Chapter 10
GLM and GAM for Absence–Presence and Proportional Data

10.1 Introduction

In the previous chapter, count data with no upper limit were analysed using Poisson generalised linear modelling (GLM) and negative binomial GLM. In Section 10.2 of this chapter, we discuss GLMs for 0−1 data, also called absence–presence or binary data, and in Section 10.3 GLM for proportional data are presented. In the final section, generalised additive modelling (GAM) for these types of data is introduced. A GLM for 0−1 data, or proportional data, is also called logistic regression.

When we wrote Chapters 8, 9, 10, and 11, we had a dilemma how to structure the material. The options were as follows:

1. First present the GLM as abstract formulae, and then show the Poisson, negative binomial, and logistic GLMs as special cases. The disadvantage of this approach is that the reader has to go through a grilling mathematical section. This approach may work for the more mathematically skilled reader, but it did not seem appropriate for our target audience.
2. Present every GLM family in detail, and explain all the procedures every time. This approach may be better for a 'GLM-only' book, but it duplicates a lot of text.
3. First present the Poisson GLM in detail, and then present logistic regression (and other GLMs) with help of a couple of examples. The disadvantage of this approach is that the reader has to read the Poisson GLM chapter, even if he or she has absence–presence data.

We decided to go for the third approach because we want to discuss not only the Poisson GLM, but also the logistic GLM, negative binomial GLM, and in Chapter 11 zero-truncated, and zero-inflated GLMs. In Chapter 9, we used a considerable number of pages explaining the Poisson GLM, and the good news is that the mathematical background for this chapter is much the same. However, this does mean you need to have read Chapters 8 and 9 before starting this chapter.

Many statistical textbooks describe logistic regression and we could fill an entire page with references. Some books are dedicated entirely to logistic regression and

A.F. Zuur et al., *Mixed Effects Models and Extensions in Ecology with R*,
Statistics for Biology and Health, DOI 10.1007/978-0-387-87458-6_10,

some only contain a chapter. Some are for medical science, some for econometrics, and some for ecology. Our favourites are McCullagh and Nelder (1989), Agresti (2002), and Fitzmaurice et al. (2004). The second reference is probably a 'must read', and the first one is a 'must cite'.

10.2 GLM for Absence–Presence Data

We illustrate the binomial GLM for absence–presence data with help of two examples. In Section 10.2.1, we model the probability that a wild boar has tuberculosis (Tb) as a function of the length of the animal (length from the nose to the tail joint along the back of the animal, expressed in centimetres). Another potential explanatory variable is age, but it is collinear with length and unbalanced. This example serves as a simple explanation of binomial GLM. A more detailed binomial GLM is presented in Section 10.2.2, which deals with the presence and absence of parasites in cod.

10.2.1 Tuberculosis in Wild Boar

Vicente et al. (2006) analysed the distribution of tuberculosis-like lesions in wild boar *Sus scrofa* to explore the potential importance of wild boar in the maintenance of tuberculosis in south central Spain. Here, we model Tb as a function of the continuous explanatory variable length (as defined above); it is denoted by LenghtCT (CT is an abbreviation of *cabeza-tronco*, which is Spanish for head-body). Tb is a vector of zeros and ones, representing absence and presence of Tb, respectively.

The first thing we do in any data analysis is a data exploration. Useful tools for most types of data are a boxplot and a scatterplot; see Fig. 10.1.

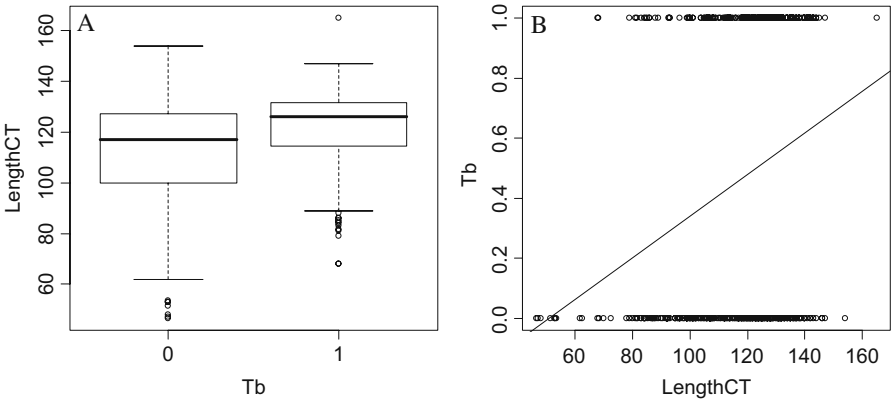

Fig. 10.1 **A**: Boxplot of LengthCT conditional on the variable TB. **B**: Scatterplot of LengthCT versus TB. A regression line was added to aid visual interpretation

The boxplot of LengthCT conditional on Tb (Fig. 10.1A) shows that animals with Tb have larger LengthCT values. The pairplot is less useful due to the 0–1 nature of Tb. When we make a pairplot, we tend to add the fit of a linear regression model. In this case:

$$Tb_i = \alpha + \beta \times CTLength_i + \varepsilon_i.$$

The question is now as follows: How sensible is it to apply linear regression on these data, and what is the interpretation of the fitted line? Let us start with the latter question. The fitted line in Fig. 10.1B suggests that an animal of LengthCT = 100 cm has approximately 0.35 Tb. However, this is a rather strange statement; an animal has Tb or it does not have Tb. It cannot have 0.35 Tb! It seems that our linear regression model is impractical.

To produce a model with fitted values that make more sense, define π_i as the probability that animal i is infected with Tb, and $1 - \pi_i$ is the probability that it is not infected. If we now imagine the vertical axis in Fig. 10.1B showing π_i, we can say that an animal of LengthCT = 100 cm has a probability of 0.35 of being infected with Tb. At least, the fitted values of the linear regression model now make a little bit more sense.

So, the vertical axis in Fig. 10.1B represents the probability that an animal is infected with Tb. Based on the line in Fig. 10.1B, this means that an animal with LengthCT = 47 cm, has a probability of -0.03 of being infected. But probabilities are supposed to be between 0 and 1! And the underlying theory of linear regression tells us that there are realisations (possible outcomes) with probabilities larger than 1 or smaller than 0. It seems we have a serious problem with the linear regression model applied on presence and absence data! The binomial GLM provides a framework to solve all these problems.

To formulate the binomial GLM in a general notation, let Y_i be 1 if animal i is infected with TB and 0 if not infected. A binomial GLM is specified with the same three steps as the Poisson and negative binomial GLMs:

1. An assumption on the distribution of the response variable Y_i. This also defines the mean and variance of Y_i.
2. Specification of the systematic part. This is a function of the explanatory variables.
3. The relationship between the mean value of Y_i and the systematic part. This is also called the link between the mean and the systematic part.

We discuss each of these points in more details next.

Step 1: We assume that Y_i is binomial distributed with probability π_i and $n_i = 1$ independent trials; see also Section 8.6. Recall that this is actually a Bernoulli distribution. As a result, the expected mean and variance of Y_i are given by: $\mathrm{E}(Y_i) = \pi_i$ and $\mathrm{var}(Y_i) = \pi_i \times (1 - \pi_i)$. The π_i plays the same role as the μ_i in Poisson regression and negative binomial regression.

Step 2: The systematic part of the model is specified by the predictor function:

$$\eta(LengthCT_i) = \alpha + \beta \times LengthCT_i$$

Step 3: In this step, we need to define the relationship between the expected value of Y_i, π_i, and the predictor function η. We already argued that the identity link (as imposed by the linear regression model) gives non-sensible results; fitted probabilities and possible realisations are smaller than 0 or larger than 1. So, we need a function that maps the values of η between 0 and 1. There are various options, e.g. the logit link, probit link, clog–log link, and log–log link, but the logit link is the default (canonical) link and is probably the most used one. We will explain it first and then quickly discuss the differences with some of the other ones.

The logit link works as follows. Recall that the problem is that π_i is bounded by a lowest value of 0 and a highest value of 1, and the fitted values obtained by the predictor function η and identity link function ignore this on both sides. Define the odds as follows:

$$O_i = \frac{\pi_i}{1 - \pi_i}$$

The odds are an unusual concept for most scientists. We are familiar with probabilities; it tells us how likely things are with a value between 0 and 1. The odds are doing the same thing, but on a different numerical scale. They are used in for example gambling offices; the odds that a race horse will win can be 9 to 1. This means that if you organise 10 races, it is likely that the horse will win 9 times and lose once. In terms of probabilities: the probability that a particular horse will win is 0.9. This is the same statement as saying that the odds are 9. The nice thing about odds is that they do not have an upper bound. Take a series of values for π_i, say 0.1, 0.2, 0.3, . . .), and 0.9, and calculate the odds; they go from something close to zero to something very large. See also Table 10.1where it shows how probabilities between 0 and 1 are transformed into odds between 0 and infinity.

So, by going from probabilities to odds, we managed to get rid of the upper boundary, but we still have the lower boundary; odds still cannot be negative. The solution is simple; take the natural logarithm of the odds, also called the log odds. The last row in Table 10.1 gives examples of log odds, which are no longer bounded

Table 10.1 Various probabilities, odds, and log odds. The table shows how odds and log odds are calculated from probabilities. The table was taken from Zuur et al. (2007)

P_i	0.001	0.1	0.3	0.4	0.5	0.6	0.7	0.9	0.999
$1 - P_i$	0.999	0.9	0.7	0.6	0.5	0.4	0.3	0.1	0.001
O_i	0.001	0.11	0.43	0.67	1	1.5	2.33	9	999
$\text{Ln}(O_i)$	−6.91	−2.20	−0.85	−0.41	0	0.41	0.85	2.20	6.91

by a lower or upper limit. In a logistic regression, we model the log odds as a linear function of the explanatory variables. This gives the following:

$$\log(O_i) = \eta(LengthCT)$$

Instead of $\log(O_i)$ we can also write $\text{logit}(\pi_i)$. The entire binomial GLM for the Tb data is now given by

$$\begin{aligned} &Y_i \sim B(1, \pi_i) \\ &E(Y_i) = \pi_i \quad \text{and} \quad \text{var}(Y_i) = \pi_i \times (1 - \pi_i) \\ &\text{logit}(\pi_i) = \alpha + \beta \times LengthCT_i \end{aligned}$$

The last line can also be written with some simple mathematics as

$$\pi_i = \frac{e^{\alpha+\beta\times LengthCT_i}}{1 + e^{\alpha+\beta\times LengthCT_i}}$$

Whatever the values of α, β and $LengthCT_i$, the fitted values for π_i are always between 0 and 1. As this model cannot produce fitted values outside the $0 - 1$ range, the binomial distribution ensures we only get sensible realisations.

10.2.1.1 R Code, Results and Fitted Values

The following R code accesses the data, applies the GLM, and presents the numerical output.

```
> library(AED); data(Boar)
> B1 <- glm(Tb ~ LengthCT, family = binomial,
            data = Boar)
> summary(B1)
```

The output is:

```
             Estimate Std. Error z value  Pr(>|z|)
(Intercept) -3.892109   0.671152  -5.799 6.67e-09
LengthCT     0.031606   0.005588   5.656 1.55e-08

Dispersion parameter for binomial family taken to be 1
Null deviance: 700.76 on 507 degrees of freedom
Residual deviance: 663.56 on 506 degrees of freedom
149 observations deleted due to missingness
AIC: 667.56
```

For the moment, we are focussing on the estimated parameters, and the consequences for the graphical interpretation of the model. Deviances, overdispersion, and AIC are discussed in the next example.

The estimated intercept and slope are -3.89 and 0.03 respectively, and are significant at the 5% level. This means that the probability that an animal is of $\texttt{LengthCT}_i$ is infected, is given by:

$$\pi_i = \frac{e^{-3.89+0.03\times LengthCT_i}}{1+e^{-3.89+0.03\times LengthCT_i}}$$

If we fill in a couple of values for LengthCT_i, we can calculate the corresponding π_i, and make a sketch of the relationship. Instead of doing this manually, we use the `predict` command in R:

```
> MyData <- data.frame(LengthCT = seq(from = 46.5,
                         to = 165, by = 1))
> Pred <- predict(B1, newdata = MyData, type = "response")
> plot(x = Boar$LengthCT, y = Boar$Tb)
> lines(MyData$LengthCT, Pred)
```

We first created a data frame `MyData` with new values for the covariate between 46.5 and 165, with steps of 1 cm. The values 46.5 and 165 are the smallest and largest values of the observed animals, and using this range prevents predictions outside the range of observed values. The resulting graph is given in Fig. 10.2. The fitted line shows the pattern of a typical sigmoid curve. Note that the fitted values are always between 0 and 1! At small lengths, the probability of sampling Tb infected animals is small, whereas the probability increases rapidly from about 70–80 cm up to about 140 cm, and then the rate of change becomes smaller again (but the probability of infection stays high).

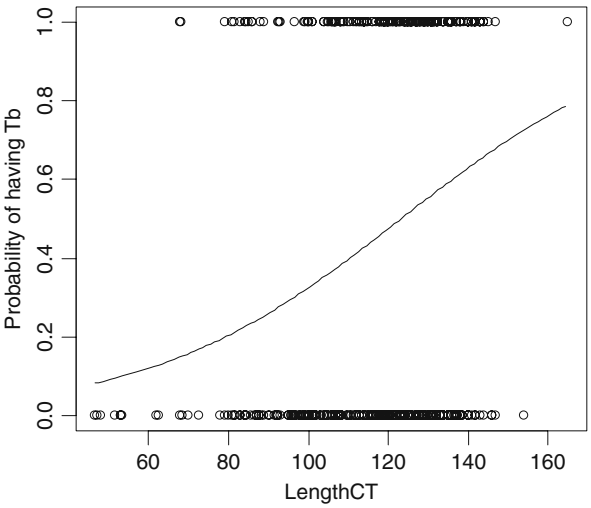

Fig. 10.2 Graph showing the fitted values (*solid line*) obtained by the binomial GLM applied on the boar data. The dots are the observed values

10.2.1.2 General Comments

For a binomial GLM applied on binary data, the mean is given by π_i and the variance by $\pi_i \times (1 - \pi_i)$. To visualise the mean–variance relationship, we plotted the estimated values of π_i versus $\pi_i \times (1 - \pi_i)$, see Fig. 10.3. Note that the variance is the largest for intermediate values of π_i.

Besides the logit link, other link functions are available and a comparison of GLMs with different link functions can be found in, for example, Hardin and Hilbe (2007). Most binary GLMs in the literature use the logit link, but the probit link is a good second choice. We will not discuss the probit link or any of the other link functions here, but the main difference is the shape of the fitted line in Fig. 10.2. In fact, we suggest that you plot the fitted curves obtained from a probit link and a clog–log link yourself. All that is needed is to modify the code in the `family` option inside the `glm` command to `family = binomial(link = "probit")` or `family = binomial(link = "cloglog")`. You will see that the fitted curve is slightly different in the lower and upper parts.

The logit and probit link functions assume that you have approximately an equal number of zeros and ones. The clog–log may be an option if you have considerably more zeros than ones, or vice versa; the sigmoidal curve is asymmetrical. If you do decide to use any of the non-standard link functions (that is other than the logit) for a binary GLM, Hardin and Hilbe (2007) give examples how you can compare different link functions and some tools are based on the AIC, BIC, and deviance.

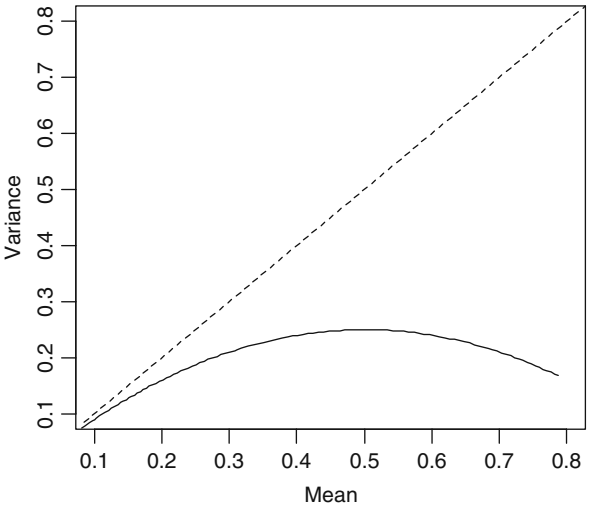

Fig. 10.3 Relationship between the mean π_i and variance $\pi_i \times (1 - \pi_i)$ is given by the *solid line*. The dotted line is the line for which the mean equals the variance. Note that the variance is the largest for intermediate values of π_i

10.2.2 Parasites in Cod

The red king crab *Paralithodes camtschaticus* was introduced to the Barents Sea in the 1960s and 1970s from its native area in the North Pacific. The leech *Johanssonia arctica* uses the carapace of this crab to deposit eggs. The leech is a vector for a trypanosome blood parasite of marine fish, including cod. Hemmingsen et al. (2005) examined a large number of cod for trypanosome infections during annual cruises along the coast of Finnmark in North Norway. These cruises covered three years and were divided in four 'stations' or areas. Full details of the research and results can be found in their paper. Their statistical analyses were carried out using Chi-square statistics and analysis of variance and are in principle all correct. Here, we use a subset of the data and repeat their analyses with GLM.

The response variable is Prevalence, which is coded as 1 if the parasite is present and 0 else. Possible explanatory variables are year, area, and the depth that fish were caught. The problem is that not all areas have the same depth; hence, purely because of the study design, depth and area are collinear (just make a boxplot of depth conditional on area, and you will see that this is indeed the case). Other explanatory variables are sex, length, weight, stage, and age of the fish. Except for sex, all these variables are highly collinear and an arbitrary choice on which one to use has to be made. However, the aim of this text is not to present a full blown analysis, but merely to explain binomial GLM. Hemmingsen et al. (2005) used a model with the main terms year, area, and length, and an interaction term year $\times$ area, and we will also use this set of covariates. We have 1254 observations, but with a few missing values in the spreadsheet. We could remove them, but we prefer using the data as they are and guide you through the problems.

This is clearly a binomial GLM as the response variable is coded as $0-1$. The following model is applied.

$$\begin{aligned} &Y_i \sim B(1, \pi_i) \\ &E(Y_i) = \pi_i \quad \text{and} \quad \text{var}(Y_i) = \pi_i \times (1 - \pi_i) \\ &\text{logit}(\pi_i) = Year_i + Area_i + Year_i \times Area_i + Length_i \end{aligned}$$

We have written down the systematic part of the model in a semi-mathematical notation because $\texttt{Year}_i$ and $\texttt{Area}_i$ are fitted as factors (each have three levels) and $\texttt{Length}_i$ is a continuous variable. The R code to fit this model is given by

```
> library(AED); data(ParasiteCod)
> ParasiteCod$fArea <- factor(ParasiteCod$Area)
> ParasiteCod$fYear <- factor(ParasiteCod$Year)
> Par1 <- glm(Prevalence ~ fArea * fYear + Length,
              family = binomial, data = ParasiteCod)
```

The `family = binomial` option and a response variable with zeros and ones is the only difference compared the GLMs used in the previous chapters.

The `summary(Par1)` command gives a lot of output due to the three levels for Area and Year. If we omit for the moment, the estimated values, standard errors, z-values, and p-values, we have

```
...
Dispersion parameter for binomial family taken to be 1
Null deviance: 1727.8 on 1247 degrees of freedom
Residual deviance: 1495.2 on 1235 degrees of freedom
6 observations deleted due to missingness
AIC: 1521.2
```

The good news is that in a Bernoulli GLM (the response variable is a vector with zeros and ones), overdispersion cannot occur. A justification for this can be found in McCullagh and Nelder (p. 125, 1989). The rest of the information is similar as for the Poisson GLM (Chapter 9). For example, the AIC can be used for model selection. And we also have the `step` function, which can be used for automatic model selection. The hypothesis testing procedures are also identical to Poisson GLM. Because we have factors with more than two levels in the model, we use `drop1(Par1, test = "Chi")` as it gives one p-value for the interaction. The output below shows that the Area $\times$ Year interaction is highly significant at the 5% level, but not the variable `Length`.

```
Single term deletions. Model:
Prevalence ~ fArea * fYear + Length
                    Df   Deviance      AIC     LRT   Pr(Chi)
<none>                    1495.16  1521.16
Length               1    1498.64  1522.64    3.47      0.06
fArea:fYear          6    1537.60  1551.60   42.44    <0.001
```

The full output from the `summary` command is not shown here, but just as in linear regression and Poisson GLM, it tells you which levels of a factor are different from the baseline level, and in the case of an interaction, which combinations are different from the baseline (Area 1 and Year 1999). To see which area−year combinations are different from *each other*, you can change these baselines to other values, and do some post-hoc testing (see Chapter 10 in Dalgaard (2002)).

The graphical model validation in a Binomial GLM with a $0-1$ response variable is some sort of an art, and Fig. 10.4 shows why. So far, we said: You should not see any patterns in the residuals. Because the observed data are zeros and ones, we now see two clear bands in these graphs. This makes it rather difficult to say anything sensible about these graphs, and one can wonder whether there is any point in using them. In cases where you have a large data set, like we have in this example (1254 observations), it may be an option to extract the residuals, put them in groups of, say 10, calculate an average of the residuals per group, and use these in graphical validation plots. The groups can be based on the order of the fitted values, or on the order of a covariate.

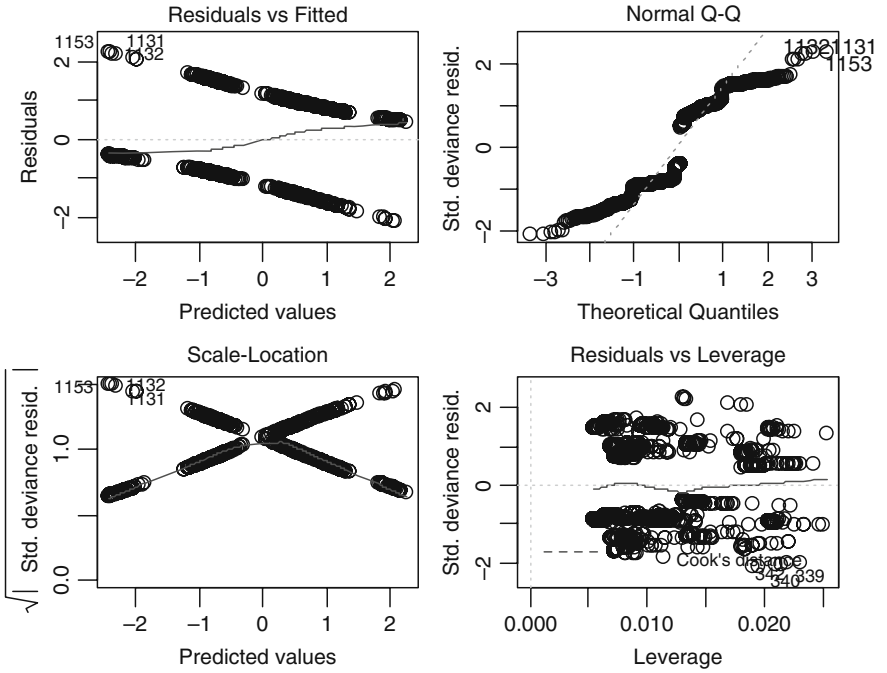

Fig. 10.4 Standard graphical validation tools for a GLM. Because we are working with a response variable that has only zeros and ones, we can see two bands of points in all but the leverage plot

The easy mistake to make with the model selection process for this data set is ignoring the missing values. Once an explanatory variable with missing values is dropped from the selection process, the new data set may have more observations, and therefore, AICs are not comparable! This also holds for analysis of deviance tables. But luckily, the `drop1` command does it properly by removing the observations with NAs, but that only works for one round. It is therefore better to remove missing values before doing the model selection process.

10.3 GLM for Proportional Data

In the previous section, the response variable Y_i was binary and a Bernoulli distribution was used. The notation for this was B(1, π_i), where π_i is the probability on 'success'.

Vicente et al. (2006) analysed data from a number of estates with wild boar and red deer in Spain. At each estate i, a group of n_i animals was sampled. The data set contains information on the tuberculosis (Tb) disease in both species, and on the parasite *Elaphostrongylus cervi*, which only infects red deer. Both variables are recorded as the number of animals that are positive for Tb or have the parasite

E. cervi. So the response variable Y_i is the number of animals that test positive out of n_i animals. There is also information on the main characteristics of the habitat and management (fencing) at each estate: The percentage of open land, scrubs and pine plantation, number of *quercus* plants per area, number of *quercus* trees per area, a wild boar abundance index, a reed deer abundance index, estate size (ha), and whether the estate is fenced (1 = yes, 0 = no).

Data like these are typically analysed using a GLM with a binomial distribution (Chapter 8). Let us focus first on *E. cervi* in deer. Define Y_i as the number of deer at estate i that have *E. cervi*, and n_i is the number of sampled deer. The binomial GLM is as follows:

$$
\begin{aligned}
&Y_i \sim B(n_i, \pi_i) \\
&E(Y_i) = \pi_i \times n_i \quad \text{and} \quad \text{var}(Y_i) = n_i \times \pi_i \times (1 - \pi_i) \\
&\text{logit}(\pi_i) = \alpha + \beta_1 \times \mathit{OpenLand}_i + \beta_2 + \mathit{ScrubLand}_i + \cdots \beta_8 \times \mathit{Fenced}_i
\end{aligned}
$$

We also assume that the n_i deer are independent and that each animal at estate i has the same probability π_i of having the parasite. If this is not the case, then you should work with the individual binary data per animal and use generalised linear mixed modelling techniques (Chapter 13). The logistic regression model can be fitted in R with the following code. The first two commands import the data. The function `corvif` is our own function that calculates variance inflation factors to detect collinearity. It is available in the `AED` package, but a similar function can be obtained from the `car` package. The variable `PinePlantation` was dropped due to collinearity. The remaining code applies the binomial GLM.

```
> library(AED); data(Tbdeer)
> Z <- cbind(Tbdeer$OpenLand, Tbdeer$ScrubLand,
             Tbdeer$QuercusPlants, Tbdeer$QuercusTrees,
             Tbdeer$ReedDeerIndex, Tbdeer$EstateSize,
             Tbdeer$Fenced)
> corvif(Z)
> DeerNegCervi <- Tbdeer$DeerSampledCervi -
                  Tbdeer$DeerPosCervi
> Tbdeer$fFenced <- factor(Tbdeer$Fenced)
> Deer1 <-glm(cbind(DeerPosCervi, DeerNegCervi)~
       OpenLand + ScrubLand + QuercusPlants +
       QuercusTrees + ReedDeerIndex + EstateSize + fFenced,
       family = binomial, data = Tbdeer)
> summary(Deer1)
```

Note that the response variable is a data frame consisting of two columns; the number of positives and the number of negatives (which is `DeerNegCervi`). It is also possible to fit the model with the following code:

```
> Tbdeer$DeerPosProp <- Tbdeer$DeerPosCervi /
                        Tbdeer$DeerSampledCervi
> Deer2 <- glm(DeerPosProp ~ OpenLand + ScrubLand +
               QuercusPlants + QuercusTrees +
               ReedDeerIndex + EstateSize + fFenced,
               family = binomial, data = Tbdeer,
               weights = DeerSampledCervi)
```

The variable `DeerPosProp` contains the proportion (as a value between 0 and 1) of animals that are positive (presence of the parasite). Both approaches give exactly the same results. The summary output from model `Deer2` is as follows.

```
                 Estimate Std. Error z value  Pr(>|z|)
(Intercept)      3.843e+00  7.772e-01   4.945 7.61e-07
OpenLand        -3.950e+00  6.383e-01  -6.187 6.12e-10
ScrubLand       -7.696e-01  6.140e-01  -1.253 0.210042
QuercusPlants   -3.633e-04  2.308e-02  -0.016 0.987439
QuercusTrees     2.290e-03  5.326e-02   0.043 0.965707
ReedDeerIndex    6.689e-02  2.097e-02   3.191 0.001419
EstateSize      -8.218e-05  2.478e-05  -3.316 0.000913
fFenced1        -2.273e+00  5.954e-01  -3.819 0.000134

Dispersion parameter for binomial family taken to be 1
Null deviance:     234.85 on 22 degrees of freedom
Residual deviance: 152.79 on 15 degrees of freedom
(9 observations deleted due to missingness)
AIC: 227.87
```

This output is similar to the output of the Poisson GLM in Chapter 9. In a binomial GLM with $n_i > 1$, we can have overdispersion. This seems to be the case here, and we have to fit a quasi-binomial model. This is doing the same thing as in a quasi-Poisson GLM by adding an overdispersion parameter ϕ to the variance of Y_i; $\text{Var}(Y_i) = \phi \times n_i \times \pi_i \times (1 - \pi_i)$. The R programming process is similar to a quasi-Poisson process; first we need to fit a model with the `family = quasibinomial` option (call the resulting object `Deer3`) and the `drop1(Deer3, test = "F")` command can be used to assess which term to drop. The final model contains only `OpenLand`:

```
> Deer4 <- glm(cbind(DeerPosCervi,DeerNegCervi) ~
               OpenLand, data = Tbdeer,
               family = quasibinomial)
> drop1(Deer4, test = "F")
```

The analysis of deviance test with the `drop1` command is not presented here, but it gives a p-value of 0.02 for `OpenLand`, and the `summary` command gives

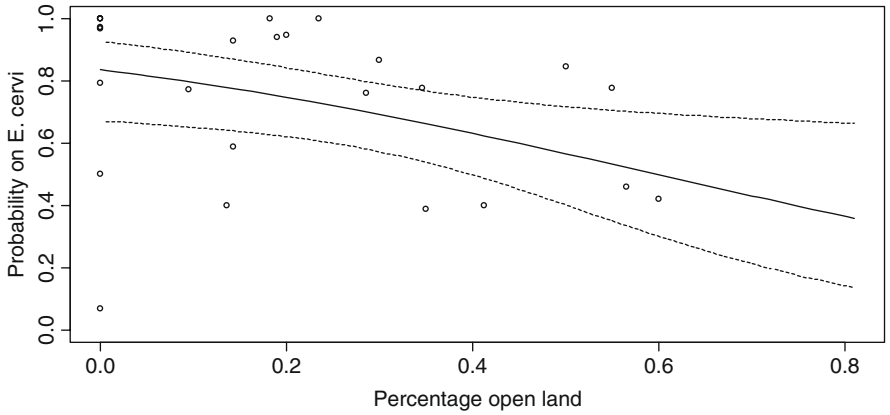

Fig. 10.5 Fitted values (*solid line*) and 95% confidence bands for the optimal binomial GLM model applied on the red deer data

a negative slope. These results suggest that the larger the percentage of open land, the smaller the probability of sampling deer with *E. cervi*. The results can also be visualised (Fig. 10.5) using the R code:

```
> MyData <- data.frame(OpenLand =
              seq(from = min(Tbdeer$OpenLand),
                  to = max(Tbdeer$OpenLand), by = 0.01))
> P1 <- predict(Deer4, newdata = MyData, type = "link",
                se = TRUE)
> plot(MyData$OpenLand, exp(P1$fit) / (1+exp(P1$fit)),
       type = "l", ylim = c(0, 1),
       xlab = "Percentage open land",
       ylab = "Probability on E. cervi")
> lines(MyData$OpenLand, exp(P1$fit+1.96*P1$se.fit) /
       (1 + exp(P1$fit + 1.96 * P1$se.fit)), lty = 2)
> lines(MyData$OpenLand, exp(P1$fit-1.96*P1$se.fit) /
       (1 + exp(P1$fit - 1.96 * P1$se.fit)), lty = 2)
> points(Tbdeer$OpenLand, Tbdeer$DeerPosProp)
```

The data frame `MyData` contains new values for the explanatory variable `OpenLand`, and we use these for the predictions. The `predict` function predicts at the level of the predictor function, and therefore, we need to transform the fitted values (and the confidence bands) with the logistic link function. This ensures that the confidence bands are between 0 and 1.

The model validation process in a binomial GLM is identical to the one in a Poisson GLM; plot the Pearson or deviance residuals against the fitted values and plot the residuals versus each explanatory variable in the model (and also against the variables that were dropped).

10.4 GAM for Absence–Presence Data

Having explained additive modelling in detail in Chapter 2 and binomial GLM in detail in all the earlier sections in this chapter, binomial GAM is just a combination of the two, and we now give a short example to illustrate the method. In Section 10.2, we analysed the presence of parasites in cod, and assumed that $Y_i \sim \text{B}(1, \pi_i)$ and

$$\text{logit}(\pi_i) = \alpha + Year_i + Area_i + Year_i \times Area_i + Length_i$$

Because all explanatory variables are nominal or continuous, the model is called a generalised *linear* model. If you are unsure that length has a linear effect, or if a plot of residuals (obtained by a GLM) against Length shows a clear pattern, we can use:

$$\text{logit}(\pi_i) = \alpha + Year_i + Area_i + Year_i \times Area_i + f(Length_i)$$

where $f(Length_i)$ is a smoothing function of *Length*. Such a model is called a generalised *additive* model. The only difference between a GLM and a GAM is that the latter contains at least one smoothing function in the predictor function. The following R code applies a GAM on the cod parasite data. Length is fitted as a smoother.

```
> library(AED); data(ParasiteCod)
> library(mgcv)
> ParasiteCod$fArea <- factor(ParasiteCod$Area)
> ParasiteCod$fYear <- factor(ParasiteCod$Year)
> P2 <- gam(Prevalence ~ fArea * fYear + s(Length),
            family = binomial, data = ParasiteCod)
```

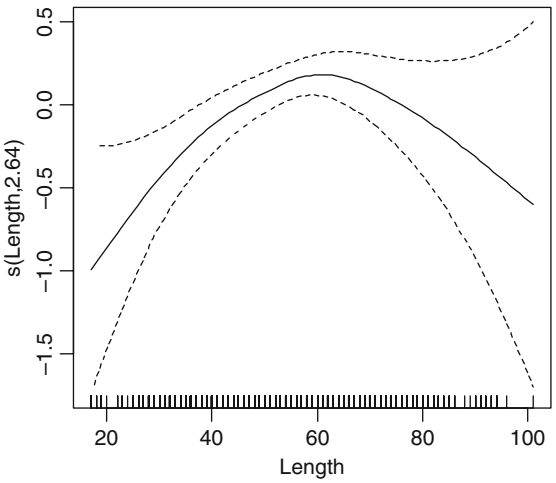

Fig. 10.6 Estimated smoother for Length obtained by the GAM applied on the cod parasite data. The solid line is the smoother, and the dotted lines are 95% confidence bands

The only difference compared to the `gam` commands in Chapter 3 is the `family = binomial` option. The same model selection and model validation steps should be applied as we did with logistic regression and discussed in previous sections. The `anova(P2)`, `summary(P2)`, and `plot(P2)` commands can be used. No numerical output is presented here, but the smoother of Length is significant at the 5% level (2.63 degrees of freedom, $X^2 = 17.08$, $p = 0.009$). The estimated smoother is presented in Fig. 10.5. Although 2.63 degrees is not strong evidence against a GLM (1 degree of freedom is identical to a GLM), the curve clearly shows a non-linear Length effect. Fish around 60 have a higher probability of having parasites than smaller and larger fishes (Fig. 10.6).

10.5 Where to Go from Here?

In Chapters 12 and 13, we extend GLM and GAM to allow for nested data, and temporal and spatial correlations, leading to the methods of generalised estimation equations, generalised linear mixed modelling, and generalised additive mixed modelling.