actually links the expected value of Y to the predictors by the function:

$$g(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \quad (13.1)$$

where $g(\mu)$ is the link function and β_0, β_1, etc., are parameters to be estimated. Three common link functions include the following.

1. Identity link, which is $g(\mu) = \mu$, and models the mean or expected value of Y. This is used in standard linear models.
2. Log link, which is $g(\mu) = \log(\mu)$, 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 $g(\mu) = \log[\mu/(1-\mu)]$, 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 and 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, zero or one), we actually model $\pi(x)$, 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 response variable (Y), which can take values of zero or one, and one 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.

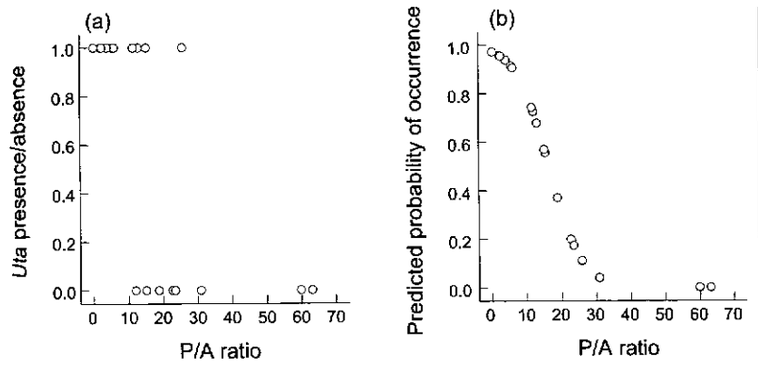

Figure 13.1 (a) Scatterplot of the presence and absence of *Uta* in relation to perimeter to area ratio on 19 islands in the Gulf of California (Polis *et al.* 1998). (b) Scatterplot of the predicted probabilities from logistic regression model of the presence of *Uta* in relation to perimeter to area ratio.

Logistic model and parameters

The logistic model is:

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \quad (13.2)$$

where β_0 and β_1 are parameters to be estimated. For the Polis *et al.* (1998) example, $\pi(x)$ is the probability that $y_i = 1$ (i.e. *Uta* is present) for a given x_i (P/A ratio). As we will see shortly, β_0 is the constant (intercept) and β_1 is the regression coefficient (slope), which measures the rate of change in $\pi(x)$ for a given change in X. This model can be fitted with nonlinear modeling techniques (Chapter 6) to estimate β_0 and β_1 but the modeling process is tedious and the output from software unhelpful.

An alternative approach is to transform $\pi(x)$ 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 $y_i = 0$:

$$\frac{\pi(x)}{1 - \pi(x)} \quad (13.3)$$

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. Then we take the natural log of the odds that $y_i = 1$:

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

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:

$$g(x) = \beta_0 + \beta_1 x_i \quad (13.5)$$

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

$$g(x) = \beta_0 + \beta_1 (\text{P/A ratio})_i \quad (13.6)$$

In model 13.6, $g(x)$ is the natural log (i.e. logit) of the odds that *Uta* is present on an island relative

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 (P/A, as a measure of input of marine detritus) for 19 islands in the Gulf of California. We modeled the presence of *Uta* (binary) against P/A as:

$$g(x) = \beta_0 + \beta_1(\text{P/A ratio})_i$$

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 13.1(a). The H_0 of main interest was that there was no relationship between the presence 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 $\beta_1 = 0$.

The maximum likelihood estimates of the model parameters were as follows.

Parameter	Estimate	ASE	Wald statistic	*P*
β_0	3.606	1.695	2.127	0.033
β_1	−0.2196	0.101	−2.184	0.029

Note that the Wald statistic is significant so we would reject the H_0 that $\beta_1 = 0$. 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*, a decrease in the odds of having *Uta* of approximately 20%. The plot of predicted probabilities from this model is shown in Figure 13.1(b), clearly showing the logistic relationship.

The other way to test the fit of the model, and therefore test the H_0 that $\beta_1 = 0$, is to compare the fit of the full model ($g(x) = \beta_0 + \beta_1 x_i$) to the reduced model ($g(x) = \beta_0$).

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 $\beta_1 = 0$. 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 χ^2 or G^2 and was not significant. Along with the low values for Pearson χ^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 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 χ^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) $\Delta\beta$. However, our conclusion for the test of whether $\beta_1 = 0$ based on the G^2 statistic (deviance) was not changed if either of these two observations were omitted.

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 $\pi(x)$ 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 β_0 and β_1 from logistic model 13.5 by maximizing the likelihood function *L*:

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

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 β_0 and β_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 (β_0) plus the product of the regression coefficient (β_1) and x_i:

$$\ln\left[\frac{\pi(x)}{1 - \pi(x)}\right] = \beta_0 + \beta_1 x_i \qquad (13.8)$$

We can compare the value of the log of the odds

$$\ln\left[\frac{\pi(x)}{1 - \pi(x)}\right]$$

for $X = x_i$ and $X = x_i + 1$, 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:

$$\text{odds ratio} = e^{\beta_1} \qquad (13.9)$$

This is telling us that β_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, β_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 $\beta_1 = 0$, 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 $\beta_1 = 0$ for a simple logistic regression model with a single predictor, we compare the fit (the log-likelihood) of the full model:

$$g(x) = \beta_0 + \beta_1 x_i \tag{13.5}$$

to the fit of the reduced model:

$$g(x) = \beta_0 \tag{13.11}$$

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 β_1 contributes little to the fit of the full model whereas if Λ is less than one, then β_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 is also called the likelihood ratio χ^2 statistic. Sokal & Rohlf (1995) called it the G statistic. It can be simplified to:

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

If H_0 ($\beta_1 = 0$) is true and certain assumptions hold (Section 13.2.4), the sampling distribution of G^2 is very close to a χ^2 distribution with one df.

Therefore, we can test H_0 that $\beta_1 = 0$ 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 $\beta_1 = 0$ 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:

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

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 [$g(x) = \beta_0 + \beta_1 x_i$] is not a saturated model, as it does not fit the data perfectly. In a simple logistic regression with two parameters (β_0 and β_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:

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

This is simply the G^2 statistic we calculated earlier. The likelihood ratio χ^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 percentage 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 be taken 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 (meters) of fragment to nearest source canyon, age (years) since the fragment was isolated by urbanization, and percentage of fragment area covered in shrubs. The analysis of these data is presented in Box 13.2.

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 (meters) to nearest source canyon (X_1), age (years) since fragment was isolated by urbanization (X_2), and percentage of fragment area covered in shrubs (X_3):

$$g(x) = \beta_0 + \beta_1(\text{distance})_i + \beta_2(\text{age})_i + \beta_3(\%\ \text{shrub})_i$$

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 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 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 $\beta_1 = 0, \beta_2 = 0$ and $\beta_3 = 0$.

The maximum likelihood estimates and tests of the parameters were as follows.

Parameter	Estimate	ASE	Wald statistic	P
β_0	−5.910	3.113	−1.899	0.058
β_1	0.000	0.001	0.399	0.690
β_2	0.025	0.038	0.664	0.570
β_3	0.096	0.041	2.361	0.018

The odds ratios were as follows.

Predictor	Distance	Age	Percentage 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 include the following.

Log-likelihood of full model: −9.679.

Reduced model	H_0	Log-likelihood	G^2	P
$\beta_0+\beta_2(\text{age})_i+\beta_3(\%\text{ shrub})_i$	$\beta_1(\text{distance})=0$	−9.757	0.156	0.693
$\beta_0+\beta_1(\text{distance})_i+\beta_3(\%\text{ shrub})_i$	$\beta_2(\text{age})=0$	−9.901	0.444	0.505
$\beta_0+\beta_1(\text{distance})_i+\beta_2(\text{age})_i$	$\beta_3(\%\text{ shrub})=0$	−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 percentage shrub cover on the probability of rodents being present, holding age and distance from nearest source canyon constant, is significant. The odds ratio for percentage shrub cover was estimated as 1.101 and the 95% CI do not include one; for a 1% 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 χ^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 ($\Delta\beta$) 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.

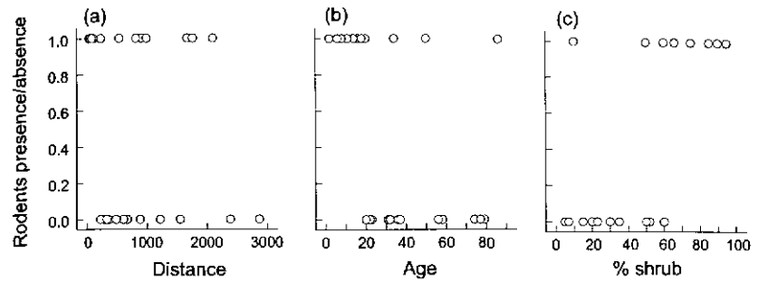

Figure 13.2 Scatterplots of the presence and absence of native rodents in relation (a) to distance to nearest source canyon, (b) age since fragment was isolated by urbanization, and (c) % of fragment area covered in shrubs. Data from Bolger *et al.* (1997).

Logistic model and parameters

The general multiple logistic regression model for p predictors is:

$$g(x)=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\ldots+\beta_px_{ip} \quad (13.16)$$

For the Bolger *et al.* (1997) data:

$$g(x)=\beta_0+\beta_1(\text{distance})_i+\beta_2(\text{age})_i+\beta_3(\%\text{ shrub})_i \quad (13.17)$$

In models 13.16 and 13.17 we find the following.

$g(x)$ 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.

β_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.

β_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 percentage shrub cover constant.

β_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 percentage shrub cover constant.

β_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 percentage 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 β_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 β_1 (distance) = 0, we compare the fit of the full model:

$$g(x)=\beta_0+\beta_1(\text{distance})_i+\beta_2(\text{age})_i+\beta_3(\%\text{ shrub})_i \quad (13.17)$$

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

$$g(x)=\beta_0+\beta_2(\text{age})_i+\beta_3(\%\text{ shrub})_i \quad (13.18)$$

with the G^2 statistic:

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

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 $SS_{Residuals}$ 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:

$$\hat{\pi}(x) = \frac{e^{b_0 + b_1 x}}{1 + e^{b_0 + b_1 x}} \tag{13.19}$$

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 χ^2 statistic based on observed (o) and expected, fitted or predicted (e) observations (Chapter 14):

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

In Equation 13.20, y_i is the observed value of Y, $\hat{\pi}_i$ is the predicted or fitted value of Y for a given value of x_i and n is the number of observations. The use of the χ^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 χ^2 statistic for goodness-of-fit is the usual χ^2 for contingency tables (Chapter 14).

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

$$\pm 2\sum_{i=1}^{n} (o.\log(o/e)) = \pm 2\left[\sum_{i=1}^{n} y_i \ln(y_i/n\hat{\pi}_i) + (n-y_i)\ln[(n-y_i)/n(1-\hat{\pi}_i)\right] \tag{13.21}$$

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 χ^2 statistic and the deviance G^2 statistic approximately follow a χ^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 ($n_i = 1$) so this assumption is not met and the Pearson χ^2 statistic and the deviance G^2 statistic will not have approximate χ^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 also 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 also that the calculation of deviance for categorical predictors depends on whether the saturated model is determined based on individual observations or groupings of observations (Siminoff 1998).

So, we cannot use the usual χ^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 of 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 χ^2 distribution with df as the number of groups minus two.

Hosmer *et al.* (1997) reviewed many goodness-of-fit tests, including the Pearson χ^2 statistic and

$\hat{C}$, for assessing logistic regression models. They found that the χ^2 statistic performed well if based on the conditional mean and variance estimate and compared to a scaled χ^2 distribution; unfortunately, the computations required to modify the usual χ^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 χ^2 statistic, and is usually expressed as a standardized residual (e_i):

$$e_i = \frac{y_i - n\hat{\pi}_i}{\sqrt{[n\hat{\pi}_i(1-\hat{\pi}_i)]}} \qquad (13.23)$$

where y_i is the observed value of Y, $\hat{\pi}_i$ is the predicted or fitted value of Y for a given value of x_i and n 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 $\Delta\beta$ (Hosmer & Lemeshow 1989), which measures the standardized change in the estimated logistic regression coefficient b_1 when an observation is deleted. The change in χ^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 χ^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:

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

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

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

where D is the number of different combinations of the categorical predictors (Larntz 1993). Models with low AICs are the best fit and if many models have similarly low AICs, 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. $\sqrt{}$), 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 log-linear model. Log-linear 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:

$$\log(\mu) = \beta_0 + \beta_1 x_i \qquad (13.26)$$

In model 13.26, μ is the mean of the Poisson distributed response variable, β_0 is the intercept (constant), β_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^{β_1} (Agresti 1996). A positive or negative value of β_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 measured on a scale of one to ten, against seven predictor variables: tree species, tree diameter, distance to nearest infested tree, distance to nearest road, percentage impermeability of ground, tree vigor 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, applies 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 $\sqrt{(\chi^2/df)}$, 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 (restricted maximum likelihood) estimate of the degree of overdispersion.

13.4 Generalized additive models

Generalized additive models (GAMs) are non-parametric modifications of GLMs where each predictor is included in the model as a non-parametric 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:

$$g(\mu) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j \qquad (13.27)$$

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:

$$g(\mu) = \beta_0 + \sum_{j=1}^{p} f_j X_j \qquad (13.28)$$

$$g(\mu) = \beta_0 + f_1 x_{i1} + f_2 x_{i2} + \cdots + f_p x_{ip} \qquad (13.29)$$

In models 13.28 and 13.29, the f_j are non-parametric functions estimated using a smoothing technique (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 south-eastern 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:

$$g(\text{mean bird abundance})_i = \beta_0 + f_1(\text{log patch area})_i + f_2(\text{years isolated})_i + f_3(\text{log nearest patch distance})_i + f_4(\text{log nearest large patch distance})_i + f_5(\text{stock grazing})_i + f_6(\text{altitude})_i \qquad (13.30)$$

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 non-parametric models determined from the observed data. A combination of the two types of models is termed semi-parametric. This is a linear model with non-parametric 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

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:

$g(\text{mean bird abundance})_i = \beta_0 + f_1(\log_{10}\text{ patch area})_i + f_2(\text{years isolated})_i + f_3(\log_{10}\text{ nearest patch distance})_i$

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 non-parametric effects for each predictor are tabulated below.

Term	Parametric df	Non-parametric df	Non-parametric *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 non-parametric 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 is as follows.

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.

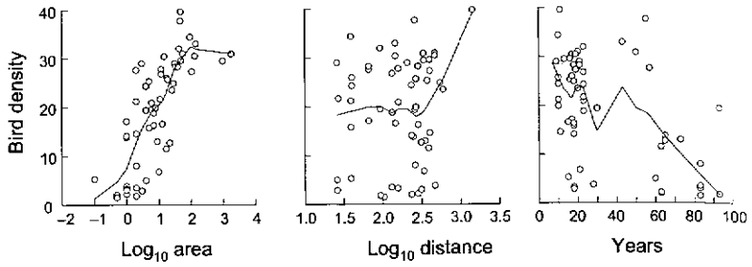

Figure 13.3 Scatterplots of bird abundance against each of three predictors ($\log_{10}$ area, $\log_{10}$ distance, years since isolation), with Loess smoothers, for the data from Loyn (1987).

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 & Tibshirani (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 χ^2 distribution. Hastie & Tibshirani (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 non-parametric fit once the linear component has been removed. Some software also provides tests of the non-parametric component for individual terms in our model. This is very useful if GAMs are used as an exploratory tool because non-significant non-parametric fits suggest that linear models are appropriate for the data.

An example of the use of GAMs in biology comes from Berteaux & 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. Berteaux & 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–recapture 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 to 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 non-linear 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 described 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 and 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 be 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:

$$\mathbf{y}_i = \beta_i \mathbf{T} + \mathbf{e}_i \qquad (13.31)$$

In model 13.31, $\mathbf{y}_i$ is the vector of response variable values for each time for unit i, $\mathbf{T}$ is a matrix representing the different times, β_i is the vector of regression coefficients (intercept and slopes, usually only one slope if $\mathbf{T}$ 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):

$$\beta_i = \gamma \mathbf{x}_i + \mathbf{u}_i \qquad (13.32)$$

In this stage two model, β_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, γ 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:

$$\mathbf{y}_i = \gamma \mathbf{I} \mathbf{x}_i + \mathbf{T} \mathbf{u}_i + \mathbf{e}_i \qquad (13.33)$$

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 multi-level models are usually fit using iterative least squares that result in 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 χ^2 distribution; for random parameters, using the χ^2 distribution will result in overly conservative tests (Burton *et al.* 1998).

Routines for fitting multi-level 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 is basically the same set 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 phosphorus 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), where the estimate of the parameter is 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 and 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

13.6.1 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 underdispersed 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 non-parametric 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).

13.6.2 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 χ^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 the chosen 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.

Chapter 14

Analyzing frequencies

The previous chapter introduced logistic regression, a generalized linear model based on a binomial distribution and logit link function for modeling binary response variables. If the response has more than two categories, then it is likely to come from a multinomial distribution, of which the binomial is a special case. We can also model the count or frequency in each category as coming from a Poisson distribution when the total count (n) across all categories is not fixed. This chapter focuses on the analysis of one or more categorical variables, particularly when we have counts of observations in each combination of the variables. When there are two or more variables, each with two or more categories, the counts form a contingency table where the observations are cross-classified by the categorical variables. Contingency tables do not specifically distinguish response and predictor variables, although such a distinction can be important in model building and interpretation.

A fundamental statistic for the analysis of categorical data is the chi-square (χ^2) statistic, also called the Pearson χ^2 statistic, which is commonly used to compare observed and theoretical (i.e. expected) frequencies in categories:

$$\sum_{i=1}^{n} \frac{(o-e)^2}{e} \qquad (14.1)$$

where o and e denote the observed and expected (or theoretical) frequencies respectively in each category or combination of categories and the summation is over all the categories. The degrees of freedom are a function of the number of categories minus one. Note that χ^2 basically measures the differences between the observed and expected values. It has a value of zero when the observed and expected values are the same. Null hypotheses in categorical analyses often imply that a sample of observations came from a population where the observed frequencies match some expected frequencies. The χ^2 statistic approximately follows a χ^2 distribution if the following assumptions hold.

1. Observations are classified into categories independently. This means that the category combination into which any observation is classified is independent of the category combination into which other observation is classified.

2. No more than 20% of the categories have expected frequencies less than about five (Agresti 1990, 1996). With smaller sample sizes, comparisons of the χ^2 statistic to a χ^2 distribution can produce misleading probabilities.

This chapter is an introduction to categorical data analyses; more detailed treatments can be found in Agresti (1990, 1996), Christensen (1997) and Tabachnick & Fidell (1996) among others. We will first illustrate some simple analyses based on the χ^2 statistic, although generalized linear models, especially log-linear models, are much more flexible for categorical data analysis. We will consider these later in the chapter.

Box 14.1 Worked example: goodness-of-fit tests for a single variable

For one of the few times in this book, we will use fictitious data. Ninety shrubs of a dioecious plant were sampled in a forest and each plant was classified as male or female. The observed counts and the predicted (expected) counts based on a theoretical 50:50 sex ratio were as follows.

	Female	Male	Total
Observed	40	50	90
Expected	45	45	

The H_0 is that this sample of plants came from a population with a sex ratio of 50:50. The expected values were derived from n and the H_0.

$\chi^2 = 1.11$, df $= 1$, $P = 0.292$.

There is no evidence that the observed sex ratio in the population is different from 50:50.

14.1 Single variable goodness-of-fit tests

A simple goodness-of-fit test is where we test whether our observations come from a population with a particular distribution of frequencies in categories of a single variable. The general data layout for these tests is usually a single categorical variable with counts or frequencies for each category (Box 14.1). The expected values (if H_0 is true) are calculated from some theoretical or predicted frequency. The H_0 is that the observed data came from a population that has the theoretical or expected frequencies. We test this H_0 by calculating a χ^2 statistic with the equation described above. We then compare the calculated χ^2 to the χ^2 distribution with the degrees of freedom being the number of categories minus one. If the probability of obtaining the calculated χ^2, or one larger, when H_0 is true, is less than our chosen significance level, then H_0 should be rejected. This is the standard logic of testing a statistical null hypothesis (Chapter 3).

An alternative goodness-of-fit test for a single variable is the Kolmogorov–Smirnov (K–S) test, which compares observed and expected cumulative frequencies (Hays 1994). The test statistic (D) is just the largest difference between the observed and expected cumulative frequencies across all possible values of the categorical variable. This test is preferred to the χ^2 when there are a large number of categories and the categories can be ordered in some way. In particular, the K–S test is suited for comparing two frequency distributions, where one distribution acts as the observed and the other the expected. As with most biostatistical analyses, the K–S test is clearly described, with formulae, in Sokal & Rohlf (1995) and is available in most statistical software.

14.2 Contingency tables

The most common form of categorical data analysis in the biological sciences is the analysis of contingency tables. These tables involve the cross-classification of sampling or experimental units by two or more variables (Table 14.1), with counts or frequencies of units in each combination of the variables, termed a cell, analogous to factorial ANOVA designs.

14.2.1 Two way tables

Tables where sampling or experimental units are cross-classified by two variables are termed two way tables. Generally, contingency tables are analyzed so that neither variable is considered as a

Table 14.1 | General data layout for a two by two contingency table

Variable 2→ Variable 1↓	1	2	Marginal totals variable 1
1	n_{11} π_{11}	n_{12} π_{12}	n_{1j} π_{1j}
2	n_{21} π_{21}	n_{22} π_{22}	n_{2j} π_{2j}
Marginal totals variable 2	n_{i1} π_{i1}	n_{i2} π_{i2}	Grand total n

Note:
Variable 1 has two levels ($I = 2$), variable 2 has two levels ($J = 2$) with observed counts or frequencies (n_{ij}) for each combination (cell) of the two variables. The probability that an observation falls in any cell is π_{ij}; marginal probabilities are π_{i+} and π_{+j}.

Table 14.2 | Observed frequencies for two way contingency tables from (a) French & Westoby's (1996) study where plant species were cross-classified by dispersal mode and regeneration mechanism and (b) Roberts's (1993) cross-classification of quadrats on a floodplain by presence/absence of dead coolibah trees and position along transect

(a)

	Dispersal mechanism		
Regeneration	Ant	Vertebrate	Total
Seed only	25	6	31
Vegetative	36	21	57
Total	61	27	88

(b)

	Dead coolibah trees		
Position along transect	With	Without	Total
Bottom	15	13	28
Middle	4	8	12
Top	0	17	17
Total	19	38	57

predictor or a response variable. For example, French & Westoby (1996) cross-classified plant species following fire by two variables: whether they regenerated by seed only or vegetatively and whether they were ant or vertebrate dispersed. These two variables could not be distinguished as response or predictor since regeneration mechanisms could just as easily "affect" dispersal mode as vice versa. This was a two by two table (Table 14.2(a)) and its analysis is in Box 14.2.

In other situations, one variable can be envisaged as a response variable and the other as a predictor. For example, Roberts (1993) sampled quadrats on a floodplain and classified them by two variables: presence/absence of dead coolibah trees (*Eucalyptus coolibah*) and position along transect (top = dunes, bottom = lakeshore, middle = intermediate). In this example, position along the transect might be considered a predictor variable and with or without dead coolibah trees as a response variable. We might expect coolibah tree mortality to be affected by position but the converse is biologically unlikely. This was a two by three table (Table 14.2(b)) and its analysis is in Box 14.3. Another example is from Clinton & Le Boeuf (1994), who looked at the association between survivorship of male northern elephant seals (*Mirounga gustirostris*) and mating success (the number of females inseminated). This was a two by two contingency table with died/survived as the response variable, zero or greater than zero females inseminated as the predictor variable and the number of male seals were the frequencies in each category.

In practice, the analysis of contingency tables is not really changed by whether we can distinguish response and predictor variables. If the response variable is binary, then we can use logistic (i.e. logit) models with categorical predictors as described in Chapter 13. However, the distinction between response and predictor variables can be important for the interpretation of log-linear models for analyzing complex contingency tables (Section 14.2.2).

Table structure

The general data layout for a two way table (cross-classification of two variables) is illustrated in Table 14.1. We will follow Agresti (1996) and use X ($i = 1$ to I categories) and Y ($j = 1$ to J categories) as labels for the two variables. For a two by two table, both I and J equal two. When one of the variables is clearly a response variable, it will be designated Y; otherwise, no particular significance should be ascribed to which variable is X and which is Y. The observed frequency in each cell is n_{ij} and the probability that an observation occurs in any cell is π_{ij}. We also have marginal totals (e.g. the total in row one is n_{1j}) and marginal probabilities (e.g. the probability that an observation occurs in row one is $\pi_{1j} = \pi_{11} + \pi_{12}$); these marginal probabilities are the probabilities that an observation occurs in a particular row or column.

Sokal & Rohlf (1995) described three different

Box 14.2 Worked example of analysis of independence in two way table: regeneration and seed dispersal mechanisms of plants

French & Westoby (1996) cross-classified plant species following fire by two variables: whether they regenerated by seed only or vegetatively and whether they were ant or vertebrate dispersed. The H_0 is that the dispersal mechanism is independent of mode of regeneration. The χ^2 statistic for testing this H_0 is 2.89 with one df and $P = 0.089$. We have no evidence to reject the H_0 of independence. The standardized residuals showed no strong patterns, although fewer species that regenerated only from seed were dispersed by vertebrates than expected by chance and the converse was true for seeds that regenerated vegetatively.

Standardized residuals are tabulated below.

	Dispersal mechanism	
Regeneration	Ant	Vertebrate
Seed only	0.757	−1.139
Vegetative	−0.559	0.840

The odds of being ant dispersed compared to being vertebrate dispersed for plants that regenerate by seed are 4.17. For plants that regenerate vegetatively, the odds are 1.71.

The sample odds ratio ($\hat{\theta}$) is:

$$\frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{25 \times 21}{36 \times 6} = 2.43$$

So the odds of being dispersed by ants is 2.43 times greater for plant species that regenerate by seed compared to those that regenerate vegetatively. We can convert $\hat{\theta}$ to logs, use Equation 14.10 to calculate the standard error of the log ($\hat{\theta}$), Equation 14.11 to calculate the 95% confidence interval for the log (θ) and back-transform this for a confidence interval for the θ.

Odds ratio	Log (odds ratio)	ASE log (odds ratio)	95% CI log (odds ratio)	95% CI (odds ratio)
2.43	0.89	0.53	± 1.04	0.86 to 6.89

The wide confidence interval includes one, indicating that odds of being dispersed by ants for plant species that regenerate by seed are not statistically different than for plant species that regenerate vegetatively.

Box 14.3 Worked example of analysis of independence in two way table: coolibah trees on a floodplain

Roberts (1993) sampled quadrats on a floodplain and classified them by two variables: presence/absence of dead coolibah trees (*Eucalyptus coolibah*) and position along transect (top = dunes, bottom = lakeshore, middle = intermediate). The H_0 is that the presence/absence of dead coolibah trees is independent of position on the floodplain. The χ^2 statistic for the test of this H_0 is 13.66 with two df and $P = 0.001$. Therefore, we reject the H_0 of independence.

Standardized residuals are tabulated below.

	Dead coolibah trees	
Floodplain position	With	Without
Bottom	1.855	−1.312
Middle	0.000	0.000
Top	−2.380	1.683

It is clear from the residuals that there were more quadrats with dead trees at the bottom of the dunes than expected and fewer quadrats with dead trees at the top of the dunes than expected.

Odds of having dead trees versus not are as follows.

Position	Odds
Bottom of floodplain	1.15
Middle of floodplain	0.50
Top of floodplain	0.00

The odds of having dead coolibah trees were greater than not having them for quadrats at the bottom of the floodplain, but the odds of having dead coolibah trees were less than not having them for quadrats at the middle of the floodplain. Because there were no quadrats with dead coolibah trees at the top of the floodplain, odds cannot be calculated for this position.

Odds ratios were calculated using the modified formula that adds 0.5 to each cell to correct for zero observed frequencies.

	Odds ratio	Log (odds ratio)	ASE	95% CI (odds ratio)
Bottom versus middle	2.17	0.77	0.69	0.59 to 8.18
Bottom versus top	40.19	3.69	1.48	2.20 to 728.36
Middle versus top	18.53	2.92	1.55	0.89 to 386.84

The 95% CI for the odds ratios of having dead coolibah trees included one the comparison of the bottom of the floodplain versus the middle and the middle versus the top. The strongest pattern is that the odds of having dead coolibah trees were greater at the bottom of the floodplain compared with the top.

models for contingency tables, based on whether the investigator predetermines the marginal totals (i.e. row and column totals).

- Model I is when none of the marginal totals are fixed, the most common situation when a number of sampling or experimental units are sampled from a population of units and each unit is classified by one or more categorical variables. An underlying Poisson distribution for the counts in each cell is assumed. The three examples described above are Model I.
- Model II is when one set of marginal totals is fixed. For example, imagine an experiment where ten rats are allocated to three different drug treatments and the survivorship of each rat in recorded at the end of the experiment. Each rat is cross-classified by treatment (fixed marginal totals of ten) and lived/died (marginal totals not fixed). In Model II tables, the variable without fixed marginal totals is usually considered a response variable (Agresti 1996).
- Model III is when both sets of marginal totals are fixed, a very uncommon situation in biology. Fisher (1935) described such a model for an experiment to test whether someone could actually tell by tasting whether or not milk had been added first to a cup of tea (see also Agresti 1990, 1996).

Null hypothesis

The H_0 is one of independence, that the sampling or experimental units come from a population of units in which the two variables (rows and columns) are independent of each other in terms of the cell frequencies. This is often expressed as no association, or interaction, between the two variables. For example, French & Westoby (1996) tested whether the mechanism of seed regeneration (seed or vegetative) was independent of dispersal mechanism (ant or vertebrate) for a number of plant species (Box 14.2, Table 14.2). Usually, the H_0 is expressed in terms of a population from which the sampling or experimental units were obtained, a population that is difficult to envisage for the French & Westoby (1996) example. Roberts (1993) wished to test whether her quadrats came from a population of quadrats on the floodplain where presence/absence of dead coolibahs was independent of position along transect (Box 14.3, Table 14.2). Clinton & Le Boeuf (1994) tested whether survivorship of male elephant seals (died or survived) was independent of whether the males had inseminated zero or more than zero females.

The H_0 can also be expressed as:

$$\pi_{ij} = \pi_{i+} \cap \pi_{+j} \tag{14.2}$$

i.e. the probability of an observation occurring in a cell equals the probability of it occurring in that row *and* that column.

We can test this H_0 using a χ^2 test by calculating the expected frequencies in each cell based on the H_0 being true and there being no association between the two variables. An expected cell frequency is simply the product of the probability of an observation occurring in that cell and the total sample size:

$$f_{ij} = n\pi_{ij} \tag{14.3}$$

We can elaborate on this as follows. If rows and columns are independent (i.e. H_0 is true), then the probability of an observation occurring in a specific cell (π_{ij}) is simply the probability of an observation occurring in the specific row (π_{i+}, estimated by row total divided by grand total) multiplied by the probability of it occurring in the specific column (π_{+j}, estimated by column total divided by grand total). Therefore, a general formula for calculating the expected frequency in each cell assuming independence of the two variables (i.e. under H_0) is:

$$[(\text{row total})(\text{column total}) \,/\, \text{grand total}] \tag{14.4}$$

We then calculate χ^2 based on Equation 14.1 where n_{ij} are observed frequencies and f_{ij} are the expected frequencies under the H_0:

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij} - f_{ij})^2}{f_{ij}} \tag{14.5}$$

We compare the χ^2 in 14.5 to a χ^2 distribution with $(I-1)(J-1)$ df. If the probability of obtaining the calculated χ^2 or one larger when H_0 is true is less than our chosen significance level, then H_0 should be rejected. For the French & Westoby (1996) example, we have no evidence to reject the H_0 that dispersal mode and regeneration mode are independent of each other (Box 14.2). For the Roberts (1993) example, we would reject the H_0

Box 14.4 Worked example of log-linear models for two way table: coolibah trees on a floodplain

We will re-analyze the contingency table from Roberts (1993) using log-linear models to test the H_0 that quadrats came from a population of quadrats where presence/absence of dead coolibahs was independent of position along transect.

Reduced model:

$$\text{Log } f_{ij} = \text{constant} + \lambda^{\text{position}} + \lambda^{\text{presence/absence}}$$

Log-likelihood for reduced model: −19.735, df = 2.

Full (and saturated) model:

$$\text{Log } f_{ij} = \text{constant} + \lambda^{\text{position}} + \lambda^{\text{presence/absence}} + \lambda^{\text{position} \times \text{presence/absence}}$$

Log-likelihood for full (and saturated) model: −10.429, df = 3.

$$G^2 = -2(\text{log-likelihood model} - \text{log-likelihood saturated model})$$
$$= -2 \times (-19.735 - (-10.429))$$
$$= 18.61, \text{df} = 1, P < 0.001.$$

Therefore we reject H_0.

that presence/absence of dead coolibah trees is independent of position on floodplain (Box 14.3).

Odds and odds ratios

Odds and odds ratios are important summary measures of association or lack of independence in contingency tables, just as they are for logistic regression models (Chapter 13). They can only be calculated for two by two tables but can also be used in larger tables by subdividing these tables into sets of two by two tables. We calculate the odds of one of the two possible categories (outcomes) of one variable for each level (j) of the other variable:

$$\frac{\pi_j}{1-\pi_j} \qquad (14.6)$$

where π_j is the probability of one of the two outcomes and one minus π_j is the probability of the other outcome.

For the French & Westoby (1996) example, the odds of being ant dispersed compared to being vertebrate dispersed for plants that regenerate by seed are 4.17 and for plants that regenerate vegetatively, the odds are 1.71 (Box 14.2). In the Roberts (1993) example, presence/absence of dead coolibah trees is the response variable and position on the floodplain is the predictor variable. The estimated odds of having versus not having a dead coolibah for the bottom of the floodplain is 1.15; this indicates that having dead coolibah trees is more likely than not having them (Box 14.3 and Box 14.4). We can calculate odds of a quadrat having dead coolibah trees for the other two floodplain positions as well. For the middle of the floodplain, the odds are 0.50 and for the top of the floodplain, the odds are zero.

The odds ratio (θ) is simply the ratio of the odds of one outcome for one level of the second variable to the odds of the same outcome for another level of the second variable. The odds ratio is a population parameter (Agresti 1996):

$$\theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} \qquad (14.7)$$

The ML estimate of this odds ratio is the sample odds ratio:

$$\hat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}} \qquad (14.8)$$

In Equation 14.8, each n_{ij} is the observed frequency in the cell based on the ith row and jth column, e.g. n_{12} is the observed frequency in the cell being the first row and first column.

Note that the odds ratio equals zero if any of the observed counts in the two by two subset table also equal zero. Agresti (1996) suggested a simple correction by adding 0.5 to each cell:

$$\hat{\theta} = \frac{(n_{11}+0.5)(n_{22}+0.5)}{(n_{12}+0.5)(n_{21}+0.5)} \qquad (14.9)$$

You can see that odds ratios are much easier to interpret for two by two tables because there is only one odds ratio. For larger tables, there will be different odds ratios for different two by two subsets. These odds ratios are not independent (Agresti 1990) and, because of this redundancy, only $(I-1)(J-1)$ odds ratios are needed to summarize the lack of independence in an I by J table.

Odds ratios are important for interpreting lack of independence in contingency tables (Agresti 1996). If the probability of one outcome (e.g. having dead coolibah trees) is the same for two floodplain positions, i.e. the presence of dead coolibah trees is independent of position, then the odds ratio will be one. If the odds ratio is greater than one, as for the bottom vs middle floodplain positions, then the odds of having dead coolibah trees is greater for one level of the other variable (bottom) than the other (middle). The converse is true if the odds ratio is less than one.

The sampling distribution of odds ratios is usually very skewed, especially for small sample sizes (Agresti 1990, 1996). To calculate a standard error and confidence interval for an odds ratio, we need to transform it to logs, which results in its sampling distribution being approximately normal. Note that an odds ratio of one (H_0 true) is a log odds ratio of zero. The asymptotic standard error for the odds ratio is:

$$\text{ASE}(\log \hat{\theta}) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} \qquad (14.10)$$

Confidence intervals for the odds ratio are best calculated on the log odds ratio and then back-transformed. The 95% CI is:

$$\pm z_{0.95}\,\text{ASE}(\text{log odds ratio}) \qquad (14.11)$$

where z is the critical value from a standard normal distribution. The antilog of these confidence limits will provide the CI for the odds ratio.

In the French & Westoby (1997) example, there is only one odds ratio because it is a two by two table. The estimated ratio is 2.43, so the odds of being dispersed by ants are 2.43 times greater for plant species that regenerate by seed than for those that regenerate vegetatively. However, the confidence interval for the odds ratio is wide (Box 14.2), which is not surprising since the test of independence was not significant. Note that the 95% CI includes one, indicating no evidence (at $\alpha = 0.05$) against the H_0 of independence.

For the Roberts (1993) two by three table, there can be three odds ratios for the presence of dead coolibah trees (bottom vs mid, bottom vs top, mid vs top). The odds of having dead coolibah trees were greater than one for all three comparisons (bottom versus middle, bottom versus top, middle versus top) but only for bottom versus top did the 95% CI for the odds ratio not include one (Box 14.3). So the major contribution to the lack of independence in Robert's (1993) data was the contrast between the bottom of the floodplain and the top of the floodplain.

Residuals

Another way of interpreting lack of independence in contingency tables is examining the pattern of the residuals, the difference between the observed and expected values ($n_{ij} - f_{ij}$). There will be a residual for each cell of the table and this is the same definition of a residual we used for linear models (e.g. Chapter 5). Absolute residuals are difficult to compare when the frequencies vary. For example, a ($n_{ij} - f_{ij}$) difference of five is more "important" when the frequencies are around ten than when the frequencies are around 100. Therefore, we usually standardize each residual by dividing by $\sqrt{(f_{ij})}$:

$$\frac{(n_{ij} - f_{ij})}{\sqrt{f_{ij}}} \qquad (14.12)$$

These are also called Pearson residuals (Agresti 1996) and are directly comparable irrespective of the absolute frequencies.

From the Roberts (1993) data, the standardized residuals showed that there were more quadrats with dead trees at the bottom of the dunes than expected and fewer quadrats with dead trees at the top of the dunes than expected (Box 14.3).

We can also calculate adjusted residuals as:

$$\frac{(n_{ij} - f_{ij})}{\sqrt{f_{ij}(1-p_{i+})(1-p_{+j})}} \qquad (14.13)$$

where p_{i+} is the proportion of the total row observations in that cell and p_{+j} is the proportion of the total column observations in that cell. Large residuals indicate large deviations from independence and the sign (+ or −) indicates more or less observed than expected under the H_0.

Small sample sizes

The χ^2 statistic, and the log-linear G^2 statistic described in Section 14.3, are based on frequencies in categories and can only be compared validly to the continuous χ^2 distribution if the sample size is big enough. We mentioned at the start of this chapter that we assume that no more than 20% of the categories have expected frequencies less than about five. What if sample sizes are smaller, i.e. we have a sparse contingency table, one with many low or zero frequencies?

Yate's correction for continuity was developed to improve the accuracy of the χ^2 test for two by two tables with small frequencies but is of debatable value and is now not regarded as necessary (Agresti 1990, Manly 1992) because of the availability of "exact" tests.

Fisher's Exact test was designed for two by two tables with fixed marginal totals. It does not use the χ^2 distribution to test the H_0 of independence but instead answers the question "Given our fixed marginal totals, what is the probability of obtaining the observed cell frequencies and all cell frequencies that are further away from the expected?". The calculations are tedious for anything but the smallest sample sizes, but it is available in most statistical software. Although Fisher's Exact test strictly should be used in situations where we have fixed marginal totals, it is commonly used more generally as a solution for small sample sizes even when both marginal totals are not fixed (e.g. Clinton & Le Beouf 1994). There are other exact tests for contingency tables more complex than two by two. These tests use resampling procedures (randomization tests – see Chapter 3) to generate an exact distribution for the χ^2 statistic rather than assuming it follows a χ^2 distribution but they require special software.

Another solution to small observed frequencies is to collapse or combine some categories. For example, when the categories are evenly spaced size classes, there might be few individuals in some of the larger classes. They can be combined into a single category that will have adequate frequencies for analysis.

14.2.2 Three way tables

An obvious extension of two way contingency tables is the addition of a third variable in the cross-classification. Again following Agresti's (1996) terminology, the three variables are labeled X (i equals 1 to I categories), Y (j equals 1 to J categories), and Z (k equals 1 to K categories) and we will use Y in cases where there is clearly one response variable. Remember that analyses of contingency tables do not usually distinguish response and predictor variables, unless the analysis uses a logit (logistic) model. However, the interpretation of the generalized linear models (log-linear models) we commonly use for complex contingency tables can depend on whether we clearly distinguish a response variable.

Two examples from the recent literature will illustrate three way contingency tables in a biological context. Sinclair & Arcese (1995) cross-classified wildebeest carcasses from the Serengeti by three variables: sex (X with I equals two: male, female), cause of death (Y with J equals two: predation, non-predation) and bone marrow type (Z with K equals three: solid white fatty, opaque gelatinous, translucent gelatinous, with the first indicating a healthy animal that is not undernourished) – Table 14.3(a). In this example, it is not clear that any of the variables could be classified as a "response" variable. We have a random sample of carcasses cross-classified by three variables, all of which can be considered responses. The analysis of these data is presented in Box 14.5.

Taulman *et al.* (1998) examined the demography of southern flying squirrels in response to experimental logging in southern Arkansas. They had a response variable: age of squirrel (Y with J equals two: adult, young). The other two variables were treatment from which squirrels were caught (X with I equals two: control, logged) and year (Z with K equals three: 1994, 1995, 1996) – see Table 14.3(b). They had pre-treatment data from 1993 but we will only consider the post-treatment data. A logit model (Section 13.2) could have been fitted to these data, with age as the response variable and treatment and year as the two categorical

Table 14.3 Observed frequencies for three way contingency tables from (a) Sinclair & Arcese's (1995) study on wildebeest carcasses cross-classified by cause of death, sex and marrow type and (b) Taulman *et al.*'s (1998) study on squirrels in logged and control stands over three years

(a)

		Marrow type			
Cause of death	Sex	SWF	OG	TG	Totals
Predation	Female	26	32	8	66
Predation	Male	14	43	10	67
Non-predation	Female	6	26	16	48
Non-predation	Male	7	12	26	45
Totals		53	113	60	226

(b)

		Age		
Treatment	Year	Adult	Juvenile	Totals
Control	1994	46	10	56
Harvest	1994	30	8	38
Control	1995	44	31	75
Harvest	1995	53	54	107
Control	1996	8	0	8
Harvest	1996	79	14	93
Totals		260	117	377

predictors. Note that there may be correlations between successive years in this study, although we will ignore these for the purposes of analysis. The analysis of these data is presented in Box 14.6.

In contrast to two way tables, there is more than one sort of (in)dependence between variables in three way tables. We can examine complete independence between all three variables (no interactions), various forms of conditional and marginal independence that we will describe in the next section, and also complete dependence where there is a three way interaction. While we can calculate expected cell frequencies and χ^2 statistics to test null hypotheses about these various forms of independence, it is more efficient to do so with log-linear models (Section 14.3.2).

Conditional independence and odds ratios

A three way table can be best interpreted by considering it as a set of partial tables, each of which is a two way table for each level of the third variable. For the wildebeest example, we can construct a partial table between sex and cause of death for each level of marrow type, i.e. partial table between X and Y for each level Z (Box 14.5). We could, of course, construct partial tables between Y and Z for each level of X and between X and Z for each level of Y. Conditional independence is where two variables are independent of each other given the level of (controlling for) the third variable, i.e. the two variables in each partial table are independent. For example, the proportions of wildebeest carcasses that suffered predation (or didn't) are independent of sex, for all marrow types. When two variables are not conditionally independent, we say they have a partial association, i.e. they are not independent for all levels of the third variable.

Odds ratios are important in the interpretation of conditional independence in three way

Box 14.5 Worked example of log-linear model for three way table: death in wildebeest (sex, predation and bone marrow type)

Sinclair & Arcese (1995) cross-classified 226 wildebeest carcasses from the Serengeti by three variables: sex (male, female), cause of death (predation, non-predation) and bone marrow type (solid white fatty, opaque gelatinous, translucent gelatinous, with the first indicating a healthy animal which is not undernourished).

We have fitted log-linear models with different combinations of terms. The fit of each model shown below is based on comparing observed and fitted cell frequencies and, equivalently, comparing the fit of each model to that of the saturated model with zero degrees of freedom. For hypothesis testing, we would fit these models hierarchically, starting with the most complex.

	Model	G^2	df	P	AIC
1	death + sex + marrow	42.76	7	<0.001	28.76
2	death × sex	42.68	6	<0.001	30.68
3	death × marrow	13.24	5	0.021	3.34
4	sex × marrow	37.98	5	<0.001	27.98
5	death × sex + death × marrow	13.16	4	0.011	5.16
6	death × sex + sex × marrow	37.89	4	<0.001	29.89
7	death × marrow + sex × marrow	8.46	3	0.037	2.46
8	death × sex + death × marrow + sex × marrow	7.19	2	0.027	3.19
9	Saturated (full) model	0	0		

The AIC chose model 7 as best fit, whereas G^2 chose model 8. The comparison of the fit of model 8 and the saturated model 9 is a test of the H_0 that there is no three way interaction. The G^2 deviance statistic results in rejection of this H_0. Standardized residuals under no three way interaction showed that more male wildebeest with SWF marrow and fewer with OG marrow were not killed by predators than expected.

		Marrow type		
Cause of death	Sex	SWF	OG	TG
Predation	Female	0.541	−0.730	0.719
Predation	Male	−0.641	0.709	−0.522
Non-predation	Female	−0.891	0.948	−0.425
Non-predation	Male	1.248	−1.088	0.364

We will also illustrate the tests for conditional independence and complete independence, although the presence of a three way interaction would usually preclude tests of two way interactions and the presence of both complete and conditional dependence would preclude testing complete independence. The relevant hierarchical comparisons of models are shown below.

Term	Models compared	G^2	df	P
Three way interaction				
death × sex × marrow	8 vs 9	7.19	2	0.027
Conditional independence				
death × sex	7 vs 8	1.28	1	0.259
death × marrow	6 vs 8	30.71	2	<0.001
sex × marrow	5 vs 8	5.97	2	0.051
Complete independence	1 vs 8	35.57	5	<0.001

This demonstrates that we would reject the H_0 of conditional independence of cause of death and marrow type.

The odds ratios for wildebeest killed by predation for each pair of marrow types separately for males and females are shown below.

	Odds ratio	95% CI
Male		
OG versus TG	0.107	0.041−0.283
SWF versus TG	0.192	0.060−0.616
SWF versus OG	0.558	0.184−1.693
Female		
OG versus TG	0.406	0.150−1.097
SWF versus TG	0.115	0.034−0.395
SWF versus OG	3.521	1.261−9.836

The conditional dependence is clearly shown by the complex pattern of odds ratios that is different for males and females. The odds of being killed by predation were less for male wildebeest with either OG or SWF marrow than TG marrow. The odds of males being killed by predators were the same for those with SWF marrow versus OG marrow. For females, the odds of being killed by predators were greater for those with SWF marrow than OG marrow but less for those with SWF marrow than TG. The odds of females being killed by predators were the same for those with OG marrow and TG marrow.

tables but are more difficult to calculate because we have three variables and odds ratios can only be calculated for two by two tables. Odds ratios can be derived for larger tables by breaking the table into two by two subsets so when the table dimensions are two by two by K, we can calculate conditional odds ratios for each set of partial tables (see Table 14.4).

One conditional odds ratio in the wildebeest study is the ratio of the odds that a male wildebeest carcass suffered predation to the odds that a female wildebeest carcass suffered predation, for one marrow type, i.e. if a carcass had marrow type SWG, are the odds of being eaten the same for males and females? Other odds ratios are the ratios of the odds that a male wildebeest carcass suffered predation to the odds that a female wildebeest carcass suffered predation for the other two marrow types. Conditional independence between Y and Z means that all the odds ratios between Y and Z equal one.

If conditional independence between two variables does not hold, then two possible patterns may occur. First, the odds ratios for two variables may all be different from one but still may be equal for all levels of the other variable, i.e. conditional dependence (association) exists between two variables but is the same for all levels of the third variable. For example, the ratio of the odds that a male wildebeest carcass suffered predation

Box 14.6 Worked example of log-linear model for three way table – demography of squirrels in response to disturbance: effects of logging and year on age

Taulman *et al.* (1998) examined the age of squirrels in relation to the treatment stand from which squirrels were caught (control, logged) and year (1994, 1995, 1996) – see Table 14.3(b). We considered age of squirrel as a response variable (treatment and year might affect the relative numbers of adult and young squirrels but not vice versa) so not all models were fitted. The interaction between treatment and year was never omitted because the investigator set these variables, so their conditional independence makes little sense.

	Model	G^2	df	P	AIC
1	treatment × year + age × year	4.13	3	0.248	0.00
2	treatment × age + treatment × year	46.27	4	<0.001	38.27
3	treatment × age + treatment × year + age × year	1.88	2	0.390	0.00
4	Saturated (full) model	0.00	0		

Either models 1 or 3 could have been chosen as best fit, with the G^2 suggesting model 3. Note that exclusion of both the three way interaction and the two way interaction between age and year results in a very poor fit. Since we have already shown that the three way interaction is not significant (model 3), this suggests that there is conditional dependence between age and year.

The relevant hierarchical comparisons of models for the tests for the three way interaction and the tests for conditional independence, with the interaction between treatment and caged years always in the models, are shown below.

Term	Models compared	G^2	df	P
Three way interaction				
treatment × age × year	3 vs 4	1.88	2	0.390
Conditional independence				
age × year	2 vs 3	44.39	2	<0.001
treatment × age	1 vs 3	2.24	1	0.134

There was no evidence to reject the hypothesis of conditional independence between age and treatment, i.e. squirrel age and treatment were independent for each year. In contrast, squirrel age and year were not independent, for control or logged treatments.

to the odds that a female wildebeest carcass suffered predation may be the same for each marrow type, even if the odds are greater for males than females consistently. This pattern is termed a homogeneous association between two variables. A homogeneous association implies no three variable interaction. Conditional independence is a special case of a homogeneous association.

Second, the pattern of dependence (association) between two variables may differ between levels of the third variable and, therefore, the odds ratios for two variables vary between the levels of the other variable. For example, the ratio of the odds that a male wildebeest carcass suffered predation to the odds that a female wildebeest carcass suffered predation are different for

Table 14.4 Partial table for $I=2$ by $J=2$ by K contingency table for $K=1$ with observed frequencies

	I	$J=1$	$J=2$
$K=1$	1	n_{11K}	n_{21K}
$K=1$	2	n_{12K}	n_{22K}

the different marrow types. This pattern indicates an interaction between all three variables and that the two variable associations will not have a simple interpretation.

The odds ratio for an I equals two by J equals two by K table, for a given level k of K, can be estimated as:

$$\hat{\theta}_{XY(k)} = \frac{n_{11k}n_{22k}}{n_{12k}n_{21k}} \qquad (14.14)$$

The odds ratios for cause of death in relation to marrow type for male and female wildebeest are presented in Box 14.5. The only odds ratio that is clearly greater than one is for female wildebeest, where the odds of a SWF marrow type animal being killed by a predator are three and half times the odds of an OG marrow type animal being killed by a predator. This indicates conditional dependence between cause of death and marrow type, where the dependence is conditional on sex.

A test for conditional independence in two by two by K tables is the Cochran–Mantel–Haenszel (C–M–H) test (Sokal & Rohlf 1995), which basically tests the null hypothesis that the conditional odds ratios between X and Y equal one for all levels of Z. It is particularly appropriate when there is no three variable (XYZ) interaction (Agresti 1996). The C–M–H statistic is converted to a χ^2 and compared to a χ^2 distribution; it is available in most statistical software. It can also be generalized for I by J by K tables where I and J are greater than two but the formulae are complex (Agresti 1990). For the squirrel example, C–M–H statistic equals 1.18 with P equals 0.530, so the ratio of the odds of a squirrel being an adult on control stands and the odds of a squirrel being an adult on logged stands were not different from one for all three years. The C–M–H test also allows a form of meta-analysis to combine the results from a number of independent two by two tables.

Marginal independence and odds ratios

Marginal tables are two way tables completely ignoring the third variable, e.g. the frequencies for X by Y pooling levels of Z. Marginal independence is independence between the two variables in the marginal table, pooling the levels of the third variable. For the squirrel example, one marginal table would be age crossed with treatment, pooling year (Box 14.6). From this marginal table, we would assess marginal independence as the independence of age and treatment combining years. We can also calculate marginal odds ratios from the marginal table. The odds of a squirrel being an adult are almost identical ($\hat{\theta}=0.996$) for control versus treatment stands, ignoring year.

Complete independence

The effects of the individual variables represent complete independence and no two or three way associations. For our two worked examples, the proportions of adult squirrels are independent of treatment and year and cause of death, sex and marrow type are completely independent of each other.

14.3 Log-linear models

The best method for analyzing contingency tables is with log-linear models. Log-linear models treat the cell frequencies as counts distributed as a Poisson random variable. Log-linear models are examples of generalized linear models (GLMs; see Chapter 13); the expected cell frequencies are modeled against the variables using the log link and a Poisson error term (Agresti 1996). As with other GLMs, we fit log-linear models and estimate their parameters using maximum likelihood techniques. ML fits for most complex log-linear models do not have simple solutions so iterative methods like the Newton–Raphson algorithm (Chapter 13) are required. The fit of the models is measured by the log-likelihood.

Log-linear models do not distinguish response and predictor variables; all the variables are considered equally as response variables. However,

there is a relationship between log-linear models and logit models (including logistic regression) discussed in Chapter 13. Logit models distinguish a response variable (with two categories in a logistic regression) and model it against predictors that can be continuous or categorical. A logit model with categorical predictors can also be analyzed as a log-linear model (Agresti 1996).

14.3.1 Two way tables

Two way tables were described in Section 14.2.1 and will be illustrated here with the example from Roberts (1993).

Full and reduced models

For a two way table (I by J), we can fit two log-linear models. The first is a saturated (full) model:

$$\log f_{ij} = \text{constant} + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY} \quad (14.15)$$

For the data from Roberts (1993), the saturated (full) model is:

$$\log f_{ij} = \text{constant} + \lambda_i^{\text{coolibah}} + \lambda_j^{\text{position}} + \lambda_{ij}^{\text{coolibah*position}} \quad (14.16)$$

In models 14.15 and 14.16:

f_{ij} is the expected frequency in cell ij, i.e. the expected number of quadrats in each combination of coolibah trees (alive, dead) and floodplain position (top, mid, low),

constant is the mean of the logs of all the expected frequencies,

λ_i^X is the effect of category i of variable X, i.e. the effect of coolibah trees being either alive or dead on the log expected frequency of quadrats in each cell,

λ_j^Y is the effect of category j of variable Y, i.e. the effect of floodplain position being top, mid or bottom on the log expected frequency of quadrats in each cell,

λ_{ij}^{XY} is the effect of any interaction between X and Y, i.e. an interactive effect of coolibah tree category and floodplain position on the log expected frequency of quadrats in each cell. The interaction measures deviations from independence of the two variables.

Models 14.15 and 14.16 fit the observed frequencies perfectly, hence the term saturated. Note that "effect" does not imply any causality, just the influence of a variable or interaction between variables on the log of the expected number of observations in a cell.

The second log-linear model represents independence of the two variables (X and Y) and is a reduced model:

$$\log f_{ij} = \text{constant} + \lambda_i^X + \lambda_j^Y \quad (14.17)$$

Again from Roberts (1993):

$$\log f_{ij} = \text{constant} + \lambda_i^{\text{coolibah}} + \lambda_j^{\text{position}} \quad (14.18)$$

The interpretation of models 14.17 and 14.18 is that the log of the expected frequency in any cell is a function of the mean of the log of all the expected frequencies plus the effect of floodplain position and the effect of the presence/absence of dead coolibah trees. Note that log-linear models do not distinguish one of the variables as a response variable, they just model the log of the expected frequencies. This is an additive linear model with no interaction between the two variables.

The parameters of log-linear models are the effects of a particular category of each variable on the expected frequencies; a larger λ means that the expected frequencies will be larger for that variable, i.e. that row or that column (Agresti 1996). These parameters are also deviations from the mean of all the log expected frequencies, just like parameters in ANOVA linear models are deviations from the overall mean. When λ is greater than zero, then the mean log expected frequency for that variable (row or column) is greater than the mean of all the log expected frequencies (Agresti 1990).

Null hypothesis of independence

The H_0 of independence in a two way table (Section 14.2.1) is also a test of the H_0 that λ_{ij}^{XY} equals zero, i.e. there is no interaction between the two variables. We can test this H_0 by comparing the fit of the model without this term (14.17) to the saturated model that includes this term (14.15). We determine the fit of each model by calculating the expected frequencies under each model, comparing the observed and expected frequencies and calculating the log-likelihood of each model. We then compare the fit of the two models with the likelihood ratio statistic (Λ), that is the ratio of the two log-likelihoods. However, sampling distribution of Λ is not well known (Sokal & Rohlf 1995), so instead we calculate the G^2 statistic (Chapter 13):

$$G^2 = -2\log\Lambda \quad (14.19)$$

G^2 follows a χ^2 distribution for reasonable sample sizes and can be generalized to:

$$G^2 = -2(\text{log-likelihood reduced model} - \text{log-likelihood full model}) \quad (14.20)$$

This is also termed the deviance and measures the difference in fit of the two models. If the H_0 of independence is true, then the reduced (no interaction) model should fit as well as the full model and the deviance (G^2) will be close to zero. If the H_0 is false, then there should be a difference in the fit of the two models and the deviance (G^2) will be greater than zero. The calculated G^2 is compared to a χ^2 distribution with $(I-1)(J-1)$ df, just like the χ^2 test of independence described in Section 14.2.1. The df $[(I-1)(J-1)]$ is the difference between the df for the full model $[(IJ-1)]$ and the df for the reduced model $[(I-1)+(J-1)]$.

Note that, for two way tables, the saturated model acts as the full model for model comparisons. This is not the case for more complex tables where many different full and reduced models can be fitted. For two way contingency tables with large sample sizes, the χ^2 test and the G^2 test will give similar results. Note that G^2 is slightly more sensitive to small sample sizes than the χ^2 statistic. In most statistical software, fitting the reduced model for a two way table will automatically provide the difference in fit between the two models.

Interpretation of lack of independence in log-linear models can be done using odds ratios and residuals, just as described in Section 14.2.1. Various types of residuals are standard output from log-linear modeling routines in most statistical software.

14.3.2 Log-linear models for three way tables

We will provide an introduction to log-linear models for three way tables. Sokal & Rohlf (1995) is also a good introduction and they provide a detailed worked example for a three way table. Agresti (1990) is a more statistically complete reference for log-linear modeling, although Agresti (1996) is a more readable version of that text for the mathematically disinclined.

Full and reduced models

For three way tables (X with I categories, Y with J categories, Z with K categories), there is a large number of full and reduced models for testing the different interactions and main effects. Like three factor ANOVA models, log-linear models for contingency tables with three variables include three main effects (X, Y, Z), three two variable interactions (XY, XZ, YZ) and one three variable interaction (XYZ). For a three way table (I by J by K), the saturated model is:

$$\log f_{ijk} = \text{constant} + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ} \quad (14.21)$$

For the wildebeest example (Sinclair & Arcese 1995), this saturated model is:

$$\log f_{ijk} = \text{constant} + \lambda^{\text{death}} + \lambda^{\text{sex}} + \lambda^{\text{marrow}} + \lambda^{\text{death}\times\text{sex}} + \lambda^{\text{death}\times\text{marrow}} + \lambda^{\text{sex}\times\text{marrow}} + \lambda^{\text{death}\times\text{sex}\times\text{marrow}} \quad (14.22)$$

In models 14.21 and 14.22:

f_{ijk} is the expected frequency in cell ijk, i.e. the expected number of carcasses in each combination of death (predation, non-predation), sex (male, female) and bone marrow type (solid white fatty, opaque gelatinous, translucent gelatinous),

constant is the mean of the logs of all the expected frequencies,

λ_i^X is the effect of category i of variable X, i.e. the effect of type of death on the log expected frequency of carcasses in each cell,

λ_j^Y is the effect of category j of variable Y, i.e. the effect of being male or female on the log expected frequency of carcasses in each cell,

λ_k^Z is the effect of category k of variable Z, i.e. the effect of bone marrow type on the log expected frequency of carcasses in each cell,

λ_{ij}^{XY} is the effect of any interaction between X and Y, i.e. an interactive effect of type of death and sex on the log expected frequency of carcasses in each cell,

λ_{ik}^{XZ} is the effect of any interaction between

Table 14.5 Some typical log-linear models fitted to a three way (*X* by *Y* by *Z*) table with their df: comparisons of models are tested with the difference between the relevant df

Log-linear model	df
$X+Y+Z$	$IJK-I-J-K+2$
$X+Y+Z+XY$	$(K-1)(IJ-1)$
$X+Y+Z+XZ$	$(J-1)(IK-1)$
$X+Y+Z+YZ$	$(I-1)(JK-1)$
$X+Y+Z+XZ+YZ$	$K(I-1)(J-1)$
$X+Y+Z+XY+YZ$	$J(I-1)(K-1)$
$X+Y+Z+XY+XZ$	$I(J-1)(K-1)$
$X+Y+Z+XY+XZ+YZ$	$(I-1)(J-1)(K-1)$
Saturated model:	
$X+Y+Z+XY+XZ+YZ+XYZ$	0

X and Z, i.e. an interactive effect of type of death and bone marrow type on the log expected frequency of carcasses in each cell,

λ_{jk}^{YZ} is the effect of any interaction between Y and Z, i.e. an interactive effect of sex and bone marrow type on the log expected frequency of carcasses in each cell,

λ_{ijk}^{XYZ} is the effect of any interaction between X, Y, and Z, i.e. an interactive effect of type of death, sex and bone marrow type on the log expected frequency of carcasses in each cell.

Models 14.21 and 14.22 include all main effects, all two way interactions and the three way interaction and fit the observed frequencies perfectly.

Because the G^2 goodness-of-fit statistic for the saturated model 14.21 is zero, then the G^2 statistic for any model represents the difference in fit of that model to the fit of the saturated model 14.21, i.e. the deviance. We can also use criteria of fit that "penalize" the model for the number of parameters, such as the Akaike Information Criterion, which for a particular model equals (Christensen 1997):

$$\begin{aligned}\text{AIC} &= G^2 - (\text{df}_{\text{Saturated model}} - 2\text{df}_{\text{Particular model}}) \\ &= G^2 - 2\text{df}_{\text{Test of model}}\end{aligned} \quad (14.23)$$

The choice of "best" model is that which minimizes either the G^2 or the AIC.

Log-linear models are usually fitted in a hierarchical fashion, i.e. the inclusion of a higher order term automatically includes all lower order terms with those variables. The model with the three variable interaction automatically includes all two way interactions and main effects. Similarly, a model which omits one or more two way interactions also must omit the three way interaction. The range of models that can be fitted for a three way table are listed in Table 14.5.

The saturated model allows for complete dependence of the three variables by including the three way interaction term. The remaining models each omit the three way interaction and one or more two way interactions. Three models omit both the three way interaction and one of the two way interactions. For example, consider the model:

$$\log f_{ijk} = \text{constant} + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ} \quad (14.24)$$

Model 14.24 implies that X and Z are conditionally independent, i.e. the odds ratios for the association between X and Z are equal to one for all levels of Y. The goodness-of-fit statistics for these models omitting a two variable interaction compare their fit to that of the saturated model and measure how much the absence of both the three way interaction and the particular two way interaction affects the fit of the model. If the three way interaction has been shown to be small, then the fit of these models really measures the effect of omitting the particular two way interaction, i.e. testing whether those two variables are conditionally independent.

In the wildebeest example from Sinclair & Arcese (1995), the model which includes death, sex, marrow, death × sex and sex × marrow is:

$$\log f_{ijk} = \text{constant} + \lambda^{\text{death}} + \lambda^{\text{sex}} + \lambda^{\text{marrow}} + \lambda^{\text{death}\times\text{sex}} + \lambda^{\text{sex}\times\text{marrow}} \quad (14.25)$$

Model 14.25 implies that there is no partial association between cause of death and marrow type for any sex. For either males or females, whether a wildebeest is taken by a predator or not is independent of which marrow type they have.

In the study on the effects of logging on squirrel demography from Taulman *et al.* (1998), the variable squirrel age (adult, young) can be viewed as a response variable and therefore all models should include the interaction between the other two variables (treatment and year). These two variables are set by the investigators and it makes no sense for the interaction between them to be zero; their conditional independence (independence of treatment and year for adult or young squirrels) has no biological meaning (see also Agresti 1996, Sokal & Rohlf 1995). Therefore, the number of models to be fitted is less than for the wildebeest example (Box 14.6).

Therefore, we test the fit of models with the relevant two way interaction terms (treatment × age and year × age) omitted. These models imply that there is conditional independence between treatment and age for each year and conditional independence between age and year for each treatment.

Note that the comparison of models that omit one of the two way interactions to the saturated model are not the best for testing the absence of two way interactions (conditional independence). This is because the reduced model has omitted both a two way interaction and the three way interaction so any difference between this model and the saturated model could be due to either the two way or the three way interaction or both. In general, the comparison of models omitting interaction terms to the saturated model should be considered an initial exploratory or screening approach to analyzing a contingency table. The exception is the valid test of the three way interaction.

Three other models omit the three way interaction and two of the two way interactions. For example, the model:

$$\log f_{ijk} = \text{constant} + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} \quad (14.26)$$

implies that X and Z are conditionally independent for each level of Y and that Y and Z are conditionally independent for each level of X. Only X and Y can be conditionally dependent. So the model that includes death, sex, marrow and death × marrow:

$$\log f_{ijk} = \text{constant} + \lambda^{\text{death}} + \lambda^{\text{sex}} + \lambda^{\text{marrow}} + \lambda^{\text{death}\times\text{sex}} \quad (14.27)$$

implies that cause of death and sex are conditionally independent for each level of marrow type and sex and marrow type are conditionally independent for each cause of death; only cause of death and marrow type are conditionally dependent.

The simplest possible model is one that assumes complete independence and excludes all interaction terms:

$$\log f_{ijk} = \text{constant} + \lambda_i^X + \lambda_j^Y + \lambda_k^Z \quad (14.28)$$

Model 14.28 implies that each variable is completely independent of the other two, e.g. the cause of death is independent of sex and marrow type.

The fit of the different possible models is a useful exploratory step in analyzing complex contingency tables and we can determine the model that provides the best fit for the fewest parameters. For the wildebeest carcasses example (Box 14.5), the two criteria (G^2 and AIC) chose different models, although the difference in fit between models 3, 5, 7 and 8 was minor. Based on the AIC, we would choose the model 7:

$$\log f_{ijk} = \text{constant} + \lambda^{\text{death}} + \lambda^{\text{sex}} + \lambda^{\text{marrow}} + \lambda^{\text{death}\times\text{marrow}} + \lambda^{\text{sex}\times\text{marrow}} \quad (14.29)$$

whereas based on the G^2, we would choose model 8:

$$\log f_{ijk} = \text{constant} + \lambda^{\text{death}} + \lambda^{\text{sex}} + \lambda^{\text{marrow}} + \lambda^{\text{death}\times\text{sex}} + \lambda^{\text{death}\times\text{marrow}} + \lambda^{\text{sex}\times\text{marrow}} \quad (14.30)$$

The AIC chose a model with fewer parameters.

In practice, however, we are usually more interested in tests of individual terms in the models. Comparisons of reduced models to the saturated model only do this in the case of the three way interaction. For the remaining models, more than one term is being omitted. Testing individual terms relates to the different forms of independence (complete, conditional, marginal) discussed in Section 14.2.2 and these tests are done by comparing the fit of full (not saturated) and reduced models.

Tests for three way interaction: complete dependence

The test of the three way interaction is a test of complete dependence. If the H_0 of no three way interaction is true, we have either conditional independence between all pairs of variables or the pattern of conditional dependence between all pairs of variables is the same for all levels of the third variable. This is similar to the interpretation of a three way interaction in an ANOVA model (Chapter 9) where the interaction between two factors depends on the level of the third factor. We test the three way interaction by comparing the fit of the saturated model, which is also the full model for the test of this term:

$$\log f_{ijk} = \text{constant} + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ} \quad (14.21)$$

to a reduced model that omits this term:

$$\log f_{ijk} = \text{constant} + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} \quad (14.31)$$

This tests the H_0 that the three way interaction term is zero. If this H_0 is true, then we have homogeneous association where each pair of variables can be conditionally dependent but this dependence is the same at each level of the third variable. If H_0 is true, we would expect models 14.21 and 14.31 to fit similarly; if the H_0 is false, we would expect the reduced model to fit significantly worse than the saturated model. We use the difference in G^2 for the reduced model and the saturated (full) model, i.e. the deviance, to test whether there is a significant three way interaction between the variables.

For the wildebeest example (Box 14.5), omitting the three way interaction term (sex × death × marrow) results in significantly worse fit so we would reject the null hypothesis of no three way interaction. The conditional dependence of cause of death and sex depends on the type of marrow. Equivalently, the conditional dependence of cause of death and marrow type depends on sex and the conditional dependence of sex and marrow type depends on cause of death. As in factorial ANOVAs (Chapter 9), interactions in log-linear models are symmetric.

For the squirrel example (Box 14.6), it is clear that omitting the three way interaction term (treatment × age × year) makes little difference to the fit of the model, so we wouldn't reject the H_0 that the three way interaction term is zero. Any conditional dependence between age of captured squirrels and treatment does not depend on year and any conditional dependence between age of captured squirrels and year does not depend on treatment.

Testing and interpreting two way interactions

Whether we test other terms depends on whether we reject the H_0 of no three way interaction between the variables. In the wildebeest carcass example, the three way interaction was significant so we could proceed in two ways. First, by examining the residuals from the model without the three way interaction term to see which cells were causing the lack of independence among the three variables (Box 14.5). The largest residuals indicate that there are more male carcasses that were not killed by predation with SWF marrow and fewer with OG marrow. None of the residuals is near two so we would not consider any observations particularly unusual. We could also examine odds ratios by breaking the table into a series of two way tables, e.g. tables of marrow type by sex for each cause of death separately. Second, we could examine dependence of pairs of variables for each level of the third variable separately, analogous to simple interaction tests in three factor ANOVA models (Chapter 9).

Although the three way interaction was significant in the wildebeest example, we will test for conditional dependence of each pair of variables to illustrate the process. Conditional independence is tested by comparing the full model

$$\log f_{ijk} = \text{constant} + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} \quad (14.31)$$

with each of the following reduced models

Test H_0: $\lambda_{ij}^{XY} = 0$ $\log f_{ijk} = \text{constant} + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$ (14.32)

Test H_0: $\lambda_{ik}^{XZ} = 0$ $\log f_{ijk} = \text{constant} + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ}$ (14.33)

Test H_0: $\lambda_{jk}^{YZ} = 0$ $\log f_{ijk} = \text{constant} + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ}$ (14.34)

For the wildebeest example, we would only reject the H_0 of conditional independence for cause of death and marrow type for each sex separately (Box 14.5). This means that cause of death and marrow type are not independent for male wildebeest and female wildebeest carcasses and the odds ratios for the association between cause of death and marrow type are different for each sex. In contrast, the odds ratios for the cause of death and sex association equal one for all marrow types and the odds ratios for the sex and marrow type association equal one for all causes of death.

In the squirrel example, the absence of a three way interaction is not rejected so there is good justification for proceeding to examine simpler models (Box 14.6). Because the treatment and year variables are set by the investigators, the independence between these two variables is not tested. There was no evidence to reject the hypothesis of conditional independence between age and treatment, i.e. squirrel age and treatment were independent for each year. This indicates that logging does not alter the relative numbers of adult and young squirrels compared to control stands in any year. In contrast, squirrel age and year were not independent in both control and logged treatments and the odds ratios for the association between age and year are different for each treatment.

We can also test for marginal independence of two variables by creating a two way table ignoring the third variable. For example, the test for marginal independence of cause of death and marrow type, ignoring sex, is done with a test of independence of the two way cause of death and marrow type table pooling the two sexes:

$G^2 = 29.52$, df = 2, $P < 0.001$

In this example, cause of death and marrow type are not marginally independent, as they are not conditionally independent, although agreement between marginal and conditional independence does not always hold (see Agresti 1996).

Test for complete independence

If none of the two way interactions are significant, we could fit the model of complete independence (no interactions) among the three variables:

$$\log f_{ijk} = \text{constant} + \lambda_i^X + \lambda_j^Y + \lambda_k^Z \quad (14.28)$$

Assuming there is no three way interaction, we can test the H_0 that all two way interactions equal zero (i.e. that the three variables are completely independent) by comparing model 14.28 to:

$$\log f_{ijk} = \text{constant} + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} \quad (14.31)$$

This comparison tests that all three variables are completely independent of each other, both conditionally and marginally. In the wildebeest example, marrow type is completely independent of cause of death and sex, sex is completely independent of cause of death and marrow type, and cause of death is completely independent of sex and marrow type (Box 14.5). There are no conditional dependencies.

We would not do this test for the wildebeest example because there is a three way interaction, nor for the squirrel example because the interaction between treatment and year should always be included because these variables are set by the investigator and independence between them makes little sense.

Analysis of deviance tables

We can create a modified analysis of deviance table, which gives the difference in G^2 between hierarchical models, showing tests for the null hypotheses that specific terms equal zero (Chapter 13). It is always better to compare the fit of full and reduced models when testing specific terms in log-linear models. Simple goodness-of-fit statistics for a given model can overestimate the importance of specific terms and should be used as an exploratory tool (except for the three way interaction). Comparing full and reduced models in a hierarchical manner is the most common method

of analyzing and presenting the results of log-linear modeling.

14.3.3 More complex tables

Log-linear models for four way and higher tables follow the logic described above, although interpretation of four way interactions is as difficult as the interpretation of four way interactions in complex ANOVA models (Chapter 9). Agresti (1990) has provided a worked example for a four way table using log-linear models.

14.4 General issues and hints for analysis

14.4.1 General issues

- Contingency tables represent a cross-classification of sampling or experimental units by two or more variables so each cell in the table contains a number of units (frequency).
- Log-linear models are GLMs that relate the log of the expected frequencies to a linear combination of the variables and their interactions.
- For two way tables, the basic χ^2 test is for independence between the two variables.
- To test H_0 that a specific term equals zero, compare the fit of the full model with that term included to the reduced model with that term omitted, using the deviance.
- Conditional independence means that two variables are independent for all levels of the third variable. Odds ratios and standardized residuals are very important tools for interpreting lack of independence in contingency tables.
- Standard significance tests can be unreliable when expected frequencies are small (less than five). Use exact tests for two way contingency tables with small sample sizes.

14.4.2 Hints for analysis

- Remember that log-linear models do not distinguish a response variable. However, when one variable is clearly a response, then some log-linear models won't make much sense. If modeling a response variable is important, consider logit models.
- As an initial analysis, it is useful to test the goodness-of-fit of a range of possible models using the deviance and AIC.
- For a complex table, breaking it into two by two by K sub-tables will allow odds ratios for conditional dependence to be calculated.

Chapter 15

Introduction to multivariate analyses

15.1 Multivariate data

A multivariate data set includes more than one variable recorded from a number of replicate sampling or experimental units, sometimes referred to as objects. If these objects are organisms, the variables might be morphological or physiological measurements; if the objects are ecological sampling units, the variables might be physicochemical measurements or species abundances. We have already considered multivariate data in linear models with two or more predictor variables, e.g. multiple regression (Chapter 6) and multifactor analysis of variance (Chapters 9–11). For these analyses, we have multiple predictor (independent) variables. The multivariate analyses we will discuss in the remaining chapters either deal with multiple response variables (e.g. MANOVA – Chapter 16) or multiple variables that could be response variables, predictor variables or a combination of both. This chapter will introduce some aspects of multivariate data and analysis that apply generally to many of the methods we will describe in the subsequent three chapters. We will illustrate these aspects with four data sets from the recent biological literature. For each data set, there are $i = 1$ to n objects with $j = 1$ to p variables measured for each object.

Chemistry of forested watersheds

In Chapter 2, we first described the study of Lovett *et al.* (2000) who examined the chemistry of forested watersheds in the Catskill Mountains in New York. They chose 39 first and second order streams (objects) and measured the concentrations of ten chemical variables (NO_3^-, total organic N, total N, NH_4^-, dissolved organic C, SO_2^{2-}, Cl^-, Ca^{2+}, Mg^{2+}, H^+), averaged over three years, and four watershed variables (maximum elevation, sample elevation, length of stream, watershed area).

Plant functional groups and leaf characters

In Chapter 9, we described the study of Reich *et al.* (1999) who examined the generality of leaf traits from different species across a range of ecosystems and geographic regions. We will use a subset of their data, Wisconsin forbs, with ten species as the objects. There were five variables measured for each species: specific leaf area, leaf nitrogen concentration, mass-based net photosynthetic capacity, area-based net photosynthetic capacity and leaf diffusive conductance at photosynthetic capacity.

Wildlife underpasses in Canada

Clevenger & Waltho (2000) reported on the effectiveness of road underpasses for wildlife in Banff National Park in Alberta, Canada. For part of their study, they quantified the human activity at the underpasses as numbers of people on bikes, on horses and on foot. The objects were the eleven underpasses and the variables were the three human activities and the data were counts.

Bats and African woodlands

Fenton *et al.* (1998) studied the effects of woodland disturbance on species richness and abundance of bats in northern Zimbabwe. They had four groups

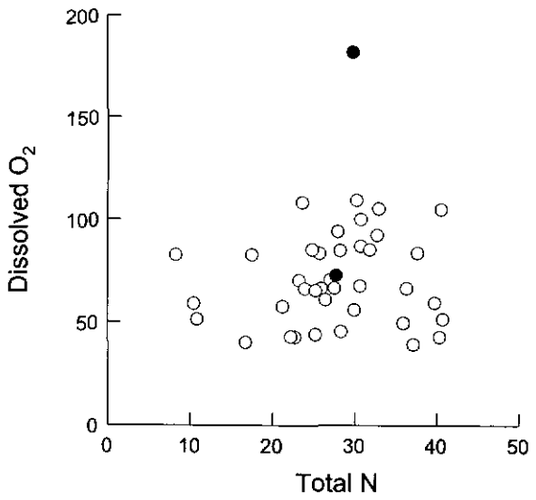

Figure 15.1 Scatterplot of dissolved oxygen against total nitrogen for 39 streams from Lovett *et al.* (2000). The centroid, the point represented by the mean of dissolved oxygen and total nitrogen, is filled. In this example, one object (grey fill) is an outlier for dissolved oxygen and also a multivariate outlier.

of sites: nine intact and nine impacted sites in Mana, six intact sites in Kanyati and six impacted sites in Matusadona. The sites within each area and disturbance category are not true replicates for assessing effects of disturbance so, like Fenton *et al.* (1998), we will combine the sites within each group. There were four objects (area and disturbance combinations) and 15 variables, species of bats. The data were numbers of each species of bat and there were numerous zero values, i.e. species absent.

15.2 | Distributions and associations

In a univariate context, we can describe the distribution of each variable and many of the parametric univariate analyses for estimating linear models and testing hypotheses about their parameters assume that the distribution of the response variable being analyzed is of a particular form (Chapters 5, 6, 8–14). For example, classical linear models assume normality (although the analyses are robust to this assumption under many circumstances), while generalized linear models allow other distributions from the exponential family (e.g. binomial, Poisson, etc.). Although the multivariate analyses we will introduce in the next three chapters are mainly descriptive, interval estimation and hypothesis tests of parameters can also be relevant and usually require the assumption of multivariate normality, where all variables and linear combinations of variables are normally distributed (Tabachnick & Fidell 1996). The simplest multivariate normal distribution is the bivariate normal distribution described in Chapter 5. Other multivariate distributions are obviously possible, although less commonly used in multivariate analyses.

One measure of the center of a multivariate distribution is the centroid. In multivariate space where each dimension is a variable, the centroid is the point represented by the univariate means of the distributions of each of the variables (Figure 15.1). The centroid is not usually estimated by a single value but is used as a description of the center of a multivariate normal distribution and for detecting multivariate outliers (Section 15.9.1).

We can summarize variation in single variables by sums-of-squares (SS) and variances (Chapter 2). When we have more than one variable, we not only have variances for each variable but also covariances between variables. To represent variation in multivariate data sets, we must use some simple matrix algebra. A data matrix (Y) for n objects by p variables is represented in Table 15.1, and illustrated using the data from Reich *et al.* (1999) for Wisconsin shrubs.

With more than one variable, we calculate both sums-of-squares for each variable and sums-of-cross-products between variables to get a p by p sums-of-squares-and-cross-products (SSCP or S) matrix (Table 15.2). The rows and columns of this matrix represent the variables ($j = 1$ to p). The main diagonal of this matrix contains the sums-of-squares for each variable. The other entries are the sums-of-cross-products, the sum of the product of the deviations of the value for each variable from its sample mean. Note that this matrix is symmetrical, i.e. the sum-of-cross-products between Y_1 and Y_2 is the same as the sum-of-cross-products between Y_2 and Y_1.

We can convert this matrix to a p by p matrix of variances and covariances (C) by dividing the

Table 15.1 Raw data matrix of p variables ($j = 1$ to p) for n objects ($i = 1$ to n), illustrated with data from Reich *et al.* (1999) for eleven species of Wisconsin forbs (objects) and five variables

$$\begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \cdots & \cdots & y_{ij} & \cdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{bmatrix}$$

	SLA ($cm^2\ g^{-1}$)	Leaf N ($mg\ g^{-1}$)	A_{mass} ($nmol\ g^{-1}\ s^{-1}$)	A_{area} ($\mu mol\ m^{-2}\ s^{-1}$)	G_s ($mmol\ m^{-2}\ s^{-1}$)
Caulophyllum thalictroides	425.0	58.2	254.0	5.9	134
Dentaria laciniate	297.0	53.0	432.0	14.2	227
Erythronium americanum	222.0	42.0	263.0	11.9	359
Silphium terebinthinaceum	133.0	14.4	175.0	13.4	615
Podophyllum peltatum	309.0	44.7	244.0	7.9	164
Baptisia leucophaea	106.3	35.9	159.0	15.0	481
Trillium grandiflora	357.0	51.6	209.0	5.8	499
Echinacea purpurea	128.5	15.0	122.9	9.8	480
Silphium integrifolium	116.3	16.6	116.0	10.0	478
Sanguinaria canadensis	321.0	53.6	255.0	7.9	208
Sarrachenia purpurea	78.1	11.4	22.8	2.9	144

Note:
SLA is specific leaf area, leaf N is leaf nitrogen concentration, A_{mass} is mass-based net photosynthetic capacity, A_{area} is area-based net photosynthetic capacity and G_s is leaf diffusive conductance at photosynthetic capacity.

sums-of-squares and sums-of-cross-products by their degrees of freedom ($n - 1$), where the main diagonal contains the variances for each variable and the other entries are the covariances between pairs of variables (Table 15.3). The covariance matrix can also be obtained directly from the raw data matrix Y, if each variable is centered (to a mean of zero), by $\mathbf{Y}'\mathbf{Y}/(n-1)$, where $\mathbf{Y}^1$ is the transpose of the centered raw data matrix.

There are two ways we can summarize the variability of a multivariate data set based on the variance–covariance matrix (Jackson 1991).

- The determinant of a square matrix is a single number summary of the matrix. The determinant of the variance–covariance matrix ($|\mathbf{C}|$) represents the generalized variance of the matrix.
- The trace of the variance–covariance matrix (Tr(C)) is the sum of the diagonal values, i.e. the sum of the variances of the centered individual variables.

Finally, we can also standardize these covariances by dividing by the standard deviations of the two variables involved to produce correlations and thus a correlation matrix (R), where r_{12} is the correlation coefficient between variables 1 and 2, etc. (Table 15.4). Note the main diagonal consists of ones because the correlation between each variable and itself is one. Covariances and correlations are measures of association between variables. Other measures of association include the χ^2 statistic, discussed in Chapter 14 as a measure of association for contingency tables.

If our objects occur in groups (e.g. experimental treatments), then we can calculate these matrices for between and within groups, analogous to analyses of variance in Chapters 8–11. Analyses based on multiple variance–covariance matrices

Table 15.2 | Sums-of-squares-and-cross-products matrix between p variables ($j = 1$ to p) for n objects ($i = 1$ to n), illustrated with data from Reich *et al.* (1999)

$$\begin{bmatrix} \sum_{i=1}^{n}(y_{i1}-\bar{y}_1)^2 & \sum_{i=1}^{n}(y_{i2}-\bar{y}_2)(y_{i1}-\bar{y}_1) & \cdots & \sum_{i=1}^{n}(y_{ip}-\bar{y}_p)(y_{i1}-\bar{y}_1) \\ \sum_{i=1}^{n}(y_{i1}-\bar{y}_1)(y_{i2}-\bar{y}_2) & \sum_{i=1}^{n}(y_{i2}-\bar{y}_2)^2 & \cdots & \sum_{i=1}^{n}(y_{ip}-\bar{y}_p)(y_{i2}-\bar{y}_2) \\ \cdots & \cdots & \sum_{i=1}^{n}(y_{ij}-\bar{y}_j)^2 & \cdots \\ \sum_{i=1}^{n}(y_{i1}-\bar{y}_1)(y_{ip}-\bar{y}_p) & \sum_{i=1}^{n}(y_{i2}-\bar{y}_2)(y_{ip}-\bar{y}_p) & \cdots & \sum_{i=1}^{n}(y_{ip}-\bar{y}_p)^2 \end{bmatrix}$$

	SLA	Leaf N	A_{mass}	A_{area}	G_s
SLA	144 120.13				
Leaf N	19 873.03	3335.73			
A_{mass}	87 160.14	15 162.00	112 204.77		
A_{area}	−1290.94	−23.86	1635.93	148.98	
G_s	−97 696.97	−14 505.68	−50 261.55	3412.31	301 594.73

Note:
Main diagonal entries are sums-of-squares, off diagonal entries are sums-of-cross-products. Variables defined in Table 15.1.

Table 15.3 | Variance–covariance matrix between p variables ($j = 1$ to p), illustrated with data from Reich *et al.* (1999)

$$\begin{bmatrix} s_1^2 & s_{12}^2 & \cdots & s_{p1}^2 \\ s_{12}^2 & s_2^2 & \cdots & s_{p2}^2 \\ \cdots & \cdots & s_j^2 & \cdots \\ s_{1p}^2 & s_{2p}^2 & \cdots & s_p^2 \end{bmatrix}$$

	SLA	Leaf N	A_{mass}	A_{area}	G_s
SLA	14 412.01				
Leaf N	1987.30	333.57			
A_{mass}	8716.01	1516.20	11 220.48		
A_{area}	−129.09	−2.39	163.59	14.89	
G_s	−9769.69	−1450.57	−5026.16	341.23	30 159.47

Note:
Main diagonal entries are variances, off diagonal entries are covariances. Variables defined in Table 15.1.

Table 15.4 | Correlation matrix between p variables ($j = 1$ to p), illustrated with data from Reich *et al.* (1999)

$$\begin{bmatrix} 1 & r_{21} & \cdots & r_{p1} \\ r_{12} & 1 & \cdots & r_{p2} \\ \cdots & \cdots & 1 & \cdots \\ r_{1p} & r_{2p} & \cdots & 1 \end{bmatrix}$$

	SLA	Leaf N	A_{mass}	A_{area}	G_s
SLA	1.00				
Leaf N	0.91	1.00			
A_{mass}	0.69	0.78	1.00		
A_{area}	−0.28	−0.03	0.40	1.00	
G_s	−0.47	−0.46	−0.27	0.51	1.00

Note:
All entries are Pearson correlations. Variables defined in Table 15.1.

nearly always have the assumption that the within-groups matrices have equal variances and covariances.

15.3 Linear combinations, eigenvectors and eigenvalues

15.3.1 Linear combinations of variables

One of the fundamental techniques in multivariate analyses is to derive linear combinations of the variables that summarize the variation in the original data set. Basically, we are "consolidating" (*sensu* Tabachnick & Fidell 1996) the variance from a data matrix into a new set of derived variables, each of which is a linear combination of the original variables. For $i = 1$ to n objects and $j = 1$ to p original variables:

$$z_{ik} = c_1 y_{i1} + c_2 y_{i2} + \cdots c_j y_{ij} + \cdots + c_p y_{ip} \quad (15.1)$$

In Equation 15.1, z_{ik} is the value of the new variable k for object i, y_{i1} to y_{ip} are the values of the original variables for object i and c_1 to c_p are weights or coefficients that indicate how much each original variable contributes to the linear combination. Depending on the analysis, these new variables are termed, variously, discriminant functions, canonical functions or variates, principal components or factors. This linear combination is analogous to a regression equation. For some analyses, the linear combination may include a constant (an intercept in regression terminology):

$$z_{ik} = \text{constant} + c_1 y_{i1} + c_2 y_{i2} + \cdots c_j y_{ij} + \cdots + c_p y_{ip} \quad (15.2)$$

The form in Equation 15.2 is common when the variables are not standardized to zero mean and unit variance; if they are, then the constant becomes zero and Equation 15.1 is appropriate.

The derived variables are extracted so the first explains most of the variance in the original variables, the second explains most of the remaining variance after the first has been extracted but is uncorrelated with the first, the third explains most of the remaining variance after the first and second have been extracted but is uncorrelated with either the first or second, etc. The new derived variables are independent of, uncorrelated with, each other. The number of new derived variables is the same as the number of original variables (p), although the variance is usually consolidated in the first few derived variables.

15.3.2 Eigenvalues

Eigenvalues, also termed characteristic or latent roots ($\lambda_1, \lambda_2, \lambda_3, \ldots \lambda_k \ldots \lambda_p$), represent the amount of the original variance explained by each of the

$k = 1$ to p new derived variables. These eigenvalues are population parameters and we estimate them using maximum likelihood (ML) to produce (l_1, l_2, l_3,...l_k...l_p) and can also determine their approximate standard errors. Note from Box 15.1 that if we use a covariance matrix and centered variables, then the sum of the eigenvalues is equal to the trace of the original covariance matrix, i.e. the sum of the variances of the original centered variables. If we use a correlation matrix and centered and standardized variables, the sum of the eigenvalues would equal the trace of the correlation matrix, i.e. the sum of the variances of the original standardized variables. We have simply rearranged the variance in the association matrix so that the first few derived variables explain most of the variation that was present (between objects) in the original variables. The eigenvalues can also be expressed as proportions or percentages of the original variance explained by each new derived variable (component).

15.3.3 Eigenvectors

Eigenvectors (characteristic vectors) are lists of the coefficients or weights showing how much each original variable contributes to each new derived variable. In general terms, the eigenvectors contain the c_j in Equation 15.1 but these coefficients can be scaled in different ways so are often represented as u_j, v_j or w_j in matrix descriptions of

Box 15.1 Deriving components (modified from Jackson 1991)

There are two different strategies for extracting eigenvectors (components) and their eigenvalues from multivariate data set of n objects by p variables. First, we can use a spectral decomposition of a p by p association matrix between variables. Second, we can use a singular value decomposition (SVD) of a n by p data matrix, with variables standardized as necessary. The SVD is more generally applicable (see Chapter 17) although most biologists are more familiar with obtaining eigenvectors and eigenvalues from a covariance or correlation matrix.

Consider the matrix (**Y**) of raw data from Clevenger & Waltho (2000) who recorded the numbers of people on bicycles, horses and on foot for eleven underpasses also used by wildlife in Alberta, Canada.

	Raw			Centered		
Underpass	Bicycle	Horse	Foot	Bicycle	Horse	Foot
1	0	6	7	−118.727	−37.273	−55.364
2	5	3	45	−113.727	−40.273	−17.364
3	6	6	14	−112.727	−37.273	−48.364
4	21	5	20	−97.727	−38.273	−42.364
5	189	42	34	70.273	−1.273	−28.364
6	8	138	77	−110.727	94.727	14.636
7	462	186	129	343.273	142.727	66.636
8	19	12	80	−99.727	−31.273	17.636
9	595	58	241	476.273	14.727	178.636
10	1	10	10	−117.727	−33.273	−52.364
11	0	10	29	−118.727	−33.273	−33.364

Spectral decomposition

We will illustrate spectral decomposition of a matrix of associations between variables (**Y'Y**). This might be a matrix of variances and covariances, **C**, among p variables based on n objects (Table 15.3).

	Bicycle	Horse	Foot
Bicycle	44 906.018		
Horse	7336.382	3862.018	
Foot	13 084.709	2205.191	4903.655

Note that we could also use a correlation matrix. Basically, we then derive two matrices, **L** and **U**, so that:

L = U'CU

U is a n by p matrix whose columns contain the eigenvectors (characteristic vectors), the coefficients of the linear combinations of the original variables. The elements of each eigenvector k are u_{jk}, the coefficient for the jth variable in the kth eigenvector. Note that we clearly need to have to some constraints imposed on the coefficients within each eigenvector, otherwise simply increasing the absolute sizes of the coefficients could increase the variance explained by each new variable. The simplest and most commonly used constraint is to restrict the sum of squared coefficients to zero, i.e. $\sum_{j=1}^{p} u_{jk}^2 = 1$. Eigenvectors that are independent and scaled to unity are termed orthonormal. Additional scaling options for the eigenvectors are available to make the variances of the eigenvectors similar (Jackson 1991), e.g. $v_{jk} = \sqrt{l_k} u_{jk}$ so the eigenvectors are in a **V** matrix and $w_{jk} = u_{jk}/\sqrt{l_k}$ so the eigenvectors are in a **W** matrix.

L is a p by p matrix whose diagonal contains the eigenvalues $l_1, l_2, \ldots l_k \ldots l_p$ (estimates of $\lambda_1, \lambda_2, \ldots \lambda_k \ldots \lambda_p$, the latent or characteristic roots) of **C**. The eigenvalues measure the variance explained by each of the eigenvectors. The number of eigenvalues is the same as the number of rows and columns in the covariance matrix and therefore the same as the number of original variables (p).

The matrix **L** for our example data set with the eigenvalues on the diagonal is:

$$\begin{bmatrix} 50\,075.681 & 0 & 0 \\ 0 & 2592.350 & 0 \\ 0 & 0 & 1003.660 \end{bmatrix}$$

The trace of this matrix, the sum of its diagonal elements, is the sum of the variances of the original centered variables. The sum of the eigenvalues from an eigenanalysis of a sums-of-squares-and-cross-products matrix or a correlation matrix would equal the sum of the variances of the original variables or the centered and standardized variables respectively. The matrix **L** represents, therefore, a reorganization of the variances of the variables from the original data matrix. Each eigenvalue is associated with each eigenvector and it is clear that the eigenvectors are extracted in order of decreasing proportions of the total variance. We often convert these eigenvalues to percentages.

Eigenvector	1	2	3
Eigenvalue	50 075.681	2592.350	1003.660
Percentage of total variance	93.300	4.830	1.870

More formally, determination of the eigenvalues involves solving the characteristic equation:

$$|\mathbf{C} - l\mathbf{I}| = 0$$

where **I** is an identity matrix of equivalent dimensions to **C**. The resulting polynomial (pth degree) in l is used to obtain $l_1, l_2 \ldots l_p$.

Based on the three human activity variables (bicycle, horse, foot) for eleven underpasses in Alberta from Clevenger & Waltbo (2000), the matrix **U** is:

	1	2	3
Bicycle	0.945	0.160	0.284
Horse	0.164	−0.986	0.011
Foot	0.282	0.036	−0.959

Each column is an eigenvector (u_k where $k = 1$ to p), the values in the eigenvector representing the coefficients or weights for that linear combination of the original variables. For example, the linear combination comprising eigenvector 1 is:

(0.945)Bicycle + (0.164)Horse + (0.282)Foot

where the values of each variable are centered because we used the covariance matrix to extract the eigenvectors. These linear equations are often termed components or factors (Chapter 17) and represent new variables derived from the original variables. Note that each variable contributes differently to each component (different coefficients or weights) and that these coefficients will depend on the units of each variable and whether standardizations are used. These linear equations can be solved to produce a component score (z_{ik}) for each object or observation for each component. For example, the score for component 1 for underpass 1:

(0.945)(−118.727) + (0.164)(−37.273) + (0.282)(−55.364) = −133.946

Singular value decomposition (SVD)

The SVD of an n by p data matrix is based on the product of the characteristic vectors of a matrix of associations between variables, the characteristic vectors of a matrix of associations between objects and their characteristic roots (eigenvalues, which are the same for both association matrices). If **Y** is a matrix of centered data (as used for the covariance matrix above), then **Y′Y** is the covariance matrix between variables (matrix **C** above) and **YY′** is the covariance matrix between objects (note these would be SSCP matrices for raw data and correlation matrices for centered and standardized data). The characteristic roots (eigenvalues) of these two matrices are the same.

The SVD of **Y** is:

$$\mathbf{Y} = \mathbf{Z}\mathbf{L}^{1/2}\mathbf{U}'$$

where **L** contains the eigenvalues, **U** is a p by p containing the eigenvectors of **Y′Y** as defined above and **Z** is an n by p matrix of eigenvectors of **YY′** and are also the principal component scores for objects scaled by the square root of the eigenvalues. Note that we now have the square root of the eigenvalues because we are dealing with the original variables rather than covariances or correlations (Jackson 1991). If **Y** contains raw data, then **L** and **U** will be the equivalent to that from the spectral decomposition of the SSCP matrix. If **Y** contains centered data, then **L** and **U** will be the equivalent to that from the spectral decomposition of the covariance matrix. If **Y** contains centered and standardized data, then **L** and **U** will be the equivalent to that from the spectral decomposition of the correlation matrix. Note that we can determine the original variables (centered and standardized if appropriate) from the matrix of component scores and vice versa when all components are extracted.

The advantage of using SVD is that extraction of eigenvectors and their eigenvalues is a one step process and SVD can also be applied to association matrices that are not square, e.g. chi-square matrices from contingency tables as used in correspondence analysis (Chapter 17). The advantage of spectral decomposition is that the choice of matrix (e.g. covariance vs correlation) will automatically center or standardize the data. As most multivariate analyses require statistical software, we rarely have to make this choice in practice.

multivariate analyses – see Box 15.1. The eigenvectors are commonly scaled so the sum of squared coefficients equals one; other forms of scaling are possible. We estimate the coefficients with maximum likelihood and can determine approximate standard errors. These linear combinations can be solved to provide a score (z_{ik}) for each object for each new derived variable. Note that there is the same number of derived variables as there are original variables (p). The new derived variables, each with an eigenvector of coefficients and an eigenvalue, are extracted sequentially so that they are uncorrelated with each other.

15.3.4 Derivation of components

We can derive the new variables (components) with matrix algebra in two ways. We can use a spectral decomposition of a p by p square matrix of associations among variables (e.g. SSCP, C or R matrices) or we can use a singular value decomposition of the n by p original data matrix. The two approaches produce equivalent results if there is a match between the association matrix used and the standardization of variables in the data matrix. One of the biggest problems facing biologists trying to become familiar with multivariate statistical techniques is the bewildering range of terminology, with different textbooks using different terms for the same property and also different labels for the relevant matrices. We have tried to summarize these two approaches for extracting components from a multivariate data set in Box 15.1, following the terminology of Jackson (1991) where possible.

The usual derivation of components is from an association matrix of covariances or correlations between variables (Box 15.1). This is sometimes termed an *R*-mode analysis and we can calculate scores for the derived variables (components) for each object (Jackson 1991, Ludwig & Reynolds 1988). We could also derive components from matrices representing covariances or correlations between objects and the derived variables (components) are linear combinations of the objects. We can calculate component scores for each variable and this is termed a *Q*-mode analysis. These two sets of component scores are related via matrix algebra and we can obtain component scores for objects from the eigenvectors of the variables and vice versa (Jackson 1991). In practice, *Q*-mode analyses comparing objects are more commonly based on dissimilarity measures (Box 15.2; Figure 15.2; Section 15.4).

The calculation of eigenvectors and their eigenvalues for new derived variables (components) from a multivariate data set is fundamental to canonical correlation analysis, principal components analysis and correspondence analysis (Chapter 17). If our data set contains groups, we can extract the components in a way that maximizes the between-group differences and this is the basis of multivariate analysis of variance and discriminant function analysis (Chapter 16).

15.4 Multivariate distance and dissimilarity measures

The methods described in the previous section deal with multivariate data sets by rearranging

Box 15.2 Measures of dissimilarity between objects for continuous variables

Consider two objects ($i = 1$ and 2), e.g. two sampling units, and a number of variables ($j = 1$ to p) recorded from each object, e.g. abundances of p species from each sampling unit. The same variables are recorded from each object (even if some variables have zero values for an object). First, we need a few definitions:

- y_{1j} and y_{2j} are the values of variable j in object 1 and object 2,
- $\min(y_{1j}, y_{2j})$ is the lesser value of each variable when it is greater than zero in *both* objects,
- p is the number of variables, and
- q is the number of variables that are zero for objects 1 and 2.

For example, y_{1j} and y_{2j} might be the abundances of species j in sampling units 1 and 2, $\Sigma\min(y_{1j}, y_{2j})$ is the sum of the lesser abundance of species j when it is present in both sampling units, p is the number of species and q is the number of species that are missing (zero values) from both samples. The formulae presented below are from Faith *et al.* (1987), except we present a more common version of the Canberra measure (see Digby & Kempton 1987) and correct their typographical error for chi-square.

Dissimilarity	Equation
Minkowski	$\left(\sum_{j=1}^{p} \lvert y_{1j} - y_{2j} \rvert^{\lambda}\right)^{1/\lambda}$
Euclidean ($\lambda = 2$)	$\sqrt{\sum_{j=1}^{p} (y_{1j} - y_{2j})^2}$
City block (Manhattan: $\lambda = 1$)	$\sum_{j=1}^{p} \lvert (y_{1j} - y_{2j}) \rvert$
Canberra	$\frac{1}{p-q} \sum_{j=1}^{p} \frac{\lvert y_{1j} - y_{2j} \rvert}{(y_{1j} + y_{2j})}$
Bray–Curtis (Czekanowski)	$1 - \frac{2\sum_{j=1}^{p} \min(y_{1j}, y_{2j})}{\sum_{j=1}^{p} (y_{1j} + y_{2j})} = \frac{\sum_{j=1}^{p} \lvert y_{1j} - y_{2j} \rvert}{\sum_{j=1}^{p} (y_{1j} + y_{2j})}$
Kulczynski	$1 - \frac{\left(\frac{\sum_{j=1}^{p} \min(y_{1j}, y_{2j})}{\sum_{j=1}^{p} (y_{1j})} + \frac{\sum_{j=1}^{p} \min(y_{1j}, y_{2j})}{\sum_{j=1}^{p} (y_{2j})} \right)}{2}$
Chi-square	$\sqrt{\sum_{j=1}^{p} \frac{\left(y_{1j} \left\lvert \sum_{j=1}^{p} y_{1j} - y_{2j} \right\rvert \sum_{j=1}^{p} y_{2j} \right)^2}{\sum_{i=1}^{n} y_i}}$

To illustrate these dissimilarity measures, we have calculated the dissimilarity between three species of Wisconsin forbs based on five leaf character variables from Reich *et al.* (1999). We have used the original variables and also variables centered and standardized to zero mean and unit variance.

Dissimilarity between	Euclidean	City block	Canberra	Bray–Curtis	Kulczynski
C. thalictroides vs *D. laciniata*:					
Raw data	238.355	82.500	0.231	0.217	0.212
Standardized data	2.992	1.143	NA	NA	NA
C. thalictroides vs *P. peltatum*:					
Raw data	121.005	34.300	0.111	0.104	0.100
Standardized data	1.337	0.498	NA	NA	NA
D. laciniata vs *P. peltatum*:					
Raw data	198.911	55.520	0.166	0.155	0.138
Standardized data	2.482	0.865	NA	NA	NA

Note that all measures show the same basic pattern, with the dissimilarity between *C. thalictroides* and *D. laciniata* the greatest and that between *C. thalictroides* and *P. peltatum* the least. Standardizing the variables to zero mean and unit variance doesn't change the relative dissimilarities although such a standardization cannot be applied to Canberra, Bray–Curtis and Kulczynski because they already include standardization as part of the calculation.

We also compared intact and impacted forest locations, based on the abundance of 15 species of bats, from Fenton *et al.* (1998). This data set allows us to include the chi-square measure, which requires integer values.

Dissimilarity between	Euclidean	City block	Canberra	Bray–Curtis	Kulczynski	Chi-square
Mana intact vs *Mana impacted*:						
Raw data	35.875	77.000	0.754	0.336	0.252	0.036
Standardized data	5.679	17.323	NA	NA	NA	NA
Range standardized data	2.720	8.255	0.835	0.770	0.435	0.428
Kanyati intact vs *Matusadona impacted*:						
Raw data	21.119	48.000	0.715	0.444	0.416	0.087
Standardized data	4.831	13.706	NA	NA	NA	NA
Range standardized data	2.390	6.663	0.792	0.719	0.703	0.491

Here the different dissimilarities produce different patterns. The intact vs impacted difference is greater for Mana than for Kanyati/Matusadona when measured with Euclidean, City block and Canberra but the reverse is true for Bray–Curtis, Kulczynski and chi-square. None of the standardizations changed the relative sizes for any of the measures except for Bray–Curtis.

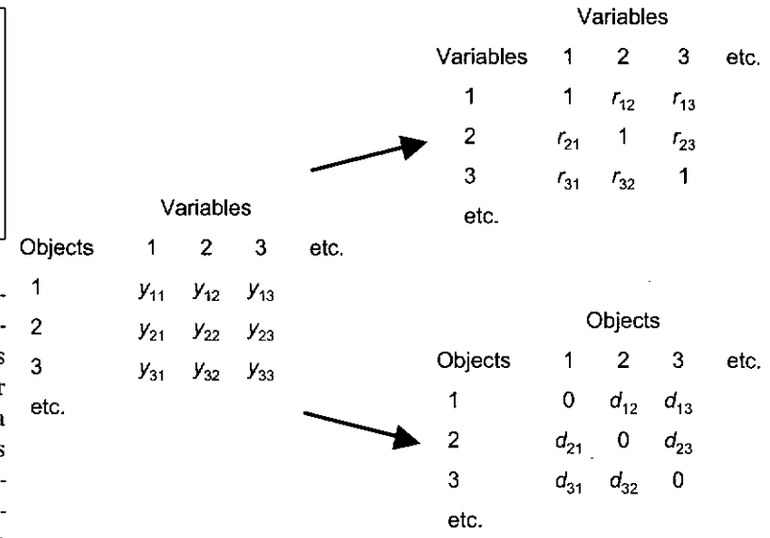

Figure 15.2 Distinction in initial steps between *R*- and *Q*-mode analyses. A data matrix of *n* rows by *p* columns is converted to a *p* by *p* matrix of associations between variables (e.g. correlations) or a *n* by *n* matrix of dissimilarities between objects.

the variance based on the association (covariances or correlations) between the variables (*R*-mode analyses). Another approach to multivariate data analyses (*Q*-mode analyses) is based on a measure of similarity or dissimilarity, sometimes termed a resemblance measure (Ludwig & Reynolds 1988), between objects.

Similarity indices measure how alike objects are, e.g. how similar sampling units are in terms of species composition or how alike specimens are in morphology. Dissimilarity indices measure how different objects are and should represent multivariate distance – if each variable is represented by an axis (or dimension) then multivariate distance is how far apart the objects are in multidimensional space. These dissimilarity indices are also called distances and are calculated for every possible pair of objects. There are numerous dissimilarity indices and the preferred ones are those that most closely represent biologically meaningful differences between objects. Particular difficulties arise when variables are measured on very different scales or when some of the variables include zero values, e.g. the variables are abundances of species of organisms and many objects have zero abundance for one or more species.

We usually represent the dissimilarities between objects as a dissimilarity matrix, converting an n rows by p columns data matrix to an n rows by n columns dissimilarity matrix. Like the covariance and correlation matrices described in Section 15.2, dissimilarity matrices are identical above and below the diagonal, which will be zeros indicating zero dissimilarity between an object and itself.

15.4.1 Dissimilarity measures for continuous variables

There is a broad range of measures of dissimilarity between objects based on continuous variables (see Digby & Kempton 1987, Faith *et al.* 1987, Legendre & Legendre 1998, Ludwig & Reynolds 1988). Their proliferation is partly due to the requirement by ecologists for measures of dissimilarity between sampling units in species composition that best represent underlying environmental gradients. We illustrate some of the commonly used measures in Box 15.2 and describe them briefly below. Legendre & Legendre (1998) provide a very thorough coverage.

Euclidean
This is based on simple geometry as a measure of the distance between two objects in multidimensional space. It is the square root of the sum, over all the variables, of the square of the difference between the values of each variable for the two objects. It is only bounded by zero for two objects with exactly the same values for all variables and has no upper limit, even when two objects have no variables in common with positive values.

City block or Manhattan
This is the sum (across variables) of the absolute differences in the value of each variable between two objects. It has properties similar to Euclidean distance and will be dominated by variables with large values.

Minkowski
Euclidean and City block are both versions of the more general Minkowski metric. Some software will, by default, "normalize" both measures by dividing by the sample size, i.e. the number of variables that contribute to the distance measure. This is only relevant if you wish to compare dissimilarities between data sets with different numbers of variables.

Canberra
This is the City block measure above, except that the difference between objects for each variable is divided by the sum of the variable values in the two objects before summing across variables. To ensure it has an upper limit of one, we standardize it by the number of variables that are greater than zero in both objects, e.g. the number of species present in at least one of the objects. This standardization is not always provided in texts (e.g. see Digby & Kempton 1988). The Canberra measure is less influenced by variables with very large values (Krebs 1989) than the City block measure.

Bray–Curtis
Developed by botanists in Wisconsin, this is also a modification of the Manhattan measure where the sum of differences between objects across variables is standardized by the sum of the variable values across objects, also summed across variables. Equivalently, it can be calculated as one minus twice the sum of the lesser value of each variable when it is greater than zero in both objects, standardized by the sum of the values of all variables in both objects. It ranges between zero (same variables and values in both objects – completely similar) and one (no variables in common with positive values – completely dissimilar) and is sometimes called percent dissimilarity (when expressed as a percentage; Ludwig & Reynolds 1988) or Czekanowski's coefficient. It is well suited to species abundance data because it ignores variables that have zeros for both objects (joint absences). Its value is determined mainly by variables with high values (e.g. species with high abundances; see Krebs 1989) because these variables are likely to be more different between the objects.

Kulczynski
This complicated measure, also termed the quantitative symmetric measure, was introduced to biologists by Faith *et al.* (1987). Like Bray–Curtis, it ranges between zero and one and has similar properties.

Chi-square
This dissimilarity measure, implicit in some multivariate analyses (e.g. correspondence analysis – Chapter 17), is only applicable when the variables are counts, such as species abundances. It is based on differences between objects in the proportional representation of each species, also adjusted for species totals.

15.4.2 Dissimilarity measures for dichotomous (binary) variables

Another group of dissimilarity coefficients has been developed for variables measured on a binary scale (e.g. presence and absence). Let a be the number of variables with non-zero values in both objects, b is the number of variables with non-zero values in object 1 and c is the number of variables with non-zero values in object 2. A simple measure of dissimilarity between two objects is Jaccard's coefficient:

$$1-\frac{a}{(a+b+c)} \qquad (15.3)$$

A slight modification is Sorensen's coefficient, which replaces a by $2a$. Sorensen's coefficient is identical to the Bray–Curtis measure for dichotomous variables.

15.4.3 General dissimilarity measures for mixed variables

Gower (1971) introduced a general dissimilarity measure that is useful for situations that include a mixture of continuous and categorical variables:

$$\frac{\sum_{j=1}^{p} w_{12j} s_{12j}}{\sum_{j=1}^{p} w_{12j}} \qquad (15.4)$$

In Equation 15.4, s_{12j} is the similarity between objects 1 and 2 based on variable j and w_{12j} equals one if the two objects can be compared for variable j and zero if they can't. So Gower's coefficient is "an average over all possible similarities" (Cox & Cox 1994) for objects 1 and 2. Gower's coefficient handles a mixture of variable types by calculating similarity for each variable separately (using appropriate coefficients for binary and continuous variables), then averaging those similarities. With all continuous variables, Gower's coefficient becomes (Cox & Cox 1994, Faith *et al.* 1987):

$$\sum_{j=1}^{p} \frac{|y_{1j} - y_{2j}|}{(\max_j - \min_j)} \qquad (15.5)$$

15.4.4 Comparison of dissimilarity measures

One characteristic of dissimilarities is whether they meet the criterion of being metric. A dissimilarity coefficient is metric if the dissimilarity between objects 1 and 2 is less than the sum of the dissimilarities between objects 1 and 3 and 2 and 3. This means that it is possible to construct a triangle whose sides match the three dissimilarities between three objects. Dissimilarity measures that meet the condition of being metric are commonly termed dissimilarity metrics. Not all dissimilarity measures are metric, e.g. Minkowski and chi-square are, but Bray–Curtis is not. If the dissimilarity is to be used in linear models (see Chapter 18), then being metric is important but otherwise the choice of dissimilarity measure for the analyses we describe in Chapter 18 is not usually based on whether it is metric or not.

Which of the many dissimilarity measures to use depends on the purpose of the analysis, the nature of the data and is closely linked to standardizations discussed in Section 15.6. When variables are measured on similar scales and have no zero values, Euclidean, City block or Canberra are good measures of dissimilarity between objects. If the scales of measurement are not consistent for different variables (e.g. the leaf characteristics from Reich *et al.* 1999), then the data need to be standardized before calculating these dissimilarities. Where the variables are species abundances (i.e. counts), an ideal dissimilarity coefficient should reach a constant maximum value when two sampling units have no species in common (i.e. it doesn't classify sampling units as similar because they have no species in common). Bray–Curtis, Kulczynski and Canberra meet this criterion, whereas Euclidean and chi-square do not. For this and other reasons, Faith *et al.* (1987) recommended the Bray–Curtis or Kulczynski coefficients for comparing objects when the variables are abundances of different species, as simulations showed these measures best matched ecological gradients. The suitability of some multivariate analyses for certain types of data is closely linked to the chosen or implicit dissimilarity measure that is used; we will discuss this further in the next two chapters.

For binary data, Kent & Coker (1992) argued that Sorenson's coefficient is preferred because it weights species (variables) in common higher than species absences (see also Krebs 1989). Remember that Sorenson's coefficient is the same as the Bray–Curtis measure with binary variables.

The general Gower dissimilarity measure is particularly useful when the data are a mixture of binary and continuous variables or when there are missing observations (but see Section 15.9.2), although Faith *et al.* (1997) showed that the version for continuous variables did not represent underlying ecological distances very well.

15.5 Comparing distance and/or dissimilarity matrices

Biologists often wish to test whether two or more matrices, or at least their corresponding elements, are correlated with each other. Such questions are particularly relevant when we are dealing with distance and/or dissimilarity matrices. For example, Sokal & Rohlf (1995) compared the matrix of genetic distances between ten villages of the Yanomama Amerindians in South America to the matrix of geographic distances between the villages. Fortin & Gurevitch (1993) emphasized the importance of examining spatial structure in field experiments, where one matrix might be differences in response of experimental units and the other might be the actual physical distances between the units.

Mantel's test is used for testing null hypotheses about correlations between matrices. It uses a randomization procedure (Chapter 3) to test whether the relationship between two matrices is more different than we would expect by chance (Manly 1997, Sokal & Rohlf 1996). We simply calculate the correlation coefficient between the corresponding elements of the two matrices, using only the lower (or upper) half of each matrix because they are symmetrical. However, the dissimilarities or distances within each matrix are not independent of each other (the dissimilarity between object 1 and 2 uses some of the same information as the dissimilarity between object 1 and 3, etc.). This is why we use a randomization test (Chapter 3) for the H_0 that the correlation between the two matrices is no different than we would expect by chance. Other statistics equivalent to the correlation coefficient for testing the H_0 in Mantel's test include Z (the sum of the products of the corresponding elements in the two matrices) and the regression coefficient (slope) for elements in one matrix regressed against elements in the other matrix. If the distances in the two matrices are standardized to zero mean and unit variance (Chapter 4), the values of the correlation coefficient, the regression slope and Z/m, where m is the number of elements in each matrix, will be the same (Manly 1997).

McCue *et al.* (1996) described genetic structure of a rare annual plant (*Clarkia springvillensis*) in California. They identified eight subpopulations and calculated Cavalli–Svorza genetic distances between subpopulations from isozyme analysis of tissue samples. They had two distance matrices – one for genetic distances between subpopulations and one for geographic distance (in meters) between subpopulations. The correlation coefficient between the two matrices was 0.632 with a randomization P-value of 0.032 and we would conclude that there is a statistically significant positive relationship between genetic and geographic distance for populations of *C. springvillensis*. Note that, in this example, the subpopulations were either really close (< 500 m) or around 8000 m apart so our interpretation of the relationship between genetic and geographic distance is constrained by the absence of data for separations between 500 and 8000 m.

The correlations can be extended to more than two matrices, using an analogue of the coefficient of multiple correlation (r^2) and partial correlations, called partial Mantel's test (Manly 1997). For example, Sklenar & Jorgensen (1999) measured floristic similarity between six mountains in Ecuador using Sorenson's index for presence–absence data. They used Mantel's test to show that there was a significant correlation between floristic similarity and differences in sampling intensity and they used a partial Mantel's test to test for a correlation between floristics and distance, holding sampling intensity constant.

15.6 Data standardization

Transformations, which change the scale of measurement of the data, were discussed in Chapter 4 in relation to meeting the normality assumption of parametric analyses and the homogeneity of variance assumption of most of these analyses. Transformations are particularly important for multivariate procedures based on eigenanalysis (e.g. principal components analysis – see Chapter 17) because covariances and correlations measure linear relationships between variables. Transformations that improve linearity will increase the efficiency with which the eigenanalysis extracts the eigenvectors.

Transformations such as log or square root will normalize positively skewed data and also reduce the influence of variables with high values (e.g. very abundant species) in multivariate procedures based on dissimilarity indices (Digby & Kempton 1987). Clarke & Warwick (1994) argued that fourth-root transformations should always be used for species abundance data before calculating dissimilarities to reduce the influence of very abundant species. One difficulty with this approach is that the effect of the transformation will depend on the underlying distributions of the variables (e.g. species) and therefore the degree of reduction of influence of very abundant species will be inconsistent. Cao *et al.* (1999) also had concerns about log transformation of water quality variables, pointing out that this transformation "indiscriminately increases the importance of a low range across all variables".

Standardizations work slightly differently

from transformations by adjusting the data so that means and/or variances or totals for each variable are the same. The following are examples (see also Table 15.5).

- Centering the data subtracts the variable mean from each observation for each variable, resulting in all variables having a mean of zero. Spectral decomposition of a covariance matrix extracts components from centered data.
- Standardizing the data divides the centered observations by the standard deviation for each variable, resulting in all variables having a mean of zero and a standard deviation (and variance) of one. Spectral decomposition of a correlation matrix extracts components from standardized data.
- Data can also be standardized so that each observation is expressed relative to the maximum value of that variable across all objects. This standardization results in observations being expressed as a proportion of the largest value for a variable, and is basically standardization based on the range within a variable.
- Cao *et al.* (1999) proposed a novel standardization for water quality data, whereby each variable is standardized in relation to the water quality standard of that variable and its range. Although acknowledging problems with their new standardization, they argued that it does allow natural variability in each variable to contribute to the results of a multivariate analysis.

These standardizations of variables are important if variables are measured in very different units or scales, because otherwise those variables with larger values or larger variances will often be more influential on the results of an analysis than variables with smaller values or smaller variances. Standardization of variables is essential if the variables are measured in very different units. For species abundances, such standardizations make all species have similar "importance" and thus "avoids a strong weighting by a few highly abundant species" (Ludwig & Reynolds 1988, p. 215). Without this standardization, rare species are often making little contribution to dissimilarities – of course, this may be the most biologically sensible interpretation.

In the same way that variables could be standardized, objects (e.g. sampling units) can also be standardized so the value for any variable for each object is expressed relative to the maximum value for that object in the whole data matrix. For species abundance data, this standardization is very important if the size of the sampling unit, and hence the total number of individuals, varies because it removes any effect of different total abundances in different sampling units, i.e. all sampling units are considered to have the same total abundance across all species.

Finally, converting abundance data to presence and absence might be considered an extreme combination of transformation and standardization. There are specific dissimilarity measures for such binary data (see Section 15.4.2).

It is often useful to analyze the same data with different standardizations, particularly in ecological research. For example, comparing the results of an analysis using raw data with one using sample-standardized data will indicate what influence different total abundances in samples have. Raw data versus species-standardized data will illustrate what influence the most abundant species have (simply leaving out different combinations of rarer species will provide similar information). Finally, to remove all effects of abundance, we can analyze just presence–absence data.

Table 15.5 Comparison of unstandardized, centered (zero mean) and standardized (zero mean and unit variance) observations for leaf N concentration for the eleven species of Wisconsin forbs from the study by Reich *et al.* (1999)

	Unstandardized	Centered	Standardized
Caulophyllum thalictroides	58.20	22.16	1.21
Dentaria laciniate	53.00	16.96	0.93
Erythronium americanum	42.00	5.96	0.33
Silphium terebinthinaceum	14.40	−21.64	−1.18
Podophyllum peltatum	44.70	8.66	0.47
Baptisia leucophaea	35.90	−0.14	−0.01
Trillium grandiflora	51.60	15.56	0.85
Echinacea purpurea	15.00	−21.04	−1.15
Silphium integrifolium	16.60	−19.44	−1.06
Sanguinaria canadensis	53.60	17.56	0.96
Sarrachenia purpurea	11.40	−24.64	−1.35
Mean	36.04	0.00	0.00
Standard deviation	18.26	18.26	1.00

15.7 Standardization, association and dissimilarity

Measures of association between variables described in Section 15.2 have implicit standardizations (see also Chapter 5). Covariances measure the linear relationships between centered variables whereas correlations measure the linear relationships between standardized (zero mean and unit variance) variables. The choice of association matrix on which to base subsequent multivariate analyses (Chapter 17) depends on whether differences in variances between variables represent important biological information that you don't wish to lose. Standardizations are also important for dissimilarity measures. Some dissimilarity measures are implicitly standardized and are unaffected by data standardizations (Faith *et al.* 1987). Some become identical after data standardization, e.g. Bray–Curtis, Kulczynski and City block are identical for count data if objects are standardized to the same total abundance. Others, e.g. Bray–Curtis and Kulczynski, produce nonsensical values when standardization is to zero mean (centering) or zero mean and unit variance (because of negative values). Standardizing by the range is a better option for these measures if you wish to reduce the influence of very abundant variables (e.g. species).

15.8 Multivariate graphics

Many of the exploratory data analysis techniques described in Chapter 4 are very applicable to multivariate data sets. In particular, describing distributions and checking for outliers for each variable separately with boxplots and examining bivariate relationships between variables with scatterplot matrices (SPLOMS) are always useful.

We may also wish to represent each observation or object in symbolic form, so that each symbol describes the relative value of all of the variables. A number of approaches have been developed to represent the different variables in a single "icon". The best known method is using Chernoff faces, where different features of the face represent different variables (Chernoff 1973; see also Everitt & Dunn 1991, Flury & Riedwyl 1988). These plots have been criticized, primarily because of the difficulty of rationally assigning variables to face features (Cox 1978), but they also have their supporters (Everitt & Dunn 1991, Flury & Riedwyl 1988). We illustrate these face plots with the Wisconsin forb data from Reich *et al.* (1999) in Figure 15.3, for both raw and standardized data. The differences between species are more noticeable for standardized variables, especially nose features representing mass-based and area-based photosynthetic capacity. Nonetheless, practice on known data sets is required to become familiar with recognizing similar and dissimilar faces.

An alternative, less "cartoonish", icon plot is to represent each object with a star, where each variable is represented by a point on the star, and the value of the variable is indicated by how far the point is from the center. There are no limits to the number of points, and therefore variables, for each star although the stars become difficult to interpret when there are too many variables. The

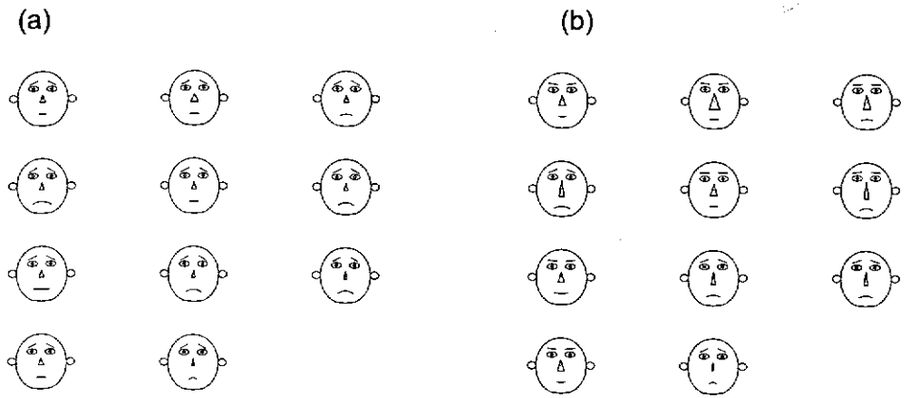

Figure 15.3 Chernoff face representation of the eleven species of Wisconsin forbs for five leaf characteristics based on raw data (a) and standardized data (b) from Reich *et al.* (1999). The features of the Chernoff faces are curvature of mouth for specific leaf area, angle of brow for leaf nitrogen concentration, width of nose for mass-based net photosynthetic capacity, length of nose for area-based net photosynthetic capacity, and length of mouth for leaf diffusive conductance at photosynthetic capacity. The species are, from left to right and row by row: *Caulophyllum thalictroides, Dentaria laciniate, Erythronium americanum, Silphium terebinthinaceum, Podophyllum peltatum, Baptisia leucophaea, Trillium grandiflora, Echinacea purpurea, Silphium integrifolium, Sanguinaria canadensis,* and *Sarrachenia purpurea.*

difference between raw and standardized variables is often very obvious on star plots. In Figure 15.4, we again illustrate the Wisconsin forb data from Reich *et al.* (1999). It is clear that *S. purpurea* is very different from the remaining species and *S. terebinthinaceum, P. peltatum, B. leucophaea* and *T. grandiflora* have larger values for leaf diffusive conductance at photosynthetic capacity, indicated by the extension of their stars to the left.

Finally, a very common method of graphing relationships between objects is to use a scatterplot where the axes represent the new derived variables from an eigenanalysis. These plots are common in the analyses described in Chapters 16 and 17, especially discriminant function analysis, principal components analysis and correspondence analysis. Alternatively, we can graphically represent a dissimilarity matrix between objects in a scatterplot, the basis of multidimensional scaling described in Chapter 18. Both types of plots are used especially by ecologists to represent the relationships between sampling or experimental units based on species composition, where they are termed "ordination" plots, the term ordination being derived from attempts to order units along some environmental gradient (Digby & Kempton 1987). Ordination is not a term familiar to most statisticians, or even non-ecological biologists, so we will call such plots of objects "scaling plots".

15.9 Screening multivariate data sets

In Chapter 4, we emphasized the importance of exploratory data analyses before proceeding with univariate statistical procedures, especially those with distributional assumptions. We also pointed out that unusual values (outliers) can have very influential effects on the conclusions from a statistical analysis, both in terms of estimation and hypothesis testing, and checking for outliers is an important precursor to any formal analysis. The need for exploratory screening of data is even more important for multivariate data sets because their complexity means that visual inspection of the raw data is likely to miss unusual patterns or observations. Additionally, the issue of missing observations is much more critical for the analyses we will describe in the next three chapters.

All of the univariate procedures we described in Chapter 4, especially graphical explorations (see previous section), can and should be used for

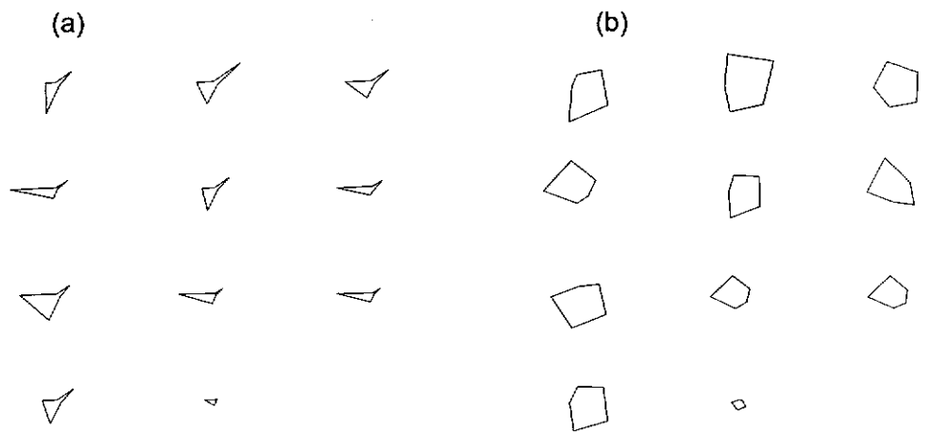

Figure 15.4 Star plot representation of the eleven species of Wisconsin forbs for five leaf characteristics based on raw data (a) and standardized data (b) from Reich *et al.* (1999). The features of the stars are, clockwise from the top, specific leaf area, leaf nitrogen concentration, mass-based net photosynthetic capacity, area-based net photosynthetic capacity, leaf diffusive conductance at photosynthetic capacity. The species are, from left to right and row by row: *Caulophyllum thalictroides, Dentaria laciniate, Erythronium americanum, Silphium terebinthinaceum, Podophyllum peltatum, Baptisia leucophaea, Trillium grandiflora, Echinacea purpurea, Silphium integrifolium, Sanguinaria canadensis,* and *Sarrachenia purpurea.*

multivariate data sets. In this section, we will focus on two particular issues: detecting multivariate outliers and dealing with missing observations.

15.9.1 Multivariate outliers

We discussed in Chapter 4 how unusually extreme values can influence the outcome of a statistical analysis. Multivariate outliers are more difficult to detect because they may not be univariate outliers for any of the individual variables (Jobson 1992). Additionally, outliers are often defined as large departures from a fitted statistical, usually linear, model to our data. For example, an observation may be an outlier from a fitted regression model (Chapters 5 and 6) and may have undue influence on the estimates of model parameters and tests of hypotheses about these parameters. In contrast, many of the multivariate techniques we will introduce in the next three chapters are more descriptive in nature, although new summary variables are often derived and can be used as response or predictor variables in subsequent linear models.

A multivariate outlier is an object with an unusual pattern of values for the variables (Tabachnick & Fidell 1996) and can be detected by measuring its distance, in multivariate space, from the centroid (Figure 15.1). The square of this distance (d_i^2 for object i) is called Mahalanobis distance (see Flury & Riedwyl 1988, Jackson 1991, Jobson 1992 for computational details) and is provided by most software in one or more of the multivariate analysis routines. If multivariate normality holds, the d_i^2 follow a χ^2 distribution with p (the number of variables) df (Manly 1994) so we can test for outliers, possibly using a strict significance level like 0.001 (Tabachnick & Fidell 1996).

Dealing with univariate outliers has been described in Chapter 4. The options for multivariate outliers are similar. If we decide that an object has such an unusual pattern of values for one or more variables that it is unlikely to be part of the population of objects we wish to describe or make inferences about, then we might delete that object from the analysis. Transformations of the variable(s) can also reduce the influence of outliers if they are extreme values in a positively skewed distribution.

15.9.2 Missing observations

Occasionally, we will have missing observations in our data set, i.e. no value was recorded for one or

more variables for one or more objects. The approaches for dealing with missing observations depend on the missing data mechanism, as introduced in Chapter 4 (see also Heitjan 1997, Little & Rubin 1987, Roth 1994). If the probability that an observation is missing is independent of the observed and missing values, the missing observations are termed missing completely at random (MCAR). This implies that the missing observations are a random subset of the data. The probability that an observation is missing might not depend on the unobserved missing value but be dependent on the values of the other variables for that object. For example, the pattern of missing data may depend on the group in which the object occurs, where another variable classifies objects into groups. This is termed missing at random (MAR). Finally, the missing values might be non-ignorable because whether an observation is missing depends on its value.

Consider the data set from Lovett *et al.* (2000) and imagine that one stream was missing a value for concentration of H^+. If the value is missing because of a random malfunction of a meter or a mistake by a researcher who forgot to write the value down then this observation might be MCAR. Our experience is MCAR is a common missing data mechanism in ecological sampling programs. If the value is missing because the stream was at a high altitude and weather conditions precluded access, then the observation might be MAR because the value of another variable (elevation), but not the unobserved H^+ value, determines the probability of it being missing. Finally, if the value is missing because the original H^+ reading was so high (e.g. Winnisook Creek) that the researcher assumed that the reading was a mistake and ignored it, the missing value is clearly non-ignorable. This situation is more common in situations when the observations depend on responses from subjects, such as in marketing surveys or clinical trials, although studies on animal behavior may suffer from this type of non-response. MCAR and MAR are much easier to deal with.

Basically, there are three approaches to dealing with missing observations (Little & Rubin 1987, Roth 1994). Our objective in this section is simply to make biologists aware that there are alternatives to simply "omitting whole rows of data", although some of the methods are sophisticated and usually require advice from statisticians experienced with their use. It is important to remember that avoiding missing data is the best solution because all of the alternatives are imperfect. We illustrate the results from some of the methods for dealing with missing observations in using a subset of the data from Reich *et al.* (1999). Our emphasis is not on the calculations, as these require appropriate software, but on the interpretation of the different methods.

Deletion

The simplest approach is to delete the entire object that has the missing value. This may be an appropriate strategy when the proportion of objects with missing values is low and the pattern is MCAR. It does result in loss of information because the non-missing values of variables for the object with the missing value are also excluded from the analysis. This is sometimes termed listwise deletion and is often the default for multivariate analyses in statistical software. If the analysis is based on pairwise associations between variables (e.g. correlations), an alternative is to use pairwise deletion. Here an object is only excluded for the calculation of the association between the two variables for which one value is missing but not excluded for the calculation of associations between other variables. This is the preferred deletion strategy when pairwise associations are the basis for the analysis.

Imputation

Imputation involves replacing (substituting) the missing values with some estimate of what the values might have been. There have been three common methods for imputing missing observations. The first is to replace the observation with the mean value of the variable calculated from the non-missing observations. Unfortunately, this tends to result in an underestimate of the true variance for that variable because these means do not contribute to the sum of squared deviations (Roth 1994). The second is to use a regression model to predict the imputed observation from other variables in the data. For example, we could determine which variable has the highest correlation with the variable with missing values from the complete objects and develop a regression model where the variable with missing values is the response variable and the other variable is the predictor. For the object with the missing value, the observed value of the predictor could then be used to predict the missing value from this regression model. Alternatively, we could use two or more predictors in a multiple regression model. Generalized linear models could be used if the assumption of normal error terms for the regressions was untenable or even generalized additive models if the shape of the relationship between the variables is not linear, although we have not seen either of these used in practice. Finally, hot-deck imputation simply replaces the missing value with the actual value from an object with similar characteristics (Roth 1994).

There are two main difficulties with these imputation methods. The first is that the imputed values are not independent of the observed data for a given variable and the precision (variances and standard errors) of the estimates of parameters based on these imputed values is generally underestimated. The second problem is that imputing a single value provides no indication of the effect that different imputed values have on the estimation of the relevant parameter (e.g. correlation), i.e. no measure of imputation uncertainty (Little 1999). Rubin (1987) developed a method termed multiple imputation as a solution to the second problem (see also Schafer 1999). Multiple imputation basically imputes a range of values for each missing observation, these values being simulated from a specific distribution for the missing values. The complete data sets (observed and imputed values) are then analyzed in the usual manner. The estimate of any parameter is simply the mean of estimates from the analyses of the imputed data sets. The standard error of this average estimate includes both the variance between imputations and the variance within each data set. Multiple imputation is clearly a sensible approach and a considerable improvement over single imputation, giving us some indication of how different imputed values affect the outcome of our analysis. The really tricky bit is developing the distribution of values from which the multiple imputations are derived. Rubin (1987) recommended a Bayesian strategy whereby the posterior distribution of missing values is conditional on the prior distribution of observed values, although the computations are complex (Schafer 1999). Multiple imputation routines are not readily available in commonly used statistical software but specialist products do exist and macros for some programs are available (see Rubin 1996 and references therein).

Maximum likelihood and EM

A different approach is to use maximum likelihood (ML) techniques to estimate the parameters of interest (e.g. means, correlation coefficients) from the observed, incomplete data (Little & Rubin 1987). Basically we use the distribution of the observed data and the conditional distribution of the pattern of missing data given the observed data. The likelihood function for any parameter can be complex with missing data so Little & Rubin (1987) also proposed methods based on factoring the likelihoods. The likelihood for a given parameter is decomposed into the sum of the likelihoods of distinct parameters given complete subsets of the data. These ML methods can estimate the missing observations once the parameters are estimated but do not use imputed values to estimate the parameters.

A combination of imputation and ML estimation is the Expectation–Maximization (EM) algorithm. This is an iterative procedure whereby the missing values are imputed, the parameters are estimated by ML, the missing values are re-estimated and imputed, the parameters re-estimated by ML, etc., until convergence of the likelihood of the parameter given the observed data is achieved. Technically, the missing values are not directly imputed using the EM method, but some function of the missing data like a predictive distribution is incorporated into the likelihood function (Little & Rubin 1987, Schafer 1999). The EM algorithm is now available in some commonly used statistical software. Multiple imputation may be more robust than EM methods for small data sets (Schafer 1999). Both straight ML and the EM method require the missing data to be at least MAR. See also Box 15.3.

Box 15.3 | Dealing with missing data

The data set on physiological variables for a range of plant species from different locations and functional groups from Reich *et al.* (1999) will be used to illustrate some of the methods for handling missing observations. We will use a subset of their data, trees from Venezuela, where there were 22 species (objects). There were five variables: specific leaf area (SLA), leaf nitrogen concentration (Leaf N), mass-based net photosynthetic capacity (A_{mass}), area-based net photosynthetic capacity (A_{area}) and leaf diffusive conductance at photosynthetic capacity (G_s). Five of the possible 110 observations were missing: SLA and A_{area} for *Eperua purpurea* and A_{mass}, A_{area} and G_s for *Micropholis maguirei*. We will assume these values are at least MAR and use listwise and pairwise deletion, regression imputation (using all other variables with complete data as predictor variables) and the EM algorithm to estimate means, standard deviations and pairwise correlations between variables. The EM algorithm converged in four iterations with −2(log-likelihood) of 650.85.

Means (standard deviations)

	SLA ($cm^2\ g^{-1}$)	Leaf N ($mg\ g^{-1}$)	A_{mass} ($nmol\ g^{-1}\ s^{-1}$)	A_{area} ($\mu mol\ m^{-2}\ s^{-1}$)	G_s ($mmol\ m^{-2}\ s^{-1}$)
Listwise	89.85 (24.04)	14.29 (4.71)	78.96 (55.23)	8.28 (3.68)	622.60 (535.76)
All values	88.20 (24.62)	14.04 (4.68)	77.82 (54.09)	8.28 (3.68)	602.90 (529.94)
EM	88.15 (24.18)	14.04 (4.68)	74.49 (55.39)	8.01 (3.67)	580.68 (535.92)
Regression	89.85 (24.04)	14.29 (4.71)	78.96 (55.23)	8.28 (3.68)	622.60 (535.76)

Correlations based on deletions

	SLA		Leaf N		A_{mass}		A_{area}		G_s	
	List	Pair	List	Pair	List	Pair	List	Pair	List	Pair
SLA	1.000	1.000								
Leaf N	0.569	0.607	1.000	1.000						
A_{mass}	0.789	0.789	0.708	0.699	1.000	1.000				
A_{area}	0.550	0.550	0.684	0.684	0.931	0.931	1.000	1.000		
G_s	0.498	0.498	0.546	0.530	0.851	0.851	0.894	0.894	1.000	1.000

Note that only the correlation between SLA and Leaf N differs much between the two methods of deletion.

Correlations based on regression imputation and EM

	SLA		Leaf N		A_{mass}		A_{area}		G_s	
	Regress	EM	Regress	EM	Regress	EM	Regress	EM	Regress	EM
SLA	1.000	1.000								
Leaf N	0.601	0.602	1.000	1.000						
A_{mass}	0.789	0.795	0.714	0.719	1.000	1.000				
A_{area}	0.555	0.563	0.681	0.685	0.931	0.932	1.000	1.000		
G_s	0.503	0.511	0.541	0.546	0.853	0.854	0.893	0.895	1.000	1.000

There are differences between the estimated correlations based on the two methods but, for these data, the differences are small.

Observed data with regression and EM imputed values (in bold)

SLA	Leaf N	A_{mass}	A_{area}	G_s
144.60	24.70	252.20	17.70	2272.00
114.30	17.90	159.30	13.80	889.00
126.40	16.50	115.50	9.10	597.00
105.40	16.40	140.40	12.80	975.00
78.10	16.90	111.50	14.00	1707.00
129.90	15.10	99.00	7.80	300.00
103.10	18.40	65.00	6.40	479.00
90.30	15.90	91.80	10.30	1009.00
82.80	6.80	46.50	5.60	490.00
75.20	7.80	47.20	6.20	693.00
86.60	8.60	34.70	4.00	321.00
82.60	10.70	52.20	6.50	411.00
82.00	17.70	67.20	8.20	381.00
67.80	9.30	38.80	5.70	241.00
76.80	15.00	44.90	5.90	329.00
67.30	13.00	53.80	8.00	378.00
86.20 (Regress)	15.20	55.10	**6.40 (Regress)**	209.00
87.10 (EM)			**6.32 (EM)**	
95.10	12.50	35.10	3.70	173.00
72.10	21.40	47.70	6.70	235.00
58.40	10.80	43.30	7.40	298.00
55.30	8.00	**4.76 (Regress)**	**4.26 (Regress)**	**114.03 (Regress)**
		20.94 (EM)	**5.01 (EM)**	**247.29 (EM)**
58.10	10.30	33.00	5.70	274.00

Note that the regression and EM imputed values are similar for *Eperua purpurea* (row 17) but very different for A_{mass} and G_s for *Micropholis maguirei* (row 21). The latter differences probably reflect the fact that only two predictor variables are available for this species for predicting the missing observations using a regression and the observed values for both of those variables are at the low end of the range for those variables. The EM imputed values are probably more reliable for this species.

15.10 | General issues and hints for analysis

15.10.1 General issues

- Variation within, and linear relationships between, two or more variables can be summarized with a sums-of-squares-and-cross-products matrix (raw data), covariance matrix (centered data) or a correlation matrix (standardized data).
- Spectral decomposition of one of these matrices produces new derived variables (components), extracted so the first explains most of the original variation, the second most of what is left, etc., and so that the new variables are uncorrelated with each other. Equivalent results are obtained from a singular value decomposition of the original data matrix, appropriately standardized.
- These new variables are linear combinations of the original variables and the coefficients

(summarized as an eigenvector) indicate the contribution of each original variable to the new variable.
- Differences between pairs of objects are measured with dissimilarities that are based on the sum of the differences for each variable between objects, often standardized so they range between zero and one.
- For measurement variables, either Euclidean or one of its modifications (City block or Canberra) are reliable dissimilarity measures, usually based on standardized data. For species abundances (counts with possible zero values), Bray–Curtis or Kulczynski are recommended.
- Graphical representations of multivariate data are available. SPLOMs display pairwise bivariate relationships and icon plots (Chernoff faces or stars) visually represent objects in terms of the relative values for the variables.
- The default for handling missing data with most software is to omit whole objects. Other approaches are generally preferred unless the sample size is large and the observations are missing completely at random.

15.10.2 Hints for analysis

- Before extracting components or determination of dissimilarities between objects when variables are measured in different scales or units, some type of standardization (based on standard deviation or range) is recommended.
- For species abundance, i.e. count, variables, different standardizations can provide useful comparative information. Standardizing objects to equal totals corrects for different sized sampling units, standardizing species to equal totals means that the most abundant species do not dominate the dissimilarity measure.
- Some standardizations can result in Bray–Curtis and Kulczynski dissimilarities not being bounded by one; standardize by range rather than by standard deviations when using these measures.
- We prefer standardizations to transformations for reducing the influence of variables with large values, although transforming variables may be relevant to improve linearity or if univariate analyses on the same variables also require transformation.

Chapter 16

Multivariate analysis of variance and discriminant analysis

In this chapter, we will examine the relationship between two or more response variables and one or more categorical predictor variables. We are primarily interested in two research questions. First, are there differences between groups based on all the response variables taken together and, second, can we successfully classify observations, particularly new observations, into the correct group.

16.1 Multivariate analysis of variance (MANOVA)

There are many situations where we record more than one response variable from each sampling or experimental unit and where these units are allocated to or occur in treatment groups. Ecologists often record the abundances of many species from each sampling or experimental unit and physiologists commonly measure more than one variable (e.g. blood pressure, heart rate, etc.) on experimental animals. For example, Peckarsky *et al.* (1993) examined the sub-lethal responses of mayfly larvae in streams to three different predator treatments (no predator and normal food, no predator and reduced food, one predatory mayfly (*Megarcys*) and normal food). There were five response variables recorded for each mayfly: body mass, egg mass, percentage of eggs, total mass, and maturation time. Botanists and zoologists also often measure many morphological variables when describing organisms from different locations or to compare organisms that may or may not be taxonomically different.

If each response variable is of inherent biological interest, our research questions might be whether there are group or treatment effects on each variable separately. Then the appropriate strategy is to analyze each variable using a separate univariate ANOVA to test for differences between groups. Some statisticians have argued that there is an inherent disadvantage to this approach. Because the response variables are measured from the same experimental or sampling units and may be highly correlated, the multiple ANOVA tests are not independent of each other and this can make interpretation difficult. Also, the number of univariate tests can get large if we have many variables so the family-wise Type I error rate may be very high for the collection of tests (Harris 1993; see also Chapter 3). A common recommendation is to adjust the significance level of each ANOVA test by using a Bonferroni-type correction so the family-wise Type I error rate stays at or below 0.05 (or whatever *a priori* significance level you choose). Unfortunately, with many response variables, this can result in unacceptably low power for each univariate test.

With multiple response variables, we might be more interested in whether there are group differences on all the response variables considered simultaneously. This is the aim of multivariate analysis of variance (MANOVA), the analogue of univariate ANOVA when we have multiple response variables for each experimental or sampling unit. Basically our hypothesis is now about group effects on a combination of the response variables and instead of comparing group means on a single variable, we now compare group