GLMs in R for Ecology

Carl Smith & Mark Warren

Preface

Our goal is to produce a set of accessible, inexpensive statistics guides for undergraduate and post-graduate students that are tailored to specific fields and that use R. These books present minimal statistical theory and are intended to allow students to understand the process of data exploration and model fitting and validation using datasets comparable to their own and, thereby, encourage the development of statistical skills. We provide a list of more comprehensive texts for those that wish to continue their development as statisticians at the end of the book. The datasets and R code used in this book can be obtained by emailing the authors.

Contents

Contributors

Carl Smith
Professor of Natural History
School of Animal, Rural & Environmental Sciences
Nottingham Trent University
Southwell NG25 0QF
UK

Department of Ecology & Vertebrate Zoology
University of Łódź
12/16 Banacha Street
90-237 Łódź
Poland

Institute of Vertebrate Biology
Academy of Sciences of the Czech Republic
Květná 8
603 65 Brno
Czech Republic

email: carl.smith02@ntu.ac.uk

Mark Warren
Data Specialist
Environment Agency
Tewkesbury GL20 8JG
UK

email: mark.warren@environment-agency.gov.uk

Cover art

The cover art is the work of Laura Andrew (www.lauraandrew.com). Laura is located in central Lincoln, UK. After studying and working as an illustrator in London, Laura returned to her roots in Lincolnshire where she produces her art, and offers courses and workshops to people looking to learn new skills such as watercolour, oil painting and printmaking. Much of her art is inspired by the natural world, particularly birds. Working professionally as both an artist and illustrator Laura sells her art worldwide and her paintings have been exhibited in galleries locally and in London.

1 Introduction to GLMs

General and Generalized Linear Models (GLMs) allow the prediction of a response (or dependent) variable by either single or multiple independent variables. Independent variables (or covariates) may be continuous, categorical or a combination of both. Statistical analyses such as t-tests, ANOVA, ANCOVA and regression are types of GLM in which the independent variables are either categorical (t-test and ANOVA), continuous (regression) or a mix of both categorical and continuous (ANCOVA). The difference between General Linear Models and Generalized Linear Models is simply the way that error (i.e. the variation in the data that is not explained by the model) is handled. In a General Linear Model, errors are assumed to be independent and follow a Gaussian (normal) distribution. In a Generalized Linear Model, other data distributions can be used as an alternative to normally distributed errors. Typical data distributions used in Generalized Linear Models are binomial, Poisson, negative binomial, beta and gamma distributions, though a wide range of distributions can potentially be used, giving great flexibility in how models can be fitted to data. We present examples of Gaussian (a General Linear Model), Poisson, negative binomial, and binomial GLMs. Hereafter we will not distinguish between General and Generalized Linear Models and will refer to both as GLMs.

GLMs are specified by three elements:

1. The distribution of error terms.
2. The predictor function; comprising a set of covariates used to predict the response variable.
3. The link function, describing the linear relationship between the mean of the response variable and the model covariates.

It is good practice to specify each of these elements in the Methods section of your paper or thesis to make explicit how you have modelled your data. Examples of model specification are presented for each of the models in this book.

1.1 Introduction to R

The advent of the statistical software package R has contributed substantially to an improvement in the quality and sophistication of data analyses performed in a range of scientific fields, including ecology. While not intuitive to use, R has become the industry standard, and time invested in learning to master R will be rewarded with an improved understanding of how to handle and model data. There are several benefits to using R. It is extremely flexible and permits exploration, analysis and visualisation of almost any type of data. R also readily permits the sharing of code with collaborators or journal reviewers and can be archived with corresponding datasets for others to use and improve upon. For this book we assume basic knowledge of running R code.

2 Gaussian GLM

A Gaussian GLM is simply a linear regression model and is widely used in ecology to model a continuous variable that is assumed to be normally distributed. Typical ecological data that can be modelled with a Gaussian GLM include growth and body size data, species distributions along environmental gradients and animal and plant densities.

2.1 River macroinvertebrate response to low flows

Macroinvertebrate communities inhabiting the substrates of rivers and streams are useful indicators of river quality. They are used extensively around the world to assess the impacts of organic and inorganic pollution, changes in physical habitat quality, sedimentation and river flow conditions. The Environment Agency in the UK uses specialist invertebrate community indices to assess the biological quality of rivers and streams throughout England and Wales. Community indices are abundance weighted using data from standardised 3-minute kick-samples. The relative abundance of different macroinvertebrate taxa in a sample can be used to provide information on environmental conditions within river and stream ecosystems. One index, called the Lotic-invertebrate Index for Flow Evaluation (LIFE) (Extence *et al.* 1999), has been specifically developed to assess the biological effects of low flows and drought. High LIFE scores indicate a macroinvertebrate community dominated by taxa associated with higher river flows (lotic) and low scores indicate dominance by taxa found in more sluggish (lentic) flow conditions.

Here we analyse data that were collected each year in spring and autumn from river sampling sites within a pre-determined network covering England and Wales. The specific aim of monitoring was to assess the effects of water extraction from rivers on biological quality whilst controlling for other environmental stressors. The prediction was that in locations with greater water extraction, and reduced river flows, biological quality will be poorer.

The data in this example are a subset of the national Environment Agency dataset from one year at 66 sites that are paired to river flow gauging stations so that recorded summer river low flow can be linked to autumn macroinvertebrate samples. The 3-minute kick samples were analysed in the laboratory, with macroinvertebrates identified to family level. Abundance weightings are assigned to each taxonomic group in the sample so that a LIFE score can be calculated. In addition to the 'observed' LIFE scores, physical habitat data are used to derive 'expected' LIFE scores to provide an indication of the macroinvertebrate community in 'reference' conditions. Dividing the 'observed' by the 'expected' gives an ecological quality ratio, this is `life` in the dataset and is used as the response variable for the analysis.

2.2 Data exploration

Before fitting a model to data, it is important to perform a data exploration. A data exploration will save time by identifying any potential problems in the data and will help in deciding what type of analysis to conduct. We adopt the protocol proposed by Zuur *et al*. (2010) for conducting data exploration. This protocol comprises 6 steps and is intended to identify:

1. Outliers in response and independent variables
2. Normality and homogeneity of the response variable
3. An excess of zeros in the response variable
4. Multicollinearity among independent variables
5. Relationships among response and independent variables
6. Independence of response variable

Here we show a basic data exploration. A fuller data exploration is presented in the R code available for the book.

Import data

Data for macroinvertebrates are saved in the tab-delimited file `invert.txt` and are imported into a dataframe in R using the command:

```
> invert <- read.table(file = "invert.txt",
                       header = TRUE, dec = ".")
```

Start by inspecting the dataframe:

```
> str(invert)
```

```
'data.frame': 66 obs. of  4 variables:
$ eco   : Factor w/ 4 levels "midland","north",
$ site  : int  54739 54740 54741 55819 ...
$ rfr   : num  9.75 9.75 9.75 4.8 4.8 0.82 ...
$ life  : num  0.99 0.97 0.98 0.72 0.77 ...
```

The dataframe comprises 66 observations of 4 variables. Each row in the dataframe represents a record for an individual kick sample from a river. The variable `site` is a unique identifier for the location where kick samples were taken in each year and season. The variable `eco` is hydro-ecological region and is a categorical variable that represents features of geology, altitude and average rainfall conditions; there are four levels of this variable. The variable `rfr` is a continuous variable and represents the proportion of natural low flow once water extraction and discharges to rivers are estimated. An `rfr` value of less than 1 indicates that low flows are less than natural levels, a value of 1 indicates natural low flow levels and greater than 1 suggests that low flow levels are higher than natural. The variable `life` is the LIFE score and is a continuous variable.

Missing data can be problematic in fitting a GLM. It is necessary to check if there are any missing values in the dataframe (missing values are designated 'NA' in the tab-delimited file).

```
>  colSums(is.na(invert))

site   eco    rfr    life
0        0      0       0
```

No missing data.

2.2.1 Outliers

Outliers in the data can be identified visually using boxplots:

```
> par(mar = c(6,6,2,2), cex.lab = 1.5)
> boxplot(life ~ eco,
        ylab = "LIFE score",
        xlab = "Hydro-ecological region",
        data = invert,
        las=1)
```

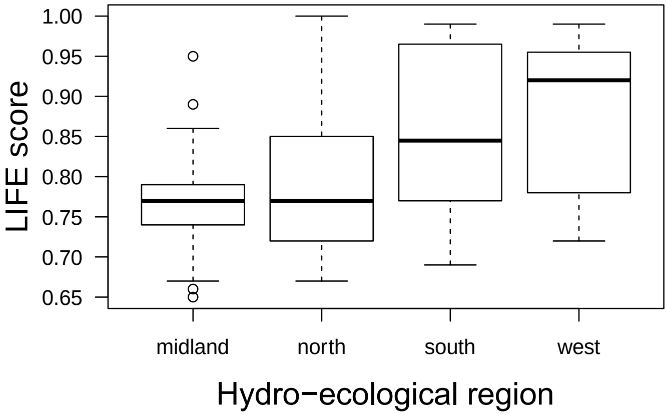

Fig. 2.1 Boxplot of LIFE scores for each hydro-ecological region.

Fig 2.1 shows that there are differences in the average LIFE scores among hydro-ecological regions. This outcome suggests there could be spatial differences in the macroinvertebrate communities related to geology, altitude and average rainfall conditions.

An alternative approach to identify outliers for continuous variables is to use multi-panel Cleveland dotplots from the `lattice` package:

```
> Names <- c("life", "rfr")
> dotplot(as.matrix(as.matrix(invert[,Names])),
        groups=FALSE,
        strip = strip.custom(bg = 'white',
        par.strip.text = list(cex = 1.2)),
        scales = list(x = list(relation = "free",
          draw = TRUE),
        y = list(relation = "free", draw = FALSE)),
                      col = 1, cex  = 1, pch = 16,
        xlab = list(label = "Value of the variable",
                      cex = 1.2),
        ylab = list(label = "Order of the data",
                      cex = 1.2))
```

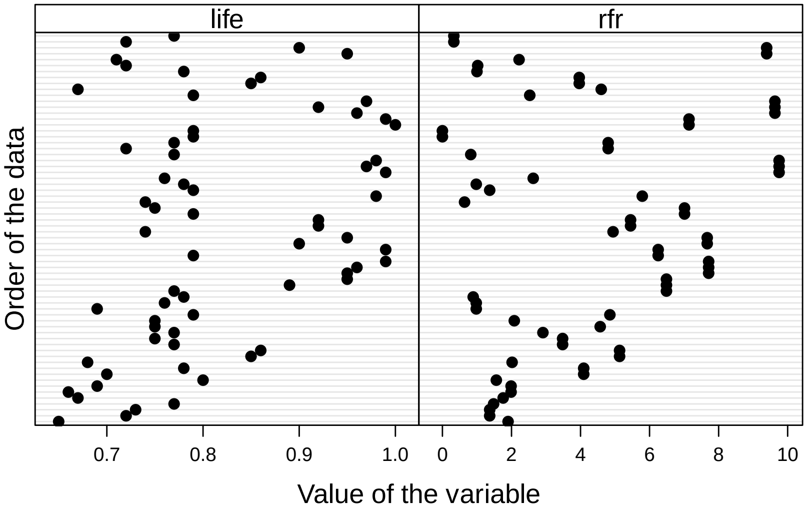

Fig. 2.2 Dotplots of the continuous variable `life` and `rfr`. Data are arranged by the order they appear in the dataframe.

Dotplots for `life` and `rfr` show no prominent outliers. However, for `rfr` there appears to be clusters of certain values. We can plot this variable on its own, split the data by hydro-ecological region and order the data by magnitude using the following R code:

```
> x <- invert[order(invert$rfr),]
> x$fEco <- factor(x$eco)
> dotchart(x$rfr,
           cex = 1,
           pch = 16,
           groups = x$fEco,
           xlab = "Proportion of natural flow")
```

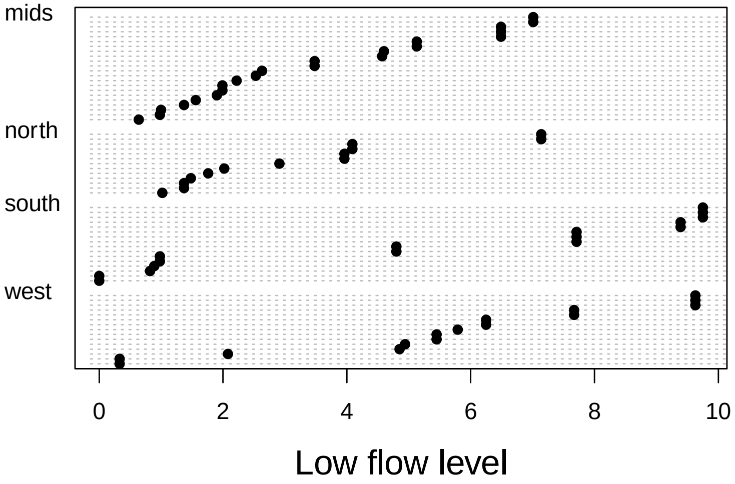

Fig. 2.3 Dotplot of `rfr` with data split by `eco` and ordered by magnitude.

There are no prominent outliers. Grubb's test can be used to test whether the value that is farthest (above or below) the mean is an outlier:

```
> grubbs.test(invert$life, type = 10))

      Grubbs test for one outlier

data:  invert$life G = 1.75465, U = 0.952, p-value = 1
alternative hypothesis: highest value 0.999 is an outlier

> grubbs.test(invert$rfr, type = 10)

        Grubbs test for one outlier
data:  invert$rfr G = 1.76860, U = 0.951, p-value = 1
alternative hypothesis: highest value 9.752 is an outlier
```

The tests indicate that there are no values that deviate significantly from the mean. Even where outliers exist, before considering dropping outliers, go on with the data exploration, but take note of the variables that have at least one outlier that may be influential in a subsequent analysis.

2.2.2 Normality and homogeneity of the dependent variable

An assumption of a Gaussian GLM is that the response variable is normally distributed at each value of the covariate values. The distribution of a continuous variable can be visualized by dividing the x-axis into “bins” and counting the number of observations in each bin as a frequency polygon using the `geom_freqpoly()` function from the `ggplot2` package:

```
> p <- ggplot()
> p <- p + ylab("Frequency")
> p <- p + xlab("LIFE score”)
> p <- p + theme(text = element_text(size=15))
> p <- p + theme(panel.background = element_blank())
> p <- p + theme(panel.border = element_rect(fill = NA,
           colour = "black", size = 1))
> p <- p + theme(strip.background = element_rect(fill =
           "white", color = "white", size = 1))
> p <- p + theme(text = element_text(size=15))
> p <- p + theme(legend.position='none')
> p <- p + geom_freqpoly(data = invert, aes(life),
           bins = 7)
> p
```

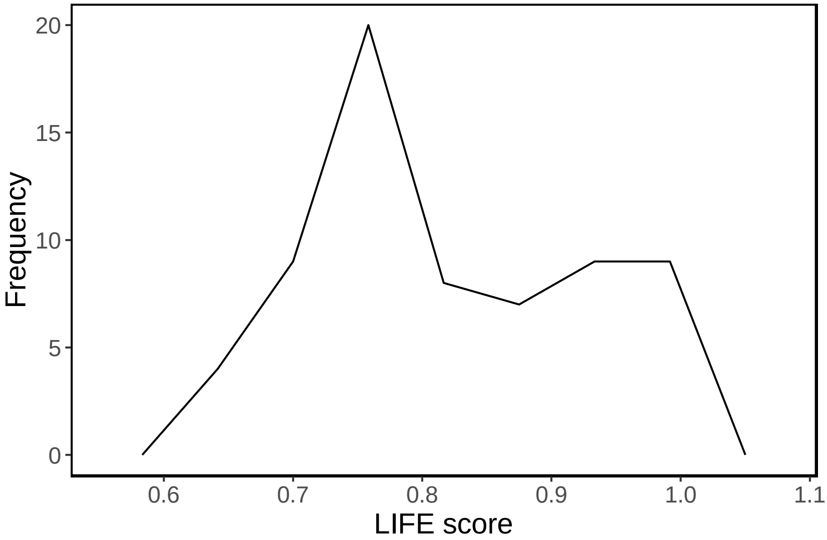

Fig. 2.4 Frequency polygon of LIFE scores for river macroinvertebrates.

The frequency polygon plot of the dependent variable (Fig. 2.4) shows potentially two distributions. However, this figure ignores the covariate values, which may explain deviation from normality. Given that we already know that the distribution of LIFE scores varies with hydro-ecological region (Fig. 2.1), it is not surprising that the data appear as they do. Low flow values and river hydro-ecological region may also affect the distribution of the dependent variable. At this stage, then, we can proceed with the data exploration bearing in mind that the raw data values for the dependent variable are not truly normally distributed. Model validation (see section 2.4) will be important to ensure the assumptions of any fitted model are met and this is more important than having normally distributed raw data values.

Homogeneity of variance is an even distribution of covariate values around the mean and is an important assumption of a Gaussian GLM. Without homogeneity of variance estimated p-values are unreliable. There are several ways to measure homogeneity of variance.

To visualise the homogeneity of the response variable in relation to a categorical covariate a boxplot is illustrative. Fig. 2.1 shows variation in spread of LIFE score data among levels of the factor `eco`, possibly indicating a lack of homogeneity. A scatterplot can be used to visualise homogeneity of variance in relation to a continuous covariate.

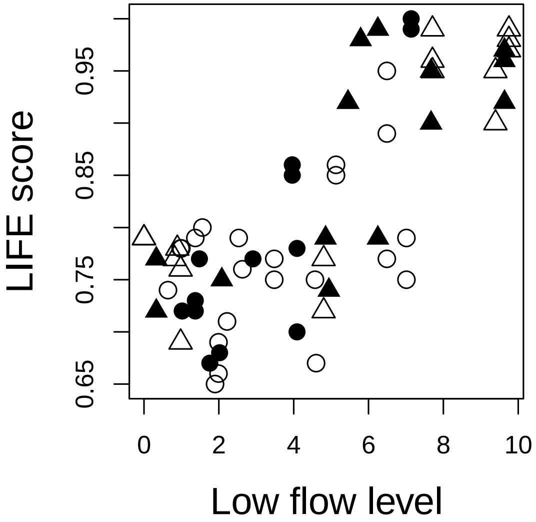

Fig. 2.5 Scatterplot of `life` scores and `rfr` for each level of `eco` (open circles = sou th, closed circles = west, open triangles = north, closed triangles = midland).

There are several tests of homogeneity of variance, such as Bartlett's Test, the F-ratio test, and Levene's test. The first two of these assume normality of the data. If your data deviate from normality they should not be used. Levene's test does not assume normality. An alternative is the Brown & Forsythe test, which uses the median rather than mean in its estimation, and is robust to departures from normality. This test is based on Levene's test and can be obtained using the `levene.test()` function from the `lawstat` package:

```
> levene.test(invert$life,
              invert$eco,
              location = c("median"),
              trim.alpha = 0.25)

Levene's Test for Homogeneity of Variance

       Df  F value  Pr(>F)
group  3   2.3415   0.0818
```

Which shows that the data do not deviate significantly from homogeneity.

2.2.3 Lots of zeros in the response variable

Zeros should not be omitted from a dataset. However, an excess of zeros in the response variable, termed 'zero inflation', can cause problems with an analysis. Fortunately, there are a number of ways of dealing with zero inflation. The first step is to identify whether there is a potential problem. The percentage of zeros in the response variable can be estimated as:

```
> sum(invert$life == 0,
      na.rm = TRUE) * 100 / nrow(invert)

[1] 0
```

There are no zeros in the response variable for this dataset but you should always check with your own datasets. If there had been zeros, how many would be too many? The question of how many zeros leads to zero inflation is often asked but cannot be answered without fitting a model and then running simulations from it to see how many zeros are predicted and then compared to the raw data. This procedure is dealt with in Section 3.4 of Chapter 3.

2.2.4 Multicollinearity among covariates

Along with normality of residuals and homogeneity of variance, an additional assumption of linear modelling is independence of the independent variables. In ecological studies it is not unusual to collect a large number of variables, which are often highly correlated. If covariates in a model are correlated, then the model may produce unstable parameter estimates with inflated standard errors that will result in an overall significant model but with no significant predictors.

Multicollinearity can be tested in several ways. The simplest is to construct a correlation matrix with corresponding pairplots. The code for this plot is available in the R file associated with this chapter.

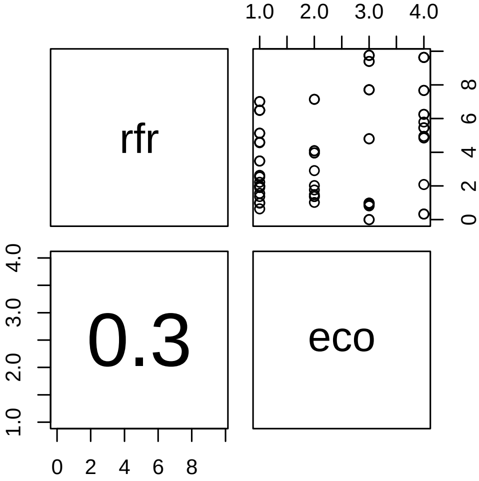

Fig. 2.6 Pairplot of `rfr` and `eco`. The lower panel shows the pairwise Pearson correlation, with font size proportional to correlation coefficient. Variables are not collinear.

Another approach to identifying multicollinearity is by calculating a variance inflation factor (VIF) for each variable. The VIF is an estimate of the proportion of variance in one predictor explained by all the other predictors in the model. A VIF of 1 indicates no collinearity. VIF values above 1 indicate increasing degrees of collinearity. VIF values exceeding 3 are considered problematic (Zuur *et al.* 2010). In this case the variable with the highest VIF should be removed from the model and the VIFs for the model recalculated.

The VIF for a model can be estimated using the `vif` function from the `car` package:

```
> vif(lm(life ~ rfr + eco,
                data = invert))

     GVIF      Df        GVIF^(1/(2*Df))
rfr  1.135876  1         1.065775
eco  1.135876  3         1.021461
```

For the macroinvertebrate model estimated VIFs are <3, so there appear to be no serious problems with multicollinearity.

2.2.5 Relationships among dependent and independent variables

Visual inspection of the data using plots is a critical step and will illustrate whether relationships are linear or non-linear and whether there are interactions between covariates.

```
> xyplot(life ~ rfr | eco,
         data = invert,
         layout = c(2,2),
         xlab = list(label = "Low flow level",
            cex = 1.2),
         ylab = list(label = "LIFE score",
            cex = 1.2),
         strip = function(bg = 'white', ...)
         strip.default(bg = 'white', ...),
         scales = list(alternating = TRUE,
            x = list(relation = "free"),
            y = list(relation = "same")),
         panel = function(x,y){
         panel.grid(h = -1, v = 2)
         panel.points(x,y, col = 1,
            pch = 16,
            cex = 1.2)
         panel.abline(lm(y~x),
            col = 1,
            lwd = 5)})
```

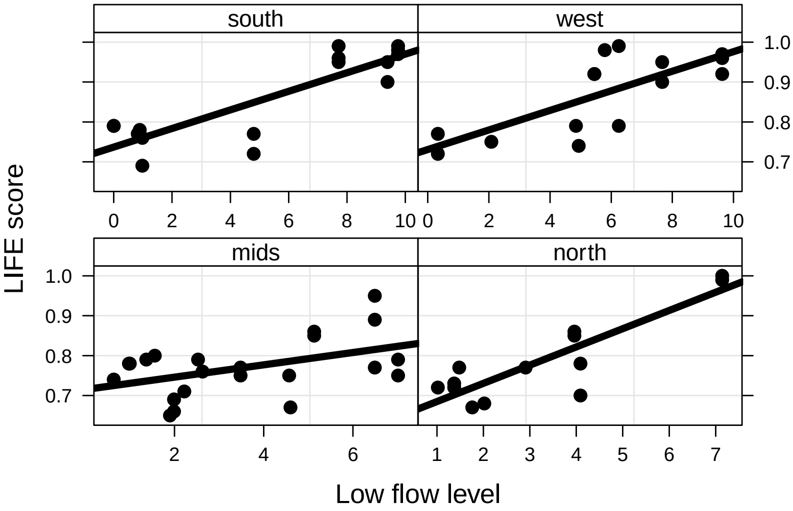

Fig. 2.7 Multipanel scatterplot of `life` scores and `rfr` across hydro-ecological regions (`eco`) with a line of best fit plotted.

The plot of the data in Fig. 2.7 do not suggest strongly non-linear patterns in the data. Fitted lines for the relationship between `life` and `rfr` indicate that the nature of this relationship is different for at least one level of `eco` (the level for 'north'), implying an interaction between low river flow and hydro-ecological region. If the relationship between `life` and `rfr` did not vary between regions; i.e. the slopes were the same in each region, the implication would be that there was no interaction with hydro-ecological region. In this case, inclusion of an interaction term in the model would not be justified.

2.2.6 Independence of response variable

A critical assumption for a GLM is that each observation in a dataset is independent of all others. For some data this assumption is difficult to confirm but the risk of non-independence can be reduced by careful sampling. Strictly randomly collected samples will tend to be independent.

Additional information, such as spatial location or time of collection, can be included in a dataset. Spatial and temporal dependency in ecological data are common and require specific modelling approaches.

For the river macroinvertebrate data, samples were collected by experienced biologists and we are only using one observation from each sampling site for one particular year. In this case, then, we can be reassured that the response variable values are independent.

2.3 Model fitting

The data exploration showed:

1. No outliers in the response variable, `life`.
2. A non-normally distributed but homogenous response variable.
3. No zeros in the response variable.
4. No serious collinearity between variables.
5. A potential interaction between `rfr` and `eco`.
6. Probable (but untested) independence of the response variable.

Given these outcomes of the data exploration the model is fitted as:

```
> Gaus1 <- lm(life ~ rfr * eco,
                     data = invert)
```

The numerical output is obtained with the `summary` function:

```
> summary (Gaus1)
              Estimate   Std. Error t value Pr(>|t|)
(Intercept)    0.715202   0.026291  27.203  <2e-16
rfr            0.015447   0.006326   2.442  0.01768
econorth      -0.076067   0.042455  -1.792  0.07840
ecosouth       0.021766   0.037257   0.584  0.56133
ecowest        0.015707   0.044434   0.353  0.72500
rfr:econorth   0.030198   0.010783   2.800  0.00692
rfr:ecosouth   0.007868   0.007496   1.050  0.29827
rfr:ecowest    0.009060   0.008432   1.074  0.28707

Residual standard error: 0.0628 on 58 degrees of freedom
Multiple R-squared:  0.6738,     Adjusted R-squared:  0.6345
F-statistic: 17.12 on 7 and 58 DF,  p-value: 4.742e-12
```

This output shows interesting patterns. However, before attempting to interpret these results it is necessary to conduct model validation.

2.4 Model validation

For the fitted model, validation requires verification of:

1. Homogeneity of variance.
2. Model misfit.
3. Normality of residuals.
4. Absence of influential observations.

2.4.1 Homogeneity of variance

Homogeneity of variance can be assessed visually by plotting model residual variance (the variance in the response variable that is not explained by the model) against model fitted values. R code to plot standardised residuals against fitted values is given by:

```
Fitted <- fitted(Gaus1)
Resid  <- resid(Gaus1, type = "pearson")
par(mfrow = c(1,1), mar = c(5,5,2,2))
plot(x = Fitted, y = Resid,
     xlab = "Fitted values",
     ylab = "Pearson Residuals")
abline(h = 0, lty = 2)
```

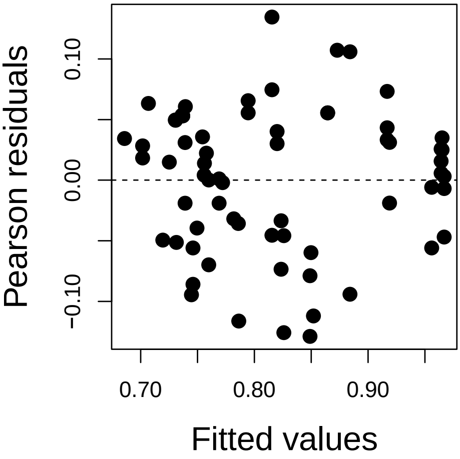

Fig. 2.8 Pearson residuals plotted against fitted values to assess homogeneity of variance. Ideally, the distribution of residuals around zero should be consistent along the horizontal axis.

The distribution of residuals is consistent along the horizontal axis (Fig. 2.8); the absolute values of the residuals are independent of the fitted values, which imply homogeneity in the model.

2.4.2 Model misfit

Model misfit occurs if key covariates (including interactions) are missing from the model, or the model departs from linearity. Model misfit can be recognised visually by plotting Pearson residuals against each covariate in the model, as well as those not included in the model.

```
> plot(x = invert$rfr,
       y = Resid,
       xlab = "Low flow level"
       ylab = "Pearson residuals",
       pch = 16, cex = 1.5)
> abline(h = 0, lty = 2)
```

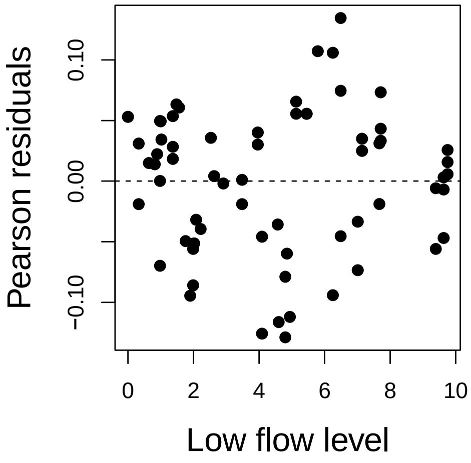

Fig. 2.9 Pearson residuals plotted against `rfr` to assess model misfit. Ideally, the distribution of residuals around zero should be consistent along the horizontal axis.

For the covariate `rfr`, the distribution of residuals is relatively consistent along the horizontal axis and shows no obvious patterns (Fig. 2.9).

```
> plot(x = invert$eco,
       y = Resid,
       xlab = "Hydro-ecological region",
       ylab = "Pearson residuals",
       pch = 16, cex = 1.5)
> abline(h = 0, lty = 2)
```

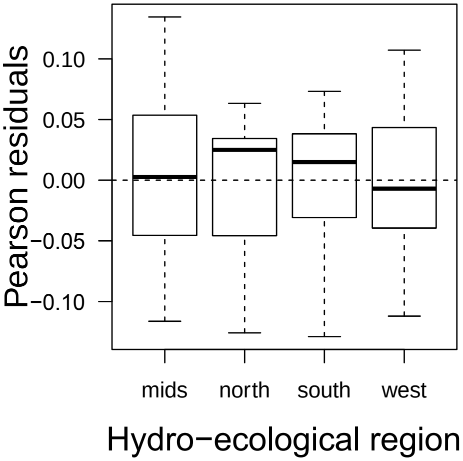

Fig. 2.10 Boxplot of Pearson residuals from different hydro-ecological regions.

For the categorical covariate `eco`, the distribution of residuals is relatively consistent across all hydro-ecological regions (Fig. 2.10).

2.4.3 Normality of residuals

The normality of residuals can be judged by plotting a histogram:

```
p <- ggplot()
p <- p + ylab("Frequency")
p <- p + xlab("Pearson residuals")
p <- p + theme(text = element_text(size=15))
p <- p + theme(panel.background = element_blank())
p <- p + theme(panel.border = element_rect(fill = NA,
               colour = "black", size = 1))
p <- p + theme(strip.background = element_rect(fill =
              "white", color = "white", size = 1))
p <- p + theme(text = element_text(size=15))
p <- p + geom_histogram(colour = "black", fill = "white",
               data = invert, aes(Resid), bins = 8)
p
```

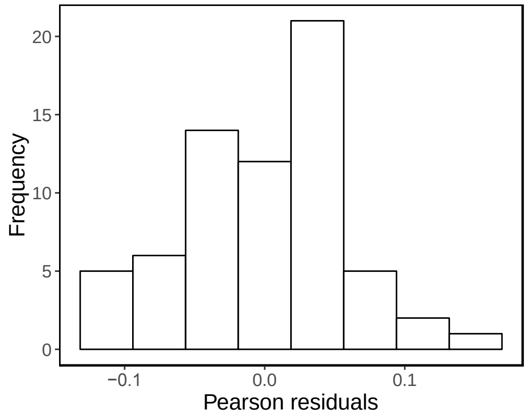

Fig. 2.11 Histogram of model Pearson residuals.

The assumption of the normality of the model residuals appears to me met (Fig. 2.11) despite the distribution of the raw data that did not follow a normal distribution (Fig. 2.4).

2.4.4 Absence of influential observations

The absence of influential observations can be tested by plotting Cook's distance. This function identifies data points with large influence. Cook's distance is estimated by systematically dropping each observation and comparing the fitted values with those when all observations are included in the model. A Cook's distance exceeding 1 indicates an influential data point. R code to plot Cook's distance for model Gaus1 is given by:

```
> par(mfrow = c(1, 1))
> plot(cooks.distance(Gaus1),
       xlab = "Observation",
       ylab = "Cook's distance",
       type = "h",
       ylim = c(0, 1.1),
       cex.lab =  1.5)
> abline(h = 1, lty = 2)
```

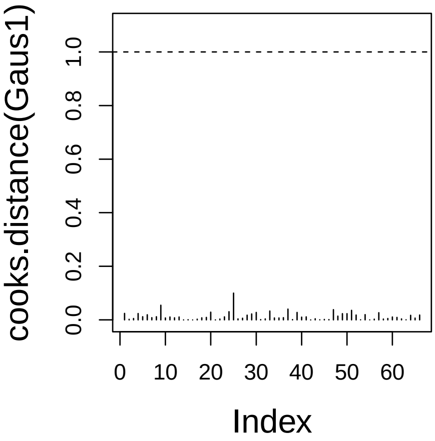

Fig. 2.12 Plot of Cook's distance for model Gaus1. A Cook's distance of 1 (indicated by a dashed horizontal line) denotes an influential observation.

There is no evidence from plotting Cook's distance for influential observation in the model (Fig 2.12).

Model validation has shown no evidence of model misfit, model residuals appear normal and there are no influential observations. However, there is some evidence for a lack of homogeneity of variance (termed heteroscedasticity) but this is for a covariate that was not included in the model.

2.5 Model presentation

We can specify the model using mathematical notation in the following way:

$$life_i \sim Gaussian(\mu_i, \sigma^2)$$

$$E(life_i) = \mu_i \text{ and } \mathrm{var}(life_i) = \sigma^2$$

$$\mu_i = \beta_1 + \beta_2 \times rfr_i + \beta_3 \times eco_i + \beta_4 \times rfr_i \times eco_i$$

Where $life_i$ is the macroinvertebrate metric for river i assuming a normal distribution with mean μ_i and variance σ^2. rfr_i is a continuous covariate corresponding with the low flow level for river i and eco_i is a categorical covariate with four levels corresponding with the hydro-ecological region in which a surveyed river was located. A full model specification should be included in the Methods section of a paper or dissertation. The numerical output of the model is obtained with:

```
> summary (Gaus1)
             Estimate   Std. Error t value Pr(>|t|)
(Intercept)   0.715202   0.026291  27.203  <2e-16
rfr           0.015447   0.006326   2.442  0.01768
econorth     -0.076067   0.042455  -1.792  0.07840
ecosouth      0.021766   0.037257   0.584  0.56133
ecowest       0.015707   0.044434   0.353  0.72500
rfr:econorth  0.030198   0.010783   2.800  0.00692
rfr:ecosouth  0.007868   0.007496   1.050  0.29827
rfr:ecowest   0.009060   0.008432   1.074  0.28707

Residual standard error: 0.0628 on 58 degrees of freedom
Multiple R-squared:  0.6738,     Adjusted R-squared:  0.6345
F-statistic: 17.12 on 7 and 58 DF,  p-value: 4.742e-12
```

These results can be more formally presented in the following way:

Table 2.1. Summary of Gaussian GLM to model the macroinvertebrate LIFE score in a set of English rivers

Model parameter	Estimate	SE	*P*
Intercept$_{\text{(midland)}}$	0.715	0.026	<0.001
rfr	0.015	0.006	0.018
eco$_{\text{(north)}}$	-0.077	0.042	0.078
eco$_{\text{(south)}}$	0.022	0.037	0.561
eco$_{\text{(west)}}$	0.016	0.044	0.725
rfr x eco$_{\text{(north)}}$	0.030	0.011	0.007
rfr x eco$_{\text{(south)}}$	0.008	0.008	0.298
rfr x eco$_{\text{(west)}}$	0.009	0.008	0.287

These results indicate a modest interaction between LIFE scores and hydro-ecological region. To understand this result it is best to visualize the model result in a figure. The R code to do so is available in the accompanying R code.

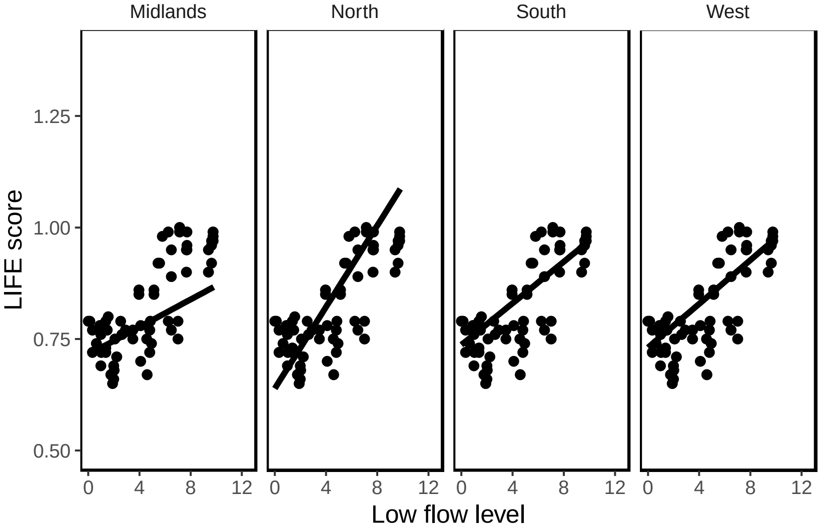

Fig. 2.13 Mean fitted LIFE scores for rivers (solid line) and 95% confidence intervals (shaded area) against low flow level in four hydro-ecological regions. Black dots are observed data.

Macroinvertebrate LIFE scores do not generally differ across all four hydro-ecological regions. However, the relationship between relative natural flow and LIFE score differs slightly between regions and is more positive for the north hydro-ecological region (Fig 2.13).

Conclusions

The Gaussian GLM predicted a positive relationship between magnitude of low river flow (`rfr`) and the macroinvertebrate index (`life`). This relationship changes slightly across hydro-ecological regions, with the northern region showing a stronger positive relationship.

References

Extence, C.A., Balbi, D.M. & Chadd, R.P., 1999. River flow indexing using British benthic macroinvertebrates: a framework for setting hydroecological objectives. *Regulated Rivers: Research & Management* 15, 545-574.

Zuur, A.F., Ieno, E.N. and Elphick, C.S., 2010. A protocol for data exploration to avoid common statistical problems. *Methods in Ecology & Evolution* 1, 3-14.

3 Poisson GLM

A Poisson GLM is suitable for ecological data in which the response variable comprises count data, such as the number of individuals or species in a specific habitat. Data must not take values below zero and the variance is assumed approximately equal to the mean.

3.1 Abundance of freshwater mussels

Unionid freshwater mussels are benthic macroinvertebrates that play a key role in the ecology of many freshwaters. They use a muscular foot and shell to burrow into the sediment and filter feed on suspended particles using cilia-generated water currents. They possess a parasitic larval stage, called a glochidia, that attach to a vertebrate host, usually a fish, and subsequently metamorphose into a juvenile mussel. Freshwater mussels are globally threatened, with declines in distribution and abundance associated with habitat modification, declines in water quality, impacts of non-native species, declines in fish hosts, and over-exploitation.

As part of a larger scale study, Smith *et al.* (2000) surveyed the abundance of freshwater mussels in a series of lakes in the Danube basin in the Czech Republic. The aim of the study was to identify which environmental variables predicted the abundance of the swan mussel (*Anodonta cygnea*). Mussels were collected by hand from 1 m^2 quadrats. In total, 21 lakes were surveyed, though data for only a single lake are presented here .

For each quadrat, water depth was measured (m) and the substrate type classified as either mud, sand or gravel. All freshwater mussels in the quadrat were collected, identified to species and counted. Four mussel species were present. In addition to the swan mussel, the duck mussel (*A. anatina*), painter's mussel (*Unio pictorum*) and swollen mussel (*U. tumidus*) were collected. The number of swan mussels is the response variable, and comprises a count that is bounded at zero. Water depth, and the abundance of duck, painter's and swollen mussels are continuous covariates. Substrate type is a categorical covariate.

It was predicted that swan mussel abundance would be positively associated with water depth and a mud substrate, but negatively with the abundance of duck mussels, to which they are closely related. The abundance of painter's and swollen mussels, which are more distantly related to swan mussels, were combined into a single covariate (unio) and predicted to have no association with the abundance of swan mussels.

3.2 Data exploration

As with a Gaussian GLM, before fitting a Poisson GLM it is necessary to perform a data exploration (see section 2.2). A Poisson GLM does not assume normality of the response variable, and homogeneity of variance will be assessed using the residuals of the model as part of model validation.

Import data

Data for mussels are saved in the tab-delimited file `muss.txt` and are imported into a dataframe in R using the command:

```
> muss <- read.table(file = "muss.txt",
                   header = TRUE, dec = ".")
```

Inspect the dataframe:

```
> str(muss)

'data.frame': 95 obs. of  5 variables:
$ depth: num  0.08 0.12 0.14 0.17 0.18 0.24 0.25...
$ subs : Factor w/ 3 levels "gravel","mud"...
$ unio : int  0 0 0 0 0 0 0 0 0 12 ...
$ duck : int  0 0 0 0 0 1 0 0 0 2 ...
$ swan : int  0 0 0 0 0 0 0 1 0 0 ...
```

The dataframe comprises 95 observations of 5 variables. Each row in the dataframe represents a separate quadrat. Substrate (`subs`) is a factor; i.e. a categorical variable, with three levels (gravel, mud, sand). Water depth (`depth`), and abundance of painter's and swollen mussels (`unio`), duck (`duck`) and swan mussels (`swan`), are all continuous covariates.

Missing data can be problematic in fitting a Poisson GLM. It is necessary to check if there are any missing values in the dataframe (missing values are designated 'NA' in the tab-delimited file).

```
>  colSums(is.na(muss))

depth   substrate   unio      duck      swan
```

```
  0           0           0           0           0
```

No missing data.

3.2.1 Outliers

Outliers in the data can identified visually using Cleveland dotplots:

```
> Var <- c("depth", "substrate", "unio", "duck", "swan")
> dotplot(as.matrix(as.matrix(muss[,Var])),
     groups=FALSE,
     strip = strip.custom(bg = 'white',
     par.strip.text = list(cex = 1.2)),
     scales = list(x = list(relation = "free", draw = TRUE),
     y = list(relation = "free", draw = FALSE)),
     col=1, cex  = 0.6, pch = 16,
     xlab = list(label = "Data range", cex = 1.5),
     ylab = list(label = "Data order", cex = 1.5))
```

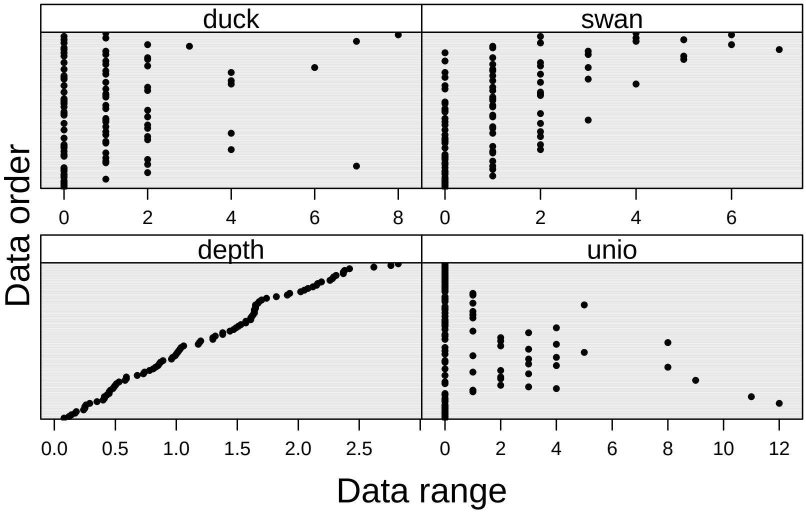

Fig. 3.1 Dotplots of duck mussel abundance (duck), swan mussel abundance (swan), depth (depth), and painter's and swollen mussel abundance (unio). Data are arranged by the order they appear in the dataframe.

There are no obvious outliers in the data (Fig. 3.1). Are the data balanced among different levels of the categorical covariate?

```
> table(muss$subs)

gravel   mud   sand
    16    49     30
```

The data are not well balanced among levels. However, if data are a random sample from the population, then a lack of balance is inevitable. In the present case, care must be taken in fitting a complex model to the data.

3.2.2 Lots of zeros in the response variable

The number of zeros in the response variable can be estimated as:

```
> sum(muss$swan == 0) * 100 / nrow(muss)
```

```
[1] 40
```

40% of quadrats contained no swan mussels. This figure is high and could cause problems.

3.2.3 Multicollinearity among covariates

Use a correlation matrix with corresponding pairplots. The code for this plot is available in the R file associated with this chapter.

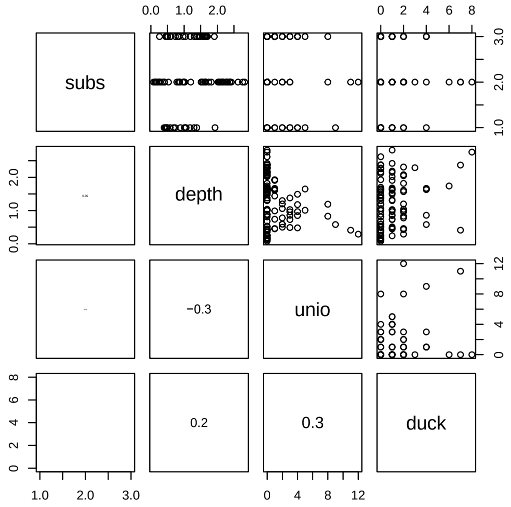

Fig. 3.2 Pairplot of covariates. The lower panel shows pairwise Pearson correlations,

with font size proportional to correlation coefficient. No covariates are collinear.

Fig. 3.2 suggests covariates are not collinear. This conclusion can be confirmed by estimating the variance inflation for the covariates using the `vif` function:

```
> vif(glm(swan ~ substrate + depth + unio + duck,
                 family = poisson,
                 data = muss))

      GVIF  Df  GVIF^(1/(2*Df))
subs  1.24  2   1.05
depth 1.56  1   1.25
unio  1.36  1   1.17
duck  1.12  1   1.06
```

Estimated VIFs are <3, so there is no problem with multicollinearity.

3.2.4 Relationships among dependent and independent variables

Plot data to examine whether data are linear or non-linear and whether there are interactions between covariates.

```
> par(mfrow=c(2,2), mar=c(5,5,1,1))
> plot(y = muss$swan, x = muss$depth,
  xlab = "Depth (m)", ylab = "Swan mussel abundance")
> plot(swan ~ unio,  data = muss,
  xlab = "Unio mussel abundance", ylab = "Swan mussel
  abundance")
> plot(swan ~ duck,  data = muss,
  xlab = "Duck mussel abundance", ylab = "Swan mussel
  abundance")
> boxplot(swan ~ subs, data = muss,
  xlab = "Substrate type", ylab = "Swan mussel
  abundance")
```

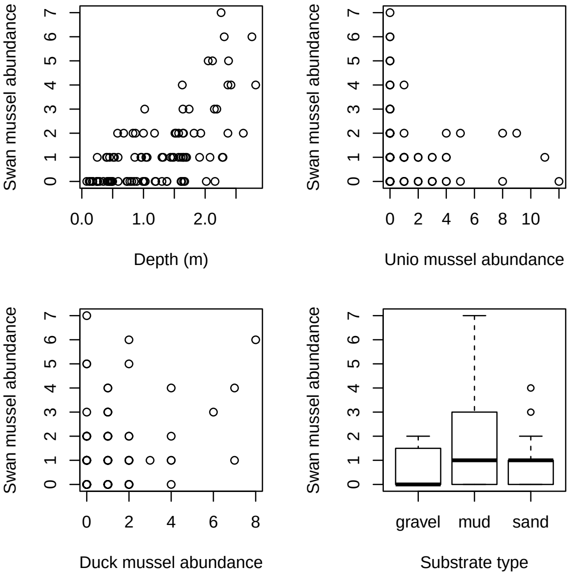

Fig. 3.3 Plots of number of swan mussels in quadrats against water depth (m), number of unio and duck mussels and substrate type.

There is a positive association between swan mussel abundance and water depth, and a weak negative association with unio abundance. There is no direct association with duck mussel abundance. Swan mussels are more abundant on a mud and sand substrate in comparison with gravel (Fig. 3.3). It is also informative to plot two covariates together using multipanel scatterplots.

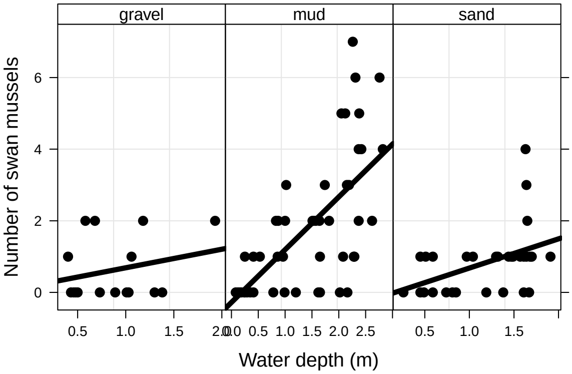

Fig. 3.4 Multipanel scatterplot of number of swan mussels in quadrats against water depth (m) on three different substrates.

The strength of relationship between the number of swan mussels in quadrats with a mud substrate in comparison with sand and gravel is greater, which suggests a possible interaction between the effects of water depth and substrate on swan mussel abundance (Fig. 3.4). The code for Figs 3.4-3.6 is available in the R file associated with this chapter.

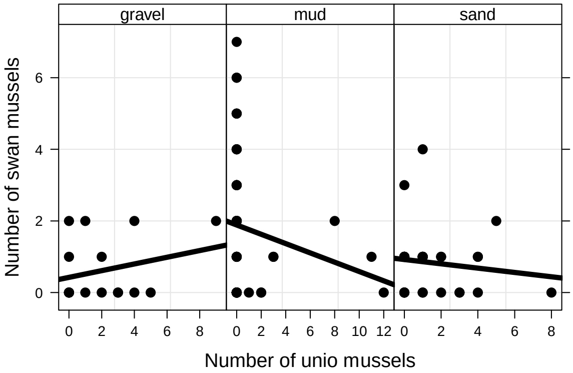

Fig. 3.5 Multipanel scatterplot of number of swan mussels in quadrats against number of unio mussels on three different substrates.

The relationship between swan and unio mussel abundance varies with substrate (Fig. 3.5). With mud and sand the relationship is negative while on a gravel substrate it is positive. Again, this pattern suggests an interaction between unio abundance and substrate on swan mussel abundance.

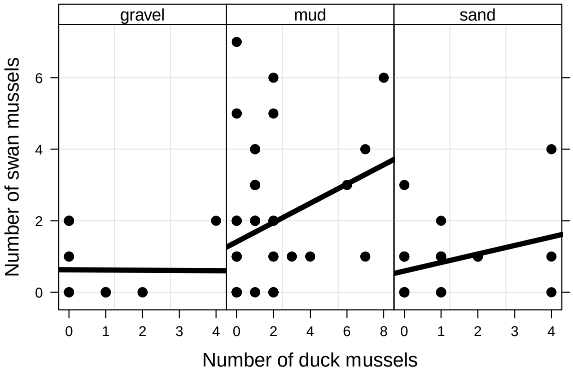

Fig. 3.6 Multipanel scatterplot of number of swan mussels in quadrats against number of duck mussels on three different substrates.

In Fig. 3.6 the strength of relationship between the number of swan mussels and duck mussels varies modestly among substrate types but is broadly consistent. This pattern suggests that there is no interaction between number of duck mussels and substrate on swan mussel abundance.

3.2.5 Independence of response variable

An assumption is that swan mussel abundances for each quadrat are independent of each other; swan mussel abundance in one quadrat should not provide be informative of swan mussel abundance in another. Data were collected to achieve independence, but insufficient data were collected to adequately test this assumption. Additional data on quadrat location could be used to test this assumption, but these were not collected. We can proceed with model fitting, but with the caveat that the assumption of response variable independence has not been tested.

3.3 Model fitting

The data exploration showed:

1. No outliers in the data.
2. A high proportion of zeros in the response variable.
3. Imbalance of data among levels of the categorical covariate 'substrate'
4. No collinearity between covariates.
5. Potential interactions between substrate type and both water depth and unio mussel abundance.
6. Probable (but untested) independence of the response variable.

Given the imbalance in the data the model will be fitted without interactions.

The Poisson distribution

The Poisson is a non-normal distribution that is effective for modelling strictly positive integer data (such as counts of mussels in quadrats). It has a single parameter (lambda, λ), which is both the mean and variance of the response variable. Sometimes you will see mu (μ) used to represent the mean. The variance in the Poisson distribution is proportional to the mean so that larger mean values have larger variation.

The predictor function

A GLM uses a predictor function (eta, η) that specifies the covariates to be used in the model. In this example for swan mussel abundance we use:

η = Intercept + Substrate + Depth + Duck mussels + Unio mussels

The link function

The link function is used to link the response variable (counts of swan mussels) and the predictor function (covariates). In the case of a Poisson GLM the default is a log link function. The link function is needed to ensure model fitted values remain positive, while allowing zeros in the data.

So, to fit the model in R, we must specify the 'family' and the link function:

```
> Pois1 <- glm(swan ~ subs + depth + duck + unio,
                      data = muss,
                      family = poisson(link = log))

> summary (Pois1)

               Estimate  Std. Error z value Pr(>|z|)
(Intercept)   -1.9556     0.5066   -3.86   <0.001
subsmud        0.8138     0.4853    1.68    0.094
subssand       0.4210     0.4951    0.85    0.395
depth          0.9547     0.1626    5.87   <0.001
duck           0.0512     0.0417    1.23    0.220
unio           0.0387     0.0531    0.73    0.466

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 176.40  on 94  degrees of freedom
Residual deviance:  96.49  on 89  degrees of freedom
AIC: 251
```

For Poisson models there is no true R^2 for the model. Instead we can calculate the explained deviance (sometimes called the pseudo-R^2). This is calculated as: 100 x (null deviance-residual deviance) / null deviance; i.e. 100 x (176.40 - 96.49) / 176.40 = 45.3% of the variation in the number of swan mussels.

The Akaike Information Criterion (AIC) is 251. The AIC is useful for comparing models with different combinations of covariates, for instance if we wish to carry out model selection.

However, before we attempt to interpret this model further, we must first carry out model validation.

3.4 Model validation

For the fitted Poisson GLM, validation is required to look for:

1. Overdispersion.
2. Model misfit.

3.4.1 Overdispersion

Poisson GLMs assume that the mean and variance of the response variable increase at the same rate (see the model summary output above and the statement `Dispersion parameter for poisson family taken to be 1`). This assumption must be confirmed. If the residual deviance of the fitted model is bigger than the residual degrees of freedom, then we have overdispersion. Overdispersion means that a Poisson distribution does not adequately model the variance and is not appropriate for the analysis.

The overdispersion statistic can be calculated with the following R code:

```
> ods <- Pois1$deviance / Pois1$df.residual
> ods

1.11
```

A value of 1.11 indicates mild overdispersion and in this case is acceptable. Values exceeding 1.2 are problematic. In Chapter 4 we explain the approach to take if a Poisson GLM shows severe overdispersion.

3.4.2 Model misfit

As with a Gaussian GLM, model misfit in a Poisson GLM is recognised by plotting Pearson residuals against fitted values, against each covariate in the model, as well as any not included in the model (in this case we included all variables in the model). The presence of influential observations can be tested by plotting Cook's distance.

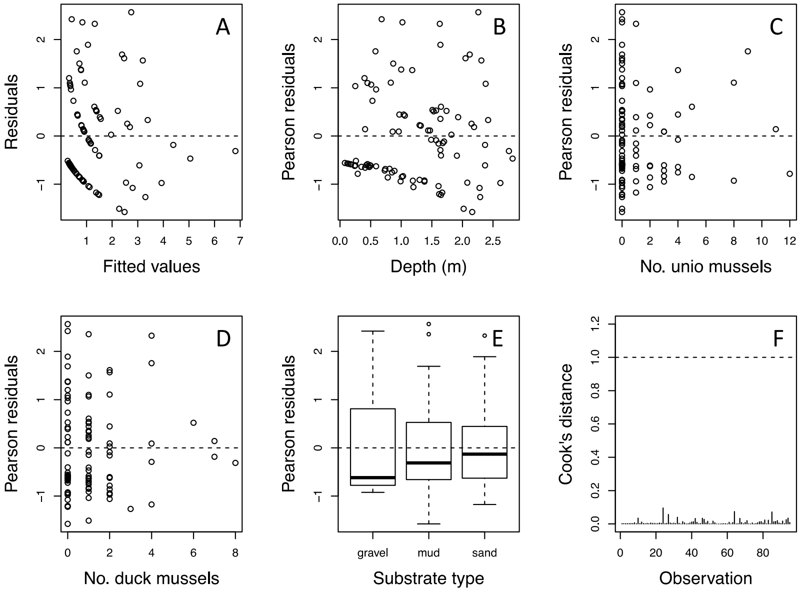

Fig. 3.7 A. Pearson residuals plotted against fitted values; B. Pearson residuals against depth; C. Number of unio mussels; D. Number of duck mussels, E. Substrate type. F. Cook's distance values for model Pois1.

Plots A-E in Fig. 3.7 show no causes for concern; residuals are distributed consistently along the horizontal axis in each case and there are no obvious patterns in the residuals. There is also no evidence from plotting Cook's distance (Fig. 3.7F) of influential observations in the model.

3.4.3 Simulating from the data

During data exploration it was observed that 40% of quadrats contained no swan mussels, and this was raised as a potential problem. As part of model validation, we can simulate data from the model and compare with the observed data to see if the number of zeros in simulated datasets matches the 40% of zeros observed.

Start by simulating 10,000 datasets using the parameters of model Pois1:

```
> Nmuss <- nrow(muss)
> Fitted<- fitted(Pois1)
> Ysim  <- matrix(nrow = Nmuss, ncol = 10000)
> Zeros <- vector(length = 10000)
> for(i in 1:10000){
  Ysim[,i] <- rpois(Nmuss, lambda = Fitted)
  Zeros[i] <- sum(Ysim[,i] == 0) / Nmuss}
```

These data are then plotted as a frequency histogram:

```
> par(mar = c(5,5,2,2), cex.lab = 1.5, mfrow = c(1,1))
> plot(table(Zeros),
     axes = FALSE,
     xlab = "Percentage of zeros",
     ylab = "Frequency",
     xlim = c(0.2, 0.6),
     ylim = c(0, 1000))
> axis(2)
> axis(1, at = c(0.2, 0.3, 0.4, 0.5, 0.6),
     labels = c("20%", "30%", "40%", "50%", "60%"))
```

Finally, the percentage of zeros in the observed data are plotted as a black diamond to indicate where in the distribution the observed data lie.

```
> points(x = sum(muss$swan == 0) / Nmuss, y = 30,
       pch = 18, cex = 5, col = 1)
```

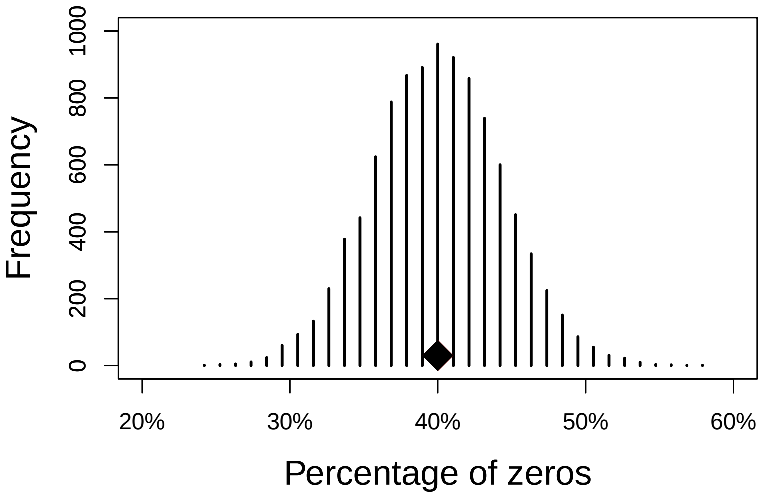

Fig. 3.8 Frequency histogram of the percentage of quadrats with no swan mussels in 10,000 simulated datasets. The black diamond is the percentage of quadrats without swan mussels in the observed data.

The number of zeros in simulated datasets corresponds well with what was observed during mussel surveys. This outcome gives us confidence that the Poisson GLM is reliably recreating a comparable pattern of data to that observed.

3.5 Model presentation

Specify model Pois1 using mathematical notation in the following way:

$$Swan_i \sim Poisson(\mu_i)$$

$$E(Swan_i) = \mu_i \text{ and } \mathrm{var}(Swan_i) = \mu_i$$

$$\log(\mu_i) = \eta_i$$

$$\eta_i = \beta_1 + \beta_2 \times Substrate_i + \beta_3 \times Depth_i + \beta_3 \times Duck_i + \beta_4 \times Unio_i$$

Where *Swan*$_i$ is the number of swan mussels in quadrat *i* assuming a Poisson distribution with mean and variance μ_i. *Depth*$_i$ is a continuous covariate corresponding with water depth of quadrat *i* (m) and *Substrate*$_i$ is a categorical covariate with three levels (gravel, sand, mud). *Duck*$_i$ is a continuous covariate corresponding with the number of duck mussels in quadrat *i* and *Unio$_i$* is the total number of painter's and swollen mussels in quadrat *i*.

The numerical output of the model is obtained with:

```
> summary(Pois1)

              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.5367     0.3987   -3.85   0.00012
subsmud       0.2934     0.3648    0.80   0.42127
subssand     -0.0440     0.3780   -0.12   0.90732
depth         0.9971     0.1668    5.98   2.3e-09
duck          0.0513     0.0419    1.22   0.22099
unio          0.0304     0.0545    0.56   0.57720
```

These results can be more formally presented in the following way:

Table 3.1. Summary of Poisson GLM to model the number of swan mussels (*Anodonta cygnea*) collected in 1 m^2 quadrats in a lake in the River Danube basin.

Model parameter	Estimate	SE	*P*
Intercept$_{(gravel)}$	-1.54	0.40	<0.001
Substrate$_{(mud)}$	0.29	0.36	0.421
Substrate$_{(sand)}$	-0.04	0.38	0.907
Depth	1.00	0.17	<0.001
Duck	0.05	0.04	0.221
Unio	0.03	0.05	0.577

Some covariates are non-significant and appear redundant in the model. Should we proceed with model selection and find an optimal model? Model selection in ecology is a contentious issue and, for now, we choose to leave the model as it was formulated to address the original model predictions.

The model can be visualized using `ggplot2`. The code for this plot is available in the R file associated with this chapter.

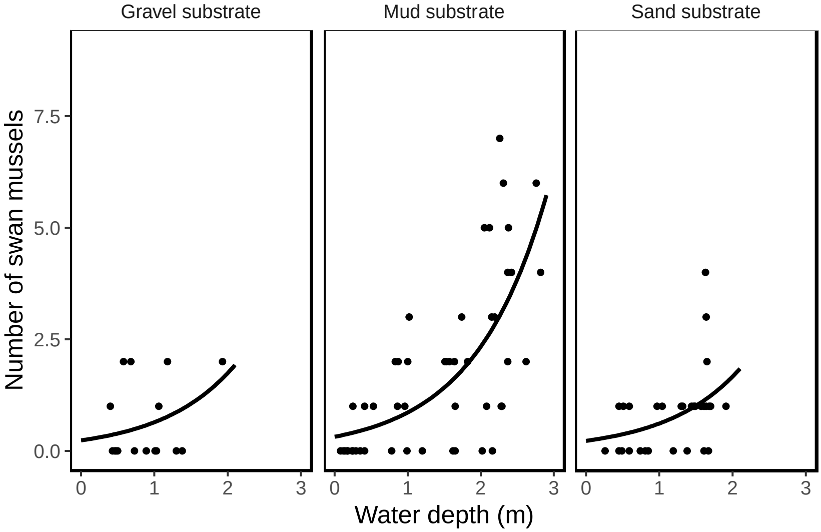

Fig. 3.9 Mean fitted number of swan mussels (*Anodonta cygnea*) (solid line) with 95% confidence intervals (shaded area) against water depth (m) on three substrate types (gravel, mud and sand). Black dots are observed data.

Conclusions

On the basis of Smith *et al.* (2000) it was predicted that the abundance of swan mussels would be positively associated with water depth and a mud substrate, negatively with duck mussel abundance, and with no relationship with *Unio* sp. mussels. The Poisson GLM fitted to these data supported the prediction for a relationship with depth, but no significant association, after controlling for the effects of depth, was demonstrated for a mud substrate. There was no support for the predicted relationship with duck mussels. As predicted, the abundance of swan mussels appeared unaffected by the abundance of *Unio* sp. mussels.

Reference

Smith, C., Reynolds, J.D., Sutherland, W.J. & Jurajda, P., 2000. The population consequences of reproductive decisions. *Proceedings of the Royal Society, London. B*. 267, 1327-1334.

4 Negative binomial GLM

A negative binomial GLM is used for the same type of ecological data that a Poisson GLM would be used to analyse; count data that does not take values below zero. However, the negative binomial GLM does not assume that the variance of the response variable is equal to its mean and, therefore, can be used to model overdispersed data (see 3.4.1), which is a common property of ecological data. Formulation of a negative binomial GLM is slightly more complex than a Poisson GLM, and a negative binomial GLM is used when a Poisson GLM is not appropriate due to overdispersion.

4.1 Species diversity of chironomids

Chironomids are a taxonomically diverse family of non-biting flies with a global distribution in freshwaters. They are capable of adapting to a wide range of environmental conditions and play a key ecological role in cycling organic matter.

A study was conducted by Leszczyńska *et al.* (2019) to analyse the structure of chironomid assemblages and identify the environmental factors that underpin variation in chironomid species richness across a set of lowland rivers. The aim of the study was to identify which environmental variables predicted chironomid species richness. Chironomid samples were collected from fourteen study sites in seven lowland rivers in central Poland. On each sampling occasion samples were collected in different months, with a total of 82 samples collected in total.

The data collected by Leszczyńska *et al.* (2019) include river name and month of sample collection. At each sampling point benthic samples containing invertebrates and particulate organic and inorganic matter were collected and the current velocity (m s^{-1}), river width (m), water depth (m), water temperature (°C), and dissolved oxygen (mg l^{-1}) were also recorded. Benthic samples were transferred to the laboratory and invertebrates were sorted from benthic sediment by hand. All chironomids in samples were identified to species level and counted. The organic content of samples was determined as benthic particulate organic matter (BPOM) (g m^{-2}). The quantity of inorganic substrate was estimated as substrate inorganic index (SI).

The number of chironomid species in each sample is the response variable, and comprises a species count that is bounded at zero. River and month are categorical variables while all the other covariates are continuous.

4.2 Data exploration

Import data

Data for chironomids are saved in the tab-delimited file `rivchir.txt` and are imported into a dataframe in R using the command:

```
> rivchir <- read.table(file = " rivchir.txt",
                        header = TRUE, dec = ".")
```

Start by inspecting the dataframe:

```
> str(rivchir)

'data.frame':     82 obs. of  7 variables:
 $ river: Factor w/ 7 levels "bzur","grab", ...
 $ vel  : num  0.61 0.61 0.28 0.26 0.28 0.33 0.31 ...
 $ si   : num  5.5 20.9 3.8 24.8 3.9 21.6 4 19.3 ...
 $ bpom : int  1700 260 2000 500 1800 400 1800 450 ...
 $ temp : num  13 13 18 21 5 0 17 19 18 18 ...
 $ oxy  : num  7.7 7.9 5 5.9 9.9 10.2 5.8 6.5 6.5 ...
 $ taxa : int  17 23 28 22 21 21 25 21 23 22 ...
```

The dataframe comprises 82 observations of 7 variables. Each row in the dataframe represents a sample collected from a different river in a different month. River (`river`) is a factor; i.e. a categorical variable. River velocity (`vel`), inorganic substrate index (`si`), benthic particulate organic matter (`bpom`), water temperature (`temp`), dissolved oxygen concentration (`oxy`), and number of chironomid species (`taxa`) are all continuous covariates.

It is necessary to check if there are any missing values in the dataframe (missing values are designated 'NA' in the tab-delimited file.

```
>  colSums(is.na(rivchir))

river   vel    si  bpom  temp   oxy  taxa
    0     0     1     0     0     1     0
```

A small number of missing values - these must be removed.

Is the categorical covariate river balanced?

```
> table(rivchir$river)

bzur grab mosz mrog mroz wart wida
  12   12   12   12   12   10   12
```

The data are well balanced.

4.2.1 Outliers

Outliers in the data can be identified visually using Cleveland dotplots.

```
> Var <- c("vel", "si", "bpom", "temp", "oxy", "taxa")
> dotplot(as.matrix(as.matrix(rivchir[,Var])),
      groups=FALSE,
      strip = strip.custom(bg = 'white',
      par.strip.text = list(cex = 1.2)),
      scales = list(x = list(relation = "free",
      draw = TRUE),
      y = list(relation = "free", draw = FALSE)),
      col=1, cex  = 1.0, pch = 16,
      xlab = list(label = "Data range", cex = 1.2),
      ylab = list(label = "Data order", cex = 1.2))
```

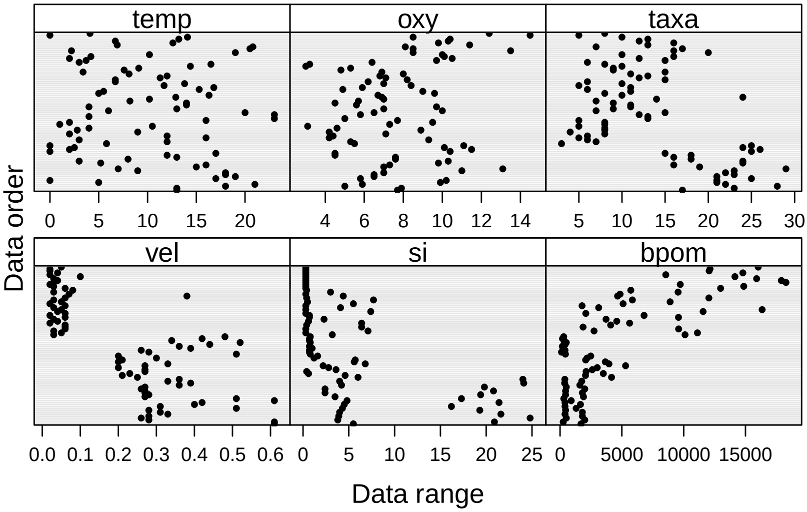

Fig. 4.1 Dotplots of river velocity (vel), inorganic substrate index (si), benthic particulate organic matter (bpom), water temperature (temp), dissolved oxygen concentration (oxy), and number of chironomid species (taxa). Data are arranged by the order they appear in the dataframe.

There are no prominent outliers in these dotplots.

4.2.2 Lots of zeros in the response variable

The number of zeros in the response variable can be estimated as:

```
> sum(rivchir$taxa == 0)
```

```
0
```

No zeros in the response variable; chironomids were found in every sample.

4.2.3 Multicollinearity among covariates

Use a correlation matrix with corresponding pairplots to visualize pairwise correlations. Code for this plot is shown in the R file associated with this chapter.

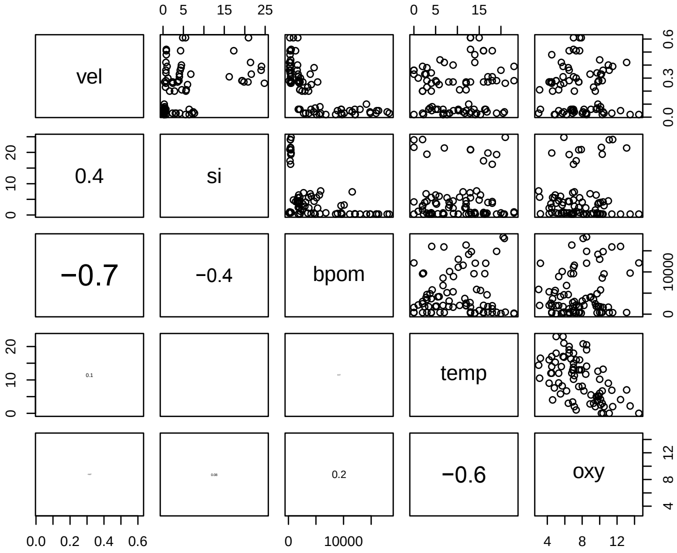

Fig. 4.2 Pairplot of covariates. The lower panel shows pairwise Pearson correlations, with font size proportional to correlation coefficient.

Some covariates appear mildly collinear from the pairplots in Fig. 4.2. Velocity (vel) is negatively collinear with benthic particulate organic matter (bpom) and water temperature is negatively collinear with dissolved oxygen. Degree of collinearity can be measured by calculating the variance inflation factors for each covariate using `vif`. For now, we will assume a Poisson GLM is appropriate for these data.

```
> vif(glm(taxa ~ vel + si + bpom + temp + oxy,
                 family = poisson,
                 data = rivchir))
```

```
vel   si    bpom  temp  oxy
2.24  1.48  2.75  1.94  2.10
```

VIF values all <3.

4.2.4 Relationships among dependent and independent variables

Visual inspection of the data using plots. The code for this plot is available in the R file associated with this chapter.

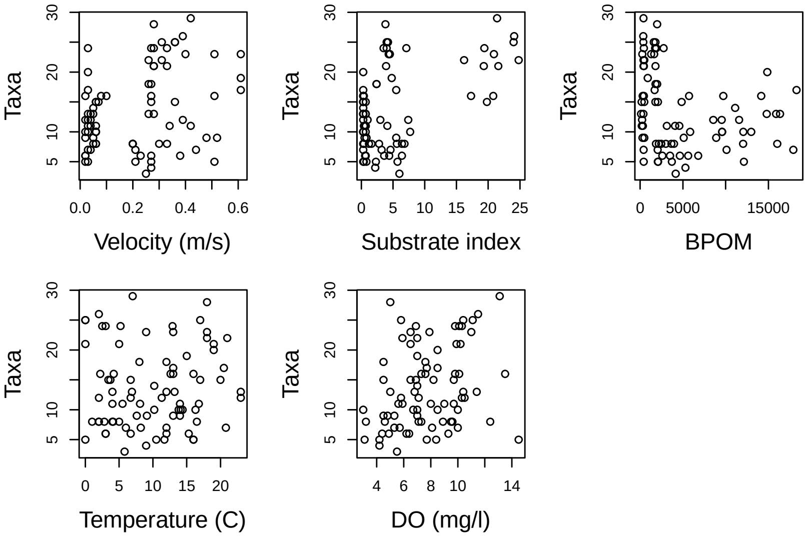

Fig. 4.3 Plots of number of chironomid taxa in benthic samples against covariates.

In Fig. 4.3 the plots of the number of chironomid taxa against covariates show no obvious patterns, with the exception of substrate index, which shows a distinctly positive relationship.

4.3 Model fitting

The data exploration showed:

1. A small number of NAs.
2. No outliers in the data.
3. No zeros in the response variable.
4. No imbalance of data among levels of the categorical covariate 'river'
5. No important collinearity between covariates.
6. Potential relationship between substrate index and number of chironomid taxa.

Before fitting a GLM, NAs must be dropped, which means the loss of a small amount of data. The categorical covariate 'river' is of no specific interest and, therefore, will not be included in the model; this decision will be discussed further at the end of the chapter. Initially a Poisson GLM will be applied to the data.

Remove NAs with

```
> rivchir1 <- rivchir[complete.cases(rivchir), ]

> dim(rivchir)

82 7

> dim(rivchir1)

80 7
```

Two rows of data have been lost.

The model is fitted as:

```
> Pois1 <- glm(taxa ~ vel + temp + si + oxy + bpom,
                      data = rivchir1,
                      family = poisson(link = log))
```

The numerical output is obtained with the `summary` function:

```
> summary (Pois1)

             Estimate Std. Error  z value Pr(>|z|)
(Intercept)  1.416e+00  1.747e-01   8.108  5.16e-16
vel          1.010e-01  2.569e-01   0.393  0.69417
temp         3.009e-02  6.422e-03   4.684  2.81e-06
si           1.730e-02  4.423e-03   3.912  9.16e-05
oxy          1.124e-01  1.732e-02   6.488  8.68e-11
bpom        -2.658e-05  1.005e-05  -2.645  0.00816
(Dispersion parameter for poisson family taken to be 1)

Null deviance: 262.62  on 79 degrees of freedom
Residual deviance: 149.68 on 74 degrees of freedom
AIC: 514
```

Before interpreting the model, we must first carry out model validation.

4.4 Model validation

For the fitted Poisson GLM, validation is required to look for:

1. Overdispersion.
2. Model misfit.

4.4.1 Overdispersion

The overdispersion statistic is calculated with:

```
> ods <- Pois1$deviance / Pois1$df.residual
> ods

2.02
```

The overdispersion statistic should take a value of 1.0. A value of 2.09 is too high; the model is overdispersed.

Overdispersion

Poisson GLMs assume the mean and variance of the response variable are approximately equal. Overdispersion can occur when this assumption is not met; variance in the data is naturally larger than the mean. This situation is termed "true overdispersion". True overdispersion is dealt with by fitting a model to the data such that the variance is greater than the mean in the response variable.

However, before we assume true overdispersion, we should consider other possible causes, which can represent underlying problems with the model. These are:

1. **Model mis-specification**. There may be key variables, including interactions, that explain a large part of the variance that are missing from the model. Model mis-specification is handled by including additional variables or adding interaction terms to the model.
2. **Too many zeros in the response variable ("zero inflation")**. If there are too many zeros a zero-inflated (e.g. a zero-inflated Poisson or ZIP model) or zero-adjusted (e.g. a zero- adjusted Poisson or ZAP) model can be used.
3. **Influential outliers**. The presence of influential observations can be tested by plotting Cook's distance and these can be dropped and the model refitted. Data dropped from the analysis must be reported in your Methods, with a justification.
4. **Non-independence of the data**. An assumption is that each observation in a dataset is independent of all others. However, there may be an underlying association between some data that results in dependency; e.g. data may have been collected by different scientists, who introduce consistent bias to the data, or data may have been collected in different months, which affects the variance structure of the data. If the source of dependency is known, it can be incorporated into the analysis as a "random" term in a Generalized Linear Mixed Model (GLMM).
5. **Wrong link function**. A GLM uses a link function to connect the response variable with the linear part of the model comprising the covariates. Trying an alternative link function to the default may solve the problem of overdispersion.
6. **Non-linearity in the data**. A GLM assumes the response variable can be modelled as a linear relationship using a link function. However, this approach may not be adequate to capture the non-linear properties of some biological systems. In this case it is necessary to switch to using Generalized Additive Models (GAMs).

As part of model validation, it is necessary to address each of these potential problems. If none prove successful in reducing overdispersion, a model with a different error structure can be applied.

1. Model mis-specification. Without additional variables to use in the model, the only option is to refit the model with interactions. Interactions must be biologically plausible; it is not satisfactory to try every possible combination of interactions between model covariates. Two plausible interactions in the case of these data are between water velocity (vel) and substrate index (si); current speed could influence the quantity of inorganic substrate, with implications for the chironomid community. A second plausible interaction is between temperature (temp) and dissolved oxygen (oxy); water temperature correlates negatively with dissolved oxygen concentration, and chironomids are adapted to low oxygen conditions.

The alternative model, then, is:

```
> Pois2 <- glm(taxa ~ vel * si + temp * oxy + bpom,
                      data = rivchir1,
                      family = poisson(link = log))

> ods2 <- Pois2$deviance / Pois1$df.residual
> ods2

2.08
```

The alternative model is still overdispersed.

2. Zero inflation. How many zeros in the response variable?

```
> sum(rivchir$taxa == 0)

0
```

Zero inflation is not the problem.

3. Influential outliers. Plot Cook's distance to identify influential observations.

```
> par(mfrow=c(1,1), mar=c(5,5,2,2))
> plot(cooks.distance(Pois1),
       xlab = "Observation",
       ylab = "Cook's distance",
        type = "h",
        ylim = c(0, 1.2),
        cex.lab =  1.5)
> abline(h = 1, lty = 2)
```

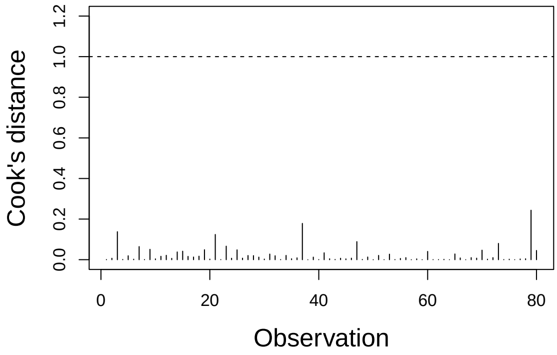

Fig. 4.4 Plot of Cook's distance for model Pois1. A Cook's distance of 1 (indicated by a dashed horizontal line) denotes an influential observation.

There is no evidence from plotting Cook's distance (Fig. 4.4) of influential observations in the model.

4. Non-independence of the data. A variable that we have hitherto ignored is the river from which samples were collected. The numbers of chironomid species in samples from the same river may be more similar to each other than they are to samples from different rivers. If the case, the assumption of independence may be violated. To investigate potential dependency in the data we can plot the numbers of chironomid species for each river. If dependency is not a problem the expectation is that the mean and variance in the number of species from samples from different rivers should be similar.

```
> par(mfrow=c(1,1), mar=c(5,5,2,2))
> boxplot(taxa ~ river,
          data = rivchir1,
          xlab = "River",
          ylab = "Number of taxa",
          cex.lab = 1.5)
```

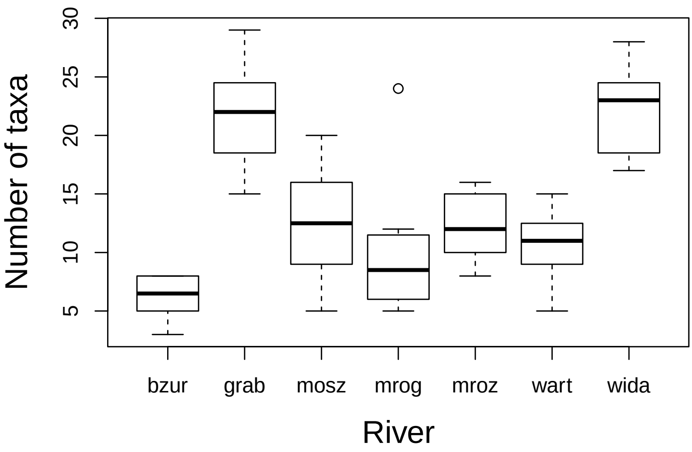

Fig. 4.5 Boxplot of number of chironomid taxa in benthic samples for river from which samples were collected.

Fig. 4.5 shows there is good evidence for dependency in the response variable; the numbers of chironomid species vary strongly among rivers. This is a potential cause of overdispersion.

5. Wrong link function. Models can be fitted with alternative link functions and the overdispersion statistic calculated to see whether there is an improvement. Two alternative link functions to a log link are an 'identity' link, which assumes a linear relationship between the response variable and covariates and a square-root link.

Identity link

```
> Pois3 <- glm(taxa ~ vel + temp + si + oxy + bpom,
               data = rivchirl,
               family = poisson(link = identity))
> ods3 <- Pois3$deviance / Pois3$df.residual
> ods3

2.00
```

An identity link does not prevent overdispersion.

Square-root link

```
> Pois4 <- glm(taxa ~ vel + temp + si + oxy + bpom,
               data = rivchir1,
               family = poisson(link = sqrt))
> ods4 <- Pois3$deviance / Pois3$df.residual
> ods4

2.00
```

And neither does a square-root link.

6. Non-linearity in the data. Non-linearities in the data can be identified by plotting the Pearson residuals of the model against each covariate and fitting a 'loess' regression through the data. A loess regression (short for 'local regression') fits a smoothed curve and is ideal for highlighting non-linear patterns. Code for these plots is available in the file associated with this chapter.

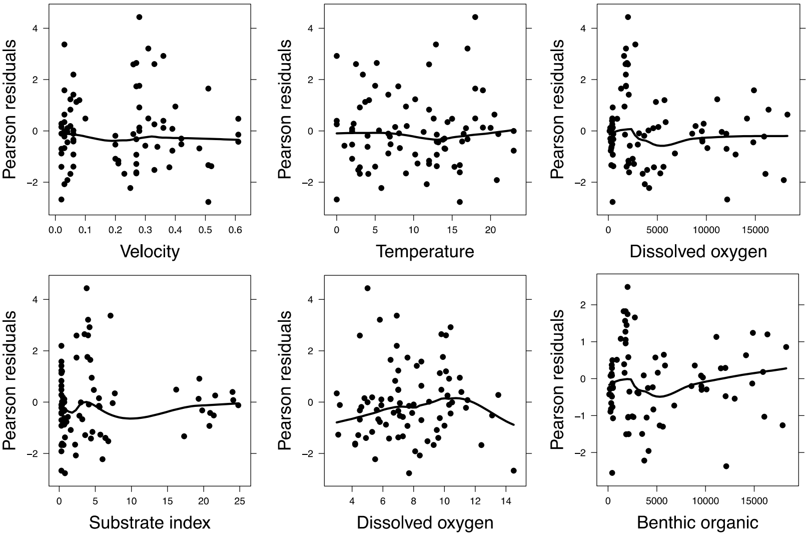

Fig. 4.6 Plots of number of chironomid taxa in benthic samples against Pearson residuals for model covariates. A loess smoother is added to aid visual interpretation.

There is no evidence from residual plots (Fig. 4.6) for non-linear patterns.

This analysis points to two sources of overdispersion: dependency due to river effects, and true overdispersion. Dependency can be addressed by fitting a GLMM and including river as a random term in the model. This approach is certainly needed here but is beyond the scope of this book. Instead, we will assume that there is true overdispersion in the data; i.e. variance in the data is naturally larger than the mean and a model with a different error structure is needed.

The model will now be fitted with a negative binomial distribution for the response variable using the `glm.nb()` function from the `MASS` package. The default link function is a log link.

```
> library(MASS)
> nb1 <- glm.nb(taxa ~ vel + temp + si + oxy + bpom,
                         data = rivchir1)
```

Assess overdispersion with:

```
> ods_nb <- nb1$deviance / nb1$df.residual
> ods_nb

1.11
```

The overdispersion statistic should take a value of 1.0. A value of 1.11 indicates mild overdispersion but is acceptable.

Before interpreting the model, we must first continue with model validation. The fitted negative binomial GLM is not overdispersed, but it is still necessary to examine model misfit.

As with a Poisson GLM, model misfit in a negative binomial GLM is recognized by plotting Pearson residuals against fitted values, against each covariate in the model, as well as any not included in the model (in this case we included all variables in the model) and the presence of influential observations is tested by plotting Cook's distance.

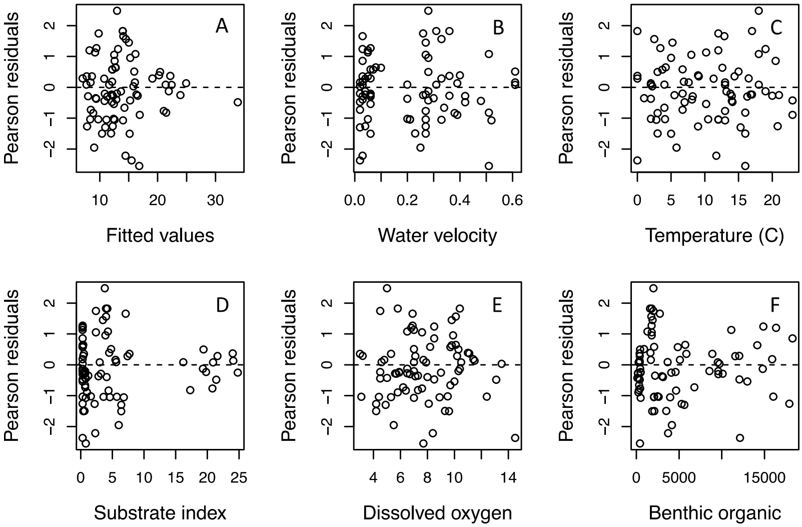

Fig. 4.7 Pearson residuals plotted against: A. Fitted values; B. Water velocity; C. Temperature (C); D. Substrate index, E. Dissolved oxygen (mg l^{-1}). F. Benthic particulate organic matter (BPOM).

Plots A-F in Fig. 4.7 show no causes for concern; residuals are distributed along the horizontal axis in each case and there are no obvious patterns in the residuals.

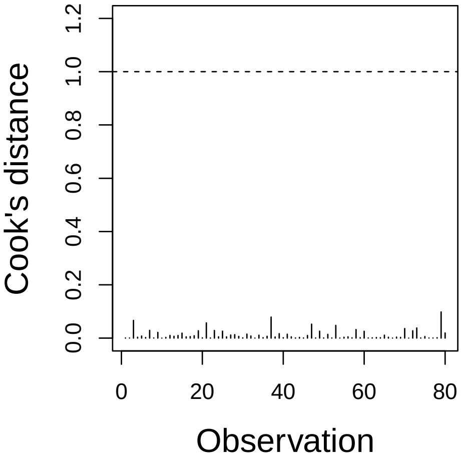

Fig. 4.8 Plot of Cook's distance for model nb1. A Cook's distance of 1 (indicated by a dashed horizontal line) denotes an influential observation.

There is no evidence from plotting Cook's distance (Fig. 4.8) of influential observations in the model.

Model comparison

It is possible to compare the performance of the GLM with Poisson error structure with the negative binomial model. This comparison can be made using the AIC (Akaike Information Criterion). AIC gives a measure of the goodness of fit of a model by "log likelihood".

The more parameters a model has, the better the fit to the data. To compensate for the inevitably better fit of models with many parameters, AIC imposes a penalty on a model as a function of the number of parameters in the model. For this reason, AIC is sometimes termed the "penalized log-likelihood". There are alternatives to AIC for measuring goodness of fit, though AIC is reliable and widely recognised.

AIC can be calculated for the two models with:

```
> AIC(Pois1, nb1)

      df   AIC
Pois1  6   508.9
nb1    7   492.5
```

The lower the AIC, the better the model; if a model has an AIC value of 2 or more lower than its rival, it is considered the better fitting model. In this case the negative binomial model (`nb1`) gives a better fit to the data. The absolute value of AIC is meaningless. Note that the negative binomial model has one more parameter than the Poisson model. This is because the negative binomial model has a dispersion parameter (*k*) that accommodates higher variance in the data, but is penalised when calculating AIC. Thus, despite having one more parameter than the Poisson model, the negative binomial model is still an improvement.

Two models with different numbers of covariates or different distributions can be compared using AIC, but they must have the same number of observations.

4.5 Model presentation

Model nb1 is specified using mathematical notation in the following way:

$$
\begin{aligned}
&Taxa_i \sim NegBin(\mu_i\,,k)\\
&E(Taxa_i) = \mu_i \;\text{ and }\; \mathrm{var}(Taxa_i) = \mu_i + (\mu_i^2 \,/\, k\,)\\
&\log(\mu_i\,) = \eta_i\\
&\eta_i = \beta_1 + \beta_2 \text{ x } Velocity_i + \beta_3 \text{ x } Temperature_i + \beta_3 \text{ x } SI_i +\\
&\qquad \beta_4 \text{ x } DO_i + \beta_5 \text{ x } BPOM_i
\end{aligned}
$$

Where $Taxa_i$ is the number of chironomid species in sample *i* assuming a negative binomial distribution with mean μ_i and variance $\mu_i + (\mu_i^2 / k\,)$. The extra parameter *k* is known as the dispersion parameter and deals with the extra variance in the data. For the model covariates $Velocity_i$ is water velocity for sample *i*, $Temperature_i$ is water temperature for sample *i*, SI_i is sample substrate index of sample *i*, DO_i is dissolved oxygen concentration for sample *i*, and $BPOM_i$ benthic particulate organic matter of sample *i*.

The numerical output of model nb1 is obtained with:

```
> summary(nb1)
```

```
              Estimate  Std. Error z value Pr(>|z|)
(Intercept)  1.360e+00  2.347e-01   5.794 6.89e-09
vel          5.946e-02  3.590e-01   0.166 0.868454
temp         3.261e-02  8.858e-03   3.681 0.000232
si           1.747e-02  6.514e-03   2.682 0.007320
oxy          1.180e-01  2.351e-02   5.020 5.17e-07
bpom        -2.770e-05  1.362e-05  -2.033 0.042024
```

These results can be more formally presented in the following way:

Table 4.1. Summary of negative binomial GLM to model the number of chironomid taxa collected in substrate samples.

Model parameter	Estimate	SE	*P*
Intercept(gravel)	1.36	0.23	<0.001
Velocity	0.06	0.36	0.868
Temperature	0.03	0.01	<0.001
Substrate index	0.02	0.01	0.007
Dissolved oxygen	0.11	0.02	<0.001
BPOM	-0.01	0.01	0.042

Water velocity is non-significant in the model. We choose to leave the model unchanged.

The model can be visualized using `ggplot2`. R code for generating the figure is available in the R code that accompanies this chapter.

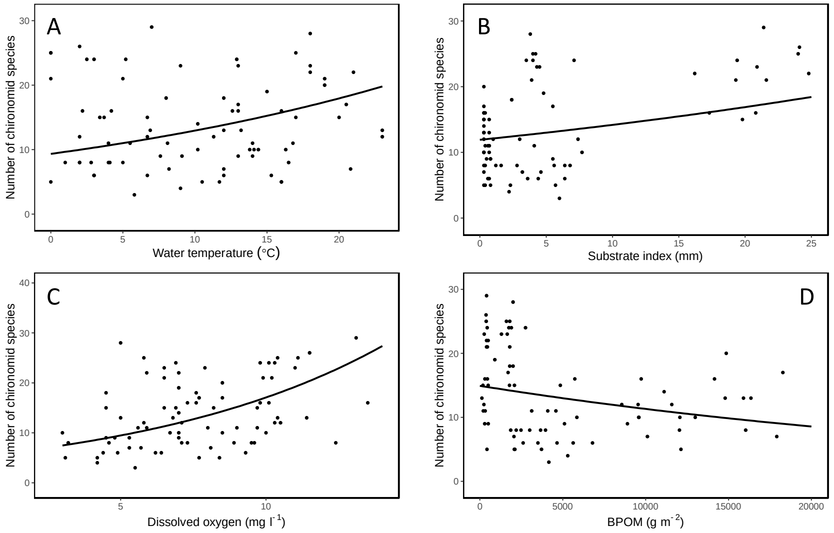

Fig. 4.8 Mean fitted number of chironomid species (solid line) with 95% confidence intervals (shaded area) against: A. water temperature (°C); B. substrate index (mm); C. dissolved oxygen (mg l^{-1}); D benthic particulate organic matter (BPOM) (g m^{-2}). Black dots are observed data.

Conclusions

The final model showed that the number of chironomid species in benthic river samples was positively associated with water temperature, inorganic substrate and dissolved oxygen, but negatively with organic matter (Fig. 2.8). There was no significant association with water velocity (Table 4.1).

Overdispersion in the Poisson model was treated as true overdispersion, with a model fitted with a negative binomial distribution controlling overdispersion. The goodness of fit of the negative binomial model, measured by AIC, was also superior to the Poisson model.

However, the negative binomial model is not optimum. As part of model validation, it was clear that there was dependency in the data due to river (Fig. 4.5). The next step in modelling these data will be to fit a GLMM to accommodate dependency in the data due to river.

Reference

Leszczyńska, J., Grzybkowska, M., Głowacki, Ł. & Dukowska, M., 2019. Environmental variables influencing chironomid assemblages (Diptera: Chironomidae) in lowland rivers of central Poland. *Environmental Entomology* in press.

5 Bernoulli GLM

A Bernoulli distribution is a discrete distribution for dealing with data with two possible outcomes such as success or failure and presence or absence. The Bernoulli GLM is for strictly binary data and is sometimes called a logistic GLM (or just "logistic regression"). In ecological studies a Bernoulli GLM is a useful tool for modelling presence/absence data.

5.1 The presence of red spots on pumpkinseed fish

In the pumpkinseed fish (*Lepomis gibbosus*) some individuals have a conspicuous red spot on their gill cover (operculum) that has been associated with behavioural dominance. Zięba *et al*. (2018) investigated the function of the red spot in populations of pumpkinseed collected from sites across Europe where the species is invasive.

Male pumpkinseed display alternative mating strategies. Some males are large-bodied and territorial. These males build nests, court females and care for the eggs that are laid in their nest. However, some males perform a 'sneaky' mating strategy, entering the nest of a territorial during spawning and fertilising eggs laid by a female courted by the nest-guarding male, with the territorial male subsequently caring for eggs and young stages; sneaker males perform no parental care.

The aim of the study was to determine whether the presence of the red operculum spot functions as a signal of sex and/or mating strategy in pumpkinseed. To do this males were categorised as territorials or sneakers and a model was fitted to test whether the probability of possessing a red spot differed between the sexes and between males adopting different reproductive strategies. The prediction was that parental males would be more likely to express a red operculum spot than sneaker males and females. However, because larger fish tend to older, the analysis needed to control for body size while simultaneously comparing the probability of red spots among individuals of different sexes/mating strategies.

The data collected by Zięba *et al*. (2018) include individual fish mating strategy (female, male territorial, male sneaker), fish length (mm), fish weight (g), and presence of a red spot. Sex and mating strategy was assigned by dissection of the gonads (see Zięba *et al*. 2018 for details). Presence of a red spot is the response variable, and the other variables are covariates; mating strategy is a categorical variable and length and weight are continuous.

5.2 Data exploration

Import data

Data for pumpkinseed are saved in the tab-delimited file `pumpkin.txt` and are imported into a dataframe in R using the command:

```
> pkin <- read.table(file = " pumpkin.txt",
                     header = TRUE, dec = ".")
```

Start by inspecting the dataframe:

```
> str(pkin)

'data.frame':     900 obs. of  6 variables:
 $ pop   : Factor w/ 14 levels "6T","BF","BP",...
 $ sex   : Factor w/ 2 levels "F","M" ,...
 $ wt    : num  6.3 8.4 6.9 8.4 9 10.1 10.7...
 $ sl    : num  60.2 64.6 64.8 66.3 70.5 74.2 74.4 ...
 $ tactic: Factor w/ 3 levels "fem","sneak",...
 $ spot  : int  0 0 0 0 0 0 0 0 0 0 ...
```

The dataframe comprises 900 observations of 6 variables. Each row in the dataframe represents an individual pumpkinseed fish collected from a different population. Population (`pop`), sex (`sex`) and mating tactic (`tactic`) are all factors; i.e. categorical variables. Fish weight (`wt`), and length (`sl`) are continuous covariates. The presence of a red spot (`spot`) is binomial and the data are coded as 0 (red operculum spot absent) and 1 (red spot present).

It is necessary to check if there are any missing values in the dataframe (missing values are designated 'NA' in the tab-delimited file.

```
>  colSums(is.na(pkin))

pop  sex  wt   sl  tactic  spot
0    0    0    0   0       0
```

No missing values.

5.2.1 Outliers

Outliers in the data can be identified visually using Cleveland dotplots:

```
> Var <- c("sl", "wt")
> dotplot(as.matrix(as.matrix(pkin[,Var])),
        groups=FALSE,
        strip = strip.custom(bg = 'white',
        par.strip.text = list(cex = 1.2)),
        scales = list(x = list(relation = "free",
        draw = TRUE),
        y = list(relation = "free", draw = FALSE)),
col = 1, cex  = 0.5, pch = 16,
xlab = list(label = "Data range", cex = 1.5),
ylab = list(label = "Data order", cex = 1.5))
```

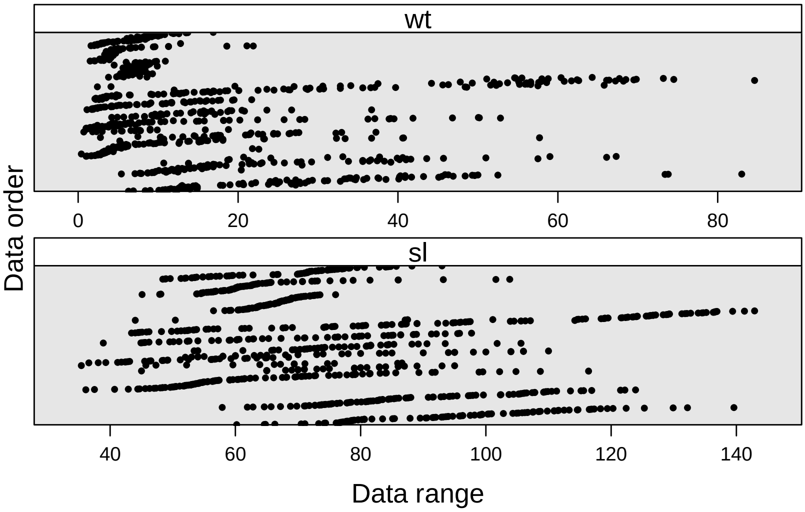

Fig. 5.1 Dotplots of pumpkinseed weight (wt) and length (sl). Data are arranged by the order they appear in the dataframe.

There are no obvious outliers in the data (Fig. 5.1). Are the data balanced between the sex of fish?

```
> table(pkin$sex)

  F   M
425 475
```

Or mating tactic?

```
> table(pkin$tactic)
```

```
fem  sneak  terr
425   95    380
```

The data are well balanced between sexes (`sex`), but less well balanced among mating tactics (`tactic`). However, if data are a random sample from the population, then a lack of balance is inevitable. However, care must be taken in fitting a complex model to these data.

An additional check is to look at a dotplot for weight and length split by sex and mating tactic.

For sex:

```
> par(mfrow = c(1,2), mar = c(5,5,1,1), cex.lab = 1.2)
> dotchart(pkin$sl, groups = pkin$sex,
  xlab = "Length (mm)")
> dotchart(pkin$wt, groups = pkin$sex,
  xlab = "Weight (g)")
```

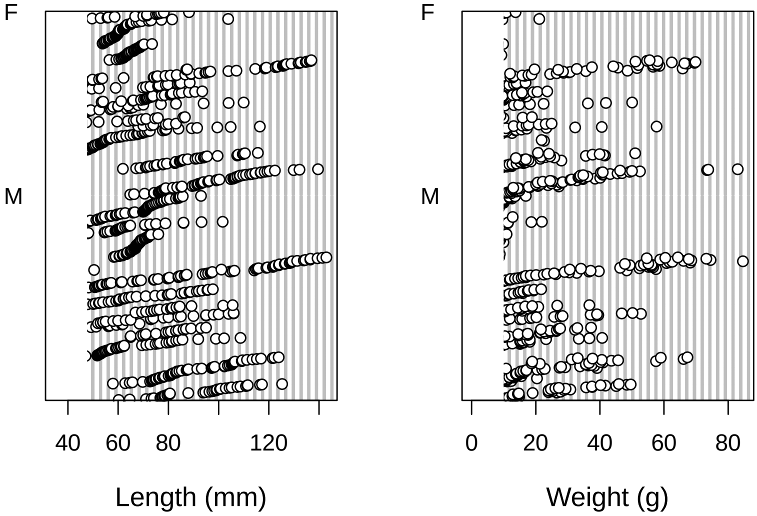

Fig. 5.2 Dotplots of pumpkinseed weight (wt) and length (sl) split by sex of fish. Data are arranged by the order they appear in the dataframe.

And mating tactic:

```
> par(mfrow = c(1,2), mar = c(5,5,1,1), cex.lab = 1.2)
> dotchart(pkin$sl, groups = pkin$tactic,
  xlab = "Length (mm)")
```

```
> dotchart(pkin$wt, groups = pkin$tactic,
  xlab = "Weight (g)")
```

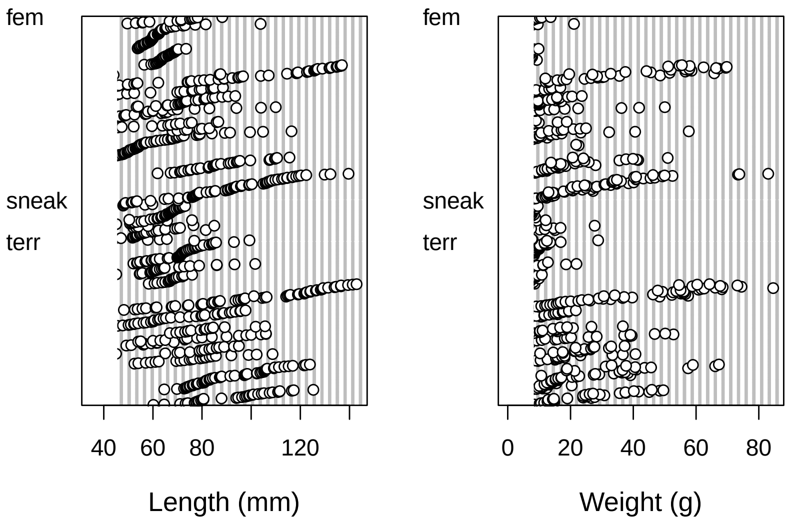

Fig. 5.3 Dotplots of pumpkinseed weight (wt) and length (sl) split by fish mating tactic. Data are arranged by the order they appear in the dataframe.

The distribution of sizes between the sexes is comparable (Fig. 5.2). Among mating tactics, however, there are differences, with males expressing the sneaker tactic tending to be smaller than territorial males and females. There may be an interaction between size and mating tactic.

5.2.2 Lots of zeros in the response variable

The number of zeros in the response variable can be estimated as:

```
> sum(pkin$spot == 0)
```

```
554
```

A total of 554 pumpkinseed did not possess a red spot. As a proportion of all fish sampled this is:

```
> sum(pkin$spot == 0) * 100 / nrow(pkin)
```

```
61.55556
```

That is 62% of fish without a red operculum spot. Note though that the model to be fitted is binomial and is expected to contain a large number of zeros, so the high proportion of zeros should not be a problem.

5.2.3 Multicollinearity among covariates

Use a correlation matrix with corresponding pairplots to visualize pairwise correlations. The code for this plot is available in the accompanying R file.

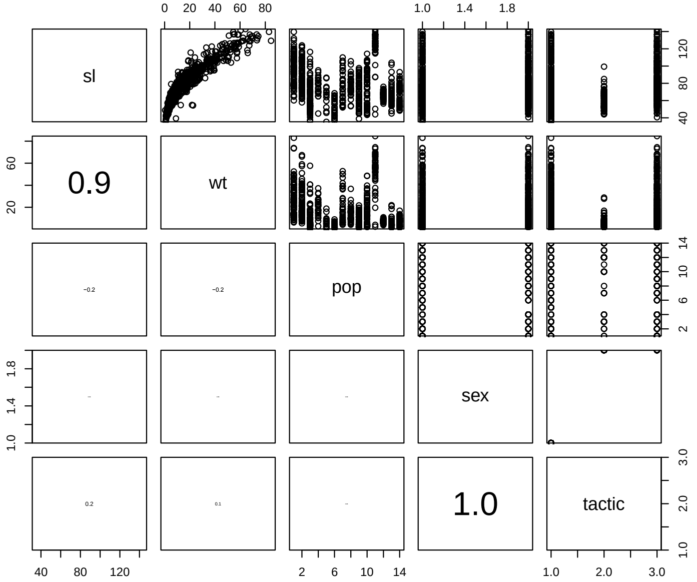

Fig. 5.4 Pairplot of covariates. The lower panel shows pairwise Pearson correlations, with font size proportional to correlation coefficient.

Two pairs of covariates appear strongly collinear from the pairplots in Fig. 5.4. Fish weight (wt) is positively collinear with length (sl) and fish sex is collinear with mating tactic. The correlation between fish weight and length is expected and one of these variables must be dropped; both cannot be included in the same model. Similarly, sex and tactic are clearly collinear; only males play the role of territorial and sneaker. To fit the model, weight and sex will be excluded.

5.2.4 Relationships among dependent and independent variables

Visual inspection of the data using plots. Code for these plots is available in the accompanying R file.

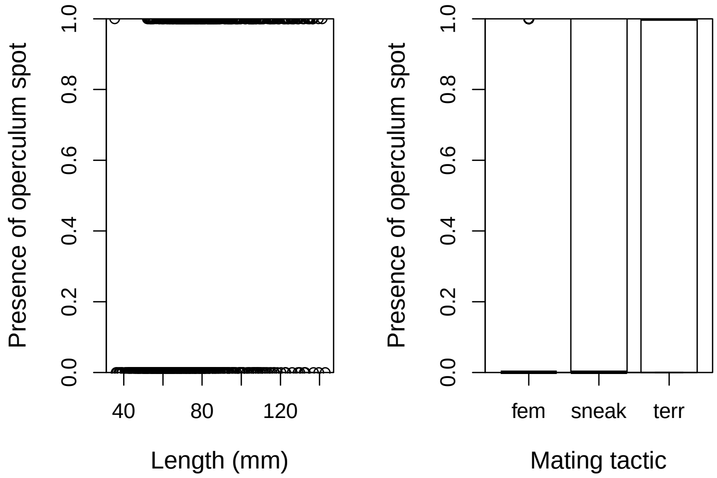

Fig. 5.5 Plots of presence of red operculum spots against fish length and mating tactic.

In Fig. 5.5 the plots of red operculum spots against covariates show no obvious patterns. However, plots of binomial data like these are not particularly informative.

5.3 Model fitting

The data exploration showed:

1. No NAs.
2. No serious outliers in the data.
3. A large number of zeros in the response variable.
4. Some imbalance of data among levels of the categorical covariate 'tactic'
5. Strong collinearity between covariates.
6. Possible interaction between size and tactic.

The model is fitted as:

```
> Bern1 <- glm(spot ~ tactic * sl,
              data = pkin,
              family = binomial(link = "logit"))
```

A second model without interaction can be fitted as:

```
> Bern2 <- glm(spot ~ tactic + sl,
               data = pkin,
               family = binomial(link = "logit"))
```

As with the Poisson and negative binomial models we use a systematic part that contains the model parameters. The link function for a Bernoulli model is a logit link. This link function ensures that the model prediction lies between 0 and 1.

The fit of models Bern1 and Bern2 can be compared using AIC:

```
> AIC(Bern1,Bern2)

       df  AIC
Bern1  6   851.0343
Bern2  4   861.9460
```

The AIC score for model Bern1 is substantially lower than Bern2, which means the model with an interaction between tactic and fish length gives a better fit to the data.

The numerical output is obtained with the `summary` function:

```
> summary (Bern1)

                 Estimate Std. Error  z value  Pr(>|z|)
(Intercept)     -5.006844   0.531217  -9.425  <2e-16
tacticsneak     -2.896331   1.963055  -1.475   0.14010
tacticterr      -0.306384   0.865575  -0.354   0.72336
sl               0.042615   0.005955   7.156   8.31e-13
tacticsneak:sl   0.081898   0.031052   2.637   0.00835
tacticterr:sl    0.030570   0.010692   2.859   0.00425

Null deviance: 1199.16  on 899  degrees of freedom
Residual deviance:  839.03  on 894  degrees of freedom
AIC: 851.03
```

Before interpreting the model, we must first carry out model validation, though this is not straightforward with a Bernoulli model.

5.4 Model validation

For the fitted model Bernoulli GLM, validation requires verification of:

1. Homogeneity of variance.
2. Model misfit.
3. Absence of influential observations.

5.4.1 Homogeneity of variance

Homogeneity of variance can be assessed visually by plotting model residual variance against model fitted values. R code to plot standardised residuals against fitted values is given by:

```
Fitted <- fitted(Bern1)
Resid  <- resid(Bern1, type = "pearson")
par(mfrow = c(1,1), mar = c(5,5,2,2), cex.lab = 1.2)
plot(x = Fitted, y = Resid,
     xlab = "Fitted values",
     ylab = "Pearson Residuals")
abline(h = 0, lty = 2)
```

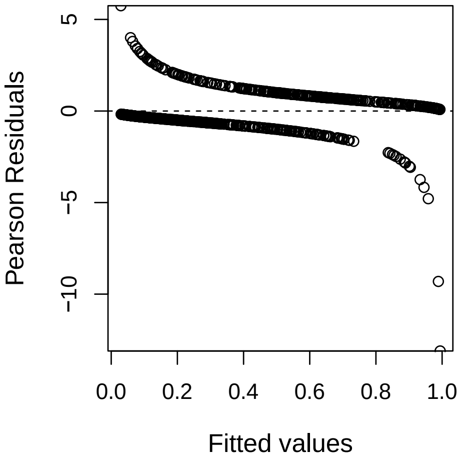

Fig. 5.6 Pearson residuals plotted against fitted values to assess homogeneity of variance. Ideally, the distribution of residuals around zero should be consistent along the horizontal axis.

The distribution of residuals is consistent along the horizontal axis, though this pattern is difficult to assess for a Bernoulli distribution.

5.4.2 Model misfit

Model misfit occurs if covariates are missing or the model departs from linearity and can be recognised visually by plotting Pearson residuals against each covariate in the model, as well as those not included in the model.

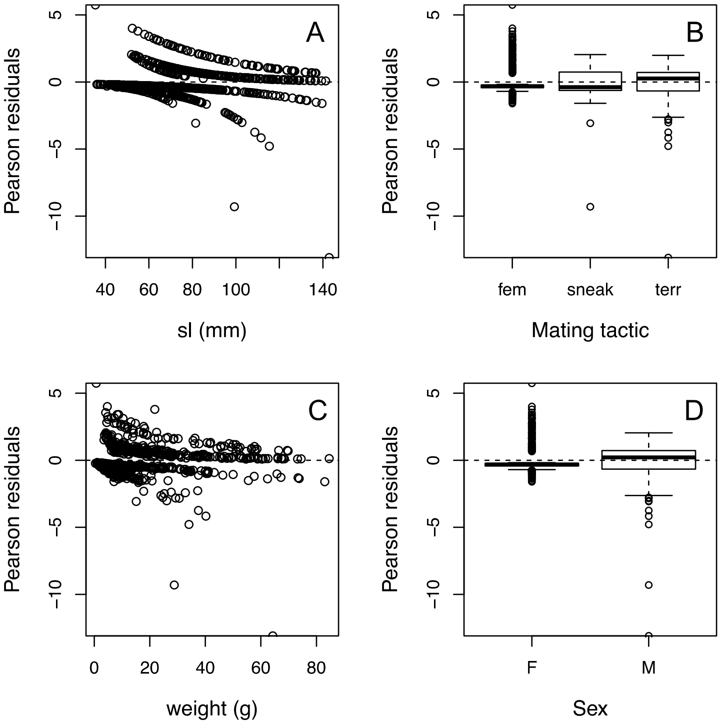

Fig. 5.7 Pearson residuals plotted against covariates to assess model misfit for covariates included in the model; length (A) and mating tactic (B), and not included in the model; weight (C) and sex (D). Ideally, the distribution of residuals around zero should be consistent along the horizontal axis or pass through the median of boxplots.

Plots A-D in Fig. 5.7 show no causes for concern; residuals are distributed consistently along the horizontal axis in each case and there are no obvious patterns in the residuals.

5.4.3 Absence of influential observations

The absence of influential observations can be tested by plotting Cook's distance. A Cook's distance exceeding 1 indicates an influential data point. R code to plot Cook's distance for model Bern1 is given by:

```
> par(mfrow = c(1, 1))
> plot(cooks.distance(Bern1),
       xlab = "Observation",
       ylab = "Cook's distance",
       type = "h",
       ylim = c(0, 1.1),
       cex.lab =  1.5)
> abline(h = 1, lty = 2)
```

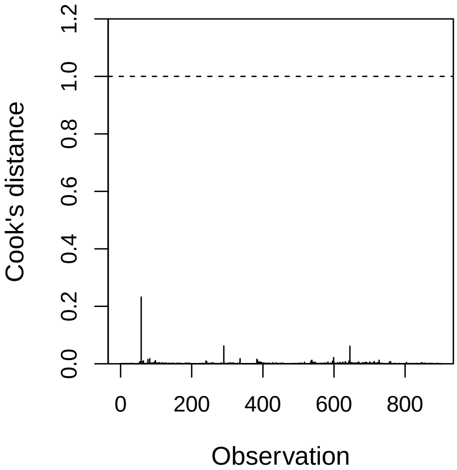

Fig. 5.8 Plot of Cook's distance for model Bern1. A Cook's distance of 1 (indicated by a dashed horizontal line) denotes an influential observation.

There is no evidence from plotting Cook's distance for influential observations in the model (Fig 5.8).

Model validation has shown no evidence of model misfit, model residuals are acceptable and there are no influential observations.

5.5 Model presentation

We can specify the model using mathematical notation in the following way:

$$Spot_i \sim Binomial(\pi_i, n_i)$$

$$E(Spot_i) \sim n_i \times \pi_i \text{ and } \mathrm{var}(Spot_i) = n_i \times \pi_i \times (1 - \pi_i)$$

$$\mathrm{logit}\ (\pi_i) = \eta_i$$

$$\eta_i = \beta_1 + \beta_2 \times SL_i + \beta_3 \times tactic_i + \beta_4 \times rfr_i \times tactic_i$$

Where $Spot_i$ is the probability of fish i having a red operculum spot, which is assumed to follow a binomial distribution with an expected probability (*E*) of expressing an operculum spot of mean $n_i\pi_i$ and variance $n_i\pi_i \times (1-\pi_i)$, with a logit link function. The logit function ensures the fitted probability of a red spot falls between 0 and 1. The variable $tactic_i$ is a categorical covariate with three levels, corresponding with fish mating tactic; female, territorial or sneaker. The model also contained a linear effect for fish length (SL_i).

The numerical output of the model is obtained with:

```
> summary (Bern1)

                  Estimate Std. Error  z value  Pr(>|z|)
(Intercept)      -5.006844   0.531217  -9.425  <2e-16
tacticsneak      -2.896331   1.963055  -1.475   0.14010
tacticterr       -0.306384   0.865575  -0.354   0.72336
sl                0.042615   0.005955   7.156   8.31e-13
tacticsneak:sl    0.081898   0.031052   2.637   0.00835
tacticterr:sl     0.030570   0.010692   2.859   0.00425

Null deviance: 1199.16  on 899  degrees of freedom
Residual deviance:  839.03  on 894  degrees of freedom
AIC: 851.03
```

These results can be more formally presented in the following way:

Table 5.1. Summary of Bernoulli GLM to model the probability of pumpkinseed expressing a red operculum spot as a function of fish length and mating tactic.

Model parameter	Estimate	SE	*P*
Intercept$_{\text{(female)}}$	-5.00	0.53	<0.001
Length	0.04	0.01	<0.001
Tactic$_{\text{(sneak)}}$	-2.90	1.96	0.140
Tactic$_{\text{(territorial)}}$	-0.31	0.87	0.723

Length x Tactic(sneak)	0.08	0.03	0.008
Length x Tactic(territorial)	0.03	0.01	0.004

These results indicate a significant interaction between fish length and mating tactic. To understand this result it is best to visualize the model result in a figure. The R code for this figure is available in the R file accompanying this chapter.

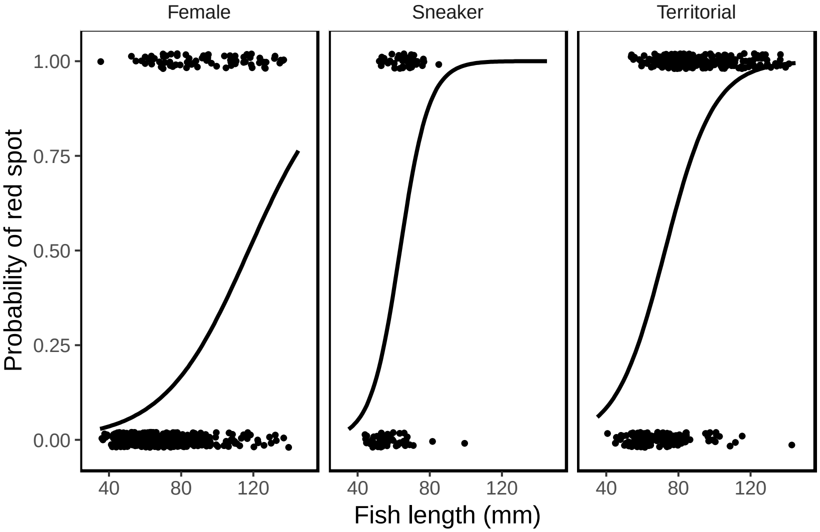

Fig. 5.9 Mean fitted probability (solid line) of pumpkinseed expressing a red operculum spot as a function of length (mm) with 95% confidence intervals (shaded area) for females, sneaker males and territorial males. Data were modelled with a Bernoulli GLM. Black dots are observed data.

The probability of female pumpkinseed expressing a red spot at a given length is lower than for sneaker and territorial males. The size range of sneaker and territorial males differs (territorial males tend to be bigger), but there appears no difference in the probability of these two groups in expressing red spots.

Conclusions

The Bernoulli GLM predicted that male pumpkinseed, either sneakers or territorial, had a significantly greater probability of expressing a red operculum spot at a given body size than females. The size range of sneakers and territorials differed making it difficult to compare directly between these male mating strategies.

The data set was quite large, but was structured by population. Fish population of origin was ignored in the analysis, but there is potential for dependency due to population; i.e. the probability of expressing a red spot may differ with population, or the interaction between length and mating tactic may vary from one population to another. If this is the case, a Generalized Linear Mixed Model (GLMM) might be more appropriate for these data, with population incorporated into the analysis as a "random" term.

Reference

Zięba, G., Smith, C., Fox, M.G., Yavno, S., Záhorská, E., Przybylski, M., Masson, G., Cucherousset, J., Verreycken, H., van Kleef, H. & Copp, G.H., 2018. Red operculum spots, body size, maturation and evidence for a satellite male phenotype in non-native European populations of pumpkinseed *Lepomis gibbosus*. *Ecology of Freshwater Fish* 27, 874-883.

Coda

We hope this book is useful in extending your understanding of GLMs. We are always interested to receive feedback; positive or negative, and would also welcome questions about your own analyses; feel free to email us.

From time-to-time we run statistics workshops, and if you think this is something that might be useful for you or your research group or institution, we are happy to discuss your requirements.

Further reading

This is not an exhaustive list, but the books and papers we have found most useful in performing statistical modelling in ecology include:

Faraway, J.J., 2016. *Linear models with R*. Chapman and Hall/CRC.

Faraway, J.J., 2016. *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. Chapman and Hall/CRC.

Zuur, A.F., Hilbe, J.M. and Ieno, E.N., 2013. *A beginner's guide to GLM and GLMM with R: A frequentist and Bayesian perspective for ecologists*. Highland Statistics Limited.

Zuur, A.F. and Ieno, E.N., 2016. A protocol for conducting and presenting results of regression-type analyses. *Methods in Ecology and Evolution* 7, 636-645.

Zuur, A.F., Ieno, E.N. and Elphick, C.S., 2010. A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution* 1, 3-14.

Zuur, A., Ieno, E.N., Walker, N., Saveliev, A.A. and Smith, G.M., 2009. *Mixed effects models and extensions in ecology with R*. Springer Science & Business Media.