11

How to separate the signal from the noise (Statistical modelling)

> *'Questions were thrown at him: How did he explain the ''life-causing property'' of the signal? How did it originate? Was it, according to him, a ''pure accident''? And, most of all – where did we get Frog Eggs from?' From* His Master's Voice *by Polish writer Stanislaw Lem (1921–2006)*

A primary task of statistics is to separate systematic from random patterns. Systematic patterns may offer scientific insights. Randomness is just a nuisance. Patterns unfold in space and time, so to detect them we need to collect repeated observations, possibly under different circumstances. Picking up the thread from Chapter 10, this chapter first asks whether hypothesis testing can be used to compare the means of multiple populations and, assuming differences between them are detected, whether they can be explained by particular attributes of the populations. The first question can be addressed by an extension of the t-test called **analysis of variance** (**ANOVA** – Section 11.1). The second question can be examined by empirically modelling the replicate observations as a function of the conditions in which they weremade. Thisgivesriseto **regression**, a statistical approach to estimating the parameters and associated uncertainty of these empirical models. The simplest such model, **linear regression**, estimates the slope and intercept of a linear relationship between the observations and a single explanatory variable (Section 11.2). Compared to ANOVA, the regression approach brings gains in predictive and inferential ability (Section 11.3), assuming that its assumptions of linearity, normality and independence are satisfied.

The rest of this chapter discusses what can be done if the **fit** of the linear model to the data is poor (Section 11.4). For example, it may be that more than one explanatory variable is required to account for the variation in the data (**multiple linear regression models** – Section 11.5). Deciding how many and which explanatory variables should be included in a multiple regression model is known as **model selection**, a problem that can be examined as a trade-off between a model's quality of fit and its predictive ability (Section 11.6).

In many cases, the variation of the observations around the mean cannot be assumed to be normal. Using distributions other than the normal gives rise to a broader class of regression models known as **generalised linear models** (**GLMs** – Section 11.7). The particularly important cases of **binomial** and **Poisson GLMs** are discussed in Sections 11.7 and 11.8. The method of **quasilikelihood** is briefly introduced as a solution to **overdispersed** and **heteroscedastic** data (Section 11.9). If the assumption of linearity is violated and the more flexible forms offered by the GLMs are still not sufficient to describe the patterns in the data, then a functional form can be designed from biological first principles and fitted to the data by **nonlinear regression** (Section 11.10). Alternatively, if such biological information is unavailable, the data may be allowed to suggest an appropriate functional form through the use of **smoothing functions** (Section 11.11).

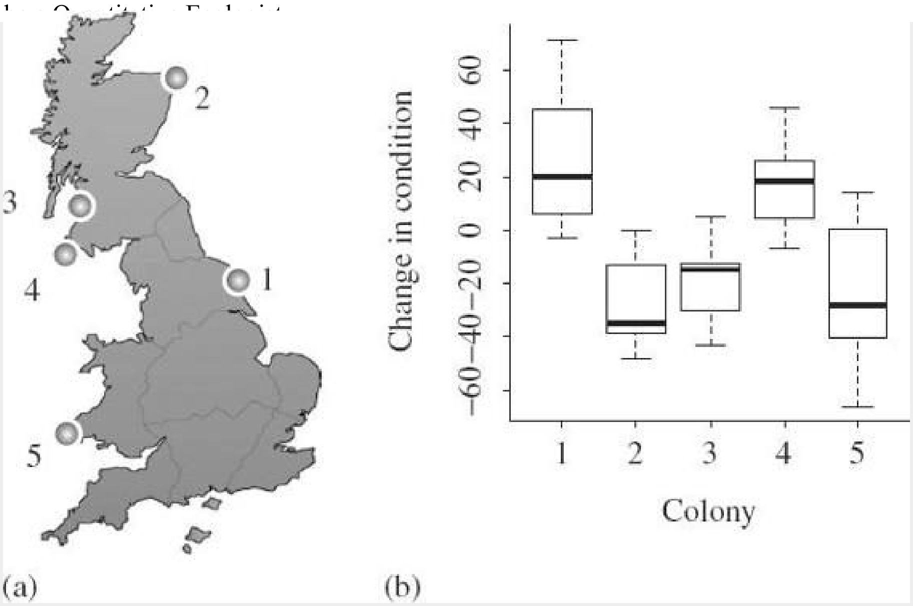
Variability around the mean prediction may be due to different identifiable sources. Real ecological data are hierarchical, often grouped according to nested or overlapping sample units (e.g. metapopulation, population, individual). The components of variability in the data can be accounted for by an extension of the linear model known as the **mixed effects model** (Section 11.12).

11.1. Comparing the means of several populations

Example 11.1: Samples of gannet condition

Change in the condition of breeding female gannets during a particular time period of the year is given by $\Delta I = I_{After} - I_{Before}$. To investigate the spatial patterns in this variable, we may track groups of females from several different colonies (Figure 11.1(a)) This will yield multiple samples of ΔI that we can use to obtain estimates of mean change in condition for each colony.

Figure 11.1: (a) Positions of five gannet colonies on mainland UK; (b) boxplots of the change in condition, incorporating ten females from each of the five colonies.



The first question is whether a spatial pattern exists at all, i.e. if mean change in condition differs between colonies. Comparisons between any two colonies can be carried out by using the t-test in Section 10.14. But how can we compare all of them at the same time? Could we, perhaps, carry out all pair-wise tests until we encounter evidence of dissimilarity? That is not a great idea, because the probability of finding a difference by chance increases with the number of tests performed (see critique of hypothesis testing in Section 10.16 and pp. 345–348 in Gotelli and Ellison, 2004).

The formal alternative to the t-test when more than two means are involved is called **analysis of variance** (**ANOVA** for short). The test examines the following two hypotheses:H_0: The means of all k populations are the same ($\mu_1 = \mu_2 = \cdots = \mu_k$) and H_1: At least two means are different. It works by comparing the variability between populations with the variability within populations. Its test statistic takes the form:

$$f = \frac{\text{Variability between populations}}{\text{Variability with in populations}}$$
(11.1)

So, larger values of f make it more likely that H_0 will be rejected. The numerator in Equation (11.1) is the following expression

$$\text{Variability between populations} = \frac{1}{k-1}\sum_{i=1}^{N} n_i(\hat{\mu}_i - \hat{\mu})^2$$
(11.2)

where k is the number of populations whose means are being compared, N is the sample size of the pooled data set, n_i is the sample size from the ith population, $\hat{\mu}$ is the average of the pooled data and $\hat{\mu}_i$ is the average of data from the ith population. This quantity, known as the **error mean square**, first carries out a comparison $(\hat{\mu}_i - \hat{\mu})^2$ between the pooled average $\hat{\mu}$) and each of the individual averages ($\hat{\mu}_i$). The multiplication with n_i inside the sum implies that individual means that come from larger samples carry a greater weight in the calculation. The error mean square has $k - 1$ degrees of freedom because, given all but one of the individual means together with the pooled mean, we can calculate the missing average.

The denominator in Equation (11.1) is given by

$$\text{Variability within populations} = \frac{1}{N-k}\sum_{i=1}^{N} (n_i - 1)s_i^2$$
(11.3)

where s^2_i is the estimated variance for the ith population. This quantity, known as the **group mean square**, adds together the sample sums of squares $(n_i - 1)s^2_i = \Sigma^{ni}_{j=i}(x_{i,j} - \hat{\mu})^2$ and then divides by the degrees of freedom. Here, we have $df = N-k$ because each sum of squares uses one degree of freedom in calculating the individual average $\hat{\mu}_i$ from the raw data.

If variabilities within and between populations are normally distributed, then the sampling distribution for this statistic is known as the **F-distribution**. Its parameters are the degrees of freedom $df_1 = k - 1$and $df_2 = N - k$.

11.1: Analysis of variance

The data presented in Figure 11.1(b) are a data frame (called `dat`) with 50 rows (ten individuals for each of the five colonies). Each row contains information about the change in individual condition (column `dat$Condition`) and colony membership (column `dat$Colony`, encoded as a factor – see R1.1 on factors). To test the H_0 that ΔI is not dissimilar between colonies, we need to define a **linear model object**.

```
mod<- lm(Condition~Colony,data=dat)
```

The reason and syntax for this will become clear in the following section. The model object, here called `mod`, is created by calling the command `lm()`. This call contains a **formula** (`Condition~Colony`) which tells R that we want to examine the effect of `Colony` on `Condition`. It also contains the name of the data frame, so that R knows where to find the data for `Condition` and `Colony`. Once this object has been created, an ANOVA can be performed very simply

```
> anova(mod)

Analysis of Variance Table



Response: Condition

          Df Sum Sq Mean Sq F value   Pr(>F)

Colony     4 24638.0 6159.5 15.893 3.508e-08 ***

Residuals 45 17439.9   387.6

---

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

Here, we have five colonies (so, $k = 5$) and a total of 50 animals ($N = 50$). The degrees of freedom listed in the column `Df` of the output are $df_1 = k - 1 = 4$ and $df_2 = N - k = 45$. The entries in column `Mean Sq` give the quantities in Equations (11.2) and (11.3) respectively. The ratio of these quantities (Equation (11.1)) is listed under F value. The p value for the null hypothesis is listed under `Pr(>F)`. In this example it is highly significant, implying that H_0should be rejected. This test therefore suggests emphatically that there are differences in ΔI between gannets in at least two of the five colonies.

The ANOVA test makes several assumptions whose violation could annul the test result.

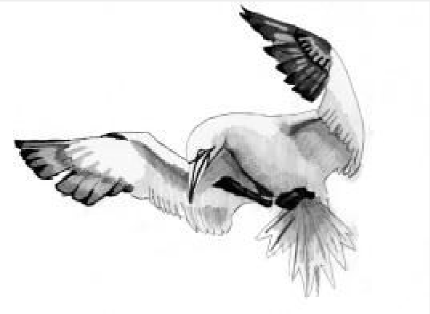
❶ *Independence:* In the case of the gannet example, this means that the measured change in condition of one female from a given colony is not affected by the change in condition of any other female (from the same, or another, colony).

❷ *Normal residuals:* For example plotting $(\Delta I - \overline{\Delta I})$ for females from a given colony should give a distribution close to a normal.

❸ *Equal standard deviations between sampling instances* (**homoscedasticity assumption**): For example, the distributions of residuals $(\Delta I - \overline{\Delta I})$ for different gannet colonies should have the same standard deviations.

If there are significant differences between populations, then perhaps a statistical summary can be estimated for each population individually, using the estimation methods of Chapter 10. You may find Example 11.2 below pedantic because it uses a lot of maths to arrive at the common-sense result that the best summary for each population is its average. However, it is useful in demonstrating a straightforward case of simultaneous estimation of many parameters and in proving the equivalence between maximum likelihood and least squares under the assumptions of ANOVA. Feel free to skip this, but come back if you have difficulties understanding how LSE and MLE are used to estimate the parameters of linear regression models in Section 11.2.

Example 11.2: Estimates of multiple population averages



Given a data set from 50 female gannets (ten from each of five colonies) we would like to find the best estimate of ΔI for each population. Intuitively, this should be the average value of ΔI for each population. Let's see if this is confirmed by least squares estimation (LSE – see Section 10.8). If the required LSE estimate for the jth population is written θ_j, then the sum of squared residuals for this data set is

$$SSR = \sum_{j=1}^{5}\sum_{i=1}^{10}(\theta_j - \Delta I_{ij})^2 \quad (11.4)$$

Here, j counts the gannet colonies and i counts the individuals in the sample from each colony. Hence, δI_{ij} is the observed change in condition in the ith animal of the jth colony. The LSE values of the parameters θ_i are those that collectively minimise the SSR. We are seeking a total of five parameters (θ_1, ..., θ_5), so we need to minimise the quantity in Equation (11.4) with respect to all five of them. Have a quick look back at Section 4.9 and Section 4.10: To minimise a function with respect to several variables, we need to minimise it with respect to every single one of them. Hence, here we need to calculate and solve the following system of equations:

$$\frac{\partial SSR}{\partial \theta_1} = 0, \ldots, \frac{\partial SSR}{\partial \theta_j} = 0, \ldots, \frac{\partial SSR}{\partial \theta_5} = 0 \quad (11.5)$$

Let's calculate the first of these (the others are done in exactly the same way):

$$\frac{\partial SSR}{\partial \theta_1} = \frac{\partial}{\partial \theta_1}\left[\sum_{j=1}^{5}\sum_{i=1}^{10}(\theta_j - \Delta I_{ij})^2\right] \quad (11.6)$$

The quantity inside the square brackets only depends on θ_1 when $j = 1$. Therefore, the partial derivatives will be zero for all $j \neq 1$. This simplifies the expression considerably.

$$\begin{aligned}\frac{\partial SSR}{\partial \theta_1} &= \frac{\partial}{\partial \theta_1}\left[\sum_{i=1}^{10}(\theta_1 - \Delta I_{i1})^2\right] = \sum_{i=1}^{10}\frac{\partial}{\partial \theta_1}(\theta_1 - \Delta I_{i1})^2 \\ &= \sum_{i=1}^{10} 2(\theta_1 - \Delta I_{i1}) = 20(\theta_1 - \overline{\Delta I_1})\end{aligned} \quad (11.7)$$

This will only be zero if $\theta_1 = \overline{\Delta I_1}$, confirming that the best estimate for each population is its average.

The same result can be reached via MLE (Section 10.9). Since we are assuming a normal distribution of independent residuals around the estimates, the likelihood can be written as a product of normal densities:

$$L(\theta_1, \ldots, \theta_5) = \prod_{j=1}^{5}\prod_{i=1}^{10}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}\frac{(\theta_j - \Delta I_{ij})^2}{\sigma^2}} \quad (11.8)$$

Note here that the variances are the same across populations (homoscedasticity assumption). The corresponding log-likelihood is

$$l(\theta_1, \ldots, \theta_5) = \sum_{j=1}^{5}\sum_{i=1}^{10}\left[\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2}\frac{(\theta_j - \Delta I_{ij})^2}{\sigma^2}\right] \quad (11.9)$$

The MLE values of the θ_i are obtainable by maximising this quantity. A few observations can quickly simplify this expression: of the two parts of the expression in the square brackets, only the second depends on the θ_i. Also, this second part is negative. Therefore, maximising Equation (11.9) is equivalent to minimising

$$l(\theta_1, \ldots, \theta_5) = \frac{1}{2\sigma^2}\sum_{j=1}^{5}\sum_{i=1}^{10}(\theta_j - \Delta I_{ij})^2 \quad (11.10)$$

The constant $1/2\sigma^2$ that has appeared in front of this expression simply scales the likelihood, so it can be removed. Having made all these simplifications, the resulting expression is exactly the same as the LSE criterion in Equation (11.4). Therefore, the assumptions of normality and homoscedasticity have ensured that the parameter estimates obtained from LSE coincide with MLE.

Having detected differences between populations, and having obtained estimates for the expected within-population values, the next question is what causes these differences. In the gannet example, what is it about some colonies that results in a deterioration of condition (negative $\overline{\Delta I_1}$) when in others, the condition of gannets is improving? The ecologist's response to this question is to list possible explanations (e.g. variable availability of food, variable predation risk, agonistic effects due to crowding, etc.). The statistician's response is to take these ecological insights and incorporate them in the modelling. If these **explanatory variables** (or **covariates**) are qualitative (e.g. factors such as the type of regional conservation plan for the species, or whether the colony is facing the Atlantic, North Sea or Irish Sea) then the approach continues to be called ANOVA. However, if the explanatory variables are quantitative, the approach is called **regression**.

11.2. Simple linear regression

Example 11.3: Density dependence of gannet condition

It is postulated that differences (ΔI) in gannet condition between colonies may be explained by density dependence. Therefore, data are collected on the density (P) of birds at the five colonies. The simplest possible relationship between these two quantities is linear, with slope a_1 and intercept a_0

$$\Delta I = a_1 P + a_0 \tag{11.11}$$

The values of the parameters $a_1 a_0$ are biologically interesting. For example, we might expect that when population density is zero, the change in condition should be positive $\Delta I = a_0 > 0$. A negative value for a_1 implies that higher densities reduce the ability of birds to improve their condition.

More generally, **simple linear regression** involves a collection of statistical methods for evaluating a linear relationship between a response (or dependent) variable Y and explanatory (or independent) variable X,

$$Y = a_1 X + a_0 \tag{11.12}$$

In the real world, such relationships are stochastic: the same value of X can give different values of Y. Hence, a **stochastic component** is often added to the end of Equation (11.12),

$$Y = a_1 X + a_0 + \varepsilon \tag{11.13}$$

In simple regression, like in ANOVA, this stochastic component is assumed to be normally distributed with a zero mean and a fixed variance for all values of $X : \varepsilon \sim N(0, \sigma^2)$. An alternative way to think of (Equation 11.13) is as a model for the mean of the data. The ith observation in the data set can be thought of as coming from the following distribution

$$Y_i \sim N(a_1 X_i + a_0, \sigma^2) \tag{11.14}$$

There are two fundamental benefits in using this approach over ANOVA: ANOVA requires us to estimate as many expectations as the number of subpopulations in the data (five parameters in the gannet example). In contrast, linear regression can model the expectations of all populations using only two parameters (the slope and intercept). Furthermore, the estimates from ANOVA only refer to the specific populations in the data. Linear regression can be used to predict the expected value of the response variable for unobserved values of the explanatory variable (see Section 11.3). These increases in estimation efficiency and predictive ability are achieved at a price: the assumption that the two variables are *linearly* related. In general, the process of estimating the best parameters for a particular curve on the basis of data is called **model-fitting**. The rationale behind the estimation of the two parameters in linear regression can again be illustrated with the specific example of gannet condition.

Example 11.4: Modelling density dependence of gannet condition

The objective is to obtain values for the parameters a_0, a_1 in Equation (11.11). This can be done by LSE or MLE. For LSE, we need to minimise the same criterion as Equation (11.6), but here the expected population-specific change in condition is replaced by the linear density-dependent model

$$SSR = \sum_{j=1}^{5} \sum_{i=1}^{10} (a_1 P_j + a_0 - \Delta I_{ij})^2 \quad (11.15)$$

Here, the change in condition is subscripted by both individual and population, and the density is subscripted by population. Equation (11.15) can be simplified by pooling all the individuals together so that change in condition (ΔI_i) for the *i*th individual is contrasted with the density (P_i) experienced by that animal

$$SSR = \sum_{i=1}^{50} (a_1 P_i + a_0 - \Delta I_i)^2 \quad (11.16)$$

The required parameter values a_0, a_1 are those that minimise this expression. MLE gives a similar result. We seek to maximise the following likelihood

$$L(a_0, a_1) = \prod_{i=1}^{50} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(a_1 P_i + a_0 - \Delta I_i)^2}{\sigma^2}} \quad (11.17)$$

Using arguments similar to those in Example 11.2, this corresponds to minimising the log-likelihood

$$l(a_0, a_1) = \frac{1}{2\sigma^2} \sum_{i=1}^{50} (a_1 P_i + a_0 - \Delta I_i)^2 \quad (11.18)$$

This expression is minimised at the same values of a_1, a_0 as Equation (11.16).

In general, therefore, estimating the parameters of the linear regression model in Equation (11.13) requires us to minimise the following quantity

$$f(a_0, a_1) = \sum_{i=1}^{n} (a_1 x_i + a_0 - y_i)^2 \quad (11.19)$$

where n is sample size and y_i, x_I are the response and explanatory data for the Δ*i*th sample unit. The partial derivatives of this expression with respect to the two parameters are

$$\frac{\partial f(a_0, a_1)}{\partial a_0} = 2\sum_{i=1}^{n} (a_1 x_i + a_0 - y_i) \qquad \frac{\partial f(a_0, a_1)}{\partial a_1} = 2\sum_{i=1}^{n} (a_1 x_i + a_0 - y_i) x_i \quad (11.20)$$

The MLE and LSE estimates of the two parameters ($\hat{a}_1$, $\hat{a}_0$) satisfy the coupled system

$$\sum_{i=1}^{n} (\hat{a}_1 x_i + \hat{a}_0 - y_i) = 0 \qquad \sum_{i=1}^{n} (\hat{a}_1 x_i + \hat{a}_0 - y_i) x_i = 0 \quad (11.21)$$

Carrying out the algebra (try it) gives the following estimates for slope and intercept:

$$\hat{a}_1 = \frac{\sum_1^n x_i y_i - n\overline{xy}}{\sum_1^n x_i^2 - n\overline{x}^2}$$

(11.22) $$\hat{a}_0 = \overline{y} - \hat{a}_1 \overline{x}$$

where $\overline{x}, \overline{y}$ are the sample averages for the explanatory and response data. Further calculations (not presented here) can be carried out to obtain standard errors and confidence intervals for these parameters. These calculations rely heavily on the underlying normality assumption about the distribution of points around the line. Standard errors can be used to construct t-tests for the slope and the intercept. A t-test for the slope is particularly informative because it tests the H_0 that $a_1 = 0$. Rejection of the H_0 implies a nonzero slope, and therefore a linear relationship between the explanatory and response variables. Thankfully, all of the above tasks can be carried out by R.

11.2: Fitting a linear model to data

ANOVA is really a special version of linear regression and that is why the command `lm()` used for fitting a linear model was also used in R11.1. The syntax is the same as before: a formula is declared within `lm()` which specifies the names of the response and explanatory variables. These names are taken from the column headings of the data frame. For example, a model of gannet condition change as a function of colony density can be created with the following line:

```
mod<- lm(Condition~Density,dat)
```

The resulting model object is stored in `mod` (or pick any other name that suits you) and the properties of this object can be explored as follows:

```
> summary(mod)



Call:

lm(formula = Condition   ~ Density, data = dat)



Residuals:

    Min      1Q   Median      3Q      Max

-40.612 -14.240   -3.571   15.169   55.241



Coefficients:

             Estimate Std. Error t value Pr(>|t| )

(Intercept) 36.48166       6.17323     5.910 3.44e-07 ***

Density      -0.72160       0.09389   -7.686 6.59e-10 ***

---

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



Residual standard error: 19.82 on 48 degrees of freedom

Multiple R-squared: 0.5517,       Adjusted R-squared: 0.5424

F-statistic: 59.07 on 1 and 48 DF, p-value: 6.59e-10
```

This rich output needs to be interpreted, one piece at a time:

- `Call:` confirms the formula of the model.

- `Residuals:` gives information about the distribution of the difference between observed and fitted values. Sometimes the linear model will be above the observations (+ve residual) and sometimes it will be below (-ve residual). In this example, the greatest overestimate of ΔI is 55.241 units over the observation for that particular bird. The inter-quartile range of the residuals can be obtained as `3Q-1Q` (= 29.409).

- `Coefficients:` a table whose rows carry information about each of the model's parameters (the model intercept and the coefficient of gannet density). The column `Estimate` gives the values of the parameters as obtained from Equations (11.22). Here, the linear model suggests a negative relationship between density and change in condition. The next four columns in this table deal with hypothesis testing about the two parameters. The estimated standard errors for each of the two parameters are used to calculate t values and then compared with a t-distribution to find *p* values. Both parameters have highly significant *p* values (as indicated by the star significance codes). Biologically, the conclusion of a negative coefficient a_1 means that breeding gannets will be worse off at crowded colonies. Also, because we are confident that the intercept is positive, we can say that at low gannet densities, birds will gain condition. The conclusion that density is negatively related with condition can be interpreted in two ways: either high densities diminish condition (via crowding and competition) or declining colonies are those whose members are failing in their effort to accumulate condition (and therefore suffer lower fecundity and survival).

The remaining output of `summary()` details how well the model fits the data. We will return to this issue in Section 11.4.

11.3. Prediction

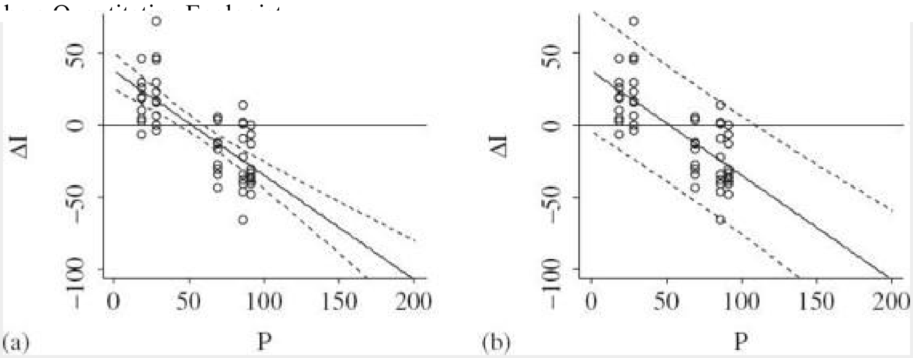
Plotting the linear model with the estimated parameters gives a **best-fit line** through a scatter plot of the data. This line is defined for all values of the explanatory variable and can therefore be used to predict the expected value of the response for values of the explanatory variable that were not originally in the data. If prediction is carried out for values within the observed range of the explanatory variable, then it is called **interpolation**. Predicting the response outside the range of observed explanatory data is called **extrapolation**. For a given value of the explanatory variable, these predictions will be subject to two types of uncertainty: first, uncertainty in mean response, as conveyed by the uncertainty in the estimated values of the regression parameters; second, variability in the actual values observed around this mean. We can visualise different types of uncertainty using **confidence and prediction intervals**. These are bands around the estimated mean, running the range of the explanatory variable.

Example 11.5: Predicting density-dependent changes in gannet condition

There are two ways to plot uncertainty in the predictions of the linear model for gannet condition changes. Confidence intervals (Figure 11.2(a)) represent uncertainty in the estimates of the model's parameters. If the regression line were to be perturbed according to this uncertainty (i.e. shifted up and down by nudging the intercept, and pivoted around its midpoint by altering the slope), the movements would chart a band that is narrower towards the middle of the observations and wider towards the extremes.

Prediction intervals (Figure 11.2(b)) represent parameter uncertainty combined with the additional stochasticity around the mean due to chance individual variation. Because variables other than density may act to determine the changes in condition observed for any one gannet, we would still get some variation around the trend even if this was known precisely. Since they account for both sources of uncertainty, prediction intervals are always wider than confidence intervals. This is also the reason why 95% confidence intervals do not generally encompass 95% of the data points (see Figure 11.2(a)).

Figure 11.2: (a) Confidence intervals for the estimated mean response; (b) prediction intervals for particular observations around the estimated mean.



Therefore, the two parts of Figure 11.2 are tackling two different questions: if you were to visit a colony of known density, and measured Δ*I* from a large sample of individuals, there is a probability of 0.95 that the average of this sample would be within the band shown in Figure 11.2(a). However, if you were to take a random animal from a colony having a given density, then the value of Δ*I* for that animal would fall within the band of Figure 11.2(b) with a probability of 0.95.

11.3: Prediction and confidence intervals

Generating predictions from a model object requires the values of the independent variables for which the predictions are sought. These values must be passed as new explanatory data to the command `predict()` as follows:

```
dens<-seq(0,200,1)# Generates a vector of densities from 0 to 200

# Create data frame with the new explanatory data

datNew<-data.frame("Density"=dens)

# Create predictions from model object mod at the positions

# of datNew and store them in the list preds

preds<-predict(mod, datNew)
```

To generate a simple plot of the data with the best fit line use the following:

```
plot(dat$Density, dat$Condition) # Scatter plot of data

lines(dens, preds) # Plots best fit line
```

To create a plot with confidence intervals (such as the one in Figure 11.2(a)), use the interval option inside `predict()`. This generates a matrix of three columns, one for the best fit line and two for its lower and upper confidence curves.

```
preds<-predict(mod, datNew, interval="confidence")

plot(dat$Density, dat$Condition)

lines(preds[,1], lty=1)

lines(preds[,2], lty=2)

lines(preds[,3], lty=2)
```

To generate a plot with prediction intervals (like the one in Figure 11.2(b)), specify the option `interval="prediction"`.

11.4. How good is the best-fit line?

Linear regression assumes that the relationship between the explanatory and the response variables is linear, that the residuals around this line are independent, normally and identically distributed. If there is a strong linear signal in the data and all other assumptions are satisfied, then the estimated parameters and confidence intervals can in good conscience be used for prediction. However, when the simple linear model fails for one or more reasons, then it may need to be extended. Therefore, this section can be treated as a springboard to the rest of the chapter because it reviews the different ways in which a linear model can fail and refers you to appropriate extensions introduced in later sections.

The first step in the diagnostic process is to decide if the model is any good at all. A measure of the goodness of fit is provided by the **coefficient of determination**, the squared correlation between the observed and estimated response values r^2 = *corr* $(y, \hat{y})^2$. This takes values between 0 and 1, with values closer to 1 indicating a good fit. Alternatively, r^2 can be interpreted as the **proportion of explained variance** (e.g. a linear model with $r^2 = 0.93$ accounts for 93% of the variance in the raw data).

11.4: Coefficient of determination

When fitting a linear model the summary of the output contains the value of the coefficient of determination under the entry `Multiple R-squared`. Have a look at the output in R11.2. The r^2 value is a relatively low 0.55, indicating that there may be more that can be done with this model.

If the r^2 value is low, then something is certainly amiss with the model, but, even if it is high, there could be problems lurking. Hence, the following diagnostics should be carried out regardless. Most of the diagnostics make use of residuals obtained by subtracting the estimated response value from the corresponding observation: $e_i = y_i - \hat{y}_i$. Any disagreements between the model and the data will be manifested in the residuals but it is hard to know what constitutes a large disagreement (and, hence, one worth worrying about). Therefore, raw residuals are often replaced by **standardised residuals**, which are designed to look like an independent sample from N (0, 1) if all the assumptions of linear regression are satisfied. This offers a good basis for diagnosing problems with each assumption:

- *Linearity:* In a plot of raw residuals against the estimated values (e.g. Figure 11.3(a)), deviations from linearity appear as consistent patterns of overestimation or underestimation. These patterns may even indicate how the linear model should be modified to better describe the data. If you think that you should be fitting a nonlinear model to your data, then Sections 11.7–11.11 are of particular relevance.

- *Homoscedasticity:* The assumption of equal variance can be examined by inspecting a plot of the residuals (raw or standardised) against fitted values (Figure 11.3(a)). If, for example, it appears that there are bottlenecks in the scatter of residuals around the mean, then this could indicate non-constant variance. A more formal diagnostic is provided by an **ncv test** (see R11.5) of the null hypothesis of homoscedasticity. If you conclude that your model suffers from this pathology, there are extensions that enable you to explicitly model changes in the variance jointly with the mean. You can find more in Section 11.9 and Faraway (2006, Section 7.3).

- *Normality:* This can be examined using the Q-Q plot of the residuals (Figure 11.3(b)). The Q-Q plot was introduced in R10.9. A Shapiro–Wilk test can be used to get a p value for the hypothesis of normality (again, see R10.9). If you know that your residuals are non-normal, then Sections 11.7–11.8 offer useful generalisations to the linear model.

- *Outliers:* Sometimes, data collection or data entry go wrong, generating outliers that have a disproportionate effect on the parameters of the resulting model. Cook's distance is a metric which measures the effect on the resulting fit of removing each point in the data. The rule of thumb is that Cook's distance values exceeding 1 should make you go back to the data and re-examine the circumstances in which that observation was collected. Plotting Cook's distances for all points (Figure 11.3(c)) is sufficient as a first measure. If outliers are present but no suspicion lies with the data collection process, then perhaps normality is also violated (see above) or the residuals are overdispersed (see Section 11.9).

- *Independence:* Dependence between observations can take many different forms. For example, spatial autocorrelation (see Example 2.4 and R2.1) may occur because neigh-bouring gannets in a given colony tend to perform similarly to each other. If the sample is obtained from a localised group of birds, then it may give a misleading impression about changes in condition across the colony. Similarly, **temporal** or **serial autocorrelation** occurs when successive observations are more similar (or dissimilar) to each other than the average similarity in the sample. For example, the position of a moving animal after 5 min will not be totally independent of its present position. If you have information on the spatial or temporal structure of the data, then plot the residuals in order of spatial or temporal proximity. Better still, plot a correlogram of the residuals as a function of time or distance. If there is evidence of serial correlation in your residuals, then Section 11.13 has some relevant material.

Example 11.6: Diagnostics for the density dependence model

Figure 11.3 shows three diagnostic plots derived from the model of R11.2.

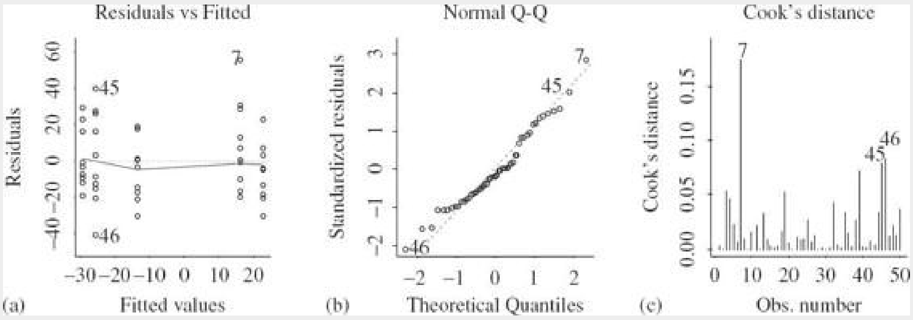
- *Linearity:* The residuals in Figure 11.3(a) do not indicate systematic biases (i.e. consistently low or high residuals in particular regions of the fitted values axis).

- *Homoscedasticity:* It is hard to tell with only five values of population density, but the scatter of residuals in Figure 11.3(a) does not appear to vary as we move from left to right.

- *Normality:* The Q-Q plot and superimposed Q-Q line in Figure 11.3(b) are typical of a sample of normally distributed residuals.

- *Outliers:* All data points have Cook's distances well below 1 (Figure 11.3(c)).

Figure 11.3: (a) Raw residuals $(y_I - \hat{y})_i$) against the fitted values $\hat{y}_i$); (b) Q-Q plot of standardised residuals against the corresponding quantiles from a normal distribution *N* (0, 1); (c) Cook's distances for all 50 observations.



As Example 11.6 illustrates, it is entirely possible for the diagnostics to reveal no problem with the assumptions and yet have poor model fit (as evidenced by r^2). This may mean that we have not offered enough information to describe the observed patterns. Specifically, it may indicate that one or more important explanatory variable has been missed out. This is the subject of the next section.

11.5: Linear model diagnostics

For a given model object `mod`, the standardised residuals are given by `rstandard(mod)`. The homoscedasticity test is carried out by the command `ncv.test(mod)`, found in the library car. The command `plot(mod, which=c(1,2,4))` will generate a sequence of plots like the ones shown in Figure 11.3.

11.5. Multiple linear regression

Three influences determine the value of a particular measurement (Section 8.1): conditions, covariates and stochasticity. Conditions are considered known and fixed. Once a covariate has been used to explain away some of the variability in the data, the remaining variability may be called stochasticity. Alternatively, it may be attributed to another covariate that operates together with the first.

Example 11.7: Combined effects of density and the environment

In addition to population density, there are a great number of other, colony-specific characteristics that may explain variations in the ability of gannets to improve their condition. Some of these may be quantitative (e.g. density of competitors in the region), others qualitative (e.g. immediate proximity to Atlantic Ocean, Irish Sea, North Sea). There is no *a priori* reason to assume that changes in condition are only affected by one covariate, so we need to model their combined effect.

The simplest form of multiple linear regression examines the effects of *n* different covariates (X_1, X_2, ..., X_n) inanadditive fashion

$$(11.23)\quad Y = a_0 + a_1 X_1 + a_2 X_2 + \cdots a_n X_n + \varepsilon$$

So, now the signal is represented by the sum of terms, each involving a covariate and its coefficient. Stochasticity is represented by the stochastic component ε, a random variable with distribution $N(0, \sigma^2)$. An alternative interpretation is that the response data come from the following distribution

$$(11.24)\quad Y \sim N(\mu, \sigma^2) \text{ where } \mu = a_0 + a_1 X_1 + a_2 X_2 + \cdots a_n X_n$$

All of the assumptions of simple linear regression are carried over and estimation follows the familiar path of ❶ writing the likelihood or least squares criterion, ❷ setting to zero all partial derivatives with respect to the parameters a_0, a_1, a_2, ..., a_n and ❸ solving the resulting system of linear equations.

Example 11.8: Combined effects of density and the environment

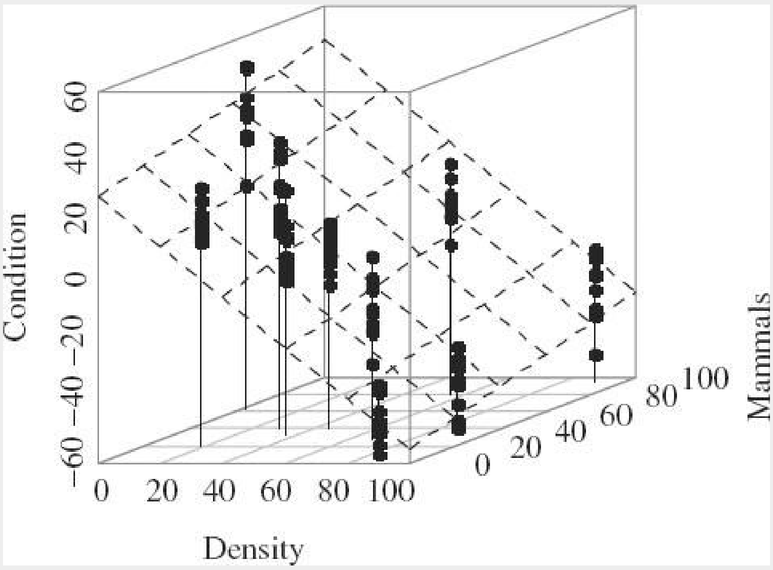
To get a better idea of what affects changes (ΔI) in the condition of gannets, data are needed from more colonies. A total of 100 birds are sampled (ten birds from each of ten colonies). For each colony, there are records of gannet density (P) and the density (M) of marine mammals (White-beaked dolphins *Lagenorhynchus albirostris* and Harbour porpoises *Phocoena phocoena*) in the feeding grounds associated with each colony. Both of these explanatory variables are included in the following multiple regression model

$$(11.25)\quad Y = a_0 + a_1 P + a_2 M + \varepsilon$$

Fitting the model to this particular data gave the parameters $a_0 = 25.96$, $a_1 =-0.81$, $a_2 = 0.23$.

Equation (11.25) describes a plane in three dimensions that can be drawn using the parameters estimated for the model (Figure 11.4) and visually compared with the data for the ten colonies. The inclination of this plane can be interpreted biologically: The declining condition with increasing conspecific density points to the detrimental effect of intra-specific competition. The positive slope with marine mammal density is more intriguing. It is explained by the observation that gannets have a commensalistic relationship with marine mammals. Porpoises and dolphins operate as 'beaters', bringing prey closer to the surface (Camphuysen and Webb, 1999). Hence, according to this data set, gannets are expected to gain condition most rapidly when they live in uncrowded colonies that are close to the foraging grounds of marine mammals.

Figure 11.4: The response and covariate data plotted as dots in 3D space. The tilted plane is the graph of the estimated model which describes how condition is expected to change in response to prevailing gannet and marine mammal densities.



11.6: Fitting multiple regression models

The material presented in R11.2 is enough to get started with fitting multiple regression models. For example, if you have a data frame (`dat`) which contains columns for bird condition (`Condition`), colony density (`Density`) and the density of marine mammals (`Mammals`), then the model can be estimated, and detailed output produced, with two lines of code:

```
> mod<- lm(Condition~Density+Mammals,dat)

> summary(mod)
```

```
Call:

lm(formula = Condition ~ Density + Mammals, data = dat)



Residuals:

       Min       1Q     Median       3Q      Max

- 25.75748  -5.39267   0.07778   5.36553 23.71727



Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept) 25.95908       2.36954     10.955 < 2e-16 ***

Density     -0.81328       0.03056   -26.617 < 2e-16 ***

Mammals      0.22949       0.03502     6.553 2.71e-09 ***

---

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



Residual standard error: 9.163 on 97 degrees of freedom

Multiple R-squared: 0.8846,     Adjusted R-squared: 0.8822

F-statistic: 371.6 on 2 and 97 DF, p-value: < 2.2e-16
```

Note how the addition of some more data and a second covariate has led to an improved r^2 value (0.88 compared to the 0.55 of the single covariate model in R11.2). The *p* values listed under the heading `Pr(>|t|)`, test the null hypothesis of no relationship between the response and each explanatory variable.

To obtain mean predictions from this model for new data, use the approach of R11.3. Plots of these predictions with respect to two covariates (e.g. Figure 11.4) can be generated using the library `scatterplot3d` as follows:

```
require(scatterplot3d)

attach(dat)

s3d <-scatterplot3d(Density,Mammals,Condition, type=“h”)

s3d$plane3d(mod)

detach(dat)
```

Finally, never forget to run diagnostics for your model (see Section 11.4). The code of R11.5 can be used intact, with models of more than one covariate.

11.6. Model selection

When trying to account for patterns in a set of observations, a good biologist will come up with a wealth of possible explanatory variables. A good modeller will try to prune these back to a minimal set.

Example 11.9: Should the beating of a butterfly's wings be used to explain gannet condition?

In Example 11.6, we found that conspecific and heterospecific density can account for 88% of the observed variability in gannet condition change. The remaining 12% could be due to any list of plausible characteristics, such as the various biotic and abiotic attributes of each colony and their surrounding habitats, or the individual characteristics (age, size, genetic makeup, etc.) of each gannet in the sample. This should help the model account for more of the variability in the data. Carrying on in this way, we may try some less plausible possibilities, such as the indirect ecosystem effects of species that are separated from gannets by two or more trophic levels. In the end, we might consider far-fetched ideas such as the effect of particular pollutants that are not toxic to seabirds. Eventually, we are guaranteed to come up with a model that accounts for 100% of the variability in the data. Surely, this is a good result?

The answer to this question is an emphatic 'no'. Philosophically, it is the wrong thing to do because it goes against the principle of parsimony (look back at Example 2.1). In the context of regression models, adding all conceivable covariates is a practical concern because most data sets carry only a limited amount of information about the study system. It is foolhardy to believe that we can perfectly explain changes in the condition of a sea bird species by observing 100 individuals from ten colonies.

Why, then, will the r^2 value continue increasing with the inclusion of covariates of ever-decreasing plausibility? The crucial fact here is that regression models are empirical: they describe patterns, not mechanisms (perhaps you have heard the saying 'correlation is not causation'?). With each additional covariate, the regression model simply becomes more flexible, and by playing one covariate against the other it can fit the data exactly, but this does not mean that all these variables are biologically important or even relevant. Such models are called **overparameterised** or **overfitted** and they have one lethal flaw: they fit the observations really well but predict appallingly in new situations.

We must, therefore, come to terms with the fact that some of the residual variability after fitting a model is due to ❶ variables that we haven't thought of, ❷ variables we don't have data for, ❸ measurement errors or ❹ violations of the model's assumptions not picked up by the diagnostics of Section 11.4. Deciding which variables to keep and which to drop is a major aspect of **model selection**.

The traditional approach to this question was to use the p values generated automatically by model-fitting software (e.g. see summary output in R11.6). If, in the summary of the model, all covariates had low p values (usually annotated with one to three stars), then the model would remain intact. If one or more covariates appeared with high p values, then one of them would be dropped and the model re-fitted. The process would continue until all covariates had stars (one, two or three depending on the significance level desired by the analyst).

This approach is a poor basis for model selection (Burnham and Anderson, 2002) because the levels of significance are arbitrary, and due to the problems associated with multiple nonindependent hypothesis tests (see Section 11.1). For these reasons, model selection by p values is liable to retain irrelevant covariates or drop important ones. So, the star system in computer output (e.g. R11.6) should be used as a rough guide during the exploratory stages of the analysis, not for model selection.

When fitting a model by LSE, a better approach looks at the balance between sample size (n) and the number of coefficients (K). It is generally poor practice to use many covariates relative to the number of observations in the data. This is easy to understand by thinking about simple linear regression: if the data have a single observation, then a linear model that contains both a slope and an intercept cannot be specified because infinite lines can be drawn through a single point. Thus, a good model is one that achieves a high r^2 using few covariates relative to the sample size. This gives rise to the **adjusted coefficient of determination**

$$\text{adjusted } r^2 = 1 - (1 - r^2)\frac{n-1}{n-K} \qquad (11.26)$$

Given two models with the same r^2, the adjusted r^2 takes higher values for the model with fewer parameters. This quantity is estimated automatically by software packages (see output in R11.2 and R11.6). To employ it for model selection you need to estimate all the models you would like to consider and select the one with the highest adjusted r^2.

This idea is carried over to a much broader class of model selection criteria by determining quality of fit on the basis of likelihood rather than residuals. The approach leads to various criteria that balance the likelihood against the number of parameters (K) in a particular model. The best known of these is the rather cryptically named **Akaike Information Criterion** or **AIC**, In which m = 2 and the term l (model | data) is the maximum log-likelihood of a model under the observations. As this quantity increases, the fitted model passes closer to the data. However, if this is achieved at the price of an increased numbers of parameters, the value of the criterion is penalised by the number of parameters (K) in the model. The negative sign in front of Equation (11.27) means that the best model in the set is the one with the smallest value of AIC. The name of the criterion is in honour of its proposer (Hirotugu Akaike) and the branch of mathematics (**information theory**) used to prove that, asymptotically (i.e. in the long run), this expression offers the best balance between model fit and predictive ability. Other information criteria, using slightly different derivations, result in different values for *m*.

$$AIC = -\{2l(\text{model} \mid \text{data}) - mK\} \quad (11.27)$$

Using information criteria is preferable to adjusted r^2 values because they are more general and reliable. This is particularly useful to remember when the two methods disagree.

Example 11.10: Model selection by adjusted r^2 and AIC

Four different models are fit to the data to examine the possible improvements in model quality brought by introducing weather variables (annual averages of temperature and rainfall) for each colony. The covariates in each model and their associated adjusted r^2 and AIC are listed in Table 11.1 (the best models according to each criterion are shown in boldface).

Table 11.1

Model	*adj r^2*	*AIC*
1 Gannet dens+Mar mam dens	0.882	**731.776**
2 Gannet dens+Mar mam dens+Rain	**0.883**	732.083
3 Gannet dens+Mar mam dens+temp	0.881	733.361
4 Gannet dens+Mar man dens+Rain+Temp	0.882	733.988

We are therefore faced with the (not atypical) situation of two model selection techniques that do not entirely agree with each other. As it happens, this example is based on made-up data in which values for temperature and rain are random numbers with no effect on condition, so AIC has got closer to the underlying truth.

Information criteria rely on asymptotic arguments, meaning that they are guaranteed to work in the majority of studies and as sample size increases. It is, however, important to check, for any particular study, whether the model selection process has yielded the best predictive model. This can be achieved by re-sampling techniques such as **cross-validation**. The idea of cross-validation is to keep aside a part of the data (e.g. data from one of the ten gannet colonies), fit the candidate models to the rest of the data and then use them to predict the missing part. Since there is no particular reason for selecting one part of the data over the others for this purpose, the whole process needs to be repeated for all parts (e.g. all ten colonies, one at a time). The model that best manages to predict the hidden data is the one that should be chosen for fitting to the entire data set. This approach certainly gets to the heart of the problem by trying to maximise the predictive power of the chosen model but it is computer-intensive and thus impractical for some studies. One possibility is to use an AIC-derived ranking to arrive at a confidence set of models (a small set of highly promising models), and then use cross-validation to select between them. A more robust alternative is to keep the best of all worlds by generating **model-averaged predictions** from this confidence set of models (see Burnham and Anderson, 2002).

One particularly thorny problem often resolved by model selection methods is **collinearity**. Imagine that you have two candidate covariates that, within the range of observed values, are closely and linearly related to each other.

Example 11.11: Collinearity in covariates of gannet condition

The ability of different gannets to acquire condition will almost certainly depend on their individual characteristics, such as age and body size. For any given bird, these are not independent of each other. Furthermore, different measures of body size such as length, girth and weight may be exactly or approximately proportional to each other. These characteristics contain similar information, so using them all together in a regression model uses up valuable parameters.

Models with collinear covariates tend to be less robust than models with linearly independent ones: consider a model $Y = a_0 + a_1X_1 + a_2X_2$ with two covariates (X_1, X_2) that are approximately proportional to each other (i.e.$X_2 \cong cX_1$). Instead of being arranged across an area in the X_1, X_2 plane, the explanatory data are arranged closely around the line $X_2 = cX_1$. We are therefore trying to fit a model with two covariates (a plane) to a set of points that are, in fact, one-dimensional. Metaphorically, it is easier to lay a flat sheet of metal on top of a bed of nails rather than trying to balance it on the teeth of a rake.

Model selection will often manage to yield a reduced set of covariates, keeping the ones with the most explanatory power. However, if you are worried that collinearity may still be a problem for your model, you may want to read more about detection with **variance inflation factors** (Fox, 2002) which can also deal with the more general problem of **multicollinearity** (i.e. one covariate being a linear function of two or more covariates).

11.7: Manual and automated model selection

Given a small set of candidate models, manual model selection is feasible. The four models in Example 11.10 can be estimated and compared (on the basis of AIC) as follows:

```
> mod1<-lm(Condition~Density+Mammals,dat)

> mod2<-lm(Condition~Density+Mammals+Rainfall,dat)

> mod3<-lm(Condition~Density+Mammals+Temperature,dat)

> mod4<-lm(Condition~Density+Mammals+Temperature+Rainfall,dat)

> AIC(mod1, mod2, mod3, mod4)

        df      AIC

mod1  4  731.7764

mod2  5  732.0837

mod3  5  733.3605

mod4  6  733.9877
```

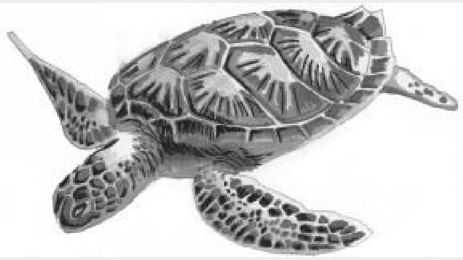
Automated model selection, an exploratory search through a large set of models, may be carried out using any of the following three methods: **forward selection** starts with an intercept-only model and examines if any of the covariates (added one at a time) improves the AIC. If so, then that model is used to construct all two-covariate models by adding the remaining covariates one at a time, and so on until none of the new models can improve on the AIC. **Backward elimination** begins with a model that contains all covariates and drops them, one at a time, using AIC to decide if these changes result in model improvement. **Stepwise selection** is a combination of the other two methods. It drops variables one at a time but, with every step, it tries to re-introduce some of the variables that were rejected in previous iterations of elimination, just in case this yields an even better (lower) AIC. To carry out stepwise selection in R, first estimate the full model and use it as the input to the command `step()`

```
mod<-lm(Condition~Density+Mammals+Temperature+Rainfall,dat)

step(mod)
```

The output of `step()` can be quite verbose because it prints a summary of every model that it has tried out. The very last model described is the one selected.

11.7. Generalised linear models

Example 11.12: Many response variables are constrained

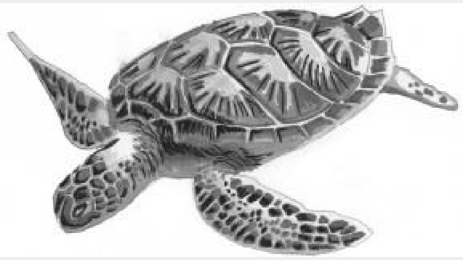


Loggerhead turtle, Caretta caretta

So far, all the examples in this chapter have modelled variations in the animals' ability to improve their condition. The ultimate goal of many population studies is to link short-term behavioural or physiological responses with long-term performance; i.e. fitness. The fitness of an individual can be broken down into the demographic components of survival and reproduction. Extended further, the concept of inclusive fitness also examines the ability of an individual's offspring to be recruited into the breeding population. However, unlike changes in condition, which can theoretically take any values from $(-\infty, \infty)$, the component variables of fitness are constrained: a turtle has a probability of survival in the range [0,1], its reproductive output cannot be less than zero and the number of its offspring recruited can only range between zero and the number of offspring it produced. The values taken by these variables within their allowed ranges will depend on various environmental influences, so it would be interesting to create regression models linking any one attribute of fitness to its environmental covariates. However, linear regression is patently inappropriate for the task because linear functions are unconstrained.

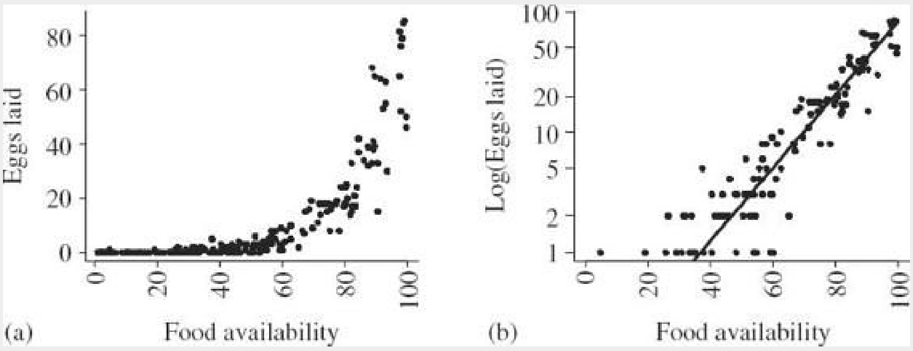
Transformations of the response data are a rather neat trick, historically used to shoe-horn nonlinear data into the constraints of the linear regression framework. If you look back at Figure 3.19 in Example 3.16, you will notice how data on population growth (a non-negative variable in that example, defined as the ratio of population size in two successive years) were converted into a linear arrangement by a log-transformation.

Example 11.13: Log-transforming fecundity data



How does a turtle's ability to reproduce at the end of a year depend on the availability of forage in the months leading up to the breeding season? If Y is a random variable describing per capita fecundity (number of eggs produced in breeding season) and $\mathbf{y} = \{y_1, ..., y_n\}$ is a data set of fecundity measurements from n turtles, then a plot of these data on a log-scale may reveal a linear pattern (Figure 11.5).

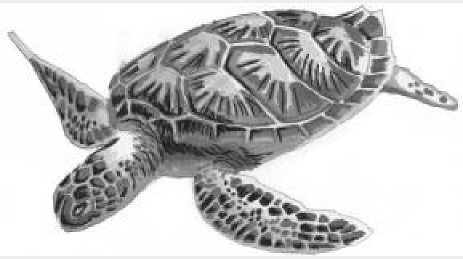
Figure 11.5: (a) Counts of eggs laid by 200 Loggerhead turtles plotted against an index of food availability and (b) the same data plotted on a log scale for the eggs axis, with an indication of the emergent linear pattern.



This example reveals two important problems with log-transforming the data. The first is that real data often contain zeros and log(0) is not defined mathematically. The simulated example in Figure 11.5 contained 76 zeroes, meaning that 38% of the data were thrown away simply by going from Figure 11.5(a) to Figure 11.5(b). The second problem is that fitting a line to the scatter plot in Figure 11.5(b) violates the homoscedasticity assumption, one of the cardinal requirements of linear regression.

One clue to the solution of the problem is to understand why zero counts occur in the first place. Is it impossible for an undernourished turtle to have any eggs at all, or is it just unlikely? Reductions in the availability of forage will diminish the expected number of eggs produced. Even though this expectation will not itself be zero, it will lead to several turtles not producing eggs. In other words, instead of transforming the data, we may transform their expected value and then model the data as arising from some random process around this expectation.

Example 11.14: Modelling count data



Y and $\overline{Y}$ are, respectively, the actual and expected number of eggs laid by a turtle living under certain conditions. In this example, where the modelled variable is a count, the Poisson distribution may be used to represent the stochasticity of Y around $\overline{Y}$. Furthermore, if X is a covariate of Y (e.g. forage availability), we may model the dependence of the mean number of eggs on X by a suitably transformed linear model. In short,

$$\begin{aligned} Y &\sim Poisson(\overline{Y}) \\ \overline{Y} &= \exp(a_0 + a_1 X) \end{aligned} \tag{11.28}$$

Here, the exponential transformation of the linear model ensures that the mean number of eggs does not become zero or negative. Equivalently, the transformation can be seen as a log-transform of the mean number of eggs ($\ln(\overline{Y}) = a_0 + a_1 X$).

This is an example of a **generalised linear model** (**GLM**). It is 'generalised' in the sense that it can deal with both constrained and unconstrained response variables. It is 'linear' because the expectation of the response variable is modelled as a transformation of a linear model. GLMs usually appear in the following form

$$\begin{aligned} Y &\sim \text{Distribution}(\overline{Y}, \theta) \\ \overline{Y} &= h(a_0 + a_1 X_1 + a_2 X_2 + \ldots) \end{aligned} \tag{11.29}$$

The GLM therefore has three components, the distribution (or **stochastic component**) which describes the stochasticity of the data, the **linear predictor** ($a_0 + a_1 X_1 + a_2 X_2 + \ldots$) which introduces the effect of covariates on the mean of the distribution and some **link function** (h) which transforms the mean of the distribution so that it can be modelled by a linear predictor. Apart from the mean ($\overline{Y}$), the distribution of the GLM may also depend on additional parameters (θ). In the case of a count variable such as the number of eggs, the Poisson is a suitable distribution and the appropriate link function is ln(). The resulting GLM is called **log-linear**.

It is helpful to acquaint yourself with the three components of the GLM (stochastic component, linear predictor, link) by recognising them in the familiar linear regression model.

Example 11.15: Seeing the linear modelasaspecialcase oftheGLM

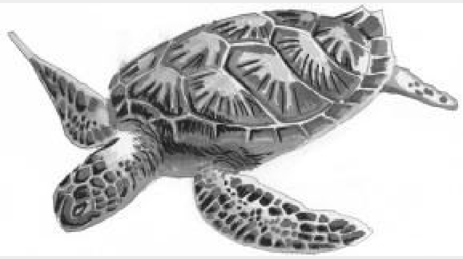
Under simple regression, we model the mean of the normal distribution directly by a linear model (see also Equation (11.14))

$$\begin{aligned} Y &\sim N(\overline{Y}, \sigma^2) \\ \overline{Y} &= a_0 + a_1 X_1 \end{aligned} \tag{11.30}$$

Here, the additional parameter is the variance σ^2, and the mean is untransformed. In GLM terminology, the link of this model is the **identity function**, i.e. the mean of the stochastic component is the same as the linear predictor.

As with all statistical modelling, the objective is to estimate the parameters of the linear predictor from a sequence of response observations. In this task, estimation by maximum likelihood really proves its usefulness.

Example 11.16: Likelihood for a log-linear GLM



For a particular expected value $\overline{Y}$, the probability of y eggs being laid is given by the PMF of the Poisson distribution (see Section 9.10)

$$f(y \mid \overline{Y}) = \frac{e^{-\overline{Y}} \overline{Y}^{y}}{y!} \tag{11.31}$$

If y $_i$ is the number of eggs laid by the *i*th turtle in the sample, and $\overline{Y}_i = \exp(a_0 + a_1 X_i)$ is the expected value for that particular turtle, considering its access to food (X_i), then the probability associated with the observation of that turtle is

$$f(y_i \mid a_0, a_1) = \frac{e^{-\overline{Y}_i}\overline{Y}_i^{y_i}}{y_i!}$$
$$\overline{Y}_i = \exp(a_0 + a_1 X_i) \tag{11.32}$$

Notice that now the probability is conditional on the parameters of the regression model. Assuming that the fecundity of different turtles in the sample is independent, the probability of the data under a set of regression coefficients is

$$f(y_1, \ldots, y_n \mid a_0, a_1) = \prod_{i=1}^{n} \frac{e^{-\overline{Y}_i}\overline{Y}_i^{y_i}}{y_i!}$$
$$\overline{Y}_i = \exp(a_0 + a_1 X_i) \tag{11.33}$$

We can now turn this around (see Section 10.9) to obtain the likelihood of any pair of parameter values being true, given the data

$$L(a_0, a_1 \mid y_1, \ldots, y_n) = \prod_{i=1}^{n} \frac{e^{-\overline{Y}_i}\overline{Y}_i^{y_i}}{y_i!}$$
$$\overline{Y}_i = \exp(a_0 + a_1 X_i) \tag{11.34}$$

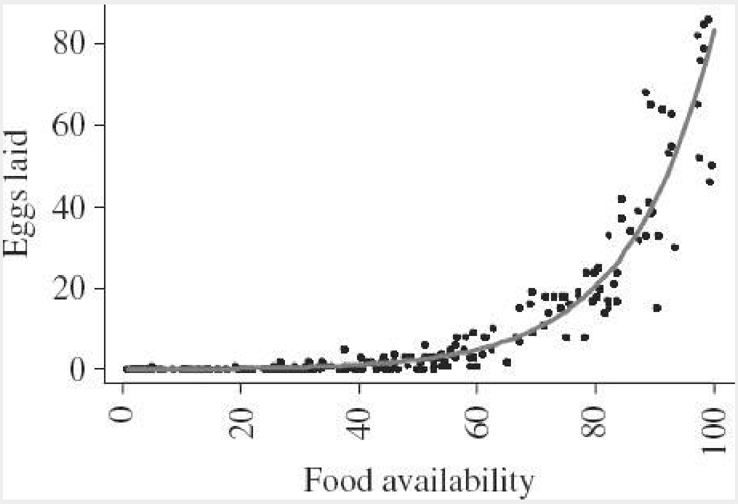
The task is now to estimate the parameters a_0, a_1 by maximising the following log-likelihood

$$l(a_0, a_1 \mid y_1, \ldots, y_n) = \sum_{i=1}^{n} \{-\overline{Y}_i + y_i \ln \overline{Y}_i - \ln(y_i!)\}$$
$$\overline{Y}_i = \exp(a_0 + a_1 X_i) \tag{11.35}$$

The awkward term $-\ln(y_i!)$ inside the sum does not depend on the parameters a_0, a_1 and can therefore be dropped from the log-likelihood because it will not affect the position of the maximum. Note that the same likelihood function can be written for more than one covariate. So, if we wanted to extend the investigation to environmental influences other than food (e.g. human disturbance near the beach, pollutants, etc.) we could just extend the linear predictor to include these additional covariates, just like we did for multiple linear regression (Section 11.5).

Maximising the likelihood (see R11.8) will yield MLE parameter values. We can place these back into the model for the mean number of eggs $\overline{Y} = \exp(\hat{a}_0 + \hat{a}_1 X)$ and plot it along with the data (Figure 11.6).

Figure 11.6: Raw data on turtle egg production and the associated best-fit log-linear model.



11.8: Fitting a and predicting from a log-linear GLM

The central command for fitting GLMs of all types is `glm()`. At a minimum, it requires the model formula, the name of the data frame from which the data are to be taken (specified via the option `data`) and the stochastic component of the GLM (specified via the option `family`). For example, given the two vectors `eggs` and `food`, containing the data in Figure 11.6, the GLM can be estimated as follows

```
dat<-data.frame(food, eggs)

mod<-glm(eggs~food, family=poisson, data=dat)
```

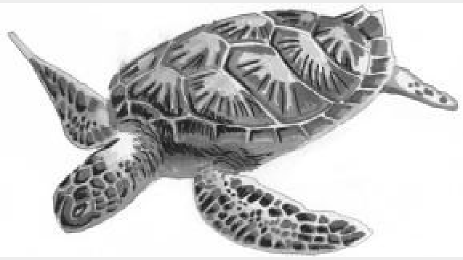
The name `mod` now holds the GLM object. You can get more information about the estimated coefficients by typing `summary(mod)`. To find out what are the fitted values for the response variable (i.e. the expected value of the response for the observed values of the explanatory variable), then type `fitted(mod)`. To predict for new values of the explanatory variable(s), use the command `predict(mod, newdata)`, where `newdata` is a data frame containing the explanatory values for which predictions are required. As an illustration, the curve in Figure 11.6 can be produced as follows

```
newdat<-data.frame(“food”=seq(1,100)) # Food availabilities

preds<-predict(mod, newdat, type = “response”)

plot(food, eggs, xlab=“Food availability”, ylab=“Eggs laid”)

lines(preds)
```

The option `type` inside the command `predict` tells R that predictions are required at the scale of the response variable, not the linear predictor. The alternative option, `type=“link”`, generates predictions on the scale of the linear predictor. This is useful if untransformed predictions are required (e.g. this is how I produced the trend line shown in Figure 11.5(b)).

Fecundity is an example of a demographic variable that is constrained below by 0. Modelling response variables that are constrained both above and below is also possible by choosing the appropriate stochastic component and link function. In the simplest case, the response (*Y*) may be a binary variable, taking only the values 1 and 0, representing success or failure (see Section 9.7).

Example 11.17: Modelling senescence in turtles

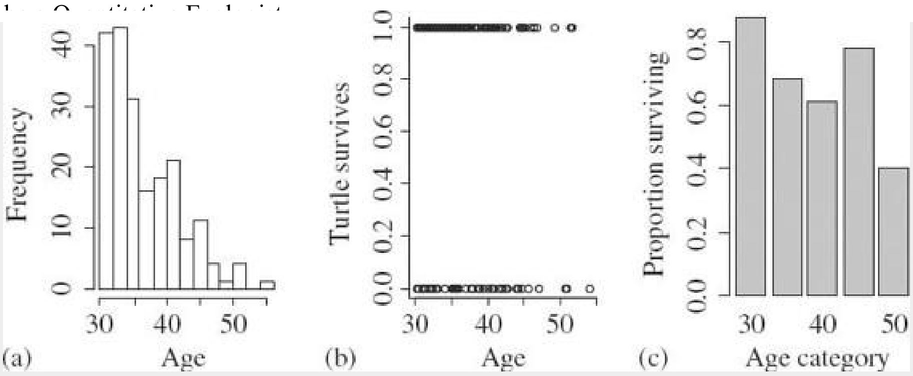


We might be interested in examining how the probability of survival of adult turtles changes with their age. The data comprise a random sample of 200 turtles whose age can be established at the beginning of the year and whose death can also be ascertained with little error.

Even if there is no senescence in the species, the frequency distribution of ages in the sample will not be uniform (Figure 11.7(a)): even if the annual probability of survival is the same for all ages (say, $p < 1$), the proportion of animals surviving to 10 years is larger than the proportion surviving to 50 ($p^{50} < p^{10}$).

The response data will be binary, taking the value 1 if a turtle survives to the end of the year and 0 if it doesn’t. The explanatory data are the ages of the turtles at the start of the year. It might be anticipated that, in the presence of senescence, we would observe more deaths at higher ages but if we plot the response against the explanatory data, we get an incredibly uninformative plot (Figure 11.7(b)). The fact that sample size declines with age means that the dots in Figure 11.7(b) become sparser as we move from left to right. This pattern applies to the survivals and deaths equally, making it hard to decide by eye whether the density of zeroes is higher than the density of ones towards the right of the plot. A better graphical approach might be to bin the observations into age categories and plot the proportion of survivors in each class as a bar plot (Figure 11.7(c)). This gives an indication of a declining trend in survival but suffers from two problems: the picture is sensitive to the arbitrary choice of bin width and the bars on the right are based on ever-decreasing sample sizes.

Figure 11.7: (a) Histogram of turtle ages in the sample; (b) plot of survivals (1) and deaths (0) against the age of turtles; (c) proportions of surviving turtles calculated in 5 yr age classes.



A binary data set like this can be modelled as follows

$$Y \sim \text{Bernoulli}(p) \tag{11.36}$$

A sigmoidal link function known as the **logit** can be used to write the success probability (p) of each trial as a function of covariates (turtle age, in Example 11.17). The logit link function enables a probability to be transformed so that it can be modelled by a linear predictor (this is written logit(p) =a_0 + a_1X). Mathematically, the transformation has the form

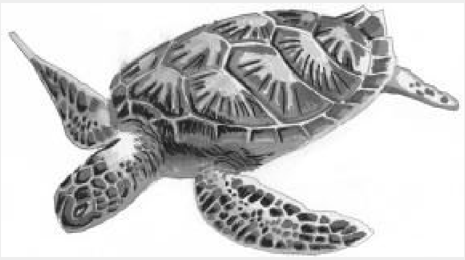
$$p = \frac{\exp(a_0 + a_1X)}{1 + \exp(a_0 + a_1X)} \tag{11.37}$$

Depending on the values of the parameters a_0, a_1, this sigmoid function either declines from 1 to 0 or it increases from 0 to 1. More covariates can be added to the linear predictor to obtain a multiple logistic regression model. You can check, by starting from the PMF of the Bernoulli (Section 9.7) that the log-likelihood of this model is

$$l(a_0, a_1 \mid y_1, \ldots, y_n) = \sum_{i=1}^{n} \{y_i \ln p_i + (1 - y_i) \ln(1 - p_i)\} \tag{11.38}$$

where the data y_i are either 1 or 0 and p_i is given by Equation (11.37) for any particular observation. The combination of binary data, Bernoulli stochastic component and logit link is a specific version of a **logistic GLM**. More generally, logistic GLMs can deal with response data that are constrained by 0 and 1 (i.e. binary responses or proportions) by modelling the associated probability with a sigmoidal link, such as the logit. As example (11.18) shows, if the data are in the form of a proportion of successes from a given number of attempts, the likelihood is derived from the binomial distribution.

Example 11.18: Estimating survival as a function of age



If we have detailed and precise data on single animals, then the analysis may be based on individual survivorship and the parameters of Equation (11.37) can be estimated by maximising the likelihood in Equation (11.38). Finally, the survival model can be plotted as a function of age (Figure 11.8(a)).

However, it may be necessary to treat animals in groups. For example, our data on turtle age may not be very precise. We may, for instance, only be able to classify age within five-year bins. In that case, the data would come in the form of Figure 11.7(c): for the ith age bin we would know the number of turtles (n_i) surviving out of the initial number (N_i) of turtles of that age. Hence, using the binomial PMF (Section 9.8) for the ith bin, the probability of the observations is

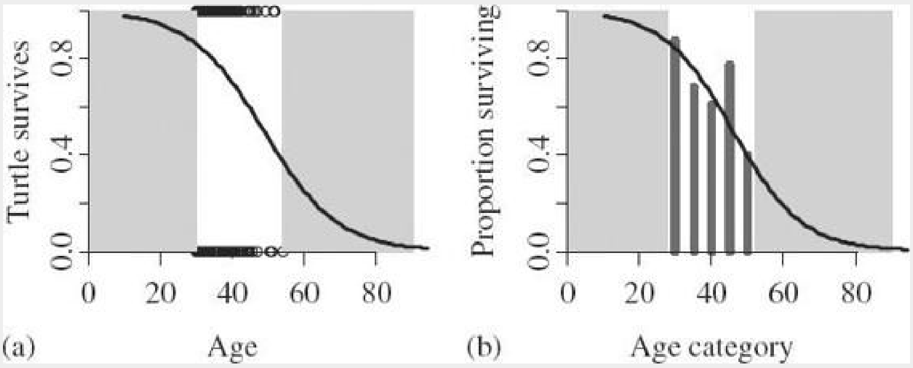
$$f(n_i \mid p_i, N_i) = \binom{N_i}{n_i} p_i^{n_i} (1 - p_i)^{N_i - n_i} \tag{11.39}$$

The age-specific probability of survival (p_i) is modelled by the logit function in Equation (11.37). A little algebra can generate the log-likelihood of the parameters for all of the m age bins combined.

$$l(a_0,a_1 \mid n_1,\ldots,n_m,N_1,\ldots,N_m) = \sum_{i=1}^{m}\{n_i \ln p_i + (N_i - n_i)\ln(1-p_i)\}$$ (11.40)

This likelihood depends on the absolute frequencies n_i of survivors, but also on the number of available animals N_i. In the terminology of logistic GLMs, this second set of numbers are called **weights**. The explanation for their name is the following: different age classes can yield the same observed proportions of survivors, even if they are populated by different numbers of individuals to start with. However, the estimated proportion obtained from an age class with more animals should be more precise. Therefore, the numbers N_i enter the likelihood so that they can correctly weight the maximum likelihood parameter estimates according to the distribution of the sample size across different age classes.

Figure 11.8: Logistic regression on binary (a) and proportional (b) response data. The solid curve shows the fitted model. Its interpretation is subtly different in the two cases. In (a) it can be thought of as the probability of survival of an animal of that age. In (b) it represents the proportion of animals of that age that are expected to survive into the next year. To illustrate the sigmoidal nature of the logistic curve, both models are extrapolated outside the range of the data (into the shaded regions).



A second point about Equation (11.40) is that it no longer contains the binomial coefficient of Equation (11.39) because, fortunately, that part of the expression does not depend on p $_{\Delta i}$ and the parameters of the model (so it does not affect the parameter estimates). Estimating the parameters of the logistic via this likelihood should give similar results to the binary case, assuming that the age classes are not too coarse (Figure 11.8(b)).

11.9: Fitting a logistic GLM to binary and proportion data

Using the specific details of Examples 11.17 and 11.18, I will deal with the binary case first. We have two vectors, the first (`y`) refers to the event of survival and contains only 1s and 0s. The second (`age`) contains individuals' ages. The model can be estimated in just two lines of code

```
dat<-data.frame(y, age) # Data frame with binary data

mod<-glm(y~age, family=binomial, data=dat)
```

Now, assume that the data come in the form of proportions (a vector `f` of relative frequencies of survival) grouped into age bins (the vector `age`). We also need to know the vector (`N`) of the initial numbers of animals in each age bin.

```
datp<-data.frame(f, N, age) # Data frame with proportions

mod<-glm(f~age, family=binomial, data=datp, weights=N)
```

Note the use of the vector `N` as the weights in this version of logistic regression. The commands `summary()`, `fitted()` and `predict()` can be used in the same way as in R11.8. For example, the curve in Figures 11.8(a) and (b) can be plotted as follows:

```
ager<-seq(10,120) #Creates new vector of age values for prediction

newdat<-data.frame("age"=ager) #Specifies single-vector data frame

plot(ager, predict(mod, newdat, type="response"), type="l")
```

11.8. Evaluation, diagnostics and model selection for GLMs

Thanks to its normality assumptions, the linear regression model was traditionally estimated by least squares (minimising the residuals of the model from the data) and so, most of its evaluation and diagnostic criteria are based on residuals. This approach does not work for GLMs which rely heavily on likelihood. We therefore need to develop an alternative criterion for goodness of fit. To do this, we require the concept of a

saturated model, one which uses up all the information in the data by having as many parameter values as there are observations. Remember that the central aim of regression is to tease apart the signal from the noise. A saturated model is one that insists that all the information in the data is signal and, therefore, requires a different parameter to describe each observation. Such a model is guaranteed to be overfit (because real data always contain noise) but it is also guaranteed to have the closest possible fit to the data. Therefore, the maximised log-likelihood of the saturated model (say l_S) will be greater than the maximised log likelihood (say l) of any other model that might be considered. A simple comparison between these two quantities, known as the **residual deviance** (or **deviance** for short) can tell us how close our model is to achieving the closest fit For all models, the deviance will be greater than zero (equal to zero if the model considered is the saturated model). Closer-fitting models have low values of deviance. But how low is low enough? Unfortunately, pseudo-r^2 criteria for GLMs, based on deviance (e.g. $1 - l/l_s$) arenot very reliable, but if you want to report on goodness-of-fit for a particular GLM, use the x^2 test described below.

$$Deviance = 2(l_S - l) \quad (11.41)$$

11.10: Null, residual deviance and diagnostic plots for a GLM

Consider the egg production case (Example 11.16). The model summary looks like this

```
Call:

glm(formula = eggs ~ food, family = poisson, data = dat)



Deviance Residuals:

     Min      1Q  Median       3Q      Max

-4.4730 -0.9846 -0.3440   0.7510   5.1735



Coefficients:

              Estimate Std. Error z value Pr(>|z|)

(Intercept) -3.115797   0.135532 -22.99     <2e-16 ***

food          0.077182   0.001552   49.72    <2e-16 ***

---

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



(Dispersion parameter for poisson family taken to be 1)



     Null deviance: 5589.98 on 199 degrees of freedom

Residual deviance: 447.37 on 198 degrees of freedom

AIC: 1007.0



Number of Fisher Scoring iterations: 5
```

The residual deviance of the model is reported close to the bottom. The null deviance is the maximised log-likelihood of the saturated model. There is nothing particularly mystical about a saturated model. You can specify such a model by asking R to treat each observation as a unique occurrence with no similarity to any of the others. In this example, we can achieve this by treating food availability as a qualitative variable:

```
glm(eggs~as.factor(food), family=poisson, data=dat)
```

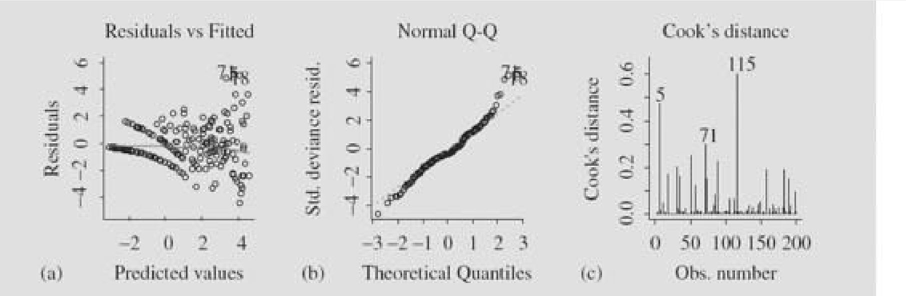
In the summary of this model, the residual deviance is practically zero.

For goodness-of-fit, the following approximate chi-square test can be used, in which a small value (e.g.< 0.05) indicates lack of model fit.

```
1-pchisq(mod$deviance, mod$df.residual)
```

The idea of deviance as a criterion of fit can be extended to single observations in the data. The **deviance residual** is the measure of deviance contributed from each observation. This can be used to generate diagnostic plots similar to the ones in Figure 11.3. Using the turtle egg production data, the command `plot(mod, which=c(1,2,4))` will produce the plots in Figure 11.9. Plotting the residuals against the predicted values of the fitted GLM (Figure 11.9(a)) has a slightly different interpretation to the linear model. Here we are dealing with a Poisson stochastic component, so the variance of the model increases with its mean. Hence the increasing spread of points from left to right. The striations observed towards the left are due to the discreteness of the count data (at low expectations, the Poisson can only yield a few distinct values). Other than that, the residuals are neither consistently over nor under the *x* -axis, so this is a perfectly healthy plot, indicating that the log-linear model is appropriate for these data.

Figure 11.9: Diagnostic plots for log-linear GLM. (a) Simple residuals plotted against the predicted values; (b) Q-Q plot using the standardised deviance residuals; (c) an approximate Cook's distance.



The modified Q-Q plot (Figure 11.9(b)) uses the fact that deviance residuals are approximately normally distributed to evaluate the appropriateness of the stochastic component of the GLM. The alignment of the central bulk of the data with the Q-Q line indicates that the Poisson is an appropriate distribution for these data. Finally, the plot of Cook's distances in Figure 11.9(c) reveals no outliers (all distances are smaller than 1).

The GLMs described above are estimated on the basis of Poisson or binomial likelihoods. It is therefore possible to define for each model a value of the Akaike Information Criterion (Section 11.6) and carry out manual or automatic model selection as outlined in R11.7. In this way, particularly when dealing with many candidate covariates at the same time, it is possible to arrive at a parsimonious model that combines quality of fit and predictive ability.

11.9. Modelling dispersion

The models in the previous sections assume a stochastic component (normal, Poisson, binomial) and use the data to estimate a trend (linear, log-linear, logistic). This approach only estimates the expected value of the response variable. It deals with the variance around this estimated expectation by making some rather restrictive assumptions. For example, in the linear model, the variance of the normal stochastic component was assumed to be constant across the range of the explanatory variables (homoscedasticity). In the case of log-linear regression, the Poisson distribution implicitly imposed equality between mean and variance. Similar implicit constraints were placed on the variance by the use of the binomial in logistic regression. All these assumptions will be violated if the data contain extra signal or extra noise than what is being accounted for by the model. Hence, omitting an important covariate or a source of stochasticity from the likelihood will usually lead to increased dispersion around the fitted model. The first step to a cure is to detect such deviations from the assumed likelihood.

11.11: Investigating violations of the dispersion assumptions

The `ncv.test()` (see R11.5) will detect heteroscedasticity in linear models. Within the GLM framework, the main diagnostic for detecting deviations from the assumed variance is the ratio of the deviance over the residual degrees of freedom. Given a model `mod`, the **dispersion coefficient** ϕ is given by

```
phi<-mod$deviance/mod$df.residual
```

If this is much greater than 1 (e.g. 1.5 or more), the model is overdispersed. Values of ϕ smaller than 1 indicate the (rarer) condition of underdispersion.

Heteroscedasticity may be due to model mis-specification. For example, an important covariate may have been omitted or a more flexible model may be required for the mean (see Sections 11.10 and 11.11 below). If there are no such improvements possible, then it may be useful to model the variance of the model together with the mean. For example, heteroscedasticity may be modelled by allowing the variance to change as a linear or power function of the mean. Alternatively, overdispersion may be modelled by inflating the model's variance. All of these tricks can be achieved by the approach of **quasi-likelihood**.

11.12: Specifying quasi-likelihood models

Consider the simple case of a Poisson likelihood with a log-linear model in which the dispersion coefficient is much greater than 1. The same model can be estimated using quasi-likelihood in a way that allows the stochastic component of the GLM to have a variance higher than its mean. Here is how,

```
mod<-glm(y~x, data=dat, family=quasipoisson)
```

For a binomial model, the option would become `family=quasibinomial`.

In both of these cases the variance is a simple scaling of what was previously stipulated by the Poisson or binomial family. The more complicated case, in which the variance increases or decreases in a nonstandard way, can also be accommodated. For example, a heteroscedastic normal model in which the variance appears to increase as the square of the mean can be specified as follows

```
mod<-glm(y~x, data=dat, family=quasi(link="identity", variance="mu^2"))
```

A difficulty with quasi-likelihood models is that they do not yield a true likelihood and therefore do not come with an associated AIC. This requires an alternative route for model selection. A manual approach using successive testing is described by Faraway (2006) pp. 149–150.

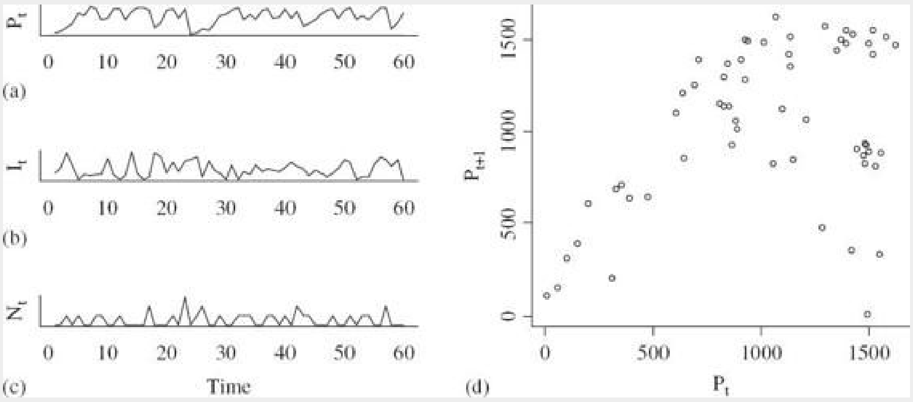
11.10. Fitting more complicated models to data: polynomials, interactions, nonlinear regression

Historically, most regression models used simple monotonic (i.e. either increasing or decreasing) functions. However, many relationships are nonlinear and nonmonotonic. There are now several different options for fitting more complicated curves to data and, to choose between them, you must first ask yourself if you have *a priori* reasons (e.g. theory, first principles or previous analyses) that suggest a particular shape for the curve to be fit. If you do, then the methods in this section will be useful. If you suspect that a nonmonotonic curve is needed but don't quite know what shape it should have, then look at the methods in Section 11.11.

Example 11.19: Density dependence and polynomial terms

Consider a biological population with a small generation time, such as a passerine bird. We have annual data on population size (P_t), an index of immigration from nearby populations (I_t) and accurate information about the number (N_t) of predators in the study area (Figure 11.10). We would like to determine which of these three (density, immigrants, predators) influence population size.

Figure 11.10: (a) Time series of population size for a passerine bird; (b) index of immigrants from nearby populations; (c) number of predators foraging in the study area; (d) population size for next year plotted against this year's population.



It is first necessary to decide on the models to be fit. The simplest possible model in discrete time is unrestricted growth $P_{t+1} = (1 + r)P_t$ (Example 3.2). This is a linear model in P_t so the corresponding regression model is $Y = aX + \varepsilon$ (where ε is the stochastic component. The intercept must be zero because no new individuals can be produced when population size is zero). However, Figure 11.10(d) indicates that the relationship between P_t and P_{t+1} is not linear. P_{t+1} decreases for high values of P_t, probably indicating the existence of density dependence. We can capture this biological feature with a model of constrained growth, such as the discrete logistic model (Example 3.3).

$$P_{t+1} = P_t + r_{\max}\left(1 - \frac{P_t}{K}\right)P_t \quad (11.42)$$

This can be expanded to give a second order polynomial in P_t

$$P_{t+1} = \left(-\frac{r_{\max}}{K}\right)P_t^2 + (1 + r_{\max})P_t \quad (11.43)$$

The general form of the corresponding regression model can be obtained by setting $Y = P_{t+1}$, $P_t = X$, $a_0 = 0$, $a_1 = (1 + r_{\max})$and $a_2 = -r_{\max}/K$

$$Y = a_2X^2 + a_1X + a_0 + \varepsilon \quad (11.44)$$

The need to fit nonlinear models to ecological data arises often. In some situations (as in the above example), the assumed ecological process gives rise to a polynomial. In other cases, a polynomial merely offers a convenient approximation to the shape of the data. Polynomial terms can be added to linear models or GLMs to increase their flexibility, but it is good practice during model selection to drop the highest order terms of each variable first.

11.13: Fitting models with polynomial terms

In the formula definition, a higher order term needs to be passed inside the expression `I()`. Below are the specifications for the linear and quadratic model in Example 11.19.

```
# Create a data frame with current (N) and

# previous (Npr) population size

dat<-data.frame("N"=n[2:tmax],"Npr"=n[1:(tmax-1))

# Fit the models

mod0_EX<-lm(N~ -1+Npr, dat) # Linear model

mod1_DD<-lm(N~ -1+Npr+I(Npr^2), dat) # Quadratic model
```

The presence of −1 in both formulae forces the models to have a zero intercept.

Example 11.20: Predation, immigration and interaction terms

The remaining information in the data may now be used to further explain the observed pattern in Figure 11.10(d). A model containing immigration and predation can be written as

$$P_{t+1} = P_t + r_{\max}\left(1 - \frac{P_t}{K}\right)P_t + \beta I_t + \alpha P_t N_t \quad (11.45)$$

Here, I have assumed that the number of immigrants is proportional to the index of immigration (I_t) and that the predators consume their prey according to a Type I functional response (see Example 4.14) with parameter α. This model now has three explanatory variables (P_t, I_t, N_t), it is quadratic in (P_t) and has a multiplication between two of its variables as part of the functional response. Setting $X_1 = P_t$, $X_2 = N_t$, $X_3 = I_t$, $a_0 = 0$, $a_2 = 0$ gives the following regression model

$$Y = a_0 + a_1X_1 + a_2X_2 + a_3X_3 + a_4X_1^2 + a_5X_1X_2 \quad (11.46)$$

This latest model includes a second order polynomial term, multiple explanatory variables and a multiplicative term between two explanatory variables. In regression models, such multiplicative terms are known as **interactions**. This name makes good sense in the above example since it refers to the trophic interaction between the predator and prey populations. More generally, an interaction term in a regression model implies that two explanatory variables do not act additively on the response (i.e. the slope of the response to one variable changes with the value of the other variable). Similar to polynomial terms, it is accepted practice that when using an interaction term, the linear terms participating in it must also be in the model (even if their coefficients are estimated close to zero).

11.14: Fitting models with interaction terms

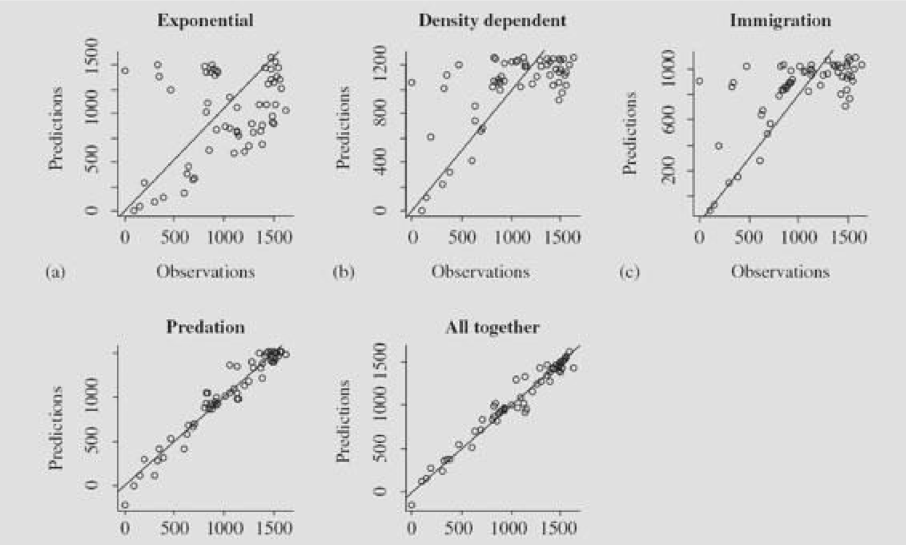
Models with interaction terms are easily specified, although it is generally harder to interpret their meaning. So, before you introduce an interaction, make sure you have a biological reason. Here are three different models examining different biological explanations for the data in Example 11.19.

```
# Model with density dependence and immigration

mod2_IM<-lm(N~-1+Npr+I(Npr^2)+I, dat)

# Model with density dependence and predation

mod3_PR<-lm(N~-1+Npr+I(Npr^2)+P+Npr*P, dat)

# Model with density dependence, immigration and predation

mod4_All<-lm(N~-1+Npr+I(Npr^2)+I+P+Npr*P, dat)
```

A comparison of the AIC values between the five models in R11.12 and R11.13 indicates that the model with density dependence, immigration and predation offers the best explanation for the observed patterns.

```
> AIC(mod0_EX, mod1_DD, mod2_IM, mod3_PR, mod4_All)

        df     AIC

mod0_EX 2 898.2920

mod1_DD 3 858.8015

mod2_IM 4 859.9041

mod3_PR 5 725.4281

mod4_All 6 709.0039
```

Figure 11.11: Plots of fitted against observed values for the five models examined here: (a) exponential growth; (b) density dependence; (c) density dependence with the effect of immigration; (d) density dependence with the effect of predation; (e) density dependence with combined effect of predation and immigration.



We can visualise the quality of fit by plotting fitted against observed values for all five models (Figure 11.11).

```
par(mfrow=c(2,3))  # Splits graphics device into 2 rows and 3 columns

plot(dat$N, fitted(mod0_EX), main="a.Exponential", xlab="Observations",

                                                  ylab="Predictions")

abline(0,1)

plot(dat$N, fitted(mod1_DD), main="b.Density dependent",

               xlab="Observations", ylab="Predictions")

abline(0,1)

plot(dat$N, fitted(mod2_IM), main="c.Immigration", xlab="Observations",

                                                  ylab="Predictions")

abline(0,1)

plot(dat$N, fitted(mod3_PR), main="d.Predation", xlab="Observations",

                                                ylab="Predictions")

abline(0,1)

plot(dat$N, fitted(mod4_All), main="e.All together", xlab="Observations",

                                                    ylab="Predictions")

abline(0,1)

par(mfrow=c(1,1))  # Returns graphics device to original
```

This latest R box illustrates the power of model selection as a method for testing scientific hypotheses. In this data set, it has been able to identify the mechanisms that drive the population dynamics of this population.

I have chosen to motivate the different types of empirical models in this section from a mechanistic viewpoint. In the above discussion, quadratic and interaction terms arise naturally from the population dynamics models examined in Chapter 3. This approach allows you to appreciate the necessity and meaning of such nonlinear terms from an ecological perspective but it is also somewhat misleading. We rarely motivate empirical models from specific mechanistic models. We may have a vague notion that predators interact multiplicatively with prey, and a population responds nonlinearly to its past density but we rarely express the regression model as a reparameterisation of a mechanistic model – it is simply too much work and not always possible.

However, if you have a mechanistic model that you feel very strongly is a good representation of reality then, assuming that it cannot be written as a linear combination of polynomial and interaction terms, there are increasingly good methods for fitting it directly to data. Such methods belong to the broader area of **nonlinear regression** that can be implemented via three routes:

➊*Nonlinear least squares* : If the residuals around the nonlinear model can be assumed to be normal, then least squares is the quickest method. Crawley (2007) and Bolker (2008) offer more details with ecological examples and R implementation.

➋*Maximum likelihood* : A custom stochastic component can be specified around the model from the selection of the distributions presented in Chapter 9. If the likelihood for this model can be written, then it may also be optimised using the methods discussed in Chapter 10. Bolker (2008) ventures quite deeply into this area.

➌*Bayesian methods* : It makes similar requirements as the maximum likelihood approach, but with the added advantages of being able to specify priors for the parameters. McCarthy (2008) and Bolker (2008) offer good introductions.

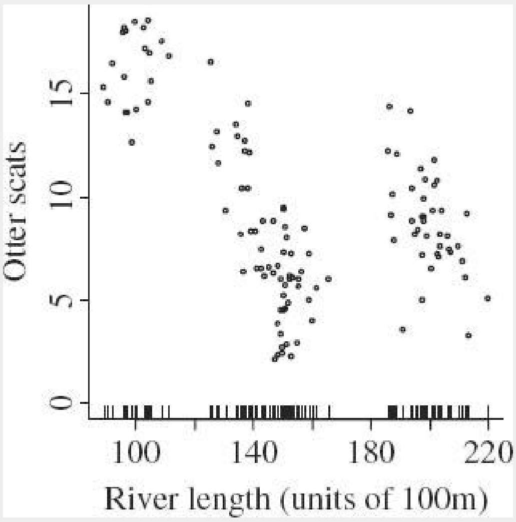
11.11. Letting the data suggest more complicated models: smoothing

Example 11.21: Distribution along a linear habitat

A particular study maps the presence of otters along a river. Researchers visit 100 m-long segments of the river bank and count the number of otter scats they find. The time intervals between visits are larger than the time it takes for scats to decompose, to ensure no double counting. Because of accessibility constraints, the sampling effort differs for different river segments. The resulting data are shown in Figure 11.12.

Assuming that the number of scats is an unbiased index of usage, these data will still be affected by sampling variation (i.e. even if usage does not change, repeat visits to the same 100 m segment will encounter different numbers of scats). So, there is certainly noise in this data, but is there also a signal? There appear to be two peaks at around the 100- and 200-segment points. Given just these data, there is no *a priori* reason to assume a curve of a particular form (as was done in Section 11.10), but it is still useful to draw a curve through the cloud, to help generate some hypotheses.

Figure 11.12: Number of otter scats found on different visits plotted against their position on the river banks. Sampling effort is indicated by the 'rug' of small vertical lines on the *x* -axis.



In situations like this, when both signal and noise are present but there is no clue for how to describe the signal mathematically, **smoothing** is a handy tool. The idea behind it is simple: if there is signal in the data, then points that are close together in the *x* -axis must also have similar *y* values. In other words, there must be some autocorrelation in the response data (see Section 2.4).

The easiest way to iron out the sampling variation is to obtain an estimate of the response variable as the average of *y* values in the vicinity of a particular *x* value ($x \pm ^{1/2} \Delta x$). Doing this for all *x* values in the range of the data yields a **moving average**, the simplest type of smoother. Defining the 'vicinity of a particular *x* value' sounds simple, but in reality the shape of the resulting curve will depend on this **smoothing window** or **bandwidth** (Δx). Furthermore, a moving average is problematic because, for any particular point *x* it uses all points inside the range $x \pm ^{1/2} \Delta x$ and ignores all points outside it. Ideally, the similarity between observations should decay with their separation along the *x*-axis.

There are several generalisations of the moving average that can account for these drawbacks. For example, a technique known as **kernel smoothing** weights the influence of neighbouring points on each-other according to their proximity along the *x* -axis. The 'kernel' in the name of the method is the weighting function that achieves this. Mathematically, for a sample of *n* observations, the kernel-smoothed estimate *f* (*x*) at a value *x* of the independent variable looks just like a weighted average

$$f(x) = \sum_{i=1}^{n} w_i y_i \Big/ \sum_{i=1}^{n} w_i \quad (11.47)$$

In this expression, the kernel is w_i. It works by allocating a portion of the observation y_i to the estimate at *x*. Many functions can serve as kernels, but we usually want functions that decay away from *x*, such as the following bell-shaped function that is related to the normal (Gaussian) PDF.

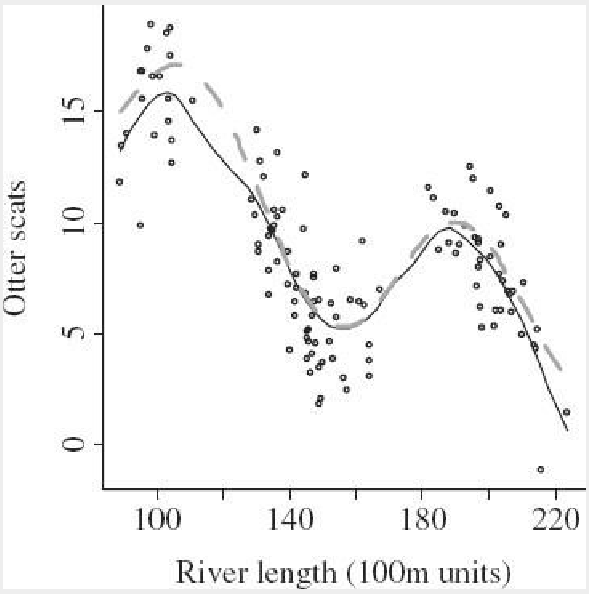
$$f(x) = e^{-\frac{(x-x_i)^2}{\lambda}} \quad (11.48)$$

This has two desirable properties: it attains its maximum value at x, so that observations close to the position being estimated drive the estimate. It also has a parameter λ which operates analogously to the variance in the normal distribution. By increasing the value of λ, the smoothing window is extended and hence imposes a greater degree of smoothness on the estimated function. Deciding just how smooth a function should be is tricky. If you have some information about this, then by all means decide for yourself what the value of λ should be. There are also automatic methods for choosing the smoothing window, based on cross-validation (see Section 11.6). They work by omitting one observation at a time, trying a range of values for λ and measuring how well each missing observation is predicted by the curves created from each λ.

Example 11.22: Otter density estimation by kernel smoothing

The data in Example 11.21 are simulated, so the true distribution of otter scats is known to us but not to the smoother. The performance of the normal kernel with automatic bandwidth selection (via cross-validation) is shown in Figure 11.13.

Figure 11.13: Kernel-smoothed estimate of otter scat density (solid line) and the underlying true density (dashed grey line).



11.15: Estimation by kernel smoothing

The built-in command `ksmooth()` will readily produce a kernel-smoothed curve for a data set with two variables `x` and `y`.

```
plot(x,y)# Scatter plot

rug(x)# Helps visualise sampling effort along x-axis

# Estimates & plots smooth

lines(ksmooth(x,y, kernel="normal", bandwidth=10))
```

Here, the options of `ksmooth` specify a normal kernel (see Equation (11.48)) and an arbitrary bandwidth of 10. The bandwidth can be determined automatically by cross-validation, via the library `sm`.

```
require(sm)# Loads library

bandw<-hcv(x, y)# Applies cross validation to estimate bandwidth

sm.regression(x, y, h=bandw)# Generates plot and adds smooth
```

Kernel smoothing can be applied with more than one explanatory variable, but requires a multivariate kernel (e.g. one based on the multivariate normal distribution, see Section 9.17). A similar method that I will not cover here is **locally weighted regression scatterplot smoothing** (**LOESS**). Both of these methods are equally flexible (or inflexible, depending on the choice of bandwidth) across the entire range of values for the independent variable. But, as we saw in Example 11.21, unbalanced sampling effort may result in more information being available for particular segments of the *x* -axis. Also, their application to more than two explanatory variables is cumbersome.

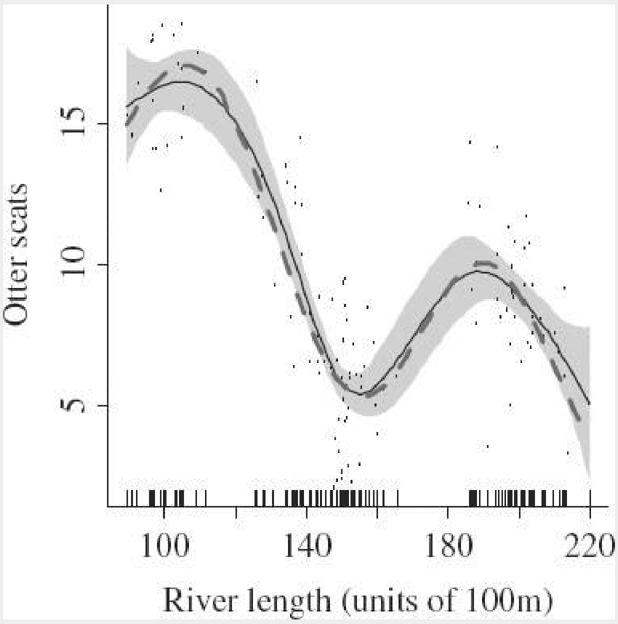
A more advanced technique, the **generalised additive model** (**GAM**), takes the concept of smoothing further by applying it in a piece-wise manner. A GAM first splits the range of *x* values into segments whose cut-off points are called **knots**. It then fits a flexible curve (usually a polynomial) to the data within each segment. The parameters of each of these piece-wise polynomials are selected in a way that satisfies two criteria: ❶ the polynomials must fit the data in each segment well and ❷ the polynomials in adjacent segments must join smoothly at the knots. By choosing the position of the knots according to the availability of the data, the GAM can describe the response with higher detail in data-rich parts of the *x* -axis.

GAMs have the full functionality of a GLM (variety of stochastic components, ability to deal with large numbers of explanatory variables, use of interaction terms, estimation of confidence intervals for predictions) combined with the flexibility of a smoother. This flexibility is both a blessing and a curse – GAMs have often been accused of overfitting the data. In the case of multiple explanatory variables in particular, a GAM needs to be parsimonious both in terms of the variables it contains and the 'wiggliness' with which it describes the response to each of them. A recent development in GAMs called a **shrinkage smoother** achieves both of these objectives simultaneously: by using cross-validation, it allows the wiggliness for each covariate to go to zero as required by the data. Such a flat-lined covariate nominally remains in the model but, practically, has no effect on model predictions (it is, in effect, selected out).

Example 11.23: Otter density estimation by GAM

One of the design features of the generalised additive model is its smoothness. By construction, GAMs are best suited to modelling response variables that change gradually with changes in the explanatory variables. In this sense, the simulated otter scat density is captured extremely well by the GAM in Figure 11.14.

Figure 11.14: GAM fit to the simulated otter density data. The thin solid line represents expected scat density as estimated by the GAM. The dashed line is the real underlying density. The confidence limits around the GAM predictions are shown as a grey band.



11.16: Estimation by GAMs

There are two different packages in R for fitting GAMs, `gam` and `mgcv`, the second being the most actively developed, documented and user-friendly. The syntax for invoking a GAM is similar to the GLM. The following example fits a GAM to the otter scat data and plots the results in a graph similar to Figure 11.14.

```
require(mgcv) # Loads library

dat<-data.frame(x,y) # Creates data frame with two variables

mod<-gam(y~s(x),data=dat, family=gaussian)# Fits the gam

plot(mod, xlab="River length (units of 100m)", ylab="Otter scats",

                              shade=T, residuals=T, shift=mean(y))
```

Notice how the explanatory variable is now enclosed in `s(x)` in the formula, indicating that the response to this variable is to be smoothed by the GAM. The option `family` here isn't strictly necessary since the default for this is `gaussian` anyway, but it does show you where to specify a `Poisson` or `binomial` stochastic component if your response data are counts or proportions. Some of the plotting options need further explanation. The confidence limits around the mean prediction can be plotted as a pair of curves or a shaded zone. The option `shade=T` gives a result like the one in Figure 11.14. The option `residuals=T` shows (as dots) the residuals of the raw data from the predictions. Also, by default, the y -axis for a GAM plot shows the residuals on the scale of the linear predictor. This being a Gaussian model, I have simply shifted the mean of the predictions up to the mean of the response data by specifying `shift=mean(y)`. As a result, here the residual dots are the raw data.

The implementation of GAMs in the `mgcv` library automates the selection of knots and the flexibility of the smooths, so all the complicated choices are transparent to the user.

11.12. Partitioning variation: mixed effects models

I started this chapter by asserting that the observed variation in real data comprises both signal and noise (although, of course, not all data sets contain detectable signal and some can be assumed to be almost noise-free). In the models examined, the signal took the form of an expectation (a point, a curve or a surface) and the noise was the remaining variation around it. All of the sections so far have looked at methods for constructing this expectation by trying out different transformations and combinations of one or more explanatory variables. In the process, much of the original variation in the data was explained away. However, in Section 11.6, while discussing model selection, I mentioned that we must resign ourselves to having some residual noise because not all of the variation in the data can be explained with a deterministic signal. This is true, but we may still have some idea of what causes the residual variation in the data and may yet be able to attribute *portions* of this variation to different properties of the sample units. This final section can be motivated by the same example that introduced the chapter.

Example 11.24: Samples of gannet condition

Change in the condition of breeding female gannets during a particular time period of the year is $\Delta i = I_{After} - I_{Before}$. This index represents the ability of birds to acquire condition. Values close to zero indicate no change, negative values indicate deterioration and positive values mean that resources are being acquired faster than they are being expended. Through slight differences in their genetic makeup and experience, different individuals will be more or less efficient at the tasks of foraging, maintenance and breeding. Additionally, through slight differences in local environmental conditions, the birds living in different colonies will be differentially efficient. We are aware that the residual variation in a model for Δi will comprise both individual and colony variation, so how can we account for these two sources in a regression approach?

Consider the linear model in one covariate

$$(11.49)\quad Y = a_1 X + a_0 + \varepsilon$$

Bundled together in the stochastic component ε are two, or perhaps more, different types of stochasticity, originating from different processes. If two such processes can be identified (stochastic components are conventionally lower-case Greek letters), then perhaps the model can be rewritten as

$$(11.50)\quad Y = a_1 X + a_0 + (\phi + \psi)$$

where $\varepsilon = \phi + \psi$. If the two errors ϕ, ψ apply equally and independently to all sample units, then this latter version of the linear model gains us nothing. However, if there are asymmetries or similarities between sample units that can be attributed to group membership, then Equation (11.50) can extract more information from the data than can Equation (11.49).

Example 11.25: Modelling individual and colony variation

In Example 11.3, the change in condition was written as a linear function of local density for 50 birds (ten from each of five colonies). I will here use subscripts to represent the *i*th bird in the *j*th colony.

(11.51) $$\Delta I_{i,j} = a_1 P_j + a_0 + \varepsilon_{i,j}$$

Here, the change in condition (Δi) and the residual (ε), refer to a particular individual so they are subscripted by both *i* and *j*. Density measurements (*P*) are the same for all birds living in the same colony, so colony density is subscripted only by *j*. Following the rationale of Equation (11.50), the stochastic component can be broken up into individual-specific ($\psi_{i,j}$ and colony-specific (ϕ_j)terms

(11.52) $$\Delta I_{i,j} = a_1 P_j + (a_0 + \phi_j) + \psi_{i,j}$$

The brackets inserted in this latest equation hint at an alternative interpretation of this model. A quick comparison between Equations (11.51) and (11.52) shows them to be structurally identical, apart from the fact that the second equation has an intercept that changes randomly from one colony to the next. Biologically, this **random-intercept** model says that the baseline ability of animals to improve their condition is affected by a number of unknown, colony-specific influences.

This model can be further extended to represent the idea that the effect of density on animals is also colony specific. Mathematically, we can achieve this by introducing a colony-specific random effect (ω_j) to the slope of the response to density, to obtain a **random slope and random intercept model**.

(11.53) $$\Delta I_{i,j} = (a_1 + \omega_j) P_j + (a_0 + \phi_j) + \psi_{i,j}$$

Although the terms ω_j, ϕ_j, $\psi_{i,j}$, in Equation (11.53) resemble model coefficients, they are, in fact, random variables with mean zero. So, what we seek to estimate for a model such as Equation (11.53) are the expectations of the **fixed effects** (a_0, a_1) and the variances of the **random effects** (ω_j, ϕ_j, $\psi_{i,j}$). Biologically, this quantifies the between-colony variation in the ability of birds to acquire condition and the impact of colony density. The mathematics used for estimating mixed effects models is beyond the scope of this textbook. It is more important for you to know when to seek the assistance of this type of model. If your data can be subdivided into discrete classes and if there is a possibility that common class membership leads to similarity between sample units, then a mixed effects model may be what you want, *assuming* you are interested in making population-level inferences that reach beyond the specific classes in your data.

Example 11.26: Why not use colony as a factor?

Consider the random intercept model in Equation (11.52). One entirely legitimate approach to dealing with colony effects would be to incorporate the colony as a factor in the model. We would then have one coefficient for the effect of density and one coefficient for each of the five levels of the factor colony. The ability of such a model to describe the data would be similar to Equation (11.52) but it would have two distinct disadvantages: first, the number of parameters in the model would increase linearly with the number of colonies in the data; second, it would only model the response of animals at the particular colonies included in the data, without giving us an estimate of between-colony variation.

11.17: Fitting a mixed effects model in lme4

Much of the original published output for mixed models with R used the `nlme` package (described in detail in Pinheiro and Bates, 2000). Users are now gradually shifting to the newer `lme4`. The command for fitting a mixed model is `lmer()`. Inside it, you have to specify similar options to what was used in `lm()` and `glm()`. The only new element is the declaration of the random effects in the model formula. A typical mixed effects formula in `lme4` looks like this:

```
response variable ~ fixed effects+(random effects | grouping factor)
```

The data frame `dat` is the same as was used in R11.2, containing a row for each individual gannet and columns for colony membership (Colony), colony density (Density) and change in condition (Condition). Below is the code for estimating the two models in Equations (11.52) and (11.53)

```
require(lme4)

# Random intercept model

mod0<-lmer(Condition~Density+(1|Colony), data=dat, family=gaussian)



# Random slope and random intercept model

mod1<-lmer(Condition~Density+(Density|Colony), data=dat, family=gaussian)
```

and here is the output generated for `mod1`

```
> summary(mod1)

Linear mixed model fit by REML

Formula: Condition ~ Density + (Density | Colony)

      Data: dat

    AIC   BIC  logLik   deviance   REMLdev

  448.8 460.3 -218.4       438.7     436.8

Random effects:

  Groups   Name        Variance   Std.Dev. Corr

  Colony   (Intercept) 125.476633 11.20164

          Density        0.016649   0.12903 -1.000

  Residual             377.198612 19.42160

Number of obs: 50, groups: Colony, 5



Fixed effects:

             Estimate Std. Error t value

(Intercept) 36.1525       9.6958   3.729

Density      -0.7189       0.1291  -5.567



Correlation of Fixed Effects:

         (Intr)

Density -0.943
```

This report is split between two main parts, referring to random and fixed effects. The output for the fixed effects is similar to what we would get out of `lm()` or `glm()` (i.e. coefficient estimates, their standard errors and significance). The output for the random effects reports their estimated variances. Given the very small value of the random effect for density (0.0166), it would appear that the estimated slope of the animals' response to density (−0.7189) does not change between colonies. It would also appear that individual variation is three times greater than between-colony variation (compare 377.199 with 125.477).

There are three important extensions of the basic mixed effects model:

❶ The response variable can be modelled with a non-normal stochastic component (giving rise to Generalised Linear Mixed Models – GLMMs) or using flexible smoothers (giving rise to Generalised Additive Mixed Models – GAMMs).

❷ The structure of the groupings can become much more elaborate. For example, sources of variation can be **nested** in multilevel hierarchies (e.g. individual, subpopulation, metapopulation) or using **crossed** classifications (e.g. male and female individuals within different populations).

❸ Mixed effects models are ideal for analysing **longitudinal data**, i.e. repeat measurements collected from the same subject at different points in time. Examples include the collection of multiple physiological and physical measurements from individuals during the course of their lives, replicate measurements of biodiversity at predetermined spatial locations, positional data on the movement of satellite-tagged animals, etc.

Longitudinal data are correlated in two different ways: first, any two observations are likely to be similar when they originate from the same subject (**within-subject correlation**) and second, two successive observations are likely to be similar because any one subject changes gradually through time (**serial correlation**). Both of these types of autocorrelation are responsible for the ailment of **pseudoreplication**, affecting all longitudinal data sets. The problem can be summarised as follows: when data are autocorrelated, the apparent sample size (i.e. the number of rows in the data frame) is misleadingly large because nonindependent data contain less information than independent ones. Assuming independence (as most elementary regression methods do) for fitting and model selection generally leads to overfitted models. In contrast, within-subject correlation is automatically taken care of by the standard mixed effects model and the addition of extra components called **autoregressive structures** can account for serial correlation.

Further reading

ANOVA is covered extensively by Sokal and Rohlf (1995), Quinn and Keough (2002), Gotelli and Ellison (2004) and Whitlock and Schluter (2009). Perhaps the most useful books on the linear model and its various extensions are by Faraway (2004, 2006), but a wealth of information can also be found in the all-purpose textbooks by Fox (1997, 2002) and Krzanowski (1998). The definitive monograph on model selection is Burnham and Anderson (2002) but the topic is covered by every textbook on multiple regression. Density estimation by smoothing is covered by Faraway (2006) and a good theoretical reference is Silverman (1986). The most comprehensive reference for GAMs is Wood (2006). Mixed effects models are covered by Pinheiro and Bates (2000) and Zuur *et al.* (2009). Most modern books on regression now offer introductory chapters to GLMs, GAMs and mixed models.

References

Bolker, B.M. (2008)*Ecological Models and Data in R*. Princeton. 408pp.

Burnham, K.P. and Anderson, D.R. (2002)*Model Selection and Multimodel Inference: A PracticalInformation-theoretic Approach*. Springer-Verlag, New York. 488pp.

Camphuysen, K. and Webb, A. (1999) Multi-species feeding associations in North Sea seabirds:

Jointly exploiting a patchy environment.*ARDEA*, **87**, 177–198.

Crawley, M.J. (2007)*The R Book*. John Wiley & Sons, Ltd, Chichester. 942pp.

Faraway, J.J. (2004)*Linear Models with R*. Chapman & Hall, Boca Raton, Florida. 240pp. Faraway, J.J. (2006)*Extending the Linear Model with R: Generalised Linear, Mixed Effects and Nonparametric Regression Models*. Chapman & Hall, Boca Raton, Florida. 301pp.

Fox, J. (1997)*Applied Regression and Analysis, Linear Models and Related Methods*. Sage, Thousand Oaks, California. 597pp.

Fox, J. (2002)*An R and S-plus Companion to Applied Regression*. Sage, Thousand Oaks, California. 311pp.

Gotelli, N.J. and Ellison, A.M. (2004)*A Primer of Ecological Statistics*. Sinauer Associates, Massachusetts. 510pp.

Krzanowski, W.J. (1998)*An Introduction to Statistical Modelling*. Arnold, London. 252pp. Pinheiro, J.C. and Bates, D.M. (2000)*Mixed-effects Models in S and S-plus*. Springer, New York. 528pp.

Quinn, G.P. and Keough, M.J. (2002)*Experimental Design and Data Analysis for Biologists*. Cambridge University Press. 537pp.

Silverman, B.W. (1986)*Density Estimation for Statistics and Data Analysis*. Chapman & Hall, Boca Raton, Florida. 176pp.

Sokal, R.R. and Rohlf, F.J. (1995)*Biometry*, 3rd edition. Freeman, New York. 887pp. Whitlock, M.C. and Schluter, D. (2009)*The Analysis of Biological Data*. Roberts & Co., Colorado. 700pp.

Wood, S. (2006)*Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC, Boca Raton, Florida. 391pp.

Zuur, A.F., Ieno, E.N., Walker, N.J., Saveliev, A.A. and Smith, G.M. (2009)*Mixed Effects Models and Extensions in Ecology with R*. Springer, New York. 574pp.