CHAPTER 7

Analysis of presence or absence of species

Analysis of presence or absence of species

In the previous chapter, we saw how species counts data can be analysed. In this chapter, we describe how data can be analysed that simply indicate whether a species is present in certain sites or absent. As for the analysis of species counts data, the data are analysed for one species at the time.

Analysis of presence or absence by cross-tabulations

As in the previous chapter, we will use a dataset that was collected in Panama, containing information on the abundance of *Faramea occidentalis*. This dataset also has observations for the environmental variables precipitation (quantitative), altitude (quantitative), age (ordinal) and geology (categorical). The dataset is provided in the previous chapter.

Imagine that you had a hypothesis that age had an influence on the chance that species *Faramea occidentalis* was present on a site. We treat age as a categorical variable – since it is an ordinal variable, we can choose whether we treat it as quantitative or categorical variable in subsequent analysis.

In a cross-tabulation analysis, you first need to count the number of sites of each category where the species occurs, and the number of sites where the species does not occur. You obtain these results by doing a cross-tabulation of species presence-absence with the age categories:

```
        c1 c2 c3
FALSE    5  6 12
TRUE     9  5  6
```

In the table, the rows indicate whether the species is absent (FALSE, based on the test whether the abundance > 0) or present (TRUE, based on the same test whether the abundance was > 0). The columns represent the three age categories. This table is a **cross-tabulation** or a **contingency table**. The cells in the table are counts of the number of observations within the specified categories of rows and columns.

We can also present these results graphically as in Figure 7.1.

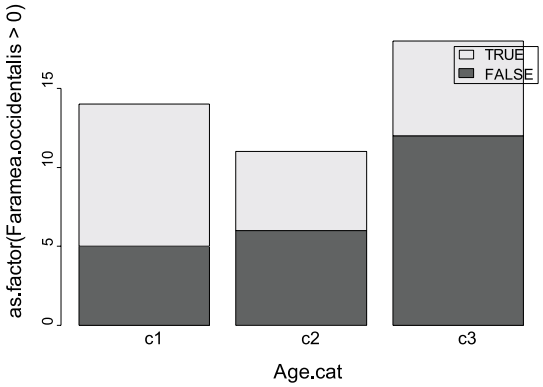

Figure 7.1 Observed frequencies for the presence and absence of *Faramea occidentalis* on three categories of age.

You can see that for sites of age category 1, the species occurs in 9 of 14 (14 = 5 +9) cases. Similarly, the species occurs in 5 of 11 sites of age category 2, and 6 of 18 sites of age category 3. For age category 1, we can use this information to calculate that the species has 9 / 14 × 100% = 64% chance of being present when the site is of age category 1. Using the same method, we can use the information to calculate a chance of 5 / 14 × 100% = 36% of being absent on sites of age category 1. From the table we can therefore calculate the chance that the species is present or absent on sites of a certain category, if sites are selected randomly. When we also calculate the chances for the other categories, we can calculate the following table:

Chances	Age category 1	Age category 2	Age category 3
Chance that the species is present (%)	64.3	45.5	33.3
Chance that the species is absent (%)	35.7	54.5	66.7

To find out whether the proportions are significantly different from each other, you can use a Chi-squared test. The result that you obtain will be:

```
        Pearson's Chi-squared test
data:  cross
X-squared = 3.0393, df = 2, p-value = 0.2188
```

This result shows that there is no evidence that differences in proportions exist between the three age categories. The significance level of P = 0.2188 indicates that there is a large chance that the differences in proportions are an effect from the random sampling of the sites from the survey area.

The Chi-squared test is limited in several ways. First of all, it is a test and is therefore not explicit in providing estimated or predicted values (the chances that were calculated earlier are not generated by the Chi-squared test). Secondly, the Chi-squared test can only be used to analyse the effect of a single categorical variable. Finally, the test is based on some assumptions that may not be reasonable.

The conditions for the Chi-squared test to be reliable are that the **expected frequencies** (expected when there is no relationship between the two variables that generated the crosstab) are not too small. How are the expected frequencies calculated and how do we evaluate whether some expected values are too small? The expected frequencies are calculated by multiplying the chance that the species is present or absent by using frequencies from the entire dataset with the number of sites of each age category. For the entire dataset of 43 sites, species presence was observed in 20 (=9 + 5 + 6) sites or 46.5% (20 / 43 × 100%) of all sites. The number of sites that are expected to contain the species for age category 1, when there is no relationship between age and presence-absence, is therefore 0.465 × 14 (=9 + 5) = 6.51. The expected frequencies can be calculated for each combination of presence-absence and age category as:

```
             c1       c2       c3
FALSE 7.488372 5.883721 9.627907
TRUE  6.511628 5.116279 8.372093
```

The condition for the Chi-squared test to provide reliable results is that all the expected frequencies should be larger than 5 (a less strict condition is that more than 20% of expected frequencies should be larger than 5, but none should be smaller than 1). Since all expected frequencies are larger than 5, we can rely on the Chi-squared test to have provided a reliable result. The Chi-squared test actually estimates the probability that the measured and expected frequencies will be the same for the survey area.

Another criticism of the Chi-squared test is that it ignores the ordering of the age categories – the table above showed that there is a trend of decreasing chance of encountering the species when the age of the plot is greater, but the Chi-squared test does not investigate the ordering of the age categories.

Analysis of presence or absence through binomial GLM

We can investigate the same hypothesis that there is an influence of age category on the presence-absence of *Faramea occidentalis* with a binomial GLM with logit link.

As we saw in the previous chapter, a GLM is defined by the variance function and the link function. When we define the variance and link functions, we assume that these functions are realistic descriptions for the dataset that we are investigating.

The logit link is defined as:

$$logit(\mu) = \log(\mu / (1 - \mu)) = a + b_1 \times x_1 + b_2 \times x_2 + b_3 \times x_3 + \ldots$$

The logit link function is one way of guaranteeing that the predicted values will be between 0 and 1, which is appropriate since we want to predict probabilities of presence of *Faramea occidentalis* which also are between 0 and 1. In the previous chapter where we analysed counts, we used a log link that ensured that the predicted values were larger than 0, but not that the predicted values were smaller than 1.

By investigating the hypothesis with a GLM, we overcome some of the shortcomings of the Chi-squared test: the GLM will provide predictions, and several explanatory variables can be analysed – including quantitative variables. Not all shortcomings of the Chi-squared test are overcome, however: the large sample assumptions are still needed for the tests of the GLM to provide realistic results.

The binomial GLM with logit link investigating the influence of age on presence-absence yields the following results:

```
glm(formula = Faramea.occidentalis > 0 ~ Age.cat, family = binomial(link = logit),
    data = faramea, na.action = na.exclude)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.5878     0.5578   1.054   0.2920
Age.catc2    -0.7701     0.8233  -0.935   0.3496
Age.catc3    -1.2809     0.7491  -1.710   0.0873 .
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 59.401  on 42  degrees of freedom
Residual deviance: 56.322  on 40  degrees of freedom
AIC: 62.322

Analysis of Deviance Table
Terms added sequentially (first to last)
        Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                       42     59.401
Age.cat  2    3.079        40     56.322     0.214
```

The results of the GLM first show the coefficients that were calculated. As for the analysis of counts data, the first category is not included explicitly in the results. The results for the first category correspond to the intercept, however.

As we saw in the previous chapter, it is a bit complicated to directly calculate the expected values from the estimations of the coefficients. The reason is that the inverse link function needs to be calculated to obtain the expected values. In the case of the logit link, the inverse logit is calculated as $y = exp(x)/(1+exp(x))$. However, the program that fits the model should be able to provide the predicted values. Here we obtain following predictions for the three categories: 0.64, 0.45 and 0.33. You could calculate these results yourself by using the inverse logit as in the box below.

You can see that you obtain the same predictions with the GLM as the chances for presence of the species that we calculated earlier. For example, the chance of presence for sites with age category 2 is calculated in both instances to be 45.5%.

The large significance level values estimated for the regression coefficients indicate that there is no evidence for significant differences between the predicted chances, however. The ANOVA table provides similar information by estimating a large significance level ($P = 0.21$). The ANOVA table also provides important information on the deviance that is explained: the model only explains 3.079 or 5.2% of total or null deviance (59.401).

The Chi-squared test of the ANOVA table is a test for the same pattern that was tested at the beginning of this chapter by the Chi-squared test for the contingency table. You could check that the explained deviance (3.079) is very similar to the Chi-squared statistic (3.039) – calculating the deviance is actually another method for analysing a cross-tabulation (called a G-test in some manuals). What is important is that both the ANOVA and the earlier Chi-squared test estimate a similar significance level ($P = 0.214$ and $P = 0.2145$), leading to the same conclusion of no evidence for an influence of age on the presence-absence. You can also see the limitations of the Chi-squared test – it is as if you ignore all other results from the GLM.

The results from the model can also be analysed graphically as in Figure 7.2. Because the observed values are either 0 (absence) or 1 (presence), the figure is not very informative and mainly shows the predicted chances for presence of the species for each age category. The wide overlap between the 95% confidence intervals shows that there is no evidence for an effect of age.

As with all regression models, the binomial GLM makes various assumptions about the data. Whether you can trust the results depends on whether these assumptions are realistic.

The quasi-binomial model was developed for situations where one of the assumptions of the binomial model does not hold – that the dispersion equals 1. As we saw in the previous chapter, for our type of data the dispersion parameter is an indication of how randomly individuals are distributed. A dispersion parameter of 1 means

```
Expected values:

Age category x: inverse logit(intercept + coefficient for age category x)

Age category 1: inverse logit(0.5878+0) =
     exp(0.5878+0) / (1 + exp (0.5878+0) ) = 0.6428602

Age category 2: inverse logit(0.5878-0.7701) =
     exp (0.5878-0.7701) / (1 + exp (0.5878-0.7701) ) = 0.4545508

Age category 3: inverse logit(0.5878-1.2809) =
     exp (0.5878-1.2809) / (1 + exp (0.5878-1.2809) ) = 0.3333438
```

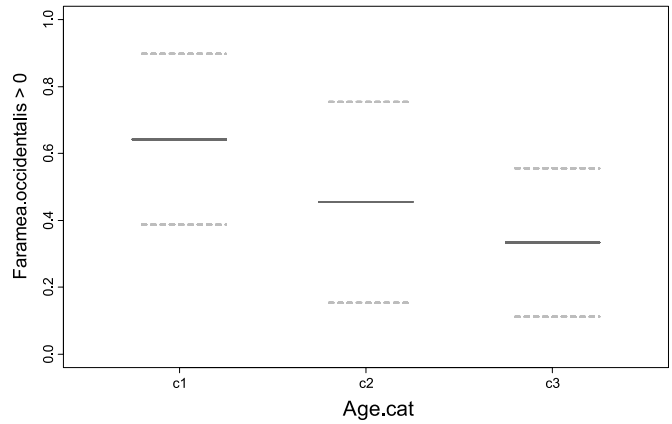

Figure 7.2 Predicted values (horizontal lines) for the binomial GLM with logit link of the presence-absence of *Faramea occidentalis* on age category. Dashed lines are 95% confidence intervals for the mean.

that the individuals are randomly distributed over the sample units. When dispersion is not 1, the GLM will estimate significance levels that are not realistic. A quasi-binomial GLM will estimate the same regression coefficients as the binomial GLM, but will estimate the dispersion parameter and use this dispersion parameter to provide different estimates of standard errors and significance levels.

The quasi-binomial GLM with logit link investigating the influence of age on presence-absence yields the following results:

```
glm(formula = Faramea.occidentalis > 0 ~ Age.cat, family = quasibinomial(link = logit),
    data = faramea, na.action = na.exclude)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4350  -1.0008  -0.9005   1.0979   1.4823

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.5878     0.5783   1.016    0.316
Age.catc2    -0.7701     0.8536  -0.902    0.372
Age.catc3    -1.2809     0.7767  -1.649    0.107

(Dispersion parameter for quasibinomial family taken to be 1.075000)

    Null deviance: 59.401  on 42  degrees of freedom
Residual deviance: 56.322  on 40  degrees of freedom
AIC: NA


Analysis of Deviance Table

        Df Deviance Resid. Df Resid. Dev      F Pr(>F)
NULL                     42     59.401
Age.cat  2    3.079        40     56.322 1.4322 0.2508
```

When we compare the results from the quasi-binomial GLM with the binomial GLM, we can see that the quasi-binomial model estimated the dispersion parameter to be 1.075. Since there is a small difference between 1.075 and 1, there will not be any substantive difference in conclusions whether the dispersion is fixed as 1 as in the binomial, or estimated as in the quasibinomial. It is possible to test whether the estimated dispersion is different from 1, and perhaps infer something on the clumpiness of the distribution on the basis of the results. However, such tests have the usual problems: they depend on the validity of the models assumed, and the results depend on both the sample size and the dispersion. Probably a better approach is to first think whether overdispersion is likely given the source of the data, and pay careful attention when it is. Then check whether the results differ in substance between the two models. If interest is really into the extent to which the distribution is clumped, regular or random, then there are better methods which focus on this aspect.

Since the dispersion parameter was estimated to be very close to one, only small differences in the estimated significance levels can be observed (for instance $P = 0.25$ in the ANOVA instead of $P = 0.21$) and both models lead to the same conclusions. Note that it is a good practice to check for the difference between the results of the binomial and the quasi-binomial GLM. A disadvantage of the quasi-binomial GLM implemented here is that it does not calculate the Akaike Information Criterion (AIC), which can be used for model selection (see previous chapter and below: binomial GLM with several explanatory variables). An advantage of the quasi-binomial GLM is that it provides better estimates of significance level when the dataset has a dispersion parameter that is more different from 1.

Binomial and quasi-binomial GLM with continuous variables

We have seen so far that analysis of a binomial or quasi-binomial GLM with logit link provides a more comprehensive analysis of frequencies than the Chi-squared test (which is only a test) and that estimation is explicit in the GLM. Another advantage of the GLM approach is that the explanatory variables can either by continuous or categorical.

To analyse a cross tabulation with continuous explanatory variables with a Chi-squared test, you need to derive a categorical variable from the continuous variable. You need to define several categories, for instance a category for altitude < 200 m and another category for altitude > 200 m. These categories need to be defined by the researcher, and there is no procedure that can tell you how these categories should be defined. With the binomial GLM, you do not need to make this decision. More importantly, when the relationship between the explanatory variable and the response variable is not a stepwise function but a gradual change, it is better to model the gradual change. The binomial GLM will accommodate gradual changes for a continuous variable, which often provides a more realistic model.

We can for example model the presence and absence of *Faramea occidentalis* based on the precipitation of a site with a quasi-binomial GLM with logit link. The results that you will obtain are shown in the box on the next page.

You could notice that we used a quasi-binomial model. The dispersion parameter is estimated to be 0.987, very close to 1. Both the binomial and the quasi-binomial GLM therefore lead to the same conclusions.

As in the regression results that we saw earlier, regression coefficients are provided for the explanatory variable. We can see that there is evidence that precipitation has an effect, since the significance level calculated for the coefficient is low ($P = 0.0172$). We can also infer that there is an effect of precipitation from the low significance level of the ANOVA table ($P = 0.005$).

```
glm(formula = Faramea.occidentalis > 0 ~ Precipitation, family = quasibinomial(link =
logit), data = faramea, na.action = na.exclude)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7303  -1.0431  -0.3289   1.1157   1.7268

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.948352   2.828347   2.457   0.0183 *
Precipitation -0.002721   0.001095  -2.484   0.0172 *

(Dispersion parameter for quasibinomial family taken to be 0.9878437)

    Null deviance: 59.401  on 42  degrees of freedom
Residual deviance: 50.561  on 41  degrees of freedom
AIC: NA


Analysis of Deviance Table

              Df Deviance Resid. Df Resid. Dev      F    Pr(>F)
NULL                              42     59.401
Precipitation  1    8.841        41     50.561 8.9494 0.004682 **

---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

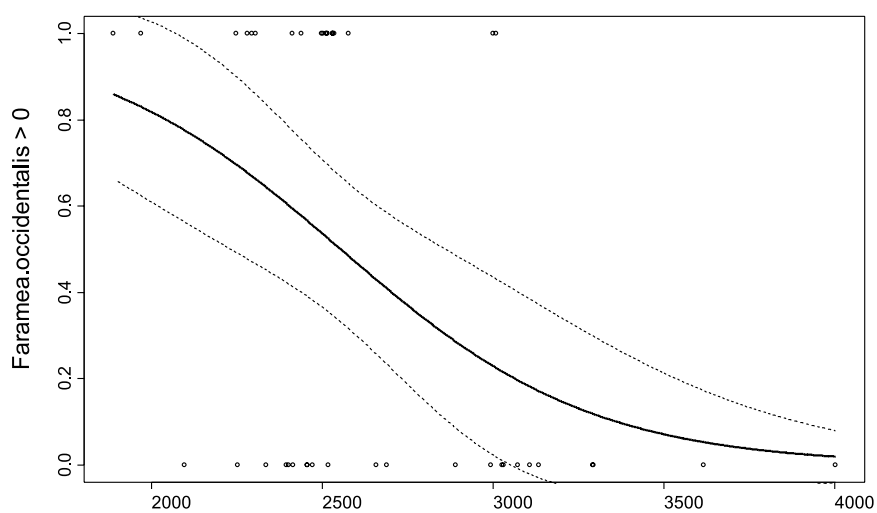

Figure 7.3 Observed values (circles) and predicted values (connected by line) for the quasi-binomial GLM model of the presence-absence of *Faramea occidentalis* on elevation. Dashed lines are 95% confidence intervals for the mean.

The predictions of the model can be shown graphically as well. Figure 7.3 provides the observed and the predicted values.

You can see the merit of the logit link in not predicting values outside of the interval with 0 and 1 as boundaries. By using the logit link, the GLM model does not predict values that we do not expect. The plot of the observations is more informative for a continuous than for a categorical variable: you could observe now that the species was never observed at the highest precipitation levels, and that it was never not observed at the lowest precipitation levels. At intermediate levels, the species was sometimes observed. This pattern is modelled as an S-shaped curve, which provides a reasonable fit to the data.

Another advantage of the binomial and quasi-binomial GLM with logit link is that several explanatory variables can be investigated. An example is provided later in this chapter.

Binomial GAM with several explanatory variables

As for the regression models for count data, you can also fit a Generalized Additive Model (GAM) using the same variance and link functions that are used in a GLM. The GAM allows estimation of a smooth relationship between the response and a quantitative explanatory variable. The smoothing function will generate a curve that can flow more freely in between the data than a straight line.

When we calculate a quasi-binomial GAM with logit link for the presence and absence of *Faramea occidentalis* using smoothing functions of precipitation and elevation, and the categorical variables geology and age category, then we obtain the following results:

```
Family: quasibinomial
Link function: logit

Formula:
Faramea.occidentalis > 0 ~ s(Precipitation) + Geology + Age.cat +
    s(Elevation)

Parametric coefficients:
              Estimate  std. err.    t ratio    Pr(>|t|)
(Intercept)    -382.04      8.204     -46.57    < 2.22e-16
  GeologyTb     30.746      314.9    0.09764    0.92287
 GeologyTbo     14.571  2.499e+04   0.000583    0.99954
  GeologyTc     891.93       20.1      44.37    < 2.22e-16
 GeologyTcm       15.3       3795   0.004032    0.9968
 GeologyTgo     7.3746      1.204      6.126    9.6953e-07
  GeologyTl     137.55  1.114e+04    0.01234    0.99023
  Age.catc2    -103.74       1434   -0.07235    0.9428
  Age.catc3    -88.072      2.189     -40.23    < 2.22e-16

R-sq.(adj) =      1   Deviance explained =  100%
GCV score = 5.0022e-06   Scale est. = 3.4996e-06  n = 43

Analysis of deviance table

Parametric Terms:
               df       chi.sq     p-value
Geology         6       1986.4     < 2.22e-16
Age.cat         2       1618.3     < 2.22e-16

Approximate significance of smooth terms:
                       edf       chi.sq     p-value
s(Precipitation)     2.917       2252.5     < 2.22e-16
    s(Elevation)         1       1499.4     < 2.22e-16
```

The quasi-binomial GAM with logit link indicates that all explanatory variables contribute to explaining the deviance in the presence-absence of the species. The `edf` indicate the estimated degrees of freedom for the smoothing functions. The fact that 3 degrees of freedom are estimated for precipitation shows that a complex pattern was modelled for this explanatory variable.

Since our dataset was quite small, all deviance was explained, and the significance level values that were estimated were ridiculously small, we need to treat the results with caution. We expect that the model overfitted the data, although it is hard to see with presence-absence data and several variables. We therefore opted to analyse the dataset further with a GLM in the next section to find out whether there was further evidence for the complex pattern between presence-absence and the quantitative variables. An alternative approach would have been to fix the degrees of freedom for the smoothing terms in a GAM, forcing the curve to be smoother and not have such a complex shape. The results with 2 degrees of freedom for precipitation and elevation look more reasonable:

```
Family: quasibinomial
Link function: logit

Formula:
Faramea.occidentalis > 0 ~ s(Precipitation, k = 2, fx = T) +
    Geology + Age.cat + s(Elevation, k = 2, fx = T)

Parametric coefficients:
              Estimate  std. err.    t ratio    Pr(>|t|)
(Intercept)    -77.189      39.05     -1.976    0.057364
  GeologyTb     1.4079      1.286      1.094    0.28250
 GeologyTbo     31.498       1023    0.03078    0.97565
  GeologyTc     24.648       9.48        2.6    0.014332
 GeologyTcm     28.945      396.8    0.07294    0.94234
 GeologyTgo    -2.0587       2.03     -1.014    0.3187
  GeologyTl     13.802      634.6    0.02175    0.9828
  Age.catc2      -16.9      88.73    -0.1905    0.85022
  Age.catc3    -4.1095       1.93     -2.129    0.041593

R-sq.(adj) =  0.695   Deviance explained = 75.8%
GCV score = 0.70325   Scale est. = 0.49064   n = 43


Analysis of deviance table

Parametric Terms:
               df       chi.sq     p-value
Geology         6       8.8875     0.21822
Age.cat         2       4.5391     0.12083

Approximate significance of smooth terms:
                       edf       chi.sq     p-value
s(Precipitation)         2       6.2923     0.057491
    s(Elevation)         2       3.8408     0.16414
```

Binomial GLM with several explanatory variables

As already shown for the GAM, it is possible to analyse several explanatory variables together in a GLM – this was also shown in the previous chapter.

Based on the results of the quasi-binomial GAM with logit link that showed curved relationships between elevation, precipitation and the presence-absence of *Faramea occidentalis*, we fitted a second-order polynomial model for the quantitative variables (see previous chapter). The results of the binomial GLM with logit link of the presence-absence of *Faramea occidentalis* on geology, age category and the second-order polynomials of precipitation and elevation are:

```
glm(formula = Faramea.occidentalis > 0 ~ Precipitation + I(Precipitation^2) +
    Geology + Age.cat + Elevation + I(Elevation^2), family = binomial(link = logit),
    data = faramea, na.action = na.exclude)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-2.031e+00  -2.753e-02  -2.107e-08   8.387e-02   1.977e+00

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.137e+02  8.737e+01  -1.302    0.193
Precipitation       1.006e-01  7.594e-02   1.324    0.185
I(Precipitation^2) -2.223e-05  1.644e-05  -1.352    0.176
GeologyTb           1.401e+00  1.718e+00   0.815    0.415
GeologyTbo          2.962e+01  1.075e+04   0.003    0.998
GeologyTc           1.266e+01  8.055e+00   1.572    0.116
GeologyTcm          2.642e+01  4.153e+03   0.006    0.995
GeologyTgo         -1.270e+00  2.709e+00  -0.469    0.639
GeologyTl           1.792e+01  7.274e+03   0.002    0.998
Age.catc2          -9.695e+00  1.024e+01  -0.946    0.344
Age.catc3          -2.983e+00  2.458e+00  -1.214    0.225
Elevation           8.701e-02  1.161e-01   0.749    0.454
I(Elevation^2)     -4.493e-04  5.293e-04  -0.849    0.396

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 59.401  on 42  degrees of freedom
Residual deviance: 17.376  on 30  degrees of freedom
AIC: 43.376

Analysis of Deviance Table

Model 1: Faramea.occidentalis > 0 ~ 1
Model 2: Faramea.occidentalis > 0 ~ Precipitation + I(Precipitation^2) +
    Geology + Age.cat + Elevation + I(Elevation^2)
  Resid. Df Resid. Dev Df Deviance      F    Pr(>F)
1        42     59.401
2        30     17.376 12   42.025 3.5021 3.298e-05 ***
```

```
Terms added sequentially (first to last)

                   Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                                    42     59.401
Precipitation       1    8.841        41     50.561     0.003
I(Precipitation^2)  1    0.722        40     49.838     0.395
Geology             6   21.144        34     28.694     0.002
Age.cat             2    7.605        32     21.089     0.022
Elevation           1    2.837        31     18.252     0.092
I(Elevation^2)      1    0.876        30     17.376     0.349

Single term deletions

                   Df Deviance    AIC    LRT   Pr(Chi)
<none>                  17.376 43.376
Precipitation       1   20.871 44.871  3.495 0.0615481 .
I(Precipitation^2)  1   21.441 45.441  4.065 0.0437749 *
Geology             6   40.674 54.674 23.298 0.0007025 ***
Age.cat             2   24.799 46.799  7.423 0.0244364 *
Elevation           1   18.020 42.020  0.645 0.4220785
I(Elevation^2)      1   18.252 42.252  0.876 0.3492853
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

The main reason for having conducted the analysis is to select variables that meaningfully contribute to explaining the deviance, with special focus on the second-order polynomial terms of precipitation^2 and elevation^2. We saw in the previous chapter that one of the criteria that can be used for selecting variables is the Akaike Information Criterion (AIC). Models with a smaller AIC are preferred over models with larger AIC. The type-II ANOVA provides a column with the AIC. We can see that the AIC for elevation^2 is smaller than the model where this variable is included (42.252 < 43.376). Based on these results, a model where elevation^2 is not included will provide a better combination of simplicity (fewer variables) and explained deviance, provided that the way that the AIC calculates the combination is the best way. The analysis of the AIC indicates what caused our problem with the first GAM results: the GAM sacrificed simplicity to explain all the deviance – this is not necessarily the best model. Because of the smaller AIC, we excluded elevation^2 from further analysis. The results of type-II ANOVA for the quasi-binomial GLM with logit link of the presence-absence of *Faramea occidentalis* on geology, age category, elevation and the second-order polynomial of precipitation now indicated that all variables could be kept in the model (see next page):

```
glm(formula = Faramea.occidentalis > 0 ~ Precipitation + I(Precipitation^2) +
    Age.cat + Geology + Elevation, family = quasibinomial(link = logit),
    data = faramea, na.action = na.exclude)


Deviance Residuals:
       Min          1Q      Median          3Q         Max
-2.253e+00  -6.919e-02  -2.107e-08   1.639e-01   1.839e+00


Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        -8.178e+01  5.183e+01  -1.578   0.1247
Precipitation       7.501e-02  4.577e-02   1.639   0.1114
I(Precipitation^2) -1.655e-05  9.835e-06  -1.683   0.1024
Age.catc2          -8.156e+00  5.234e+00  -1.558   0.1293
Age.catc3          -1.890e+00  1.363e+00  -1.387   0.1753
GeologyTb           1.846e+00  1.324e+00   1.395   0.1730
GeologyTbo          2.551e+01  9.106e+03   0.003   0.9978
GeologyTc           9.729e+00  4.800e+00   2.027   0.0513 .
GeologyTcm          2.531e+01  3.519e+03   0.007   0.9943
GeologyTgo         -3.601e-01  1.696e+00  -0.212   0.8333
GeologyTl           1.839e+01  6.435e+03   0.003   0.9977
Elevation          -1.179e-02  1.330e-02  -0.887   0.3820
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

(Dispersion parameter for quasibinomial family taken to be 0.7169455)

    Null deviance: 59.401  on 42  degrees of freedom
Residual deviance: 18.252  on 31  degrees of freedom
AIC: NA


Analysis of Deviance Table

Model 1: Faramea.occidentalis > 0 ~ 1
Model 2: Faramea.occidentalis > 0 ~ Precipitation + I(Precipitation^2) +
    Age.cat + Geology + Elevation
  Resid. Df Resid. Dev Df Deviance      F    Pr(>F)
1        42     59.401
2        31     18.252 11   41.149 5.2178 0.0001320 ***

Terms added sequentially (first to last)


                   Df Deviance Resid. Df Resid. Dev       F    Pr(>F)
NULL                                   42     59.401
Precipitation       1    8.841        41     50.561 12.3310 0.0013893 **
I(Precipitation^2)  1    0.722        40     49.838  1.0076 0.3232568
Age.cat             2    0.991        38     48.847  0.6911 0.5085612
Geology             6   27.758        32     21.089  6.4528 0.0001718 ***
Elevation           1    2.837        31     18.252  3.9577 0.0555415 .
```

```
Single term deletions
                  Df Deviance F value     Pr(F)
<none>                 18.252
Precipitation      1   21.469  5.4648 0.0260342 *
I(Precipitation^2) 1   21.969  6.3138 0.0173944 *
Age.cat            2   28.606  8.7932 0.0009443 ***
Geology            6   40.748  6.3681 0.0001904 ***
Elevation          1   21.089  4.8193 0.0357555 *

---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

We showed the results for the quasi-binomial model as this model estimated dispersion to be 0.71. As we saw before, the regression coefficients and deviances of the ANOVA tables are the same for the binomial and quasi-binomial models, but the estimated significance levels will be different.

One important pattern that you can observe in the results is that none of the significance values for the regression coefficients are small. Only the coefficient for the category of geology *Tc* has a significance level (P=0.0513) that is smallish. Despite the fact that there is no evidence from the significance levels of the regression coefficients, the model explains most of the deviance. Moreover, we used the AIC to select those variables that meaningfully contributed to explaining the deviance.

What is going on? The reason for the difference between the regression and ANOVA results is that we have few observations for most categories of geology. This pattern is depicted in Figure 7.4. As we saw at the beginning of this chapter with analyses for cross-tabulations, we will not get reliable results when the number of observations is very small for one category (technically, when the expected frequencies are very small). For category *Tbo* there was only one observation, whereas only two categories had more than 5 observations. Moreover, most categories are dominated either by presence or absence, which made it easier to obtain correct predictions. For example, by predicting for *Tbo*, *Tcm* and *Tl* that the species is present, all predictions will be correct. One of the problems with these data is that the model does not allow that predicted values are exactly 1 or exactly 0. Another problem is that since we only have 1 observation for *Tbo*, 2 observations for *Tl* and 5 observations for *Tcm*, we do not have any information to infer that the same predictions will be valid for the entire survey area – the sample size is simply too small. The sample size is even too small for *pT*, although it is dominated by sites where the species is absent.

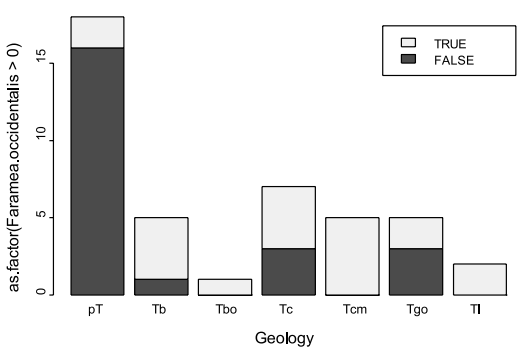

Figure 7.4 Observations of the presence-absence of *Faramea occidentalis* categorized by geology.

The fact that the small sample sizes and the similar observations within the categories of geology lead to a majority of correct predictions (and also a large amount of explained deviance since predicted and observed probabilities for presence were close), the other explanatory variables only needed to explain the odd cases such as the two sites with presence of the species on geology *pT*. This implies that sample size was also small for the other explanatory variables, so that significance levels for the regression coefficients were large.

In conclusion, we saw that the various models that we fitted explained most of the deviance, but that we do not have sufficient information to infer that similar patterns would be observed for the entire survey area. The sample size is simply too small. One possible solution could be to collapse some categories of geology into a smaller number of categories. This can only meaningfully be done if there is a logical method for combining the various categories. To further analyse the influence of geology on presence-absence with the present categories, we need to add new sites to the dataset.

Always think whether your analysis objective is realistic given the data you have. With a small number of presence-absence observations, can you expect to be able to detect and estimate the effects of 7 geology types as well as the complex, curved relationships with precipitation and elevation?

Choice of the best model

As seen at the end of the previous chapter, the most important criterion that you could use to choose between different models is the level to which the assumptions of the models are realistic. We investigated the residuals of the models that were used for count data to check the reliability of the regression models. With presence-absence data, the residuals are more difficult to investigate given that the observations were either 0 or 1. There is actually no standard method for investigating the residuals for presence-absence data.

You may also favour models with a good balance between explanatory power and simplicity. Some tests (such as the AIC) may be used to help you in selecting the model with the best balance.

However, it is not possible to provide tests or a procedure that will always select the one and only best model. The ecologist or biodiversity scientist needs to check (partially on non-statistical grounds) whether a particular model provides the best answer for the research hypothesis. The model with the largest AIC is not always the best in explaining a particular pattern. For example, when results of previous research efforts are also considered, a model with a slightly smaller AIC may be judged to be better for the particular study. All statistical models are approximations and simplifications of ecological patterns, rather than 'correct' descriptions of biological processes. So the purpose of modeling has to be considered along with statistical evidence when choosing between alternatives. For example, if average rate of response to a continuous explanatory variable is needed, a straight line model may be appropriate even if a curvature is statistically significant.

By having analysed the same dataset using the species counts as response variable in the previous chapter and transforming the counts to presence-absence in this chapter, we saw that the analysis lead to different conclusions. Since a different response variable was used, this was not a complete surprise. The different response variable was a transformation of the other response variable, however. The different results therefore showed that the transformation had an important effect, and that a different pattern prevailed. This simply means that count and presence-absence data do not necessarily follow the same pattern. If you are interested in investigating both patterns, then you need to record species counts and not simply presence or absence. Again this should follow from the initial hypotheses that you had.

References

Condit R, Pitman N, Leigh EG, Chave J, Terborgh J, Foster RB, Nuñez P, Aguilar S, Valencia R, Villa G, Muller-Landau HC, Losos E and Hubbell SP. 2002. Beta-diversity in tropical forest trees. *Science* 295: 666–669. (dataset used as example)

Fowler J, Cohen L and Jarvis P. 1998. *Practical statistics for field biology.* Chichester: John Wiley and sons.

Hastie TJ and Pregibon D. 1993. Generalised Linear Models. In: Chambers, JM and Hastie TJ. *Statistical models in S.* London: Chapman and Hall.

Jongman RH, ter Braak CJF and Van Tongeren OFR. 1995. *Data analysis in community and landscape ecology.* Cambridge: Cambridge University Press.

Legendre P and Legendre L. 1998. *Numerical ecology.* Amsterdam: Elsevier Science BV.

Pyke CR, Condit R, Aguilar S and Lao S. 2001. Floristic composition across a climatic gradient in a neotropical lowland forest. *Journal of Vegetation Science* 12: 553-566. (dataset used as example)

Quinn GP and Keough MJ. 2002. *Experimental design and data analysis for biologists.* Cambridge: Cambridge University Press. (recommended as first priority for reading)

Doing the analyses with the menu options of Biodiversity.R

Load the datasets Panama species.txt and Panama environmental.txt, and make them the species and environmental datasets, respectively. Give them the names "spec" and "faramea".

Data > Import data > from text file… (Panama species.txt)

→Enter name for data set: spec

Data > Import data > from text file… (Panama environmental.txt)

→Enter name for data set: faramea

Biodiversity > Community Matrix > Select community data set…

→Data set: spec

Biodiversity > Environmental Matrix > Select environmental data set…

→Data set: faramea

These are the original datasets, to use the reduced datasets that will be analysed, remove the sites where there is missing information on the variable "Analysed".

Biodiversity > Community matrix > Remove NA from environmental data set…

→Select variable: Analysed

As an alternative, load the dataset Faramea.txt, and make it both the species and environmental dataset (as both the species and environmental information is in the same dataset).

Data > Import data > from text file… (Faramea.txt)

→Enter name for data set: faramea

Biodiversity > Community Matrix > Select community data set…

→Data set: faramea

Biodiversity > Environmental Matrix > Select environmental data set…

→Data set: faramea

To analyse presence or absence by cross-tabs:

Biodiversity > Analysis of species as response > Species presence-absence as response…

→Model options: crosstab

→Response: Faramea.occidentalis

→Explanatory: Age.cat

To calculate a generalized linear regression model (GLM):

Biodiversity > Analysis of species as response > Species presence-absence as response…

→Model options: binomial model

→Response: Faramea.occidentalis

→Explanatory: Age.cat

→print summary

→print anova

→Plot options: diagnostic plots

→Plot variable: Age.cat

→Plot options: term plot

→Plot options: effect plot

→Model options: quasi-binomial model

To calculate a generalized additive regression model (GAM):

Biodiversity > Analysis of species as response > Species presence-absence as response…

→Model options: binomial model

→Response: Faramea.occidentalis

→Explanatory: s(Precipitation) + Geology + Age.cat + s(Elevation)

→print summary

To calculate a GLM with several explanatory variables:

Biodiversity > Analysis of species as response > Species presence-absence as response…

→Model options: binomial model

→Response: Faramea.occidentalis

→Explanatory: Precipitation + I(Precipitation^2) + Geology + Age.cat + Elevation + I(Elevation^2)

→print summary

→print anova

Doing the analyses with the command options of Biodiversity.R

Load the datasets Condit species.txt and Condit environmental.txt. Give them the names "spec" and "faramea". Alternatively, load the dataset Faramea.txt and give it the name "faramea".

```
spec <- read.table(file="D://my files/Condit species.txt")
attach(spec)
faramea <- read.table(file="D://my files/Condit environmental.
  txt")
faramea <- read.table(file="D://my files/Faramea.txt")
attach(faramea)
```

To analyse presence or absence by cross-tabs:

```
faramea$Faramea.occidentalis<<- spec$Faramea.occidentalis
table1 <- table(Faramea.occidentalis>0, Age.cat)
Presabs.1 <- chisq.test(table1)
Presabs.1
Presabs.1$observed
Presabs.1$expected
```

To calculate a generalized linear regression model (GLM):

```
Presabs.model2 <- glm(formula = Faramea.occidentalis>0 ~
  Age.cat, family = binomial(link=logit), data = faramea,
  na.action = na.exclude)
summary(Presabs.model2)
anova(Presabs.model2,test='F')
predict(Presabs.model2, type='response', se.fit=T)
null.model <- glm(formula = Faramea.occidentalis>0 ~ 1,
  family = binomial(link=logit) , data = faramea, na.action =
  na.exclude)
anova(null.model, Presabs.model2, test='Chi')
plot(Presabs.model2)
termplot(Presabs.model2, se=T, partial.resid=T, rug=T,
  terms='Age.cat')
plot(effect('Age.cat', Presabs.model2))
Presabs.model3 <- glm(formula = Faramea.occidentalis>0 ~ Age.
  cat, family = quasibinomial(link=logit) , data = faramea,
  na.action = na.exclude)
Presabs.model4 <- glm(formula = Faramea.occidentalis>0 ~
  Elevation, family = quasibinomial(link=logit) , data =
  faramea, na.action = na.exclude)
```

To calculate a generalized additive regression model (GAM):

```
Presabs.model5 <- gam(formula = Faramea.occidentalis>0 ~
  s(Precipitation) + Geology + Age.cat + s(Elevation), family
  = quasibinomial(link=logit) , data = faramea, na.action =
  na.exclude)
summary(Presabs.model5)
```

To calculate a GLM with several explanatory variables:

```
Presabs.model6 <- glm(formula = Faramea.occidentalis > 0 ~
  Precipitation + I(Precipitation^2) + Geology + Age.cat +
  Elevation + I(Elevation^2), family = binomial(link = logit)
  , data = faramea, na.action = na.exclude)
summary(Presabs.model6)
anova(Presabs.model6,test='Chi')
drop1(Presabs.model6, test='Chi')
```