13 GENERALIZED LINEAR MODELS AND LOGISTIC REGRESSION

So far, most the analyses we have described have been based around linear models that assume normally distributed populations of the response variable and of the error terms from the fitted models. Most linear models are robust to this assumption, although the extent of this robustness is hard to gauge, and transformations can be used to overcome problems with non-normal error terms. There are situations where transformations are not effective in making errors normal (e.g. when response variable is categorical) and in any case, it might be better to model the actual data rather than data that are transformed to meet assumptions. What we need is a technique for modeling that allows other types of distributions besides normal. Such a technique was introduced by Nelder & Wedderburn (1972) and further developed by McCullough & Nelder (1989) and is called generalized linear modeling (GLM). In this chapter, we will examine two common applications of GLMs: logistic regression, used when the response variable is binary, and Poisson regression, when the response variable represents counts. In the next chapter, we will describe log-linear models when both response and predictor variables are categorical and usually arranged in the form of a contingency table.

13.1 Generalized linear models

Generalized linear models (GLMs) have a number of characteristics that make them more generally applicable than the general linear models we have considered so far. One of the most important is that least squares estimation no longer applies and maximum likelihood methods must be used (Chapter 2).

A GLM consists of three components. First is the random component, which is the response variable and its probability distribution (Chapter 1). The probability distribution must be from the exponential family of distributions, which includes normal, binomial, Poisson, gamma and negative binomial. If *Y* is a continuous variable, its probability distribution might be normal; if *Y* is binary (e.g. alive or dead), the probability distribution might be binomial; if *Y* represents counts, then the probability distribution might be Poisson. Probability distributions from the exponential family can be defined by the natural parameter, a function of the mean, and the dispersion parameter, a function of the variance that is required to produce standard errors for estimates of the mean (Hilbe 1994). For distributions like binomial and Poisson, the variance is related to the mean and the dispersion parameter is set to one. For distributions like normal and gamma, the dispersion parameter is estimated separately from the mean and is sometimes called a nuisance parameter.

Second is the systematic component, which represents the predictors (*X* variables) in the model. These predictors might be continuous and/or categorical and interactions between predictors, and polynomial functions of predictors, can also be included.

Third is the link function, which links the random and the systematic component. It actually links the expected value of *Y* to the predictors by the function:

$$g(m) = b_0 + b_1 X_1 + b_2 X_2 + \ldots \qquad (13.1)$$

where $g(m)$ is the link function and b_0, b_1 etc are parameters to be estimated. Three common link functions include:

1. Identity link, which is $g(m) = m$, and models the mean or expected value of *Y*. This is used in standard linear models.

2. Log link, which is $g(m) = \log(m)$, and models the log of the mean. This is used for count data (that cannot be negative) in log-linear models (Chapter 14).

3. Logit link, which is $g(m) = \log[m/(1-m)]$, and is used for binary data and logistic regression (Section 13.2).

GLMs are considered parametric models because a probability distribution is specified for the response variable and therefore for the error terms from the model. A more flexible alternative is to use quasi-likelihood models that estimate the dispersion parameter from the data rather than constraining it to the value implied by a specific probability distribution, such as one for a binomial and Poisson. Quasi-likelihood models are particularly useful when our response variable has a binomial or Poisson distribution but is over or under dispersed, i.e. the probability distribution has a dispersion parameter different from one and therefore a variance greater or less than expected from the mean.

GLMs are linear models because the response variable is described by a linear combination of predictors (Box 5.1). Fitting GLMs and maximum likelihood estimation of their parameters is based on an iterative reweighted least squares algorithm called the Newton-Raphson algorithm. Linear regression models (Chapters 5 & 6) can be viewed as a GLM, where the random component is a normal distribution of the response variable and the link function is the identity link so that the expected value (the mean of *Y*) is modeled. The OLS estimates of model parameters from the usual linear regression will be very similar to the ML estimates from the GLM fit.

Readable introductions to GLMs can be found in, among others, Agresti (1996), Christensen (1997), Dobson (1990), and Myers & Montgomery (1997).

13.2 Logistic regression

One very important application of GLMs in biology is to model response variables that are binary (e.g. presence/absence, alive/dead). The predictors can be either continuous and/or categorical. For example, Beck (1995) related two response variables, the probability of survival (survived or didn’t survive) and the probability of burrowing (burrowed or didn’t burrow), to carapace width for stone crabs (*Menippe* spp.). Matlack (1994) examined the relationship between the presence/absence of individual species of forest shrubs (response variables) against a number of continuous predictors, such as stand area, stand age, distance to nearest woodland etc. In both examples, logistic regression was required because of the binary nature of the response variable.

13.2.1 Simple logistic regression

We will first consider the case of a single continuous predictor, analogous to the usual linear regression model (Chapter 5). When the response variable is binary (i.e. categorical with two levels), we actually model $p(x)$, the probability that *Y* equals one for a given value of *X*. The usual model we fit to such data is the logistic regression model, a nonlinear model with a sigmoidal shape (Figure 13-1). The change in the probability that *Y* equals one for a given change in *X* is greatest for values of *X* near the middle of its range, rather than for values at the extremes. The error terms from the logistic model are not normally distributed; because the response variable is binary, the error terms have a binomial distribution. This suggests that ordinary least squares (OLS) estimation is not appropriate and maximum likelihood (ML) estimation of model parameters is necessary. In this section, we will examine a situation with one binary dependent variable (*Y*), which can take values of zero or one, and one independent continuous predictor (*X*).

Lizards on islands

Polis *et al*. (1998) studied the factors that control spider populations on islands in the Gulf of California. Potential predators included lizards of the genus *Uta* and scorpions (*Centruroides exilicauda*). We will use their data to model the presence/absence of lizards against the ratio of perimeter to area for each island. The analysis of these data is presented in Box 13-1.

Logistic model and parameters

The logistic model is:

$$p(x) = \frac{e^{b_0 + b_1 x}}{1 + e^{b_0 + b_1 x}} \qquad (13.2)$$

where b_0 and b_1 are parameters to be estimated. For the Polis et al. (1998) example, $p(x)$ is the probability that y_i equals one (i.e. *Uta* is present) for a given x_i (P/A ratio). As we will see shortly, b_0 is the constant (intercept) and b_1 is the regression coefficient (slope), which measures the rate of change in $p(x)$ for a given change in X. This model can be fitted with non-linear modeling techniques (Chapter 6) to estimate b_0 and b_1 but the modeling process is tedious and the output from software unhelpful.

An alternative approach is to transform $p(x)$ so that the logistic model closely resembles a familiar linear model. First, we calculate odds that an event occurs (e.g. $y_i = 1$ or *Uta* is present), which is the probability that an event occurs relative to its converse, i.e. the probability that $y_i = 1$ relative to the probability that $y_i = 0$:

$$\frac{p(x)}{1 - p(x)} \qquad (13.3)$$

If the odds are >1, then the probability that $y_i = 1$ is greater than the probability that $y_i = 0$; if the odds are <1, then the converse is true. If we take the natural log of the odds that $y_i = 1$:

$$\ln\left[\frac{p(x)}{1 - p(x)}\right] \qquad (13.4)$$

This is the logit transformation or link function, that we will term $g(x)$, and which can be modeled against our predictor much more easily as:

$$g(x) = b_0 + b_1 x_i \qquad (13.5)$$

For the example from Polis *et al*. (1998):

$$g(x) = b_0 + b_1(\text{P / A ratio})_i \qquad (13.6)$$

In model 13.6, $g(x)$ is the natural log (i.e. logit) of the odds that *Uta* is present on an island relative to being absent. We now have a familiar linear model, although the interpretation of the coefficients is a little different (see below). The logit transformation does two important things. First, $g(x)$ now ranges between $-\infty$ and $+\infty$ whereas $p(x)$ is constrained to between zero and one. Linear models are much more appropriate when the response variable can take any real value. Second, the binomial distribution of errors is now modeled.

The logistic regression model is a GLM. The random component is Y with a binomial probability distribution; the systematic component is the continuous predictor X; and the link function that links the expected value of Y to the predictor(s) is a logit link.

Now we use maximum likelihood (ML) techniques to estimate the parameters b_0 and b_1 from logistic model 13.5 by maximizing the likelihood function L:

$$L = \prod_{i=1}^{n} p(x_i)^{y_i} [1 - p(x_i)]^{1 - y_i} \qquad (13.7)$$

It is mathematically much easier to maximize the log likelihood function ln(L) (Chapter 2). ML estimation is an iterative process requiring appropriate statistical software that will also provide standard errors of the ML estimates of b_0 and b_1. These standard errors are asymptotic because they are based on a normal distribution of the parameter estimates that is only true for large sample sizes. Confidence intervals for the parameters can also be calculated from the product of the asymptotic standard error and the standard normal z distribution. Both the standard errors and confidence intervals should be considered approximate.

We earlier defined the odds of an event occurring, which is the probability an event occurs relative to its converse, i.e. the probability that $y_i = 1$ relative to the probability that $y_i = 0$ or the probability that *Uta* occurs on an island relative to it not occurring. Our logistic regression model is that the natural log of the odds equals the constant (b_0) plus the product of the regression coefficient (b_1) and x_i:

$$\ln\left(\frac{p(x)}{1-p(x)}\right) = b_0 + b_1 x_i \qquad (13.8)$$

We can compare the value of the log of the odds, $\ln\left(\frac{p(x)}{1-p(x)}\right)$, for X equals x_i and X equals x_i plus one, i.e. for the predicted Y values in a logistic regression model for X values one unit apart. For the Polis *et al*. (1998) data, this is comparing the log of the odds of *Uta* occurring on an island for P/A ratios that differ by one unit. The ratio of these two odds is called the odds ratio and it is a measure of how the odds of *Uta* occurring change with a change in P/A ratio. Some simple arithmetic produces:

$$\text{odds ratio} = e^{b_1} \qquad (13.9)$$

This is telling us that b_1 represents the change in the odds of an outcome for an increase in one unit of X. For the Polis *et al*. (1998) data, the estimated logistic regression coefficient (b_1) is an estimate of how much the odds of *Uta* occurring on an island (compared to not occurring) would change for an increase in P/A ratio of one unit. A positive value of b_1 indicates that the odds would increase and a negative value indicates the odds would decrease.

The constant, b_0, is the value of $g(x)$ when $x_i = 0$ and represents the intercept of the logistic regression model; its interpretation is similar to the intercept of the linear regression model (chapter 5) and it is usually of less biological interest.

Null hypotheses and model fitting

The H_0 of main interest when fitting a simple logistic regression model is that b_1 equals zero, i.e. there is no relationship between the binary response variable and the predictor variable. In the Polis et al. (1998) study, the H_0 is that there is no relationship between the presence/absence of *Uta* and the P/A ratio of an island). Equivalently, the H_0 is that the log of the odds of *Uta* occurring on an island relative to not occurring is independent of the P/A ratio of the island.

There are two common ways of testing this H_0. The first is to calculate the Wald statistic, a ML version of a *t* test, which is the parameter estimate divided by the standard error of the parameter estimate:

$$\frac{b_1}{s_{b_1}} \qquad (13.10)$$

Note that the standard error (s_{b_1}) is asymptotic (often written as ASE), which means the distribution of b_1 approaches normality for large sample sizes, so the standard error should be considered approximate for small sample sizes. The Wald statistic is sometimes called the Wald *t* (or *t* ratio) statistic because of its similarity to a *t* statistic (Chapter 3). The Wald statistic is traditionally compared to the standard normal *z* distribution (Agresti 1996, Neter *et al*. 1996).

The Wald statistic is most reliable when sample sizes are large so an alternative hypothesis testing strategy that is more robust to small sample sizes and provides a link to measuring the fit of GLMs would be attractive. The approach is similar to that described for OLS regression models in Chapters 5 and 6 where we compare full and reduced models, except that we use log likelihood as a measure of fit rather than least squares. To test the H_0 that b_1 equals zero for a simple logistic regression model with a single predictor, we compare the fit (the log likelihood) of the full model:

$$g(x) = b_0 + b_1 x_i \tag{13.5}$$

to the fit of the reduced model:

$$g(x) = b_0 \tag{13.11}$$

To compare likelihoods, we use a likelihood ratio statistic (Λ), which is the ratio of the log likelihood of reduced model to the log likelihood of full model. Remember from Chapter 2 that larger log likelihoods mean a better fit, so if Λ is near one, then b_1 contributes little to the fit of the full model whereas if Λ is less than one, then b_1 does contribute to the fit of the full model. To test the H_0, we need the sampling distribution of Λ when H_0 is true. The sampling distribution of Λ is messy so instead we calculate a G^2 statistic:

$$G^2 = -2\ln(\Lambda) \tag{13.12}$$

This also called the likelihood ratio c^2 statistic. Sokal & Rohlf (1995) called it the G statistic. It can be simplified to:

$$G^2 = \text{-2(log likelihood reduced - log likelihood full)} \tag{13.13}$$

If H_0 (b_1 equals zero) is true and certain assumptions hold (Section 13.2.4), the sampling distribution of G^2 is very close to a c^2 distribution with one df.

Therefore, we can test H_0 that b_1 equals zero with either the Wald test or with G^2 test comparing the fit of reduced and full models. In contrast to least squares model fitting (Chapter 5), where the t test and the F test for testing b_1 equals zero are identical for a simple linear regression, the Wald and G^2 tests are not the same in logistic regression. The Wald test tends to be less reliable and lacks power for smaller sample sizes and the likelihood ratio statistic is recommended (Agresti 1996, Hosmer & Lemeshow 1989).

The G^2 statistic is also termed the deviance when the likelihood ratio is the likelihood of a specific model divided by the likelihood of the saturated model. The deviance therefore is:

$$\text{-2(log likelihood specific model - log likelihood saturated model)} \tag{13.14}$$

The saturated model is a model that explains all the variation in the data. In regression models, the saturated model is one with as many parameters as there are observations, like a linear regression through two points (Hosmer & Lemeshow 1989). Note that the full model [$g(x) = b_0 + b_1x_i$] is not a saturated model, as it does not fit the data perfectly. In a simple logistic regression with two parameters (b_0 and b_1), we can compare the deviance of the full and reduced models, i.e. the G^2 statistics for each model compared to a saturated model. The difference between the deviances tells us whether or not the two models fit the data differently. We do not actually fit a saturated model in practice because the log-likelihood of the saturated model is always zero (the maximum value of a log-likelihood because the model is a perfect fit), so the deviance for a given model is simply the log-likelihood of that model. Therefore, the difference in deviances equals:

$$\text{-2(log likelihood reduced - log likelihood full)} \tag{13.15}$$

This is simply the G^2 statistic we calculated earlier. The likelihood ratio c^2 statistic (G^2) therefore equals the difference in deviance of the two models. This concept becomes much more important when we have models with numerous parameters (i.e. multiple predictors) and therefore we have lots of possible reduced models (Section 13.2.2).

The other reason the deviance is a useful quantity is because it is the GLM analogue of $SS_{Residual}$, i.e. it measures the unexplained variation for a given model and therefore is a measure of goodness of fit (Section 13.2.5). In the same way that we could create analysis of variance tables for linear models by partitioning the variability, we can create an analysis of deviance table for GLMs. Such a partitioning of deviance is very useful for GLMs with numerous parameters, especially complex contingency tables (Chapter 14).

13.2.2 Multiple logistic regression

Logistic regression can be easily extended to situations with multiple predictor variables. The model fitting procedure is just an extension of the log-likelihood approach described in the previous section. For example, Wiser *et al*. (1998) studied the invasion of mountain beech forests in New Zealand by the exotic perennial herb *Hieracium lepidulum*. They modeled the probability of the exotic occurring on approximately 250 plots in relation to a number of predictor variables measured for each plot, including richness of plant species, the % of total species in the tall herb guild, the distance to the nearest non-alpine open land, other physical variables such as annual potential solar radiation, elevation etc., and chemical characteristics of the soil (Ca, K, Mg, P, pH, N and C:N). Hansson *et al*. (2000) modeled the probability of predation by avian predators on artificial eggs in nests of the Great Reed Warbler in Sweden. Their predictor variables included experimental period (early and late in year) and attractiveness of the territory in which nest occurred, as well as the interaction between these two variables. Our worked example will from a study of the ecology of fragmentation in urban landscapes.

Fragmentation and native rodents

Bolger *et al*. (1997) recorded the number of species of native rodents (except *Microtus californicus*) on 25 canyon fragments in southern California. These fragments have been isolated by urbanization. We will use their data to model the presence/absence of any species of native rodent in a fragment against three predictor variables: distance (m) of fragment to nearest source canyon, age (yr) since the fragment was isolated by urbanization, % of fragment area covered in shrubs. The analysis of these data is presented in Box 13-2.

Logistic model and parameters

The general multiple logistic regression model for *p* predictors is:

$$g(x) = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_p x_{ip} \qquad (13.16)$$

For the Bolger et al. (1997) data:

$$g(x) = b_0 + b_1(\text{distance})_i + b_2(\text{age})_i + b_3(\%\ \text{shrub})_i \qquad (13.17)$$

In models 13.16 and 13.17:

$g(x)$ is the natural log of the odds ratio of $y_i = 1$ versus $y_i = 0$, i.e. the log of the odds of a species of native rodent occurring relative to not occurring in a fragment.

b_0 is the intercept or constant, i.e. the log of the odds of a species of native rodent occurring relative to not occurring in a fragment when all predictors equal zero.

b_1 is the partial regression coefficient for X_1, holding the remaining predictors constant, i.e. the change in the log of the odds of a species of native rodent occurring relative to not occurring in a fragment for a single unit increase in distance to nearest source canyon, holding canyon age and % shrub cover constant.

b_2 is the partial regression coefficient for X_2, holding the remaining predictors constant, i.e. the change in the log of the odds of a species of native rodent occurring relative to not occurring in a fragment for a single unit increase in canyon age, holding distance to nearest source canyon and % shrub cover constant.

b_3 is the partial regression coefficient for X_3, holding the remaining predictors constant, i.e. the change in the log of the odds of a species of native rodent occurring relative to not occurring in a fragment for a single unit increase in % shrub cover, holding distance to nearest source canyon and canyon age constant.

Just like in multiple linear regression models, we can firstly test the significance of the overall regression model by comparing the log-likelihood of the full model (13.16 and 13.17) to the log-likelihood of the reduced model (constant, or b_0, only). We calculate a G^2 statistic

[-2(log likelihood reduced - log likelihood full)] to test the H_0 that at least one of the regression coefficients equals zero.

To test individual coefficients, we can calculate Wald statistics, each one being the estimated regression coefficient divided by standard error of estimated coefficient. These Wald statistics are the equivalent of *t* tests for partial regression coefficients in multiple linear regression (Chapter 6) and can be compared to the standard normal (*z*) distribution. Our reservations about Wald tests (lack of power with small sample sizes) described in Section 13.2.1 apply equally here.

A better approach is to fit a series of reduced models and compare their fit to the full model. To test H_0 that b_1 (distance) equals zero, we compare the fit of the full model:

$$g(x) = b_0 + b_1(\text{distance})_i + b_2(\text{age})_i + b_3(\%\ \text{shrub})_i \qquad (13.17)$$

to the fit of a reduced model based on H_0 being true:

$$g(x) = b_0 + b_2(\text{age})_i + b_3(\%\ \text{shrub})_i \qquad (13.18)$$

with the G^2 statistic:

$$[-2(\text{log likelihood reduced - log likelihood full})] \qquad (13.15)$$

If the G^2 test is significant, we know that the inclusion of distance as a predictor makes the full model a better fit to our data than the reduced model and therefore H_0 is rejected. We can do a similar model comparison test for the other predictors.

The difference between the full and reduced models is also the difference in the deviances of the two models. Remember that the deviance is a measure of the unexplained variability after fitting a model so comparing deviances is just like comparing $SS_{\text{Residuals}}$ for linear models. Neter *et al.* (1996) called this the partial deviance and we can present the results of a multiple logistic regression as an analysis of deviance table.

Other aspects of multiple linear regression described in Chapter 6 also apply to multiple logistic regression. In particular, including interactions between predictors and polynomial terms might have great biological relevance and these terms can be tested by comparing the fit of full model to the appropriate reduced models.

13.2.3 Categorical predictors

Categorical predictor variables can be incorporated in the logistic modeling process by converting them to dummy variables (Chapter 5). Logistic regression routines in most statistical software will do this automatically. We described two sorts of coding for turning categorical predictors into continuous dummy variables for OLS regression in Chapter 5. It is important that you know which method your statistical software is using, as the interpretation of the coefficients and odds ratios is not the same for the two methods. Most programs use reference cell coding where one group of a categorical predictor is used as a reference and the effects of the other groups are relative to that reference group. Alternatively, effects coding could be used, where each group logit is compared to the overall logit (Hosmer & Lemeshow 1989).

A model with a binary response variable and one or more categorical predictors is usually termed a logit model (Agresti 1990, 1996), to distinguish it from classical logistic regression. If all the predictors are categorical, then log-linear modeling (Chapter 14) is a more sensible procedure because the data are in the form of a contingency table. However, log-linear modeling does not automatically distinguish one of the variables as a response variable. For different log-linear models, there are equivalent logit models that identify a response variable (see Agresti 1996, p.165; Chapter 14).

13.2.4 Assumptions of logistic regression

Like all GLMs, logistic regression assumes that the probability distribution for the response variable, and hence for the error terms from the fitted model, is adequately described by the random component chosen. For logistic regression, we assume that the binomial distribution is appropriate, which is likely for binary data. The reliability of the model estimation also depends on the logistic model being appropriate and checking the adequacy of the model is important (Section 13.2.5).

When there are two or more predictors in the model, then absence of strong collinearity (strong correlations between the predictors) is as important for logistic regression models as it was for OLS regression models (Chapter 6). While not necessarily reducing the predictive value of the model, collinearity will inflate the standard errors of the estimates of the model coefficients and can produce unreliable results (Hosmer & Lemeshow 1989, Menard 1995, Tabachnick & Fidell 1996). Most logistic regression routines in statistical software do not always provide automatic collinearity diagnostics, but examining a correlation matrix between the continuous predictors or a contingency table analysis for categorical predictors will indicate if there are correlations/associations between predictors. Tolerance, the r^2 of a regression model of a particular variable as the response variable against the remaining variables as predictors, can also be calculated for each predictor by simply fitting the model as a usual OLS linear regression model. Because tolerance only involves the predictor variables, its calculation is not affected by the binary nature of the response variable.

13.2.5 Goodness-of-fit and residuals

Checking the adequacy of the regression model is just as important for logistic models as for general linear models. One simple and important diagnostic tool for checking whether our model is adequate is to examine the goodness of fit. As with linear models fitted by least squares, the fit of a logistic model is determined by how similar the observed *Y*-values are to the expected or predicted *Y*-values. The predicted probabilities that $y_i = 1$ for given x_i are:

$$\hat{p}(x) = \frac{e^{b_0+b_1x}}{1+e^{b_0+b_1x}} \tag{13.19}$$

In model 13.19, b_0 and b_1 are the estimated coefficients of the logistic regression model. A measure of fit of a particular model is the difference between the observed and fitted values, i.e. the residuals. Residuals in GLMs are similar to those for linear models, the difference between the observed probability that $y_i = 1$ and the predicted (from the logistic regression model) probability that $y_i = 1$.

There are two well-known statistics for assessing the goodness-of-fit of a logistic regression model. These statistics can be used to test that the observed data came from a population in which the fitted logistic regression model is true. The first is the Pearson c^2 statistic based on observed (*o*) and expected, fitted or predicted (*e*) observations (Chapter 14):

$$\sum_{i=1}^{n}\frac{(o-e)^2}{e} = \sum_{i=1}^{n}\frac{(y_i - n_i\hat{p}_i)^2}{n_i\hat{p}_i(1-\hat{p}_i)} \tag{13.20}$$

In equation 13.20, y_i is the observed value of *Y*, $\hat{p}_i$ is the predicted or fitted value of *Y* for a given value of x_i and n_i is the number of observations. The use of the c^2 statistic for logistic regression models is best visualized by treating the data as a two (binary response, *Y*) by *n* (different values of *X*) contingency table. The c^2 statistic for goodness of fit is the usual c^2 for contingency tables (Chapter 14).

The other is the G^2 statistic, which is:

$$\pm 2\sum_{i=1}^{n}(o.\log(o/e)) = \pm 2\left\{\sum_{i=1}^{n} y_i \ln(y_i / n_i\hat{p}_i) + (n_i - y_i)\ln[(n_i - y_i)/n_i(1-\hat{p}_i)]\right\}$$

(13.21)

The terms in equation 13.21 are as defined as in equation 13.20. The G^2 statistic is also the deviance for a given model, defined in Section 13.2.1.

In both cases, low values indicate that the model is a better fit to the data, i.e. the observed and fitted values are similar. The Pearson c^2 statistic and the deviance G^2 statistic approximately follow a c^2 distribution under certain assumptions. The most important assumption is that the minimum predicted frequency of either of the binary outcomes is not too small (see Chapter 14). When the predictors are continuous, however, there will usually be one or few observations of Y for each combination of values of the predictor variables (n_i equals one) so this assumption is not met and the P values associated with the Pearson c^2 statistic and the deviance G^2 statistic will not have approximate c^2 distributions. The statistics themselves are still valid measures of goodness of fit; it is just their P values that are unreliable (Hosmer *et al.* 1997). Note that when we have multiple observations for each combination of X-values, such as when the predictors are categorical, we will have a contingency table in which the expected frequencies are more likely to be reasonable (see Section 13.2.3 and Chapter 14) and the P values associated with these statistics will be much more reliable. Note that the calculation of deviance for categorical predictors depends on whether the saturated model is determined based in individual observations or groupings of observations (Siminoff 1998).

So, we cannot use the usual c^2 or G^2 statistics to test null hypotheses about overall goodness-of-fit of a model when the predictors are continuous, although they are still useful as comparative measures of goodness-of-fit. Hosmer & Lemeshow (1989) developed a solution to the problem testing goodness-of-fit for continuous predictors in logistic regression by grouping observations so that the minimum expected frequency of either of the binary outcomes is not too small. The Hosmer-Lemeshow statistic, also termed the deciles of risk (DC) statistic, is derived from aggregating the data into ten groups. The grouping is based on either each group having one tenth of the ordered predicted probabilities so the groups have equal numbers of observations, or the groups being separated by fixed cutpoints (e.g. first group having all probabilities ≤ 0.10 etc.). Both grouping methods produce a statistic ($\hat{C}$) which approximately follows a c^2 distribution with df as the number of groups minus two.

Hosmer *et al.* (1997) reviewed many goodness-of-fit tests, including the Pearson c^2 statistic and $\hat{C}$, for assessing logistic regression models. They found that the c^2 statistic performed well if based on the conditional mean and variance estimate and compared to a scaled c^2 distribution; unfortunately, the computations required to modify the usual c^2 statistic are not straightforward. They also recommended $\hat{C}$, as it is available in most statistical software and is powerful and we support their recommendation.

There has also been work on analogues of r^2 used as a measure of explained variance in OLS regression. Menard (2000) discussed a range of measures like r^2 for logistic regression and tentatively recommended:

$$r_L^2 = \frac{[\ln(L_0) - \ln(L_M)]}{\ln(L_0)} = 1 - \frac{\ln(L_M)}{\ln(L_0)} \qquad (13.22)$$

In equation 13.22, L_0 is the likelihood for the model with only the intercept and L_M is the likelihood for the model with all predictors (one in the case of simple logistic regression).

13.2.6 Model diagnostics

As well as assessing the overall fit of the model, it is also important to evaluate the contribution of each observation, or group of observations, to the fit and deviations from the fit. In OLS linear models, we have emphasized the importance of residuals, the difference between each observed and fitted or predicted value. There are two types of residuals from logistic regression models. The first is the Pearson residual for an observation, which is the contribution of the difference between the observed and predicted value for an observation to the Pearson c^2 statistic, and is usually expressed as a standardized residual (e_i):

$$e_i = \frac{y_i - n_i\hat{p}_i}{\sqrt{\left[n_i\hat{p}_i(1-\hat{p}_i)\right]}} \qquad (13.23)$$

where y_i is the observed value of Y, $\hat{p}_i$ is the predicted or fitted value of Y for a given value of x_i and n_i is the number of observations. The second is the deviance residual for an observation, which is the contribution of the difference between the observed and predicted value for an observation to the total deviance.

The Pearson and deviance residuals approximately follow a normal distribution for larger sample sizes when the model is correct and residuals greater than about two indicate lack of fit (Agresti 1996, Hosmer & Lemeshow 1989, Menard 1995). When predictor variables are continuous and there is only a single value of Y for each combination of values of the predictor variables, then the large sample size condition will not hold and single residuals will be difficult to interpret. When the predictor variables are categorical and we have reasonable sample sizes for each combination of predictor variables, then residuals are easier to interpret and we will examine such residuals in the context of contingency tables in Chapter 14.

Diagnostics for influence of an observation, i.e. how much the estimates of the parameters change if the observation is deleted, are also available and are similar to those for OLS models (Chapter 5; see also Hosmer & Lemeshow 1989, Menard 1995). These include (i) leverage, which is measured in the same way as for OLS regression, and (ii) an analogue of Cook's statistic standardized by its standard error called *Dfbeta* (Agresti 1996) or Δb (Hosmer & Lemeshow 1989), which measures the standardized change in the estimated logistic regression coefficient b_1 when an observation is deleted. The change in c^2 or deviance when an observation is deleted can also be calculated. These diagnostics are standard output from many logistic regression routines in statistical software. Influential observations should always be checked and our recommendations from Chapters 4 and 5 apply here.

13.2.7 Model selection

As with OLS multiple linear regression, we often wish to know which of the two or more predictor variables in the logistic regression model contributes most to the pattern in the binary response variable. A related aim is to find the "best" model, one that provides the maximum fit for the fewest predictors. The criteria for assessing different models include the Pearson c^2 or deviance (G^2) statistics, r_L^2 and information criteria like Akaike's (see Chapter 6). The Akaike Information Criterion (*AIC*) adjusts ("penalizes") the G^2 (deviance) for a given model for the number of predictor variables:

$$AIC = G^2 - n + 2p \qquad (13.24)$$

where n is the number of observations and p is the number of predictors. For categorical predictors:

$$AIC = G^2 - D + 2p \qquad (13.25)$$

where D is the number of different combinations of the categorical predictors (Larntz 1993). Models with low *AIC*s are the best fit and if many models have similarly low *AIC*s, you should choose the one with the fewest model terms. For both continuous and categorical predictors, we prefer comparing full and reduced models to test individual terms rather than comparing the fit of all possible models to try and select the "best" one.

We will not discuss stepwise modeling for multiple logistic regression or more general logit models. Our reservations about stepwise procedures (see also James & McCulloch 1990) have been stated elsewhere (Chapter 6).

13.2.8 Software for logistic regression

Logistic regression models can be fitted using statistical software in two main ways. Most programs provide logistic regression modules, often as part of a general regression module. It is assumed that the response variable is binary and that a GLM is fitted with a binomial distribution for the error terms and a logit link function. Some software offers GLM routines and the error distribution and link function might need to be specified. The range of diagnostics is usually extensive but it is always worth running a known data set from a text like Christensen (1997) or Hosmer & Lemeshow (1989). Tabachnick & Fidell (1996) have provided an annotated comparison of output from four common programs.

13.3 Poisson regression

Biologists often deal with data that are in the form of counts (e.g. number of organisms in a sampling unit, numbers of cells in a tissue section) and we commonly wish to model a response that is a count variable. Counts usually have a Poisson distribution, where the mean equals the variance and therefore linear models based on normal distributions may not be appropriate. One solution is to simply transform the response variable with a power transformation (e.g. √), which tends to remove any relationship between the mean and variance. An alternative is to use a GLM with a Poisson error term and a log link function that is called a loglinear model. Loglinear models are commonly used to analyze contingency tables (Chapter 14) but can also be used effectively when the predictors are continuous and the response variable is a count to produce a Poisson regression model:

$$\log(m) = b_0 + b_1 x_i \qquad (13.26)$$

In model 13.26, m is the mean of the Poisson distributed response variable, b_0 is the intercept (constant), b_1 is the regression coefficient and x_i is the value of a single predictor variable for observation *i*. The model predicts that a single unit increase in *X* results in *Y* increasing by a factor of e^{b_1} (Agresti 1996). A positive or negative value of b_1 represents *Y* increasing or decreasing respectively as *X* increases. Such models can be easily extended to include multiple predictors. For example, Speight *et al.* (1998) described the infestation of a scale insect *Pulvinaria regalis* in an urban area in England. They modeled egg code, the level of adult/egg infestation measure on a scale of one to ten, against seven predictor variables: tree species, tree diameter, distance to nearest infested tree, distance to nearest road, % impermeability of ground, tree vigour and distance from nearest building.

Nearly all the discussion in previous sections related to logistic regression, including estimation, model fitting and goodness of fit, and diagnostics, apply similarly to Poisson regression models. One additional problem that can occur when modeling count data is that we are assuming that the response has a Poisson distribution where the mean equals the variance. Often, however, the variance is greater than the mean, which is termed overdispersion (Agresti 1996). In GLMs, the dispersion parameter is now less than or greater than one (see Section 13.1). Standard errors of estimated regression coefficients will be smaller than they should and tests of hypotheses will have inflated probabilities of Type I error. Overdispersion is usually caused by other factors, which we have not measured, influencing our response variable in heterogeneous ways. For example, we might model number of plant species per plot against soil pH in a forest; if unmeasured nutrient levels also vary greatly between plots, then variance in the number of species may be greater than the mean. There are at least three possible ways of dealing with overdispersion:

- We can correct the standard errors of the parameters by multiplying by $\sqrt{(c^2/\mathrm{df})}$, as suggested by Agresti (1996). Gardner et al. (1995) provide a complex adjustment based on an estimate of the dispersion parameter.
- We could use a more appropriate probability distribution, such as the negative binomial (Chapter 2, Gardner et al. 1995).
- We could use quasi-likelihood models where the dispersion parameter is estimated from the data rather than restricted to the value defined by a Poisson distribution.

Criteria for assessing the fit of GLMs, such as the likelihood ratio statistic and AIC, are also sensitive to overdispersion. Fitzmaurice (1997) suggested that such criteria could be simply scaled by a REML estimate of the degree of overdispersion.

13.4 Generalised additive models

Generalized additive models (GAMs) are nonparametric modifications of GLMs where each predictor is included in the model as a nonparametric smoothing function (Hastie & Tibshirani 1990). In general terms, with a response variable and j = 1 to p predictor variables, a GLM can be written as:

$$g(m) = b_0 + \sum_{j=1}^{p} b_j X_j \tag{13.27}$$

Note that we have summarized the systematic component representing the predictor variables as a sum of products between regression coefficients and predictors.

A GAM fits a more flexible model:

$$g(m) = b_0 + \sum_{j=1}^{p} f_j X_j \tag{13.28}$$

$$g(m) = b_0 + f_1 x_{i1} + f_2 x_{i2} + ... + f_p x_{ip} \tag{13.29}$$

In models 13.28 and 13.29, the f_j are nonparametric functions estimated using a smoothing techniques (Chapter 5). These smoothing functions, which are commonly Loess or cubic splines for GAMs, are usually estimated from exploratory scatterplots of the data (Yee & Mitchell 1991).

For example, recall the data from Loyn (1987) described in Chapter 6. These data were the abundances of birds from 56 forest patches in southeastern Australia. Six predictor variables were recorded for each patch: area, distance to nearest patch, distance to nearest largest patch, grazing intensity, altitude and years since isolation. A GAM with all predictors (area and the two distances transformed to logs), using a normal probability distribution and identity link function and based on Loess smoothing functions for each predictor, would be:

g(mean bird abundance)$_i$ = b_0 + f_1(log patch area)$_i$ + f_2(years isolated)$_i$ + f_3(log nearest patch distance)$_i$ + f_4(log nearest large patch distance)$_i$ + f_5(stock grazing)$_i$ + f_6(altitude)$_i$ (13.30)

where f_j is a Loess smoothing function. Note that there is no requirement for the same criteria to be used for each smoothing function, e.g. Loess smoothers for X_1 and X_2 may use different smoothing parameters, or even for the same type of smoothing function to be used for each predictor, e.g. a Loess could be used for X_1 and a cubic spline for X_2. The smoothing function for each predictor is derived from the data separately from the smoothing function for any other predictor. We will illustrate the fit of a GAM to a subset of these data from Loyn (1987), incorporating only three predictors (log patch area, log nearest patch distance, years isolated), in Box 13-3.

The main difference between GLMs and GAMs is that the former fits models that are constrained to a parametric (linear) form whereas the latter can fit a broader range of nonparametric models determined from the observed data. A combination of the two types of models is termed semi-parametric. This is a linear model with nonparametric terms included for at least one but not all of the predictors. GAMs are termed additive because the response variable is modeled as the sum of the functions of each predictor with no interactions.

Like GLMs, GAMs need a link function defined and a probability distribution for the response variable that implies a probability distribution for the error terms from the model. The difficulty in specifying a probability distribution for the response variable and error terms is often overcome in GAMs by using quasi-likelihood models where only a relationship between mean and variance is specified and the dispersion parameter (i.e. the variance) is derived from the data (Section 13.1). The fit of a GAM is based on something called the local scoring algorithm, an extension of the Newton-Raphson algorithm used for fitting GLMs. Details of both can be found in Hastie & Tibshiranie (1990) but basically local scoring uses a backfitting algorithm that iteratively fits a smoothing function, determines the partial residuals, and smooths these residuals. The details are complex and understanding them is not necessary to appreciate GAMs.

The important point is that we can measure the fit of a particular GAM, using measures like deviance and AIC, and also compare the fit of models with and without particular terms or combinations of terms. This allows us to assess the contribution of each predictor, modeled with its specific smoothing function, to the pattern in the response variable based on the usual analysis of deviance as used for GLMs. The difference in deviance between two hierarchical models (one with and one without the term being tested in the H_0) can be compared asymptotically to a c^2 distribution. Hastie & Tibshiranie (1990) also suggested that deviance statistics can be converted to approximate *F*-ratio statistics when the dispersion parameter is unknown and *F* tests are common output from software that fits GAMs. In summary, GAMs can be analyzed using the same framework as linear and generalized linear models.

There are some complexities when using GAMs for inference that we do not find in linear and generalized linear models. The use of smoothing functions means that the degrees of freedom will usually not be an integer (Yee & Mitchell 1991). Additionally, the degrees of freedom for a smoothing term can be split into two components, that due to the parametric linear fit and that due to the nonparametric fit once the linear component has been removed. Some software also provides tests of the nonparametric component for individual terms in our model. This is very useful if GAMs are used as an exploratory tool because nonsignificant nonparametric fits suggest that linear models are appropriate for the data.

An example of the use of GAMs in biology comes from Bertaux & Boutin (2000) who modeled the breeding behavior of female red squirrels against 13 possible predictor variables, including minimum age of females, food abundance in same year as female behavior observed, food abundance in previous year, minimum number of middens owned by female, number of juveniles at weaning, and year of study. Their response variable was categorical, values being one, two or three: one was females keeping their territory and excluding juveniles after breeding, two was females sharing their territories with juveniles and three was females bequeathing their territories to juveniles. Bertaux & Boutin (2000) fitted GAMs with different combinations of predictors and with cubic splines as the smoothing functions. They used a quasi-likelihood model to estimate the variance in their response variable because a Poisson distribution was not quite appropriate. They also used the Akaike Information Criterion (AIC) to select the best model, which turned out to be the one with the predictors listed above but not including the remaining seven predictors. They also used logistic regression to model a binary response (females disperse or not disperse after breeding) against these previously described predictor variables; pretty much the same set of variables as for the GAM had the best fit in the logistic model.

Bjorndal *et al*. (2000) also used GAMs to model the growth rates (from mark-recpature data) of immature green turtles in the Bahamas against five predictor variables (sex, site, year, mean size and recapture interval). They used a similar modeling procedure to Bertaux & Boutin (2000), with quasi-likelihood models and cubic spline smoothing functions. However, they sensibly did not try and select a single best model, but rather estimated the fit and parameters for a model with all predictors, including specific contrasts between sexes (male vs female and male versus unknown) and between the three sites. They also tested for nonlinear effects for some of the predictors (see also Yee & Mitchell 1991).

Although GAMs are very flexible models that can be fitted for a wide range of distributions of the response variable, especially exponential distributions, their application is not straightforward. First, we must choose a smoothing function for each predictor and also a smoothing parameter for each smoothing function. Second, we must make the same decisions as for GLMs: which probability distribution and link function combination is appropriate or use quasi-likelihood models. Third, we must have appropriate software and routines for fitting GAMs are not available in most commercial programs, although S-Plus is a noteable exception. With these limitations in mind, GAMs can be very useful, both as an exploratory tool that extends the idea of smoothing functions, and as a more formal model fitting procedure that lets the data determine many aspects of the final model structure.

13.5 Models for correlated data

One of the most challenging data analysis tasks for biologists is dealing with correlated data. For example, repeated observations on the same sampling or experimental units, either under sequential treatment applications or simply through time, cause difficulties for analysis. All the linear and additive models we have describe so far assume independence of observations. If observations are correlated, then the variances and standard errors of estimated model parameters will be inappropriate. For example, positive correlations between observations will result in standard errors of parameter estimates being too low and increased Type I error probabilities for hypothesis tests and negative correlations will result in the converse effect (Dunlop 1994; see also Chapters 5 and 8 for discussion of effects of non-independence in linear regression and ANOVA models).

We have already described methods for dealing with correlated observations that are based on adjusting estimates and hypothesis tests depending on the degree of correlation. For example, the ANOVA models we used for repeated measures designs in Chapters 10 and 11 are basically standard partly nested models where we adjust the tests of significance in a conservative fashion to correct for inflated Type I errors resulting from the correlated observations. While allowing reliable significance tests for repeated measures designs, we would really like a method that fits predictive models that incorporate a mixture of continuous and categorical predictors in a general modeling framework. We will briefly describe two relatively recently developed modeling techniques that specifically address correlated data. Details of the methods are beyond the scope of this book, and our expertise. Their main application seems to have been in the medical literature, especially various types of clinical trials, and in education, although they clearly have potential application in biology given the prevalence of repeated measures designs in the literature. Our aim is simply to make biologists aware that there are methods based on linear and generalized linear models for dealing with correlated data, and to provide references to the literature that will help biologists wishing to investigate these methods further.

These two modeling approaches are just some of the many methods for dealing with correlated data, especially longitudinal data where we have repeated observations of sampling or experimental units. As well as the adjusted ANOVA models described in Chapters 10 & 11, there are growth models, structural equation models, Markov models, transition models and more formal nonlinear time series analyses (see Chapter 5). These techniques, and the two described below, are reviewed by Bijleveld & van der Kamp (1998) and Diggle et al. (1994).

13.5.1 Multi-level (random effects) models

We often deal with observations from sampling or experimental units that are arranged hierarchically. In Chapter 9, we described nested ANOVA models for situations where we had categorical predictors (factors) that were nested within other factors. In those analyses, we used a single model that incorporated the top level factor plus a second level factor nested within the top level factor and so on. One assumption was that observations at the lowest level (“replicates”) were independent of each other. Longitudinal, repeated measures, data can also be viewed as hierarchical with the repeated measurements being nested within an individual sampling or experimental unit and those units being nested within some other (between unit) factors. The difference from the classical nested design described in Chapter 9 is that the measurements nested within each unit are not independent of each other. Laird & Ware (1982) proposed using multi-level linear models with random effects for analyzing longitudinal data, including repeated measures designs. In fact, these models include both fixed and random effects and are therefore best described as multi-level mixed models (Bijleveld & van der Kamp 1998, Ware & Liang 1996).

Consider a fictitious study on growth rates of animals where we use a repeated measures design with a single between subjects factor (sex) and time as the within subjects factor. The subjects or units might be individual animals and the response variable might body size. The basic idea is that we fit a model in two stages; we will mainly follow the terminology of Bijleveld & van der Kamp (1998). In the first stage, we model the response variable for the observations within each unit, against whichever predictor variables are represented by the different times. For example, the predictors may be simply time (in days, months or years) and/or some polynomial of time, or may represent successively applied treatments. With usual linear or generalized linear modeling techniques, we estimate the fixed model parameters for the time effects within each unit and the random error terms:

$$\mathbf{y}_i = b_i\mathbf{T} + \mathbf{e}_i \tag{13.31}$$

In model 13.31, $\mathbf{y}_i$ is the vector of response variable values for each time for unit *i*, **T** is a matrix representing the different times, b_i is the vector of regression coefficients (intercept and slopes, usually only one slope if **T** contains only a single time variable) and $\mathbf{e}_i$ is the vector of random error terms. In the second stage, we treat the regression coefficients as random effects allowing the coefficients (slopes and/or intercepts) of the regressions against time to vary from unit to unit. We are assuming the observed regression coefficients for each unit are a sample from some probability distribution of coefficients. We now model these random coefficients against the predictor variables measured at the between-unit (or subject) level, which will be the between-subjects factor(s):

$$b_i = g\mathbf{x}_i + \mathbf{u}_i \tag{13.32}$$

In this stage two model, b_i is the vector of regression coefficients from stage one, $\mathbf{x}_i$ is the matrix of between unit predictor variables, such as the between subjects design structure, *g* is the vector of coefficients relating the original regression coefficients to the between subjects factor and $\mathbf{u}_i$ is the vector of random error terms.

These two stages can be combined into a single mixed model:

$$\mathbf{y}_i = g\mathbf{T}\mathbf{x}_i + \mathbf{T}\mathbf{u}_i + \mathbf{e}_i \tag{13.33}$$

There are two sets of random effects, the error term from the first level model (within units) and those from the second level model (between units). Different formulations of this model for situations where we allow the slopes or the intercepts or both to vary between units are provided by Burton *et al*. (1998), Cnaan *et al*. (1997) and Omar *et al*. (1999). These models can also be extended to three and more levels.

These multilevel models are usually fit using iterative least squares that result in restricted maximum likelihood (REML) estimates of parameters. The random effects are often estimated as variance components. Tests of particular terms in the model are based on comparing models with and without the term of interest with likelihood ratio (deviance) tests. For the fixed parameters, these deviances can be compared to a c^2 distribution; for random parameters, using the c^2 distribution will result in overly conservative tests (Burton et al. 1998).

Routines for fitting multilevel mixed models are becoming available, both as stand-alone programs (Burton et al. 1998) and in more general use statistical software (e.g. S-Plus). These multi-level mixed models are complex, the literature replete with slightly different formulations of what are basically the same sets of model for a given number of levels. They are particularly useful if the relationship between the response variable and time for each sampling or experimental unit is of interest because this pattern can be modeled, allowing for different slopes and/or intercepts for each unit, against between unit (between subject) predictors (factors).

13.5.2 Generalized estimating equations

Generalized estimating equations (GEEs) were introduced by Liang & Zeger (1986) as an extension of GLMs to model correlated data. To understand the basics of GEEs, we need to examine how we fit GLMs in a little more detail. GLMs are fitted, and therefore parameters of the model are estimated, by solving complex likelihood equations using the iterative Newton-Raphson algorithm. If the response variable has a probability distribution from the exponential family, then the likelihood equations can be viewed as estimating equations (Agresti 1990), equations that are solved to produce ML estimates of model parameters. The normal equations that are solved to produce OLS estimates of linear regression models (Chapter 5) can also be considered as estimating equations. The estimating equations for GLMs are characterized by a covariance (or correlation) matrix that comprises zeros except along the diagonal, i.e. correlations between observations are zero (Dunlop 1994). Liang & Zeger (1986) generalized these estimating equations to allow for covariance matrices where correlations between observations on the same sampling or experimental unit ("subject") are not zero. Solving the GEEs results in estimates of model parameters with variances (and standard errors) that are robust to correlations between observations (Burton et al. 1998). GEEs are not restricted to situations where the response variable has a probability distribution from the exponential family. In fact, quasi-likelihood methods are used where we only need to specify a relationship between the mean and variance for *Y* and we estimate the variance from the data (Section 13.1).

GEEs fit marginal models, where the relationship between the response variable and predictor variables is modeled separately from the correlation between observations within each experimental or sampling unit (Diggle et al. 1994). For example, imagine a data set where we have *n* sampling units (e.g. permanently marked plots in a forest) and we record a response variable (e.g. growth rate of plants) and a predictor variable (e.g. soil phosphorous concentration) at a number of times. Our main interest is probably the relationship between plant growth and soil P, but we want to estimate the parameters of a regression model between these variables accounting for the correlation between observations through time for the same plot. The GEE method will estimate the regression separately from the within-unit correlation. In a repeated measures design, we might have experimental units within a number of treatment groups but these units are observed repeatedly through time. A GEE approach to the analysis would estimate the correlation structure within units separately and use this when fitting a linear model of the response variable against the treatment variable. The correlation structure is treated as a nuisance parameter used to adjust the variance and standard errors of the parameter estimates (Omar *et al.* 1999).

Burton *et al.* (1998) summarized the steps in fitting a GEE. First, a GLM is fitted to all observations and the residuals calculated. These residuals are used to estimate the correlation between observations within each unit. The GLM is refitted but now incorporating the correlation matrix just estimated into the estimating equations. The residuals from this new fit are used to re-estimate the correlation structure and the steps repeated until the estimates stabilize. Hypothesis tests for individual parameters of the model are usually done with Wald tests (Section 13.2.1), the estimate of the parameter divided by its robust standard error estimated from the GEE model.

Besides finding software that will fit GEEs, the main difficulty is that the structure of correlations between observations (i.e. the covariance matrix) needs to be specified *a priori*. Burton *et al.* (1998) and Horton & Lipsitz (1999) suggested a range of working correlation structures:

- Independence, where there are no correlations between observations. Clearly, this is not a sensible choice when we have repeated observations.
- Exchangeable, where the correlations between different observations are identical, no matter how close they are in a time sequence. This is the equivalent of compound symmetry, described for analyses of repeated measures designs with ANOVA models in Chapters 10 and 11.
- Unstructured, where the correlations between pairs of observations can vary and are estimated from the data.
- Fixed, where we fix the correlations rather than estimating them from the data.

- Autoregressive, where correlations between observations closer together in a time sequence are more correlated than observations further apart. This is the situation we anticipate in repeated measures designs and why we usually need to adjust significance tests when fitting partly nested ANOVA models to repeated measures data (Chapters 10 & 11). This choice of correlation structure is used when the residuals from a linear model fit are used to estimate the correlations between observations.

All choices except an unstructured correlation matrix will constrain the pattern of estimated correlations between observations within the same unit. Horton & Lipsitz (1999) recommended an unstructured correlation matrix if the data set is balanced (no missing values) and the number of observations within a unit is small. It turns out that one of the strengths of GEEs is that, although correct specification of the correlation structure makes estimation more efficient, parameter estimates are usually consistent even if the wrong correlation structure is used, i.e. the estimates of model parameters are not very sensitive to the choice of correlation structure. Omar *et al*. (1999) showed this for real data, where estimates and standard errors of between subject treatment differences from a repeated measures design with repeated observations within subjects were similar for unstructured, exchangeable and autoregressive correlation structures.

While GEEs may not work as well for small sample sizes (Ware & Liang 1996), all model fitting methods have difficulties in this situation. GEEs can handle missing data effectively as long as the observations are missing completely at random (Chapters 4 and 15), and therefore provide a real alternative to classical ANOVA type models for repeated measures designs that do not handle missing observations very effectively (Chapters 10 and 11). GEEs can be used for any combination of categorical and continuous response variables and predictors and can make use of the GLM framework of specifying a link function, so that the GEEs can resemble logistic and log-linear models.

In a comparison of different methods for analyzing repeated measurement data, Omar *et al*. (1999) argued that GEEs are most applicable when the pattern of observations through time for sampling or experimental units is not the main research question. For example, in a repeated measures design, GEEs might be suitable when the main factor of interest was between subjects and the within subjects component represents repeated observations through time. If the within subjects component is a factor of specific interest, GEEs are less useful. GEEs are really best for estimating regression models where we have a mixture of repeated and independent observations or when the focus is on comparisons of groups where the units are independent between groups, even if there are also repeated observation within units.

13.6 General Issues and hints for analysis

General issues

- Generalized linear models (GLMs) provide a broad framework for testing linear models when the distribution of model error terms, and the response variable, is from the exponential family (e.g. normal, binomial, Poisson etc.).
- Logistic regression is a GLM for modeling binary response variables against categorical or continuous predictors.
- GLMs such as logistic regression are parametric analyses. Choosing the correct probability distribution, and therefore mean and variance relationship, is important. Quasi-likelihood models are more flexible if you are not sure about the probability distribution or you have data that are under- or overdispersed.
- Poisson regression is a GLM for modeling Poisson response variables (e.g. counts) against categorical or continuous predictors.
- Generalized additive models (GAMs) increase the flexibility of GLMs by permitting a range of nonparametric smoothing functions, rather than just linear relationships.

- For modeling correlated data, generalized estimating equations (GEEs) can provide estimates of parameters and robust standard errors that account for the correlations but are most suited to situations where the pattern through time is not of much interest.
- Multi-level mixed models fit linear models through time for each sampling and experimental unit (stage one) and then model the coefficients from those stage one models against between unit predictor variables (stage two).

Hints for analysis

- Goodness of fit tests for logistic models with continuous predictors are difficult to interpret. The Hosmer-Lemeshow $\hat{C}$ statistic is recommended; do not rely on *P* values from standard c^2 or G^2 statistics.
- Always compare GLMs with multiple predictors in a hierarchical fashion. If an interaction term is included, also include all lower order terms. Check for collinearity if you have two or more predictor variables.
- Overdispersion in binomial or Poisson distributions (where the variance is greater than would be expected based on chosen the probability distribution) can affect parameter estimates and significance tests. Adjustments can be made or use quasi-likelihood models.
- When both the response variable and predictor variable(s) are categorical, log-linear models are easier to interpret if distinguishing a response variable is not essential.

Box 13-1 Worked example of logistic regression: presence/absence of lizards on islands

Polis *et al*. (1998) studied the factors that control spider populations on islands in the Gulf of California. We will use part of their data to model the presence/absence of lizards (*Uta*) against the ratio of perimeter to area (as a measure of) for 19 islands in the Gulf of California. We modeled the presence/absence of *Uta* (binary) against P/A as:

$$g(x) = b_0 + b_1(\text{P/A ratio})_i$$

where $g(x)$ is the natural log of the odds of *Uta* occurring on an island. *Uta* occurred on ten of the 19 islands and the data are plotted in Figure 13-1a. The H_0 of main interest was that there was no relationship between the presence/absence of *Uta* (i.e. the odds that *Uta* occurred relative to not occurred) and the P/A ratio of an island. This is the H_0 that b_1 equals zero.

The maximum likelihood estimates of the model parameters were:

Parameter	Estimate	ASE	Wald statistic	*P*
b_0	3.606	1.695	2.127	0.033
b_1	-0.202	0.101	-2.184	0.029

Note that the Wald statistic is significant so we would reject the H_0 that b_1 equals zero. The odds ratio for P/A was estimated as 0.803 with 95%CI from 0.978 to 0.659. For a one unit increase in P/A, an island has a 0.803 chance of having *Uta* compared to not have *Uta*, an decrease in the odds of having *Uta* of approximately 20%. The plot of predicted probabilities from this model is shown in Figure 13-1b, clearly showing the logistic relationship.

The other way to test the fit of the model, and therefore test the H_0 that b_1 equals zero, is to compare the fit of the full model ($g(x) = b_0 + b_1x_i$) to the reduced model ($g(x) = b_0$):

Full model log-likelihood = -7.110

Reduced model (constant only) log-likelihood = -13.143

G^2 = -2(difference in log-likelihoods) = 12.066, df = 1, P = 0.001. This is also the difference in deviance of the full and reduced models. This test also results in us rejecting the H_0 that b_1 equals zero. Note that the Wald test seems more conservative (larger *P* value).

Goodness of fit statistics were calculated to assess the fit of the model. The Hosmer-Lemeshow statistic was more conservative than either Pearson c^2 or G^2 and was not significant. Along with the low values for Pearson c^2 or G^2, there was no evidence for lack of fit of the model. The logistic analogue of r^2 indicated that about 46% of the uncertainty in the presence/absence of *Uta* on islands could be explained by P/A ratio.

Statistic	Value	df	*P*
Hosmer-Lemeshow ($\hat{C}$)	2.257	5	0.813
Pearson c^2	15.333	17	0.572
Deviance (G^2)	14.221	17	0.651
r_L^2	0.459		

Analysis of diagnostics showed that two islands, Cerraja and Mitlan, were more influential than the rest on the outcome of the model fitting. They had the largest Pearson and deviance residuals and also unusually large values for the logistic regression equivalent of Cook's measure of influence, Hosmer & Lemeshow's (1989) Δb. However, our conclusion for the test of whether b_1 equals zero based on the G^2 statistic (deviance) was not changed if either of these two observations were omitted.

Box 13-2 Worked example of logistic regression: presence/absence of rodents in habitat fragments

Using the data from Bolger *et al*. (1997), we will model the presence/absence of any species of native rodent (except *Microtus californicus*) against three predictor variables: distance (m) to nearest source canyon (X_1), age (yr) since fragment was isolated by urbanization (X_2), % of fragment area covered in shrubs (X_3):

$$g(x) = b_0 + b_1(\text{distance})_i + b_2(\text{age})_i + b_3(\%\ \text{shrub})_i$$

where $g(x)$ is the natural log of the odds of a species of native rodent occurring in a fragment. The scatterplots of the presence/absence of rodents against each predictor are shown in Figure 13-2. The H_0s of main interest were that there was no relationship between the presence/absence of native rodents (i.e. the odds that native rodents occurred relative to not occurred) and each of the predictor variables, holding the others constant. These H_0s are that b_1 equals zero, b_3 equals zero and b_3 equals zero.

The maximum likelihood estimates and tests of the parameters were:

Parameter	Estimate	ASE	Wald statistic	*P*
b_0	-5.910	3.113	-1.899	0.058
b_1	0.000	0.001	0.399	0.690
b_2	0.025	0.038	0.664	0.570
b_3	0.096	0.041	2.361	0.018

The odds ratios were:

Predictor	Distance	Age	% shrub cover
Odds ratio	1.000	1.025	1.101
95% CI	0.999 – 1.002	0.952 – 1.104	1.016 – 1.192

Model comparisons:

Log likelihood of full model: -9.679

Reduced model	H_0	Log likelihood	G^2	*P*
$b_0 + b_2(\text{age})_i + b_3(\%\ \text{shrub})_i$	$b_1(\text{distance}) = 0$	-9.757	0.156	0.693
$b_0 + b_1(\text{distance})_i + b_3(\%\ \text{shrub})_i$	$b_2(\text{age}) = 0$	-9.901	0.444	0.505
$b_0 + b_1(\text{distance})_i + b_2(\text{age})_i$	$b_3(\%\ \text{shrub}) = 0$	-14.458	9.558	0.002

The conclusions from the Wald test and from the G^2 tests from the model fitting procedure agree. Only the effect of % shrub cover on the probability of rodents being present, holding age and distance from nearest source canyon constant, is significant. The odds ratio for % shrub cover was estimated as 1.101 and the 95% CI do not include one; for a one % increase in shrub cover, a fragment has a 1.101 more chance of having a rodent than not, so even though the effect is significant, the effect size is small. The odds ratios for the other two predictors clearly include one, indicating that increases in those predictors do not increase the probability of a rodent being present in a fragment.

Goodness of fit statistics were calculated to assess the fit of the model. The Hosmer-Lemeshow statistic was not significant indicating no evidence for lack of fit of the model.

Statistic	Value	df	*P*

Hosmer-Lemeshow ($\hat{C}$)	6.972	6	0.323
Pearson c^2	20.823	21	0.470
Deviance (G^2)	19.358	21	0.562
r_L^2	0.441		

The model diagnostics suggested that the only fragment that might be influential on the results of the model fitting was Spruce, with a dfbeta (Δb) and Pearson and deviance residuals much greater than the other observations. Unfortunately, we could not get the algorithm to converge on ML estimates when this observation was deleted, so we could not specifically examine its influence on the estimated regression coefficients.

Box 13-3 Worked example of generalized additive models: bird abundances in habitat fragments

We will use the data from Loyn (1987), first introduced in Chapter 6, to illustrate a simple application of GAMs. We will model the abundance of birds in 56 forest patches against three predictors: $\log_{10}$ patch area, $\log_{10}$ distance to nearest patch and years since patch isolation. The boxplot of bird abundance is symmetrical so we will use a normal (Gaussian) probability distribution and an identity link function. We will also use a Loess smoothing function for each of the predictors and keep the smoothing parameter the same for all three functions. We fitted the models using S-Plus 2000 for Windows software.

Full model:

$$g(\text{mean bird abundance})_i = b_0 + f_1(\log_{10}\text{ patch area})_i + f_2(\text{years isolated})_i + f_3(\log_{10}\text{ nearest patch distance})_i$$

Deviance for null model:	6337.929 with 55 degrees of freedom
Residual deviance from fitted model:	1454.314 with 40.529 degrees of freedom

Degrees of freedom and *F*-ratios for nonparametric effects for each predictor:

Term	Parametric df	Nonparametric df	Nonparametric *F*-ratio	*P*
Intercept	1			
$\log_{10}$ patch area	1	4.2	1.817	0.142
years isolated	1	3.3	0.618	0.620
$\log_{10}$ nearest patch distance	1	4.1	2.576	0.051

None of the terms had significant nonparametric components, suggesting that the linear model we fitted in Chapter 6 was appropriate, at least for these three predictors. This is clear from the Loess fits to scatterplots of bird abundance against each predictor (Figure 13-3) with only $\log_{10}$distance suggesting some nonlinearity.

Test of $\log_{10}$ patch area:

Model	$df_{Residual}$	$Deviance_{Residual}$
$\log_{10}$ patch area + years isolated + $\log_{10}$ nearest patch distance	40.529	1454.314
years isolated + $\log_{10}$ nearest patch distance	45.683	3542.574

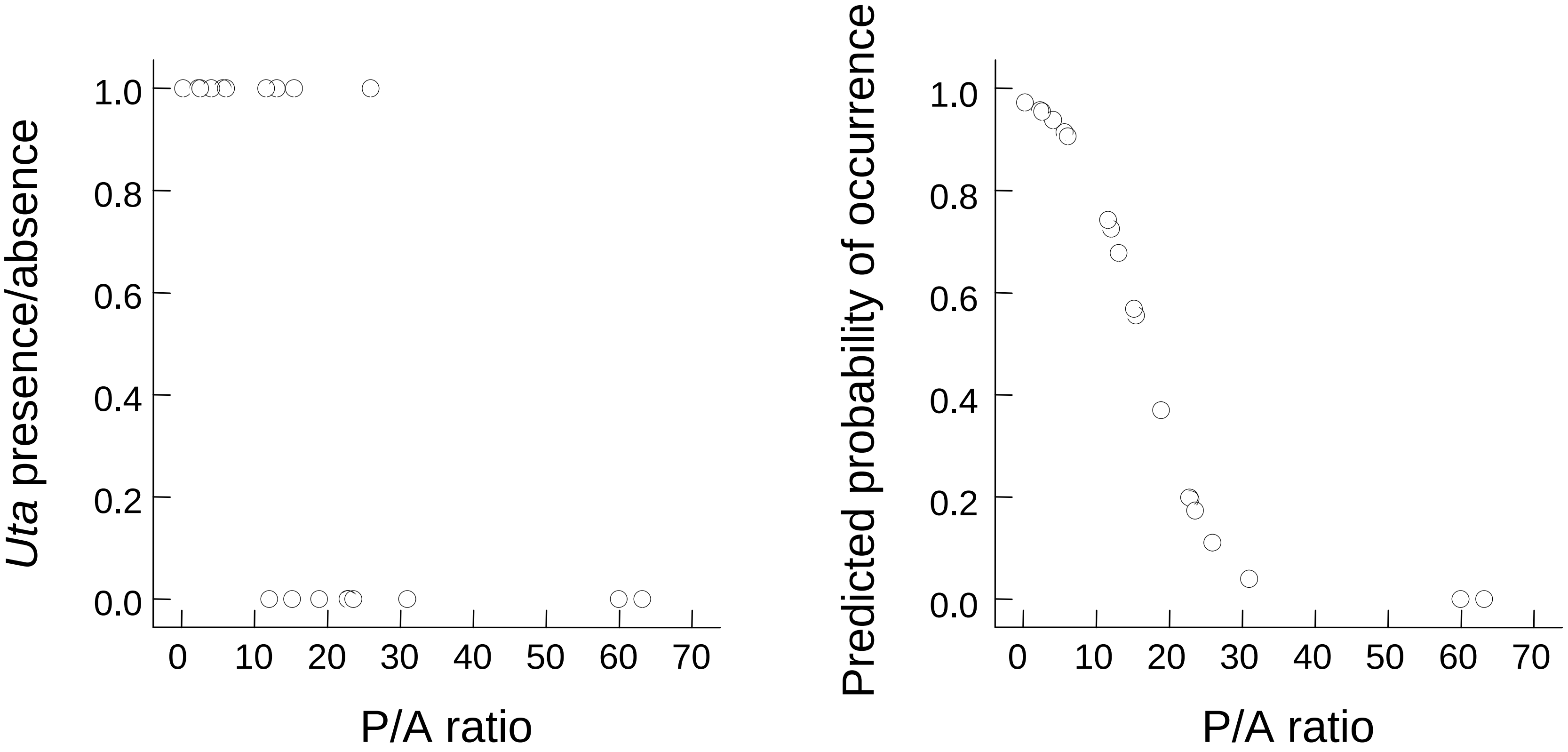
Difference in deviance = -2088.26, df = -5.154, approximate *F*-ratio = 11.291, $P < 0.001$

Clearly, a model that includes $\log_{10}$ patch area was a significantly better fit than a reduced model that doesn't. Equivalent model comparisons could be done for the remaining two predictors.

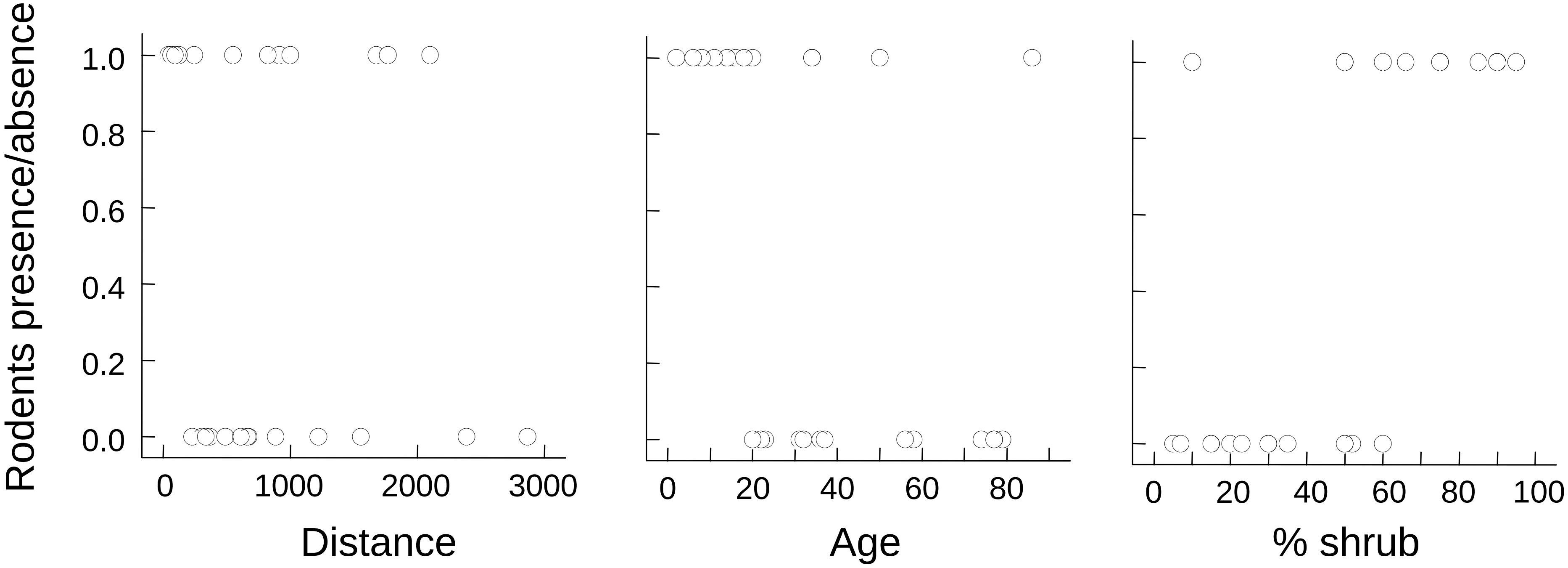
Figure 13-1 (a) Scatterplot of the presence and absence of Uta *in relation to perimeter to area ratio on 19 islands in the Gulf of California (Polis et al. 1998). (b) Scatterplot of the predicted probabilities from logistic regression model of the presence of* Uta *in relation to perimeter to area ratio.*

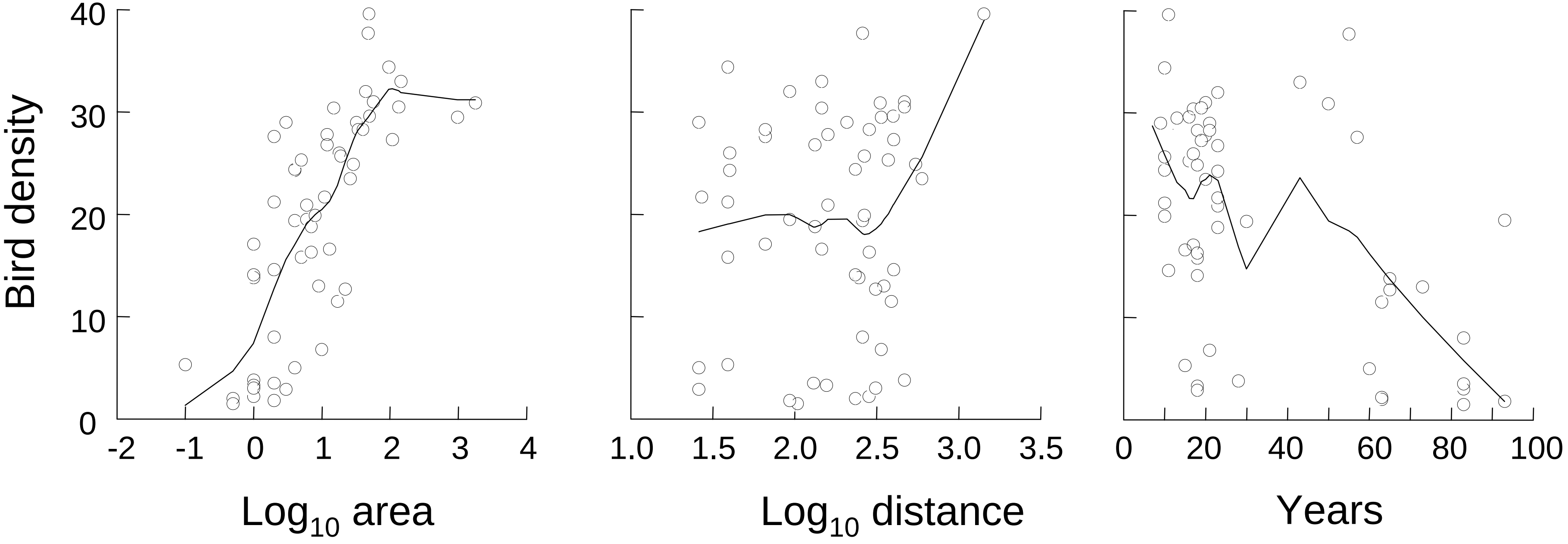
Figure 13-2 Scatterplots of the presence and absence of native rodents in relation (a) to distance to nearest source canyon, (b) age since fragment was isolated by urbanization, (c) % of fragment area covered in shrubs. Data from Bolger et al (1997).

Figure 13-3 Scatterplots of bird abundance against each of three predictors (log_{10} area, log_{10} distance, years since isolation), with Loess smoothers, for the data from Loyn (1987).



Uta presence/absence
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
50
60
70
P/A ratio
Predicted probability of occurrence
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
50
60
70
P/A ratio


Rodents presence/absence
1.0
0.8
0.6
0.4
0.2
0.0
0
1000
2000
3000
Distance
0
20
40
60
80
Age
0
20
40
60
80
100
% shrub


Bird density
40
30
20
10
0
-2
-1
0
1
2
3
4
Log$_{10}$ area
1.0
1.5
2.0
2.5
3.0
3.5
Log$_{10}$ distance
0
20
40
60
80
100
Years