6 Generalised linear modelling

6.1 Poisson regression

In the linear regression chapter, we analysed the RIKZ data and identified various problems:

- Some fitted values for the response variable were close to negative, but the response variable (species richness) can only take positive values. In this case, we were just lucky that they were positive.
- The Gaussian density curves suggest that, in theory, some values could be negative.
- There was a violation of the homogeneity assumption because the spread in the residuals increased for the larger fitted values indicating heterogeneity.

A data transformation might solve the heterogeneity problem, but this would not avoid the negative fitted values. We have even more problems if the data are of the form 0/1 (0 = absence and 1 = presence), proportions between zero and one, or percentages (between 0% and 100%). However, using Poisson or logistic regression can solve these problems, with Poisson regression used with count data and logistic regression used with presence–absence or proportional data.

Good starting points for generalised linear modelling are Chambers and Hastie (1992), Pampel (2000), Crawley (2002), Dobson (2002), Quinn and Keough (2002), Fitzmaurice et al. (2004), and especially for logistic regression Kleinbaum and Klein (2002). Note that some of these references do not use ecological data. At a more mathematical level we recommend McCullagh and Nelder (1989).

The Poisson regression provides the mathematical framework. First, it assumes that the values for the response variable Y_i are Poisson distributed with expectation μ_i. The notation for this is $Y_i \sim P(\mu_i)$, and as a direct consequence of this distributional assumption, the expectation of Y_i is equal to its variance: $E[Y_i] = \mu_i = \text{var}(Y_i)$. The probability density function is

$$P(x \text{ occurences}) = e^{-\mu} \times \frac{\mu^x}{x!}$$

Using this formula, it is easy to calculate P(x = 0), P(x = 1), P(x = 2), etc. for a given μ. Four different Poisson density curves are shown in Figure 6.1. Panels A and B show the Poisson density function for μ = 1 and μ = 5, where both curves

are skewed to the right. Panels C and B show the probabilities for $\mu = 15$ and $\mu = 25$, where the skew is reduced. In panel C the density function is only slightly skewed to the right, and in Panel D, the second curve is approximately symmetrical. Indeed, for larger values of the expected value, the shape of density curve of a Poisson distribution becomes similar to that of the Gaussian density curve. The width of the curve in Panel D is the largest of the four curves, and this is inherent to the Poisson distribution; the larger the mean, the larger the variance. Note that the Poisson distribution is only used for discrete data (response variable). Rounding the data to the nearest integer will achieve this, but be careful if the data values of your response variable are all between a and $a + 1$; using the Poisson distribution is not sensible in this case.

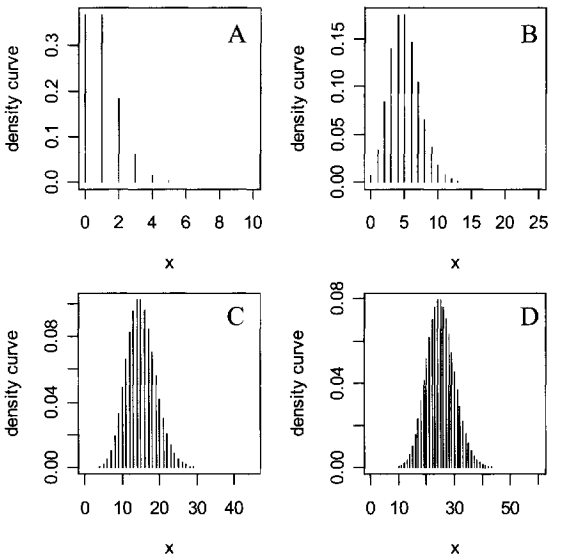

Figure 6.1. Poisson distributions for $\mu = 1$ (A), $\mu = 5$ (B), $\mu = 15$ (C) and $\mu = 25$ (D). In each panel the horizontal value shows possible values of x (the response variable) and the vertical axis gives the corresponding probability that this value is observed, given that the mean is μ.

In Poisson regression, Poisson density curves replace the Gaussian density curves used in linear regression. This step allows for some increase in the spread, and avoids density curves that suggest possible negative realisations. However, this does not stop the model from giving negative fitted values. It is now useful to introduce some mathematical notation. Let $g(x)$ be the so-called predictor function:

$$g(x_i) = \alpha + \beta_1 X_{1i} + \ldots + \beta_p X_{pi}$$

In the linear regression model, we have the distributional assumption $Y_i \sim N(\mu_i,\sigma^2)$, the model itself: $Y_i = g(x_i) + \varepsilon_i$, and as a consequence: $E[Y_i] = \mu_i = g(x_i)$. Negative fitted values occur if $g(x_i)$ becomes negative. In Poisson regression, we use a slightly different relationship between the expectation μ and linear predictor function $g(x_i)$:

$$\mu_i = e^{g(x_i)} \quad \text{or} \quad \log(\mu_i) = g(x_i) \tag{6.1}$$

Because the mean (or fitted value) is modelled as an exponential model, it is always positive. The Poisson regression model can be summarised as

$$R_i \sim P(\mu_i) \quad \text{and} \quad E[Y_i] = \mu_i = e^{g(x_i)} = e^{\alpha + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}}$$

To illustrate this model, we applied a Poisson regression to the RIKZ data. For details on these data, see Chapter 31. The species richness at site i is denoted by R_i and is used as response variable and NAP as the explanatory variable. The model is

$$R_i \sim P(\mu_i) \qquad \text{and} \qquad E[R_i] = \mu_i = \exp(\alpha + \beta_1 NAP_i)$$

The fit and the graphical interpretation of this model are shown in Figure 6.2. Dots represent the observed values in the R-NAP space, and R is the response variable. The fitted line, which is now an exponential curve, is plotted in the same space.

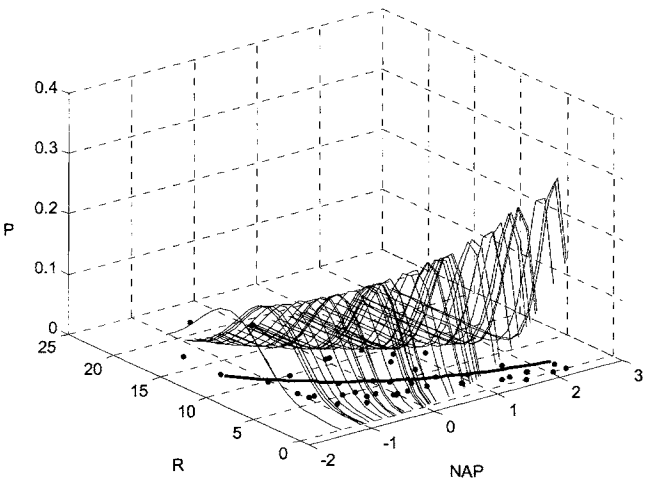

Figure 6.2. Fit of Poisson regression model for RIKZ data. R is the response variable (species richness), and NAP is the explanatory variable. The third axis identifies the (smoothed) probability of the density curves. Dots are the observed values, and the line in the R-NAP space is the fitted Poisson regression curve. The density curves show the probability of other realisations at the same NAP value.

On top of this curve, we plotted the Poisson density curves, and these define the most likely values of other potential values for R. Technically, these density curves are discrete, but for visualisation purposes, we drew a line. Note the fitted values are non-negative, Poisson density curves show that negative realisations are not possible, and only one observation is in the tail of the distribution.

Figure 6.3 shows the same graph, but from a different angle. It clearly shows the change in the Poisson density curves from small non-symmetric curves to wide and approximately symmetrically curves.

Dispersion

Sometimes the increasing spread in count data is even larger than can be modelled with the mean-variance relationship of the Poisson distribution. A possible solution is to introduce a dispersion parameter ρ such that $E[R_i] = \mu_i$, and the variance of R_i is modelled as $\rho\mu_i$, For $\rho > 1$, this allows for more spread than the standard Poisson mean-variance relationship and is called overdispersion. If $\rho < 1$, it is called underdispersion.

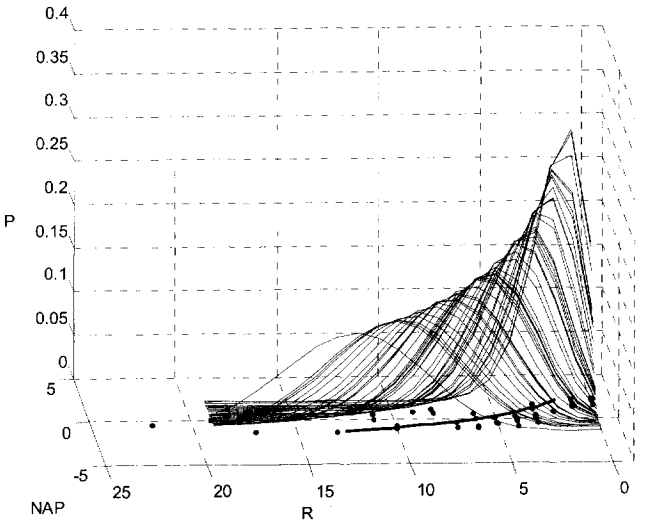

Figure 6.3. As Figure 6.2, but from a different angle.

Estimation of a Poisson regression model with a dispersion parameter is called quasi-likelihood (or quasi-Poisson), but technically it is no longer a Poisson model. A consequence of introducing an overdispersion parameter ρ is that all estimated standard errors are multiplied with the square root of ρ, and therefore ignoring overdispersion can lead to wrong conclusions. We demonstrate the solutions and effects of overdispersion using the RIKZ data. The Poisson regression

model using richness as the response variable and NAP, as an explanatory variable, gave the following.

	Estimate	Std. Error	t-value	p-value
Intercept	1.79	0.06	28.30	<0.001
NAP	−0.56	0.07	−7.76	<0.001

The null deviance is 179.75 on 44 degrees of freedom, the residual deviance is 113.18 on 43 degrees of freedom and the AIC = 259.18. The null deviance is the equivalent of the total sum of squares in linear regression, whereas the residual deviance is the equivalent for the residual sum of squares. The estimated parameters are highly significant. The dispersion parameter can be estimated from the residuals of the Poisson regression model, or one can simply apply the quasi-Poisson model. The output after correcting for overdispersion (using the quasi-Poisson model) is:

	Estimate	Std. Error	t-value	p-value
Intercept	1.79	0.10	16.23	<0.001
NAP	−0.56	0.12	−4.45	<0.001

The overdispersion parameter was $\rho = 3.04$. We explain later how to estimate ρ. The standard errors are automatically corrected, and are now considerably larger, because they were multiplied by the square root of the dispersion parameter. This results in the estimated regression parameters being 'less' significant (the t-values are divided by the square root of the dispersion parameter). However, in this example, they are still significant at the 5% level. Note that for the Poisson regression model, there was just one observation in the tail (Figure 6.3), and this observation probably caused the overdispersion. Ignoring overdispersion can easily result in wrongly deciding a parameter is significant.

The Poisson regression model is also called a generalised linear model with Poisson distribution and log link function; the model $E[Y_i] = \exp(\alpha + \beta_1 \text{NAP}_i)$ can be written as $\log(E[Y_i]) = \alpha + \beta_1 \text{NAP}_i$, hence, the name log-link. Most model selection and validation techniques for Poisson regression are similar to those used in linear regression (ANOVA-tables, t-values, AIC, hat values, Cook's distances), but a few are different, and these are discussed next.

Deviance

The (residual) deviance D is the GLM equivalent of the residual sum of squares. Technically, it is the difference between the log-likelihood of the saturated model (using as many parameters as observations) and the log-likelihood for the fitted model. It is useful for model comparisons. A small deviance value indicates a good fit, and a large value a poor fit. To decide whether D is small or large, you can use a Chi-square distribution. The degrees of freedom for this distribution is $n - p - 1$, where n is the number of observations and p is the number of parameters (slopes) in the model. This test can only be used for reasonably large n, and

many authors warn about its approximate nature (McCullagh and Nelder 1989; Hosmer and Lemeshow 2000). A safer use of the deviance is for model comparison of two nested models. Suppose we are fitting a Poisson model to the RIKZ data of the form:

$$R_i \sim P(\mu_i) \quad \text{and} \quad E[R_i] = \mu_i = \exp(\alpha + \beta_i NAP_i + Exposure_i + Week_i)$$

where exposure (3 levels) and week (4 levels) are fitted as nominal variables. The question is whether a model with NAP, Exposure and Week is better than a model with NAP and Week. The deviance of the model containing all three variables is $D_1 = 47.80$, and the deviance for the model with only NAP and Week is $D_2 = 53.47$ (the smaller the deviance the better). The difference between D_1 and D_2, which is asymptotically Chi-square distributed with $p_1 - p_2$ degrees of freedom (p_1 and p_2 are the number of parameters in the two models), is 5.67 ($p = 0.02$). This test assumes there is no overdispersion. What we are testing here is whether the null hypothesis that all the regression parameters for Exposure (all levels) are equal to zero, and the results suggest there is evidence to reject this assumption. Another way to test whether Exposure can be dropped from the model is to look at the *t*-values:

	Estimate	Std. Error	*t*-value	*p*-value
Intercept	2.53	0.13	19.68	<0.001
NAP	−0.49	0.07	−6.57	<0.001
factor(week)2	−0.76	0.35	−2.16	0.03
factor(week)3	−0.51	0.21	−2.40	0.02
factor(week)4	0.12	0.23	0.55	0.58
factor(exposure)10	−0.43	0.19	−2.24	0.03
factor(exposure)11	−0.65	0.33	−1.96	0.05

Note that there are no entries for week 1 and exposure level 8 as these are nominal explanatory variables. This was discussed in Chapter 5.

The *t*-value is asymptotically normal distributed with expectation zero and variance one; hence, the 95% confidence interval for the NAP parameter is $-0.49 \pm 1.96 \times 0.07$. Since 0 is not in this interval the NAP parameter is significantly differently from 0 at the 5% level. One of the exposure levels is also significantly different from 0 at the 5% level. Instead of doing this process manually, several statistics programmes provide automatic 'drop 1 variable' tools. The output of this function is of the form:

	df	Deviance	AIC	LRT	*p*-value
<none>		47.80	203.80		
NAP	1	93.46	247.46	45.66	<0.001
factor(week)	3	58.37	208.37	10.57	0.01
factor(exposure)	2	53.47	205.46	5.67	0.06

If all explanatory variables are used, the deviance is 47.8. Each explanatory variable is deleted in turn. For example, if exposure is dropped the deviance increases to 53.47, which is an increase of 5.66. This difference follows a Chi-

square distribution with 2 degree of freedom (the nested model has 2 parameters fewer than the full model). The associated p-value is 0.06, indicates that you cannot reject the null hypothesis at the 5% level, and that the regression parameters for exposure are zero. This, together with all other variables being significantly different from 0 at the 5% level, indicates that exposure can be dropped from the model. Clearly, NAP is the most important variable; leaving it out results in the highest deviance, indicating the poorest fit.

For small datasets the t-statistic may give a different message compared with the deviance test. In this case, you should use the deviance test, as it is more reliable that the t-test for small datasets. For large datasets, p-values obtained by the deviance test and t-test are likely to be similar (provided there is no strong collinearity). The deviance test can also be used to compare nested models where the difference is more than one explanatory variable. Instead of the deviance test, you can also use the AIC.

The deviance test described above assumes there is no overdispersion. However, as shown in the dispersion paragraph above, the RIKZ data might be overdispersed, and in this case, the difference between the deviances of two nested models will not be Chi-square distributed. If the dispersion parameter is estimated (as in Quasi-Poisson), the two models can be compared using:

$$\frac{D_2 - D_1}{\rho(p-q)} \sim F$$

where ρ is the overdispersion parameter, and p + 1 and q + 1 are the number of parameters in models 1 and 2, respectively. The '+1' is for the intercept. D_1 is the deviance of the full model and D_2 of the nested model. Under the null-hypothesis the regression parameters of the omitted explanatory variables are equal to zero, and the F-ratio follows an F-distribution with $p - q$ and $n - p$ degrees of freedom (n is the number of observations). For the model with all explanatory variables, we have $D_1 = 47.803$, and for the model without exposure $D_2 = 53.466$. The difference in number of explanatory variables is 2, and $n = 45$. The F-statistic is obtained from an analysis of deviance table:

Model	Resid. df	Resid. Dev	df	Deviance	F-statistic	p-value
1	38	47.80				
2	40	53.46	2	5.66	2.39	0.10

Model 1 contains the explanatory NAP, week (as a factor) and exposure (as a factor). Model 2 has NAP and week. This shows that by considering overdispersion, exposure is not significant at the 5% level. Repeating the analysis without exposure may change the estimated parameters, the overdispersion parameter, and even the order of importance of the explanatory variables.

Validation plots

Using GLM validation plots for the first time can be confusing because the fitted values can be plotted on either the predictor scale or on the original scale. In the first case, the function $g(x)$ is plotted, but in the second situation, we use $\exp(g(x))$. If observed values are close to zero, then $\exp(g(x))$ is hopefully close to zero as well, but this means that $g(x)$ has negative values, and these negative values can sometimes be rather large.

For the linear regression model, we used the ordinary, standardised and Studentised residuals. In non-normal GLM models (the Gaussian model is a GLM model as well) other types of residuals are defined. The two most important types of residuals are the deviance (E^D) and the Pearson (E^P) residuals. For the Poisson model they are defined by

$$E_i^D = sign(Y_i - \hat{\mu}_i)\sqrt{d_i}$$

$$E_i^P = \frac{Y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$$

where d_i is the contribution of the i^{th} observation to the residual deviance D, and $\hat{\mu}_i$ the fitted value. The deviance is calculated from the sum of the squared deviance residuals. Hence, deviance residuals show which observations have a large contribution to the deviance. The underlying idea of Pearson residuals is as follows. The Poisson distribution allows for larger spread for larger fitted values and therefore it doesn't make sense to inspect observed values minus fitted values. Therefore, we scale these differences by the square root of the variance.

Both types of residuals are useful for detecting residuals with large influences. For non-Poisson and non-Binomial (see below) distributions, the dispersion parameter can be estimated from the sum of squared Pearson residuals divided by $n - q$ (n is the number of observations and q is the number of parameters in the model). An alternative estimator is $D/(n - q)$. For the normal distribution, the dispersion parameter is the variance.

Useful validation plots are as follows:

1. A scatterplot of the observed data versus the fitted values (Y_i versus μ_i). These should lie as much as possible on a straight line with slope 1.
2. Absolute deviance residuals versus the linear predictor $g(x)$. The discrete nature of the data might show a distinct pattern.
3. Individual explanatory variables versus residuals.
4. Hat (leverage) values and Cook's distances.
5. Influential observations (in terms of overall fitting criteria and estimated parameters).
6. Residuals versus each explanatory variable. If one can see a pattern in these graphs, generalised additive modelling is a possible follow-up analysis. Alternatives are including interaction terms or quadratic terms (Crawley 2005).

As we are now using techniques that do not need a normal distribution, a normal distribution of the residuals is no longer of concern. Therefore, histograms

and QQ-plots of the residuals should be interpreted in terms of how well the model fits the data rather than the normality of the residuals. Several examples of these model validation techniques are shown in the case study chapters.

Overdispersion revisited

The overdispersion parameter in a Poisson GLM is estimated using Pearson residuals. The exact formula is

$$\rho = \frac{1}{n-p} \frac{\sum_{i=1}^{n} (Y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

There are various reasons why the model might give large Pearson residuals, for example:

1. There are observations with large values that are a poor fit with the model: outliers. Particularly observations with fitted values close to zero.
2. There is a model misspecification. An important explanatory variable is missing, interaction terms are not added, or there is a non-linear effect of an explanatory variable. In the latter case, the solution is to apply a transformation on the explanatory variable, add quadratic terms, or use smoothing components (e.g., GAM).
3. Violation of the independence assumption. There is correlation between the observations. In this case, random effect models (Chapter 8) or generalised linear mixed models (Fitzmaurice et al. 2004) can be used.
4. There is clustering of samples.
5. The wrong link function has been used.

The overdispersion correction gives the following model:

$$E[Y_i] = \mu_i \quad \text{and} \quad Var[Y_i] = \rho\mu_i$$

Although it is common in ecology to have datasets with a large overdispersion, it is unclear how large a value of ρ is acceptable. Some authors report that $\rho = 5$ or 10 is large, and other authors use overdispersion parameters of 50 or more. We suggest tackling overdispersion by trying to improve the systematic part of the model using GAM or by adding interactions. However, this is not always the best solution. A GLM modelling approach that allows for correlation is generalised estimation equations (GEE). These were introduced by Liang and Zeger (1986), and a useful introductory paper is Hardin and Hilbe (2003). GEE is similar to generalised least squares (GLS) and is an extension of regression models where a non-diagonal error covariance matrix is used. The underlying principle in GLS can be found in Greene (2000). An alternative to GEE is generalised linear mixed modelling (GLMM), which is a combination of GLM and mixed effects modelling. Currently, software for GLMM and its generalised additive modelling equivalent (GAMM) is being developed and improved by the scientific community (Wood 2006). It is likely that GLMM and GAMM will become a standard tool in the

ecologist's statistical toolbox. Yet, another way to deal with overdispersed count data is Zero Inflated Poisson (ZIP) models (Tu 2002). The underlying model is

$$Y_i \sim \begin{cases} 0 & \text{with prob. } p_i \\ P(\mu_i) & \text{with prob. } 1 - p_i \end{cases}$$

Both components of the model are modelled in terms of the explanatory variables and are fitted simultaneously. Yet, another possible solution for overdispersion is to use the negative Binomial distribution (Lindsey 2004) instead of the Poisson distribution. It is also possible to use models of the form:

$$E[Y_i] = \mu_i \quad \text{and} \quad Var[Y_i] = \rho\mu_i^2$$

It can be used if the variance is considerably larger than the mean.

6.2 Logistic regression

In the previous section, we used the species richness as the response variable. This diversity index measures the number of different species per site giving a response variable with a vector of length 45 (number of observations) containing non-negative integers. Although Poisson regression is suitable for modelling count data, a different approach is needed for 0–1 and proportional data. An example of such data is where the response variable is a vector of ones and zeros representing presence and absence of a particular species at a site. For this type of data we need to consider logistic regression. The next few paragraphs are largely based on ideas from Pampel (2000).

Later in this book we look at a case study using the flatfish *Solea solea* measured in the Tagus estuary in Portugal. The data for this study include the abundances of *S. solea* and several explanatory variables, such as salinity and mud content measured at 61 sites in the estuary. The data were noisy and to compensate for the noise they were transformed to presence–absence data. Figure 6.4 shows a scatterplot of *S. solea* versus salinity. Define P_i as the probability that *S. solea* is present at site i, and $1 - P_i$ as the probability that it is not present at the site. A more formal notation is $P_i = P(Y_i = 1)$. We will model P_i as a function of the explanatory variables. However, P_i is always between 0 and 1 as it is a probability. The regression line in Figure 6.4 is a first attempt to estimate the probabilities P_i, but the line takes values larger than one for small salinity values. Additionally, the Gaussian density curves on top of the fitted values (not plotted here) suggest that realisations outside the 0–1 range are possible. Therefore, we need to apply a series of transformations on P_i such that the transformed values are not restricted to be in a certain interval. We can then back-transform so that the original probabilities are between 0 and 1. So, in logistic regression the transformed values (which can take any value) are modelled as a function of the explanatory variables.

The odds of an event occurring (i.e., the presence of *S. solea*) can be defined as

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

Table 6.1 shows the value of the odds for various P_i values. If the probability that *S. solea* is measured at site i is equal to $P_i = 0.5$, then $1 - P_i$ is also 0.5 and the odds are 1. For larger values of P_i, the odds become larger as well, whereas for smaller values of P_i, the odds become small, but positive. An odds of nine means that it is nine times more likely that *S. solea* will be recorded than not recorded. Or stated slightly differently, at that level of salinity you would expect to have nine records of *S. solea* being present for every ten samples. To compare two odds with each other, their ratio can be used. This is the odds ratio. If the odds ratio of two samples is close to zero, then the odds of the second sample is much higher.

Note that the odds are always larger than zero. By taking the natural logarithm of the odds, also called the log odds, we can get values that are not restricted to lie between 0 and 1, and negative values are possible. The log odds are also symmetrically distributed around 0. Note that a small change in probabilities for P_i close to 0.5 has a different change on the log odds compared with the same change for P_i close to 1 or 0. We will visualise this later.

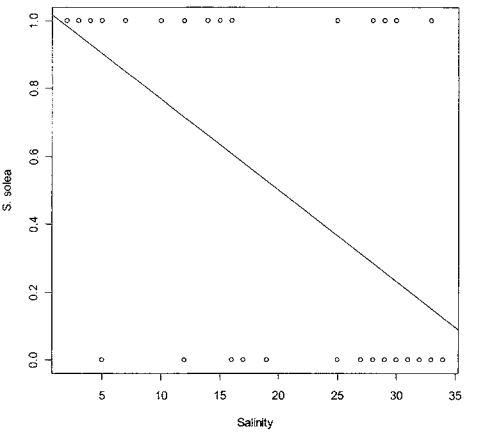

Figure 6.4. Scatterplot of *S. solea* data versus salinity. A regression line was added as a first attempt to estimate the probability of recording *S. solea*.

Table 6.1. Various probabilities, odds and log odds. The table shows how log odds are calculated from probabilities.

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
$\mathrm{Ln}(O_i)$	−6.91	−2.20	−0.85	−0.41	0	0.41	0.85	2.20	6.91

In logistic regression, the log odds are modelled as a linear function of the explanatory variables

$$\ln(O_i) = \ln(\frac{P_i}{1-P_i}) = g(x_i)$$

where $g(x_i) = \alpha + \beta_1 X_{1i} + \cdots + \beta_p X_{pi}$ is a linear function of the p explanatory variables. Using simple algebra, it follows that:

$$O_i = \frac{P_i}{1-P_i} = e^{g(x_i)} \Rightarrow P_i = \frac{e^{g(x_i)}}{1+e^{g(x_i)}} \tag{6.2}$$

It is easily verified that P_i is always between 0 and 1, whatever the value of the function $g(x_i)$. This provides a framework that gives fitted probabilities between 0 and 1. However, compared with linear regression and Poisson regression, we also need to replace the Gaussian (or Poisson) density curves by something more appropriate. More 'appropriate' means the realisations should be equal to 0 or 1, or between 0% and 100% for proportional data and the Bernoulli and Binomial distribution should be used. Choosing between the binomial or the Bernoulli distribution depends on the number of samples per X value. First, we assume there is only one observation per sample: $n_i = 1$. In this case the logistic regression model is of the form:

$$Y_i \sim \mathrm{B}(1,P_i) \quad \text{and} \quad E[Y_i] = P_i = \mu_i = \frac{e^{g(x_i)}}{1+e^{g(x_i)}}$$

$\mathrm{B}(1,P_i)$ is a Bernoulli distribution, and the variance of Y_i is given by $P_i(1-P_i)$. Before discussing the interpretation of the regression parameters, we look at an example of logistic regression using the *S. solea* data. The following model was used:

$$Y_i \sim \mathrm{B}(1,P_i) \text{ and } E[Y_i] = P_i = \mu_i = \frac{e^{g(x_i)}}{1+e^{g(x_i)}} \text{ where } g(x) = \alpha + \beta \times \text{Salinity}_i.$$

Figure 6.5 shows the model fit of the logistic regression model and the numerical output is given below.

	Estimate	Std. Error	t-value	p-value
(Intercept)	2.66	0.90	2.95	0.003
sal	−0.12	0.03	−3.71	<0.001
Null deviance: 87.49 on 64 degrees of freedom				
Residual deviance: 68.56 on 63 degrees of freedom. AIC: 72.56				

The t-values (or p-values) indicate that Salinity is significantly different from 0 at the 5% level. The fitted curve in Figure 6.5 shows the typical S-shape of a logistic regression curve. Note that at P_i values around 0.5, the rate of change in probabilities is larger than the P_i values close to 0.9 and 0.2. Stated differently, for salinity values between 15 and 25, the rate of changes in the probability of observing

S. solea is larger compared with samples with smaller and larger salinity values. At extreme values of the salinity gradient, a change in salinity has less effect of the probability compared with average salinity values.

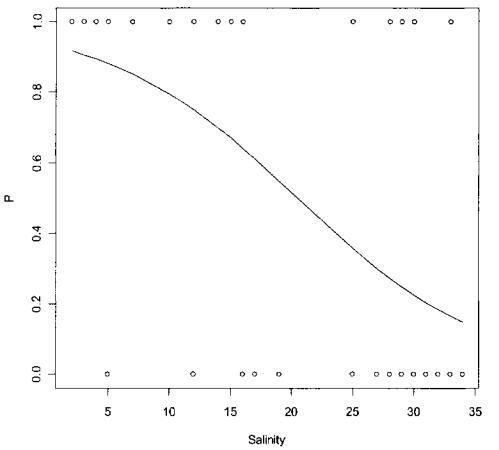

Figure 6.5. Observed and fitted values for *S. solea* data obtained by a logistic regression model.

In linear regression, interpreting a regression coefficient is relatively simple; it represents the one-unit change in *Y* for a one-unit change in X while keeping all other parameters constant. In logistic regression, this is slightly more complicated. The model we have applied to the *S. solea* data is of the form:

$$O_i = \frac{P_i}{1-P_i} = e^{\alpha+\beta*Salinity_i} = e^{2.66-0.13*Salinity_i} = e^{2.66} * e^{-0.13*Salinity_i}$$

So, the relationship between the odds and salinity is modelled in a non-linear way. If the regression parameter β is estimated as 0, then the exponentiated value is l, and has no effect on the odds. Therefore, a logistic regression parameter β that is zero has no effect on the odds. A positive regression parameter corresponds to an increase in the odds and a negative value to a decrease. For the *S. solea* data, a one-unit change in salinity matches a change in the odds of $e^{-0.13} = 0.88$. So, a one-unit change in salinity means the probability of recording *S. solea* at a site, divided by the probability of not recording it, will change by 0.88. As $e^0 = 1$ represents the effect of no change, the following formula gives the percentage increase or decrease in the odds due to a one-unit change in the explanatory variable: $(e^{\beta} - 1) \times 100$. In this case, for a one-unit change in salinity there is a 12% decrease in the odds of recording *S. solea*.

Validation graphs are shown in Figure 6.6. Panel A shows the residuals versus fitted values. Although this graph shows a distinct pattern, this is because of the

presence–absence nature of the data and it does not indicate a lack of fit. So, one string of points corresponds to the samples with a 1, and the other to the samples with a 0. This graph is still useful as it allows you to check that there are no samples with a score of 1 in the string of points for the 0's, and vice versa. This is best done by using the values of the response variable as labels (0 or 1). The QQ-plot (often printed by default) can be ignored for presence/absence data as the residuals will never be normally distributed.

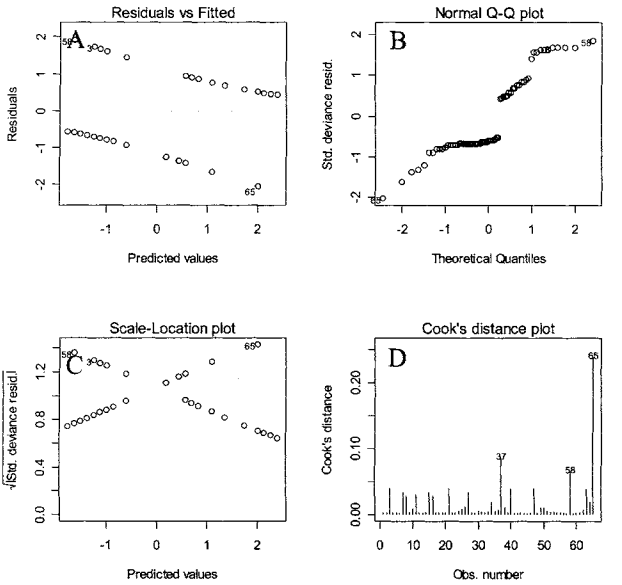

Figure 6.6. Validation graphs for *S. solea* data obtained by a logistic regression model.

Data on proportions can be expressed as Y_i successes out of n_i trials with probability P_i. To be more precise, if Z_j is Bernoulli distributed with B(1,P_{ji}), then the sum of Z_1, Z_2...Z_{ni} is Binomial distributed: B(n_i, P_i). The expectation and variance of Y_i is equal to

$$E[Y_i] = n_i P_i$$
$$Var[Y_i] = n_i P_i (1 - P_i)$$

This assumes that the n_i individuals are independent of each other, and that they all have the same probability of success, namely P_i. So, if we take 20 samples with a salinity of 15, and find 5 samples with *S. solea* present, this can be modelled as a binomial distribution with n_i = 20, a probability P_i (which is hopefully close to 5/20 = 0.25) and Y_i = 5. If similar data are available from stations with other salin-

ity values, the probability P_i can be modelled in terms of the explanatory variable(s). The logistic regression model is now given by

$$Y_i \sim \mathrm{B}(n_i, P_i) \quad \text{and} \quad E[Y_i] = P_i n_i = \mu_i = \frac{e^{g(x_i)}}{1 + e^{g(x_i)}} \quad \text{and} \quad \mathrm{Var}(Y_i) = n_i P_i (1 - P_i)$$

where n_i is the number of samples at site i. Summarising, the Binomial distribution at a site i is specified by the probability P_i and number of observations n_i. The expected value of Y_i is given by $P_i n_i$, and its variance by $P_i n_i$ $(1 - P_i)$. The link between the expectation and the linear predictor function is called the logistic link function. Just as in Poisson regression, there can be over- and under-dispersion in logistic regression with the Binomial distribution ($n_i > 1$). One of the reasons for overdispersion is if the n_i individuals are not independent (if they are positively correlated). Overdispersion can be treated in the same way as in Poisson regression, by using the quasi-likelihood method and introducing an overdispersion parameter. However, for the logistic regression model with a Bernoulli distribution ($n_i = 1$), overdispersion does not exist, and therefore, one should not apply a correction for overdispersion.

Testing the significance of regression parameters

There are two options: (i) to use the ratio of the estimated parameter and its standard error or (ii) to use the maximum likelihood ratio test. For large sample size, the ratio of the estimated parameter and its standard error follows a z-distribution. Some programmes give the Wald statistic instead, which is the square of this ratio and it follows a Chi-square distribution. There is some criticism on the use of this test (McCullagh and Nelder 1989) and it should be used very carefully. Raftery (1995) used a BIC value to test the significance of regression parameters. For each regression parameter the following value is calculated:

$$\mathrm{BIC} = z^2 - \log(n)$$

The null hypothesis is H_0: β=0, where z is the z-value (z^2 is the Wald statistic) and n the sample size. If the BIC is smaller than zero, there is no evidence to reject the null hypothesis. A BIC value between 0 and 2 indicates a weak relationship, values between 2 and 6 indicate a relationship, a strong relationship is indicated by values between 6 and 10, and a very strong relationship is indicated by values greater than 10.

Maximum likelihood

The regression parameters in ordinary regression models are estimated using the least squares method. In logistic regression, this is done with maximum likelihood, and a short discussion is presented next. Readers not interested in the statistical background can skip this section. Suppose we toss a coin 10 times and obtain 3 heads and 7 tails. Let P be the probability that a head is obtained, and $1 - P$ the

probability for a tail. Using basic probability rules, the probability for 3 heads and 7 tails is

$$\text{P(3 heads and 7 tails)} = \frac{10!}{3!7!}P^3(1-P)^7$$

If the coin is fair, you would expect P=0.5. However, suppose we do not know whether it is fair and want to know the most likely value for P. What value of P makes the probability of getting 3 heads and 7 tails as large as possible? Table 6.2 shows the value of the probability of 3 heads and 7 tails for various values of P. For P = 0.3 the probability has the highest value, hence P = 0.3 gives the highest probability for 3 heads and 7 tails. This is the underlying principle of maximum likelihood estimation.

Table 6.2. P(3 heads and 7 tails) for various values of P.

P	P(3 Heads and 7 Tails)	P	P(3 Heads and 7 Tails)
0.1	0.057	0.6	0.042
0.2	0.201	0.7	0.009
0.3	0.267	0.8	0.001
0.4	0.215	0.9	0.000
0.5	0.117	1.0	0.000

In logistic regression, we can also formulate a probability for finding the data:

$$L = \prod_{i=1}^{n} P_i^{Y_i}(1-P_i)^{1-Y_i}$$

where Y_i takes binary values translated to 0 or 1. In the coin tossing example, a head is Y_i = 1 and tail is Y_i = 0. Tossing the coin 10 times might give the Y sequence of 1 0 1 0 0 0 0 0 1 0 (3 heads and 7 tails). Except for the factorial values, this is the same probability as in the previous paragraph. Assuming the probability P(head) does not vary, the question is identical as above; what is the value of P such that the probability of 3 heads and 7 tails is the highest? Recall that in logistic regression P is a function of the regression parameters:

$$P_i = \frac{e^{g(x_i)}}{1+e^{g(x_i)}}$$

The new question is now: what are the values of the regression parameters such that the probability L of the observed data is the highest? Note that L is always between 0 and 1 because probabilities are multiplied with each other. To avoid problems with the 0–1 range (regression lines going outside this range), the natural log is taken of L, resulting in

$$L' = \log(L) = \sum_{i=1}^{n} Y_i \log(P_i) + (1 - Y_i)\log(1 - P_i)$$

Note that $\log(L)$ is now between minus infinity and 0, and the upper bound corresponds to a high probability for the data (which is what we want). To find the regression parameters that produce the highest value for the log likelihood, fast non-linear optimisation routines exist, called the iteratively weighted least squares (IWLS) algorithm. This algorithm is described in McCullagh and Nelder (1989), Garthwaite et al. (1995) or Chambers and Hastie (1992), and it is not discussed here. Some of these textbooks are mathematically oriented, but of the three, Chambers and Hastie is the least mathematically demanding.

To assess the significance of regression parameters, the log likelihood value L' itself is not useful. It is more useful to compare L' with the log likelihood value obtained by a reference model. Two candidates for this reference model are the null model and the saturated model. In the null model, we only use the intercept α in the linear predictor function $g(x_i)$. So, this is a poor model and the difference between both L' values only shows how much better our model performs compared with the worst-case scenario. The alternative is to compare L' with the log likelihood of a saturated model that produces an exact fit. The difference between the two log likelihood values show us how much worse our model is compared with the 'perfect' model. The deviance is defined as

$$D = 2 \times (L' \text{ saturated model} - L' \text{ model})$$

This value will always be positive, and the smaller the better. It can be shown that D is asymptotically Chi-square distributed with $n - p$ degrees of freedom (p is the number of parameters and n is the number of observations). Just as in Poisson regression, you can calculate the deviance of two nested models. The difference between the two deviances is again an asymptotically Chi-square distribution, but now with $p_1 - p_2$ degrees of freedom where p_1 and p_2 are the number of parameters in the two nested models. For small datasets ($n < 100$), the deviance test gives more reliable results compared with the z- or Wald test. Just as for Poisson regression, the Chi-square test needs to be replaced by an F-test in case of overdispersion.

Example

Returning to the *S. solea* data we can use the following logistic regression model:

$$Y_i \sim \mathrm{B}(1, P_i) \quad \text{and} \quad E[Y_i] = P_i = \frac{e^{g(x_i)}}{1 + e^{g(x_i)}}$$

where $g(x) = \alpha + \beta_1 \times \text{Salinity}_i + \beta_2 \times \text{Temperature}_i$. This model states that the probability that *S. solea* is measured at a particular site i follows a binomial distribution with probability P_i. This probability is a function of salinity and temperature. The numerical output is given by

	Estimate	Std. Error	*t*-value	*p*-value
(Intercept)	5.21	3.52	1.48	0.13
temp	−0.10	0.13	−0.75	0.44
sal	−0.14	0.03	−3.60	<0.001
Null deviance: 87.49 on 64 degrees of freedom				
Residual deviance: 67.97 on 62 degrees of freedom. AIC: 73.97				

The deviance is $D = 67.97$ (called the residual deviance in the printout). The *z*-value indicates the regression parameter for temperature is not significantly different from 0. The BIC value is −3.58, which confirms the non-significance. For salinity, the BIC value of 8.87 suggests a strong relationship between salinity and *S. solea*. The deviance of the full model is 67.97, and the deviance of the model without temperature is 68.56 (see above). The difference is 0.58, and this is asymptotically Chi-square distributed with 1 degree of freedom. This is not significant at the 5% level (the *p*-value is 0.44[1]), indicating that temperature can be dropped from the model. Just as in Poisson regression, the AIC can be used and in this instance gives the same result for both of these models.

[1] The R command 1-pchisq(Statistic,df) can be used for this. For an *F*-distribution (in case of overdispersion), use 1-pf(Statistic,df_1,df_2).