> Fit.AM3 <- fitted(AM3)
> plot(x = Fit.AM3, y = E.AM3, xlab = "Fitted values",
       ylab = "Residuals")
```

Again, it is important to plot the residuals against each individual explanatory variable! If any of these graphs show a pattern, you have to find a solution.

The final thing we need to ask is whether the GAM was necessary. We ended up with the same set of explanatory variables, and one can imagine a straight line within the 95% confidence bands in Fig. A.4. The estimated degrees of freedom of 2.79 also indicate a nearly linear L.AREA effect. In fact, we can test whether the GAM is any better than the linear regression model because both models contain the same set of explanatory variables. The R code, and the output, is

```
> M3 <- lm(ABUND ~ L.AREA + fGRAZE, data = Loyn)
> AM3 <- gam(ABUND ~ s(L.AREA, bs = "cs") + fGRAZE,
             data = Loyn)
> anova(M3, AM3, test ="F")

Analysis of Variance Table
Model 1: ABUND ~ L.AREA + fGRAZE
Model 2: ABUND ~ s(L.AREA, bs = "cs") + fGRAZE
   Res.Df     RSS      Df Sum of Sq      F  Pr(>F)
1 50.0000 1730.40
2 48.2211 1528.39  1.7789    202.01 3.5827 0.04034
```

The underlying null-hypothesis is that both models are the same or formulated more mathematically that the smoother is a straight line (1 df). In this case, we can reject this null hypothesis as the more complicated model from the GAM; it is significantly better at the 5% level, even though it has a fairly unconvincing *p*-value of 0.04. But, we also prefer the GAM as it shows no residual patterns. However, the non-linear L.AREA effect is mainly due to two large patches. It would be useful to sample more of this type of patch in the future.

A.5 Further Extensions

Using the knowledge from Chapter 4, you may try extending the linear regression model with different variances per GRAZE level using the `gls` function from the `nlme` package (part of the base installation) with the `varIdent` variance structure. The same can be done for the additive model using the `gamm` function from the `mgcv` package. Examples and R code is given in Chapter 4.

Other model extensions you can try are interactions with a continuous variable and a smoother via the `by` option in the `gam` function (see also the bioluminescent case study chapter) and 2-dimensional smoothers (which models the interaction between two continuous terms). However, these extensions may cause numerical problems for this particular data set as there are only 56 observations.

A.6 Information Theory and Multi-model Inference

There has been an increasing movement away from the 'all or nothing' use of null hypotheses and *p*-values in the statistical community. Rather than thinking simply in terms of significance or non-significance with respect to an arbitrarily chosen threshold (usually 5%), it would seem more appropriate in many situations to think in terms of effects sizes and corresponding uncertainty. This has been one of the main reasons for the growth in popularity of Bayesian methods, which focus on parameter distribution (expressed in the posterior) as opposed to assessing whether a parameter is 'significant' or not.

This rejection of hypothesis testing has concentrated attention on alternative approaches. One such framework advocated by, amongst others, Burnham and Anderson is based on so-called information theory. All inference in this paradigm stems from a result derived by Kullback and Leibler (the Kullback–Leibler information, KL for short) such that (for continuous data – an alternative expression exists for count data)

$$I(f, g) = \int f(x) \log \left(\frac{f(x)}{g(x/\theta)} \right) dx$$

This quantity expresses the information lost when a chosen model 'g' is used to approximate absolute reality (denoted 'f' in the above formulation). Of course, absolute reality is not within our reach; so this quantity is not attainable in a

statistical setting. However, it is this concept of information loss which leads to the derivation of Akaike's information criterion (AIC $= -2 \times \log L(\theta; x) + 2 \times K$). This expresses the *relative expected KL information* between competing models.

So AIC quantifies the relative proximity to absolute reality amongst a candidate set of models, ideally chosen a priori. A number of refinements to the AIC have been proposed since Akaike first published this result in 1973: AIC_c (second-order AIC), advocated when sample size is relatively low; BIC (sometimes called SIC), the Bayesian information criterion; QAIC based on quasi-likelihood and preferred when overdispersion is observed in the response; and DIC (deviance information criterion), which is again used in a Bayesian setting.

Proponents of information-theory emphasise the importance of proposing a handful of scientifically meaningful models a priori. It is important to bear in mind that no statistical model will exactly represent reality. But a carefully chosen model based on sound science can lead to a close enough approximation of absolute reality to afford valuable insight. In fact, it may be that better understanding can be achieved by considering the relative performance of some or even all of the candidate models (this re-emphasises the previous point that we don't want to specify too many models at the outset, most of which may in any case be wholly implausible).

It is computationally straightforward to estimate AIC, and there are a number of other useful quantities (similarly easy to derive) that follow. Δ_i is simply the difference between AIC for model i and the lowest AIC from the set (so 0 for the 'best' model and all others referenced to). The Akaike weight w_i is a very useful quantity; it is a number between 0 and 1 reflecting the relative strength of model i relative to the other candidates. This could be interpreted in a frequentist sense as the estimated expected proportion of times that this model would turn out to be the 'best' according to the criteria chosen (AIC etc). Thus, the higher this value, the more 'weight' we put on the associated model in comparison to the others. This leads to the concept of the 'evidence ratio' (w_i/w_j) expressing the relative weight of models i and j.

A nice result of this multi-model approach is that we can base parameter inference on some or all of the models based on their relative performance (or weight). This idea becomes particularly compelling if the 'best model' is not obvious (from consideration of, for example, Akaike weights) – especially if parameter estimate θ is very different in competing models. This also leads to different quantification of error than if we were to look at one model in isolation.

This information-theoretic approach lends itself to bootstrapping too, although this is more computationally intensive.

It is not within the scope of this book to dissect this subject in detail (a good text with nice examples is *Model Selection and Multi-model Inference: A Practical Information-theoretic Approach* by Burnham and Anderson), but it is important that the reader is aware of this increasingly popular approach to statistical inference and the accompanying rationale, which represents an appealing alternative to straightforward null-hypothesis testing.

A.7 Maximum Likelihood Estimation in Linear Regression Context

Before considering this, it is worth for a moment re-visiting the mathematical form for the Normal distribution, which is used to model the random component in linear regression, namely,

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{\frac{-(y_i-\mu)^2}{2\sigma^2}}$$

The first thing to do to put this in a maximum likelihood setting is to express this as a likelihood function as a product of all the individual observations. For mathematical convenience, this is then transformed to the log-likelihood which is easier to work with.

Linear regression turns out to be a special case of GLM, namely, we can obtain so-called *closed-form solutions for the maximum likelihood* estimates for the regression parameters. For other distributions this is not possible, hence the need to express in terms of a likelihood function and solve the equations iteratively. But it is not necessary in the linear regression context and we spare the reader the full forms of the likelihood and log-likelihood.

After deriving first-order equations for the 2 parameters in the linear regression equation $y_i = \alpha + \beta \times x_i + \varepsilon_i$, where $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$, we have the following optimal solutions for both α and β (the constant and slope in the linear regression equation, respectively)

$$\hat{\alpha} = \overline{y} - \hat{\beta} \times \overline{x}$$
$$\hat{\beta} = \frac{\Sigma(x_i - \overline{x}) \times (y_i - \overline{y})}{\Sigma(x_i - \overline{x})^2}$$

These are the optimal estimates through the least squares algorithm, and if we apply a maximum likelihood approach, we will default back to these very same solutions. An estimate of variance can be obtained by dividing the total sum of residual squares by $n - 2$, i.e.

$$\hat{\sigma}^2 = \frac{\sum(y_i - \hat{y})^2}{n - 2}$$

Note that it is $n - 2$ because we have estimated 2 parameters. In matrix notation, this equates to maximum likelihood estimates of all model parameters of

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \times \mathbf{X})^{-1} \times \mathbf{X}^T \times \mathbf{y}$$

The divisor for the estimate of variance will depend on how many parameters have been estimated.

References

Agarwal DK, Gelfand AE, Citron-Pousty S (2002) Zero-inflated models with application to spatial count data. Environmental and Ecological Studies 9(4):341–355(15), Springer

Agresti A (2002) Categorical Data Analysis. Second Edition. Wiley Series in Probability and Statistics

Akaike H (1973) Information theory as an extension of the maximum likelihood principle. Pages 267–281 in BN Petrov, Csaki F, editors. Second International Symposium on Information Theory. Akadémiai Kiadó, Budapest, Hungary

Anderson DR, Burnham KP, Thompson WL (2000) Null hypothesis testing: problems, prevalence, and an alternative. Journal of Wildlife Management 64:912–923

ANZECC (1998) National koala conservation strategy. Environment Australia, Canberra, Australia

Austin MP, Meyers AJ (1996) Current approaches to modelling the environmental niche of Eucalypts: implications for management of forest biodiversity. Forest Ecology and Management 85:95–106

Barbraud C, Weimerskirch H (2006) Antarctic birds breed later in response to climate change. Proceedings of the National Academy of Sciences of the USA 103:6048–6051

Barry SC, Welsh AH (2002) Generalized additive modelling and zero inflated count data. Ecological Modelling 157:179–188

Bates D, Sarkar D (2006) lme4: Linear mixed-effects models using S4 classes. R package version 0.9975–10

Bradner H, Bartlett M, Blackinton G, Clem J, Karl D, Learned J, Lewitus A, Matsuno S, O'Connor D, Peatman W, Reichle M, Roos C, Waters J, Webster M, Yarbrough M (1987) Bioluminescence profile in the deep Pacific Ocean. Deep-Sea Research 34:1831–1840

Bjørnstad ON, Falck W (2001) Nonparametric spatial covariance functions: estimation and testing. Environmental and Ecological Statistics 8:53–70

Bjornstad ON (2008). ncf: spatial nonparametric covariance functions. R package version 1.1-1. http://onb.ent.psu.edu/onb1/R

Booth GD, Niccolucci MJ, Schuster EG (1994) Identifying proxy sets in multiple linear regression: an aid to better coefficient interpretation. Research paper INT-470. United States Department of Agriculture, Forest Service, Ogden, USA

Bowman A, Azzalini A (1997) Applied smoothing techniques for data analysis: the Kernel approach with S-Plus illustrations. Oxford, UK: Oxford University Press

Broström G (2008) glmmML: Generalized linear models with clustering. R package version 0.81–2

Brown H, Prescott R (2006) Applied Mixed Models in Medicine. Second edition. Wiley/Blackwell.

Burnham KP, Anderson DR (2001) Kullback-Leibler information as a basis for strong inference in ecological studies. Wildlife Research 28:111–119

Burnham KP, Anderson DR (2002) Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Second edition. Springer-Verlag, New York, USA

Cameron AC, Trivedi PK (1998) Regression Analysis of Count Data. Cambridge University Press, Cambridge

Carrivick PJW, Lee AH, Yau KKW (2003) Zero-inflated Poisson modeling to evaluate occupational safety interventions. Safety Science 41, 53–63
Chambers JM, Hastie TJ (1992) Statistical Models in S. Wadsworth & brooks/cole Computer Science Series
Chatfield C (2003) The Analysis of Time Series: An Introduction. Sixth edition. Chapman and Hall, Ltd, London
Chatterjee S, Price B (1991) Regression Analysis by Example. Second edition. John Wiley & Sons, New York, USA
Clarke KR, Warwick RM (1994) Change in Marine Communities: An Approach to Statistical Analysis and Interpretation. 1st edn, Plymouth, UK: Plymouth Marine Laboratory, 144 pp, Plymouth, UK: PRIMER-E, 172 pp
Cleveland WS (1993) Visualizing Data, 360 pp, Hobart Press, Summit, NJ, USA
Cliff AD, Ord JK (1981) Spatial processes. Models and applications. Pion, London, UK
Cloern, JE (2001) Our evolving conceptual model of the coastal eutrophication problem. Marine Ecology Progress Series 210:223–253
Collet D (2003) Modelling Binary Data, Texts in Statistical Science Series, Chapman and Hall, New York
Congdon P (2005) Bayesian models for categorical data, Wiley, Chichester, 446 pages
Crawley MJ (2002) Statistical computing. An introduction to data analysis using S-Plus. Wiley, New York
Crawley MJ (2005) Statistics. An introduction using R. Wiley, New York
Cressie NAC (1993) Statistics for Spatial Data. Wiley, New York
Cronin MA (2007) Abundance, habitat use and haul out behaviour of harbour seals (Phoca vitulina vitulina L.) in southwest Ireland 2007. Unpublished PhD thesis. University College Cork, Ireland. 263 pp
Cronin M, Duck C, O'Cadhla O, Nairn R, Strong D, O'Keefe C (2007) An assessment of population size and distribution of harbour seals (Phoca vitulina vitulina) in the Republic of Ireland during the moult season in August 2003. Journal of Zoology, 273:131–139
Cruikshanks R, Laursiden R, Harrison A, Hartl MGH, Kelly-Quinn M, Giller PS, O'Halloran J (2006) Evaluation of the use of the Sodium Dominance Index as a potential measure of acid sensitivity (2000-LS-3.2.1-M2) 26 pp Synthesis Report, Environmental Protection Agency, Dublin
Dalgaard P (2002) Introductory Statistics with R. Springer
Davison AC, Hinkley DV (1997) Bootstrap Methods and their Applications. Cambridge University Press, Cambridge, UK
Diggle PJ (1990) Time series: a biostatistical introduction, Oxford University Press, London
Diggle PJ, Heagerty P, Liang KY, Zeger SL (2002) The Analysis of Longitudinal Data. Second Edition. Oxford University Press, Oxford, England
Diggle PJ, Ribeiro PJ Jr (2007) Model-Based Geostatistics. Springer, New York
Dique DS, Thompson JH, Preece J, de Villiers DL, Carrick FN (2003) Dispersal patterns in a regional koala population in south-east Queensland. Wildlife Research 30:281–290
Dobson AJ (2002) Introduction to Generalized Linear Models. Second Edition. Chapman & Hall/CRC Press
Draper N, Smith H (1998) Applied Regression Analysis. Third edition. Wiley, New York
Efron B, Tibshirani RJ (1993) An Introduction to the Boostrap. Chapman and Hall, New York
Efron B, Tibshirani R (1997) Improvements on cross-validation: The .632+ bootstrap method. Journal of the American Statistical Association 92:548–560
Eilers PHC, Marx BD (1996) Flexible smoothing with B-splines and penalties, Statistical Science 11(2):89–121
Ellis-Iversen J, Smith RP, Van Winden S, Paiba GA, Watson E, Snow LC, Cook AJC (2008) Farm practices to control E. coli O157 in young cattle – A randomised controlled trial. Veterinary Research 39:03
Elphick CS, Oring LW (1998) Winter management of California rice fields for waterbirds. Journal of Applied Ecology 35:95–108

Elphich CS, Oring LW (2003) Effects of rice field management on winter waterbird communities: conservation and agronomic implications. Agriculture, Ecosystems & Environment 94:17–29

Elphich CS, Zuur AF, Ieno EN, Smith GM (2007) Investigating the effects of rice farming on aquatic birds with mixed modelling (Chapter 23) In: Analysing Ecological Data (Zuur AF, Ieno, EN, Smith GM, eds) Springer. pp. 417–434

Emmerson MC, Raffaelli DG (2000) Detecting the effects of diversity on measures of ecosystem function: experimental design, null models and empirical observations. Oikos 91(1):195–203

Emmerson MC, Solan M, Emes C, Paterson DM, Raffaelli D (2001) Consistent patterns and the idiosyncratic effects of biodiversity in marine ecosystems. Nature 411:73–77

Fahrig L (2003) Effects of habitat fragmentation on biodiversity. Annual Review of Ecology Evolution and Systematics 34:487–515

Fahrig L, Pedlar JH, Pope SE, Taylor PD, Wegner JF (1995) Effect of road traffic on amphibian density. Biological Conservation 73:177–182

Faraway JJ (2005) Linear models with R. Chapman & Hall/CRC, FL, p 225

Faraway JJ (2006) Extending the Linear Model with R. Chapman & Hall/CRC, London

Fitzmaurice GN, Laird NM, Ware J (2004) Applied longitudinal analysis. Wiley-IEEE

Flather CH, Bevers M (2002) Patchy reaction-diffusion and population abundance: the relative importance of habitat amount and arrangement. American Naturalist 159:40–56

Forman RT, Sperling T, Bissonette D, Clevenger J, Cutshall A, Dale CV, Fahrig L, France R, Goldman C, Heanue K, Jones J, Swanson F, Turrentine T, Winter T (2002) Road ecology: science and solutions. Island Press, Washington, DC

Fox J (2000) Nonparametric Simple Regression. Smoothing Scatterplots. Sage Publications, Thousand Oaks, CA

Fox J (2002) An R and S-Plus companion to applied Regression. Sage publications, Thousand Oaks, CA

Gelfand AE, Smith AFM (1990) Sampling-based approaches to calculating marginal densities. JASA 85:398–409

Gelman A (1996) Inference and monitoring convergence. In: Markov chain Monte Carlo in practice (Wilks WR, Richardson S, Spiegelhalter DJ, eds), Chapman and Hall, London, pp. 131–143

Gelman A, Carlin JB, Stern HS, Rubin DB (2003) Bayesian Data Analysis. Second Edition. Chapman and Hall. 668 pp

Gelman A, Hill J (2007) Data Analysis Using Regression and Multilevel/Hierarchical Models, Cambridge

Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence 6:721–741

Genersch E, Forsgren E, Pentikainen J, Ashiralieva A, Rauch S, Kilwinski J, Fries I (2006) Reclassification of *Paenibacillus larvae* subsp. *pulvifaciens* and *Paenibacillus larvae* subsp. *larvae* as *Paenibacillus larvae* without subspecies differentiation. International Journal of Systematic and Evolutionary Microbiology, 56: 501–511

Gilks WR, Richardson S, Spiegelhalter DJ (1996) Markov chain Monte Carlo in practice. Chapman & Hall. 512p

Gillibrand EJV, Bagley P, Jamieson A, Herring PJ, Partridge JC, Collins MA, Milne R, Priede IG (2006) Deep Sea Benthic Bioluminescence at Artificial Food falls, 1000 to 4800m depth, in the Porcupine Seabight and Abyssal Plain, North East Atlantic Ocean. Marine Biology 149: doi: 10.1007/s00227-006-0407-0

Gillibrand EJV, Jamieson AJ, Bagley PM, Zuur AF, Priede IG (2007) Seasonal development of a deep pelagic bioluminescent layer in the temperate Northeast Atlantic Ocean. Marine Ecology Progress Series 341:37–44

Goldstein H (2003) Multilevel Statistical Models. Third edition. Arnold

Graham MH (2003) Confronting multicollinearity in ecological multiple regression. Ecology 84:2809–2815

Greene WH (1997) Econometric Analysis, Third edition. Prentice-Hall, Inc, Upper Saddle River, NJ

Guthery FS, Brennan LA, Peterson MJ, Lusk JJ (2005) Information theory in wildlife science: critique and viewpoint. Journal of Wildlife Management 69:457–465

Hansen H, Brødsgaard CJ (1999) American foulbrood: a review of its biology, diagnosis and control. Bee World 80:5–23

Hanski I (1998) Metapopulation dynamics. Nature 396:41–49

Hardin JW, Hilbe JM (2002) Generalized Estimating Equations Chapman & Hall/Crc

Hardin JW, Hilbe JM (2007) Generalized Linear Models and Extensions. Second Edition. Stata Press, Texas

Harrell FE Jr (2007) Design: Design Package. R package version 2.1-1. http://biostat.mc.vanderbilt.edu/s/Design, http://biostat.mc.vanderbilt.edu/rms

Harvey AC (1989) Forecasting, structural time series models and the Kalman filter. Cambridge, UK: Cambridge University Press

Hastie T (2006) gam: Generalized Additive Models. R package version 0.98

Hastie T, Tibshirani R (1990) Generalized additive models. Chapman and Hall, London

Hastings WK (1970) Monte Carlo sampling methods using Markov Chains and their Applications. Biometrika 57:97–109

Heger A, Ieno EN, King NJ, Morris KJ, Bagley PM, Priede IG (2008) Deep-sea pelagic bioluminescence over the Mid-Atlantic Ridge. Deep-Sea Research 55:126–136

Hemmingsen W, Jansen PA, MacKenzie K (2005) Crabs, leeches and trypanosomes: an unholy trinity? Marine Pollution Bulletin 50(3):336–339

Hilbe JM (2007) Negative Binomial Regression. Cambridge

Hilborn R, Mangel M (1997) The ecological detective. Confronting models with data. Princeton University Press, Princeton, USA

Hindell MA, Lee AK (1987) Habitat use and tree preferences of koalas in a mixed eucalypt forest. Australian Wildlife Research 14:349–360

Hornitzky MAZ, Karlovskis S (1989) A culture technique for the of *Bacillus larvae* in honeybees. Journal of Apicultural Research 28(2):118–120

Hosmer DW, Lemeshow S (2000) Applied Logistic Regression. Second edition. John Wiley & Sons, New York, USA

Ieno EN, Solan M, Batty P, Pierce GJ (2006) Distinguishing between the effects of infaunal species richness, identity and density in the marine benthos. Marine Ecology Progress Series 311:263–271

Jackman S (2007) pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory, Stanford University. Department of Political Science, Stanford University. Stanford, California. URL http://pscl.stanford.edu/

Jiang J (2007) Linear and Generalized Linear Mixed Models and Their Applications Series, Springer, New York, 257 pages

Johnson DH (1999) The insignificance of statistical significance testing. Journal of Wildlife Management 63:763–772

Keele LJ (2008) Semiparametric Regression for the Social Sciences Wiley. John wiley & Sons Ltd. Chichester. 230 pages

Keitt TH, Bjørnstad ON, Dixon PM, Citron-Pousty S (2002) Accounting for spatial pattern when modeling organism-environment interactions. Ecography 25:616–625

Kuhnert PM, Martin TG, Mengersen K, Possingham HP (2005) Assessing the impacts of grazing levels on birds density in woodland habitats: a Bayesian approach using expert opinion. Environmetrics 16:717–747

Lambert D (1992) Zero-inflated Poisson Regression, With an Application to Defects in Manufacturing. Technometrics 34:1–14

Lampitt RS, Bett BJ, Kiriakoulakis K, Popova EE, Ragueneau O, Vangriesheim A, Wolff GA (2001) Material supply to the abyssal seafloor in the Northeast Atlantic. Progress in Oceanography 50:27–63

Landwehr JM, Pregibon D, Shoemaker AC (1984) Graphical methods for assessing logistic regression models. Journal of the American Statistical Association 79:61–71
Langton AES (2002) Measures to protect amphibians and reptiles from road traffic. In: Sherwood B, Cutler D, Burton J. Wildlife and Roads. The ecological impact. Imperial College Press, London. P. 223–248
Legendre P (1993) Spatial autocorrelation: trouble or new paradigm? Ecology 74:1659–1673
Legendre P, Legendre L (1998) Numerical Ecology. Second English edition. Amsterdam, The Netherlands: Elsevier, 853 pp
Liang K, Zeger S (1986) Longitudinal data analysis using generalized linear models. Biometrika 73:13–22
Lichstein JW, Simons TR, Shriner SA, Franzreb KE (2002) Spatial autocorrelation and autoregressive models in ecology. Ecological Monographs 72:445–463
Ligges U, Mächler M (2003) Scatterplot3d – an R Package for Visualizing Multivariate Data. Journal of Statistical Software 8(11), 1–20
Lindstrom A (2006) Distribution and transmission of American Foulbrood in honey bees. Doctoral thesis. Swedish University of Agricultural Sciences, Uppsala. (In English)
Longhurst AR (1998) Ecological Geography of the Sea. Academic Press, San Diego
Loyn RH (1987) Effects of patch area and habitat on bird abundances, species numbers and tree health in fragmented Victorian forests. In: Nature Conservation: the role of remnants of native vegetation (Saunders DA, Arnold GW, Burbidge AA, Hopkins AJM, eds), Surrey Beatty & Sons, Chipping Norton, NSW, pp. 65–77
Lukacs PM, Thompson WL, Kendall WL, Gould WR, Doherty PF, Burnham KP, Anderson DR (2007) Concerns regarding a call for pluralism of information theory and hypothesis testing. Journal of Applied Ecology 44:456–460
Luke DA (2004) Multilevel Modeling. SAGE Publications, Newbury Park
Luque PL (2008) Age determination and interpretation of mineralization anomalies in teeth of small cetaceans. Unpublished PhD-thesis. University of Vigo, Spain
Lunn DJ, Thomas A, Best N, Spiegelhalter D (2000) WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. Statistics and Computing, 10:325–337
Luque PL (2008) Age determination and interpretation of mineralization anomalies in teeth of small cetaceans
Maindonald J, Braun J (2003) Data Analysis and Graphics using R. Cambridge University Press, Cambridge
Martin TG, Wintle BA, Rhodes JR, Kuhnert PM, Field SA, Low-Choy SJ, Tyre AJ, Possingham HP (2005) Zero tolerance ecology: improving ecological inference by modeling the source of zero observation. Ecology Letters 8:1235–1246
Matthews A, Lunney D, Gresser S, Maitz W (2007) Tree use by koalas (Phascolarctos cinereus) after fire in remnant coastal forest. Wildlife Research 34:84–93
McAlpine CA, Rhodes JR, Callaghan JG, Bowen ME, Lunney D, Mitchell DL, Pullar DV, Possingham HP (2006) The importance of forest area and configuration relative to local habitat factors for conserving forest mammals: a case study of koalas in Queensland, Australia. Biological Conservation 132:153–165
McCarthy MA (2007). Bayesian Methods for Ecology. Cambridge University Press
McCullagh P, Nelder J (1989) Generalized Linear Models. Second edition. Chapman and Hall, London, UK
McCulloch CE, Searle SR (2001) Generalized, linear, and mixed models. John Wiley & Sons, New York, USA
McKay M, Beckman R, Conover W (1979) A comparison of three methods for selecting values of input variables in the analysis of output from computer code. Technometrics 21:239–245
Mendes S, Newton J, Reid R, Zuur A, Pierce G (2007) Teeth reveal sperm whale ontogenetic movements and trophic ecology through the profiling of stable isotopes of carbon and nitrogen. Oecologia 151:605–615

Melzer A, Carrick F, Menkhorst P, Lunney D, John BS (2000) Overview, critical assessment, and conservation implications of koala distribution and abundance. Conservation Biology 14:619–628

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equations of state calculations by fast computing machines. Journal of Chemical Physics 21:1087–1091

Miller J, Franklin J, Aspinall R (2007) Incorporating spatial dependence in predictive vegetation models. Ecological Modelling 202:225–242

Minami M, Lennert-Cody CE, Gao W, Roman-Verdoso M (2007) Modelling shark bycatch: the zero-inflated negative binomial regression model with smoothing. Fisheries Research 84:210–221

Montgomery DC, Peck EA (1992) Introduction to linear regression analysis. Wiley, New York, 504 pp

Moore BD, Wallis IR, Marsh KJ, Foley WJ (2004) The role of nutrition in the conservation of the marsupial folivores of eucalypt forests. In: Conservation of Australia's forest fauna. (Lunney D, ed), Royal Zoological Society of New South Wales, Mosman, Australia. P 549–575

Naves J, Wiegand T, Revilla E, Delibes M (2003) Endangered species constrained by natural and human factors: the case of brown bears in northern Spain. Conservation Biology 17:1276–1289

Neter J, Wasserman W, Kutner MH (1990) Applied linear statistical models. Regression, analysis of variance, and experimental design. Irwin, Homewood, USA

Nordström S, Forsgren E, Fries I (2002) Comparative diagnosis of American foulbrood using samples of adult honey bees and honey. Journal of Apicultural Science 46:5–12

O'Neil RV (1989) Perspectives in hierarchy and scale. In: Perspectives in ecological theory. (Roughgarden J, May RM, and Levin SA, eds), Princeton University Press, Princeton, USA, pp. 140–156

Pan W (2001) Akaike's information criterion in generalized estimation equations. Biometrics 57:120–125

Parmesan C (2006) Ecological and evolutionary responses to recent climate change. Annual Review of Ecology Evolution and Systematics 37:637–669

Pearce J, Ferrier S (2000) Evaluating the predictive performance of habitat models developed using logistic regression. Ecological Modelling 133:225–245

Pebesma EJ (2000) Gstat User's manual (100+pp; PDF available from http://www.gstat.org/)

Pebesma EJ (2004) Multivariable geostatistics in S: the gstat package. Computers & Geosciences, 30:683–691

Penston MJ, Millar CP, Zuur AF, Davis IM (2008) Spatial and temporal distribution of Lepeophtheirus salmonis (Krøyer) larvae in a sea loch containing Atlantic salmon, *Salmo salar* L., farms on the north-west coast of Scotland. Journal of Fish Diseases 31:361–371

Phillips SS (2000) Population trends and the koala conservation debate. Conservation Biology 14:650–659

Phillips S, Callaghan J (2000) Tree species preferences of koalas (Phascolarctos cinereus) in the Campbelltown area south-west of Sydney, New South Wales. Wildlife Research 27:509–516

Phillips S, Callaghan J, Thompson V (2000) The tree species preferences of koalas (Phascolarctos cinereus) inhabiting forest and woodland communities on Quaternary deposits in the Port Stephens area, New South Wales. Wildlife Research 27:1–10

Pierce DA, Schafer DW (1986) Residuals in generalized linear models, Journal of the American Statistical Association 81:977–986

Pierce GJ, Santos MB, Smeenk C, Saveliev A, Zuur AF (2007) Historical trends in the incidence of strandings of sperm whales (Physeter macrocephalus) on North Sea coasts: an association with positive temperature anomalies. Fisheries Research 87:219–228

Pinheiro J, Bates D (2000) Mixed effects models in S and S-Plus. Springer-Verlag, New York, USA

Pinheiro J, Bates D, DebRoy S, Sarkar D and the R Core team (2008) nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1–88

Plummer M, Best N, Cowles K, Vines K (2007) coda: Output analysis and diagnostics for MCMC. R package version 0.13-1

Plummer M, Best N, Cowles K, Vines K (2008) coda: Output analysis and diagnostics for MCMC. R package version 0. 13–3.
Potts JM, Elith J (2006) Comparing species abundance models. Ecological Modelling 199: 153–163
Priede IG, Bagley PM, Way S, Herring PJ, Partridge JC (2006) Bioluminescence in the deep sea: Free-fall lander observations in the Atlantic Ocean off Cape Verde. Deep-Sea Research I. 53:1272–1283
Quinn GP Keough MJ (2002) Experimental design and data analysis for biologists. Cambridge University Press
R Development Core Team (2008) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org
Raudenbush SW, Bryk AS (2002) Hierarchical Linear Models: Applications and Data Analysis Methods. Second edition. Sage, Newbury Park, CA
Raya CP, Balguerías E, Fernández-Núñez MM, Pierce GJ (1999) On reproduction and age of the squid Loligo vulgaris from the Saharan Bank (north-west African coast). Journal of the Marine Biological Association of the United Kingdom 79:111–120
Reed JM, Elphick CS, Zuur AF, Ieno EN, Smith GM (2007) Time series analysis of Hawaiian waterbirds. In: Analysing Ecological Data. Zuur, AF, Ieno, EN and Smith, GM. (2007). Springer. New York
Rhodes JR, McAlpine CA, Lunney D, Possingham HP (2005) A spatially explicit habitat selection model incorporating home range behavior. Ecology 86:1199–1205
Rhodes JR, Wiegand T, McAlpine CA, Callaghan J, Lunney D, Bowen M, Possingham HP (2006) Modeling species' distributions to improve conservation in semiurban landscapes: koala case study. Conservation Biology 20:449–459
Ribeiro PJ, Diggle PJ (2001) geoR: a package for geostatistical analysis R-NEWS, 1(2):15–18
Ricketts TH (2001) The matrix matters: effective isolation in fragmented landscapes. American Naturalist 158:87–99
Ridout M, Demetrio CGB, Hinde J (1998) Models for count data with many zeros. International biometric conference, Cape Town
Roulin A, Bersier LF (2007) Nestling barn owls beg more intensely in the presence of their mother than their father. Animal Behaviour 74:1099–1106
Ruppert D, Wand MP, Carroll RJ (2003) Semiparametric Regression. Cambridge Univerity Press
Sarkar D (2008) Lattice: Lattice Graphics. R package version 0.17–2
Schabenberger O, Pierce FJ (2002) Contemporary Statistical Models for the Plant and Soil Sciences, CRC Press, Boca Raton, FL
Schimek MG (ed) 2000. Smoothing and Regression: Approaches, Computation and Application, New York: Wiley
Seabrook L, McAlpine CA, Phinn S, Callaghan J, Mitchell D (2003) Landscape legacies: koala habitat change in Noosa Shire, south-east Queensland. Australian Zoologist 32: 446–461
Semlitsch RD, Bodie JR (2003) Biological criteria for buffer zones around wetlands and riparian habitats for amphibians and reptiles. Conservation Biology 17:1219–1228
Shimanuki H (1997) Bacteria. In: Honey Bee Pests, Predators, and Diseases. (Morse RA, Flottum K, eds), 3:rd ed. A.I. Root Company, Medina, Ohio, U.S.A 718 pp
Sikkink PG, Zuur AF, Ieno EN, Smith GM (2007) Monitoring for change: Using generalised least squares, non-metric multidimensional scaling, and the Mantel test on western Montana grasslands. In: Analysing Ecological Data (Zuur, AF, Ieno, EN and Smith GM). Springer
Smith JM, Pierce GJ, Zuur AF, Boyle PR (2005) Seasonal patterns of investment in reproductive and somatic tissues in the squid Loligo forbesi. Aquatic Living Resources 18:341–351
Snijders T, Bosker R (1999) An Introduction to Basic and Advanced Multilevel Modelling. SAGE Publications Ltd, Thousand Oaks, CA
Sokal RR, Rohlf FJ (1995) Biometry. Third edition. Freeman, New York, 887p

Solan M, Ford R (2003) Macroalgal-induced changes on the distribution, composition and diversity of benthic macrofauna: implications for ecosystem functioning. In: The Estuaries and Coasts of north-east Scotland. (Raffaelli D, Solan M, Paterson D, Buck AL, Pomfret JR, eds), Coastal Zone Topics 5: Proceedings of ECSA local meeting, Aberdeen, 2001. Estuarine and Coastal Sciences Association, UK, pp. 77–87

Spiegelhalter DJ, Best NG, Carlin BP, Van der Linde A (2002) Bayesian measures of model complexity and fit (with discussion). Journal of the Royal Statistical Society, Series B (Statistical Methodology) 64:583–639

Spiegelhalter DJ, Thomas A, Best N, Lunn D (2005) WinBugs User Manual Version 2.10. Cambridge, UK: MRC Biostatistics Unit

Stephens PA, Buskirk SW, Hayward GD, Del Rio CM (2005) Information theory and hypothesis testing: a call for pluralism. Journal of Applied Ecology 42:4–12

Stone M (1974) Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society. Series B (Methodological) 36:111–147

Thomas A, O'Hara B, Ligges U, Sturtz S (2006) Making BUGS Open. R News 6(1), 12–17

Thompson SK (1992) Sampling. John Wiley & Sons, New York, USA

Tobler W (1979) Cellular geography, In: Philosophy in Geography (Gale S, Olsson G, eds), pp. 379–386

Trzcinski MK, Fahrig L, Merriam G (1999) Independent effects of forest cover and fragmentation on the distribution of forest breeding birds. Ecological Applications 9:586–593

Underwood AJ (1997) Experiments in Ecology: Their Logical Design and Interpretation Using analysis of Variance. Cambridge university press, Cambridge, UK

Vaughan IP, Ormerod SJ (2005) The continuing challenges of testing species distribution models. Journal of Applied Ecology 42:720–730

Venables WN, Ripley BD (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York, ISBN 0-387-95457-0

Ver Hoef JM, Boveng PL (2007) Quasi- Poisson Vs. negative binomial regression: How should we model overdispersed count data. Ecology 88(11):2766–2772

Ver Hoef JM, Jansen JK (2007). Space-time Zero-inflated count models of harbor seals. Environmetrics 18:697–712

Verbeke G, Molenberghs G (2000) Linear Mixed Models for Longitudinal Data. New York: Springer, 568 pp

Verzani J (2005) Using R for Introductory Statistics. CRC Press, Boca Raton

Vicente J, Höfle U, Garrido JM, Fernández-de-Mera IG, Juste R, Barralb M, Gortazar C (2006) Wild boar and red deer display high prevalences of tuberculosis-like lesions in Spain. Veterinary Research 37:107–119

Villard MA, Trzcinski MK, Merriam G (1999) Fragmentation effects on forest birds: relative influence of woodland cover and configuration on landscape occupancy. Conservation Biology 13:774–783

Welsh AH, Cunningham RB, Donnelly CF, Lindenmayer DB (1996). Modelling the abundance of rare species. Statistical models for counts with extra zeros. Ecological Modelling, 88: 297–308

West B, Welch KB, Galecki AT (2006) Linear Mixed Models: A Practical Guide Using Statistical Software Chapman & Hall/CRC

Widder EA, Johnsen S, Bernstein SA, Case JF, Neilson DJ (1999) Thin layers of bioluminescent copepods found at density discontinuities in the water column. Marine Biology 134(3):429–437

Wood SN (2004) Stable and efficient multiple smoothing parameter estimation for generalized additive models. Journal of the American Statistical Association 99:673–686

Wood SN (2006) Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC

Wooldridge JM (2006) Introductory Econometrics: A Modern Approach. Third edition. Thomson. South-Western College Publishing

Woodroffe R, Ginsberg JR (1998) Edge effects and the extinction of populations inside protected areas. Science 280:2126–2128

Xie M, He B, Goh TN (2001) Zero-inflated Poisson model in statistical process control Computational statistics and data analysis, Volume 38, Number 2, 28, 191–201(11), Elsevier
Yan J (2002) geepack: Yet Another Package for Generalized Estimating Equations R-News, 2/3, pp. 12–14
Yan J, Fine JP (2004) Estimating Equations for Association Structures Statistics in Medicine, 23, pp. 859–880
Yee TW, Wild CJ (1996) Vector generalized additive models. Journal of the Royal Statistical Society, Series B, Methodological 58:481–493
Yee TW (2007) VGAM: Vector Generalized Linear and Additive Models. R package version 0.7-5 http://www.stat.auckland.ac.nz/∼yee/VGAM
Zar JH (1996) Biostatistical analysis. Third edition. Prentice-Hall, Upper Saddle River, USA
Zar JH (1999) Biostatistical analysis. Fourth edition. Prentice-Hall, Upper Saddle River, USA
Zeileis A, Hothorn T (2002) Diagnostic Checking in Regression Relationships. R News 2(3), 7–10. URL http://CRAN.R-project.org/doc/Rnews/
Zeileis A, Kleiber C, Jackman S (2008) Regression Models for Count Data in R, Journal of Statistical Software 27(8)
Ziegler A, Kastner C, Gromping U, Blettner M (1996) The Generalized Estimating Equations in the Past Ten Years: An Overview and A Biomedical Application ftp://ftp.stat.uni-muenchen.de/pub/sfb386/paper24.ps.Z
Zuur AF, Ieno EN, Smith GM (2007) Analysing Ecological Data. Springer. 680 p
Zuur AF, Ieno EN, Walker NJ, Saveliev AA, Smith GM (2009) Mixed effects models and extensions in ecology with R. Springer.

Index