1

An introduction to computer-intensive methods

What are computer-intensive data methods?

For the purposes of this book, I define computer-intensive methods as those that involve an iterative process and hence cannot readily be done except on a computer. The first case I examine is maximum likelihood estimation, which forms the basis of most of the parametric statistics taught in elementary statistical courses, though the derivation of the methods via maximum likelihood is probably not often given. Least squares estimation, for example, can be justified by the principle of maximum likelihood. For the simple cases, such as estimation of the mean, variance, and linear regression analysis, analytical solutions can be obtained, but in more complex cases, such as parameter estimation in nonlinear regression analysis, whereas maximum likelihood can be used to define the appropriate parameters, the solution can only be obtained by numerical methods. Most computer statistical packages now have the option to fit models by maximum likelihood but they typically require one to supply the model (logistic regression is a notable exception).

The other methods discussed in this book may have an equally long history as that of maximum likelihood, but none have been so widely applied as that of maximum likelihood, mostly because, without the aid of computers, the methods are too time-consuming. Even with the aid of a fast computer, the implementation of a computer-intensive method can chew up hours, or even days, of computing time. It is, therefore, imperative that the appropriate technique be selected. Computer-intensive methods are not panaceas: the English adage "you can't make a silk purse out of a sow's ear" applies equally well to statistical analysis. What computer-intensive methods allow one to do is to apply a statistical analysis in situations where the more "traditional" methods fail. It is important to remember that, in any investigation, great efforts should be put

into making the experimental design amenable to traditional methods, as these have both well-understood statistical properties and are easily carried out, given the available statistical programs. There will, however, inevitably be circumstances in which the assumptions of these methods cannot be met. In the next section, I give several examples that illustrate the utility of computer-intensive methods discussed in this book. Table 1.1 provides an overview of the methods and comments on their limitations.

Why computer-intensive methods?

A common technique for examining the relationship between some response (dependent) variable and one or more predictor (independent) variables is linear and multiple regression. So long as the relationship is linear (and satisfies a few other criteria to which I shall return) this approach is appropriate. But suppose one is faced with the relationship shown in Figure 1.1, that is highly nonlinear and cannot be transformed into a linear form or fitted by a polynomial function. The fecundity function shown in Figure 1.1 is typical for many animal species and can be represented by the four parameter (M, k, t_0, b) model

$$F(x) = M(1 - e^{-k(x-t_0)})e^{-bx} \tag{1.1}$$

Using the principle of maximum likelihood (Chapter 2), it can readily be shown that the "best" estimates of the four parameters are those that minimize the residual sums of squares. However, locating the appropriate set of parameter values cannot be done analytically but can be done numerically, for which most statistical packages supply a protocol (see caption to Figure 1.1 for S-PLUS coding).

In some cases, there may be no "simple" function that adequately describes the data. Even in the above case, the equation does not immediately "spring to mind" when viewing the observations. An alternative approach to curve fitting for such circumstances is the use of local smoothing functions, described in Chapter 6. The method adopted here is to do a piece-wise fit through the data, keeping the fitted curve continuous and relatively smooth. Two such fits are shown in Figure 1.2 for the *Drosophila* fecundity data. The loess fit is less rugged than the cubic spline fit and tends to de-emphasize the fecundity at the early ages. On the other hand, the cubic spline tends to "over-fit" across the middle and later ages. Nevertheless, in the absence of a suitable function, these approaches can prove very useful in describing the shape of a curve or surface. Further, it is possible to use these methods in hypothesis testing, which permits one to explore how complex a curve or a surface must be in order to adequately describe the data.

Table 1.1 *An overview of the techniques discussed in this book*

Method	Chapter	Parameter estimation?	Hypothesis testing?	Limitations
Maximum likelihood	2	Yes	Yes	Assumes a particular statistical model and, generally, large samples
Jackknife	3	Yes	Yes	The statistical properties cannot generally be derived from theory and the utility of the method should be checked by simulation for each unique use
Bootstrap	4	Yes	Possible[a]	The statistical properties cannot generally be derived from theory and the utility of the method should be checked by simulation for each unique use. Very computer-intensive.
Randomization	5	Possible	Yes	Assumes difference in only a single parameter. Complex designs may not be amenable to "exact" randomization tests
Monte Carlo methods	5	Possible	Yes	Tests are usually specific to a particular problem. There may be considerable debate over the test construction.
Cross-validation	6	Yes	Yes	Generally restricted to regression problems. Primarily a means of distinguishing among models.
Local smoothing functions and generalized additive models	6	Yes	Yes	Does not produce easily interpretable function coefficients. Visual interpretation difficult with more than two predictor variables
Tree models	6	Yes	Yes	Can handle many predictor variables and complex interactions but assumes binary splits.
Bayesian methods	7	Yes	Yes	Assumes a prior probability distribution and is frequently specific to a particular problem

[a] "Possible" = Can be done but not ideal for this purpose.

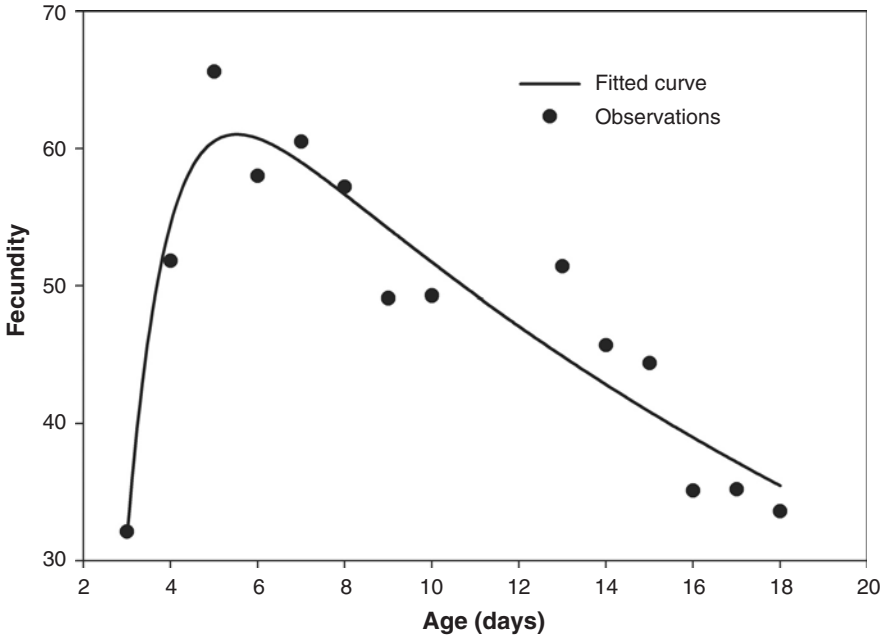

Figure 1.1 Fecundity as a function of age in *Drosophila melanogaster* with a maximum likelihood fit of the equation $F(x)=M(1-e^{k(x-t_0)})e^{-bx}$. Data are from McMillan *et al.* (1970).

Age (x)	3	4	5	6	7	8	9	10	13	14	15	16	17	18
F	32.1	51.8	66	58	60.5	57.2	49.1	49.3	51.4	45.7	44.4	35.1	35.2	33.6

S-PLUS coding for fit:

```
# Data contained in data file D
# Initialise parameter values
   Thetas <- c(M=1, k=1, t0=1, b=.04)
# Fit model
   Model <- nls(D[,2]~M*(1-exp(-k*(D[,1]-t0)))*exp(-b*D[,1]), start=Thetas)
# Print results
   summary(Model)
OUTPUT
Parameters:
        Value   Std. Error    t value
 M  82.9723000  7.52193000  11.03070
 k   0.9960840  0.36527300   2.72696
t0   2.4179600  0.22578200  10.70930
 b   0.0472321  0.00749811   6.29920
```

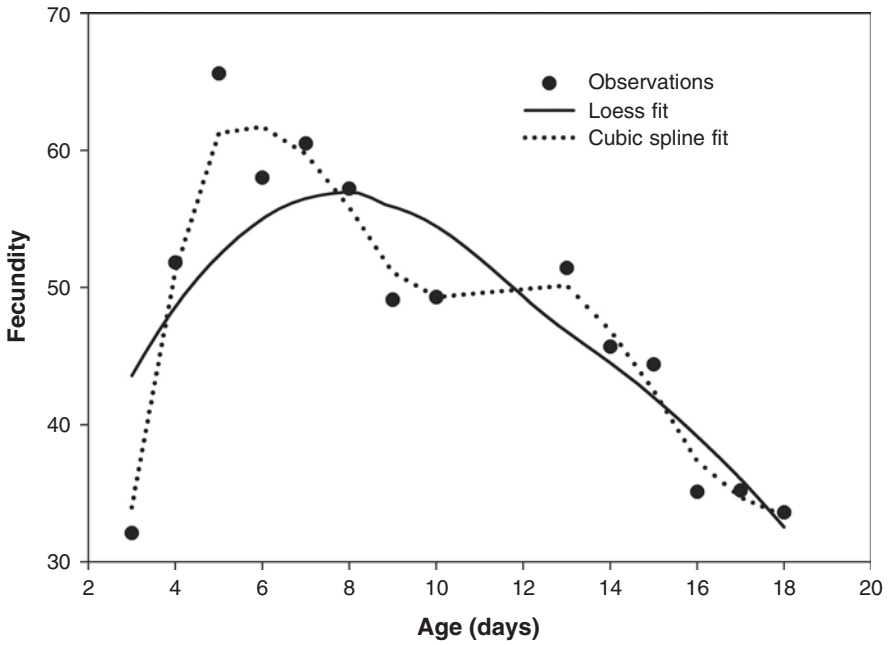

Figure 1.2 Fecundity as a function of age in *Drosophila melanogaster* with two local smoothing functions. Data given in Figure 1.1.
S-PLUS coding to produce fits:

```
# Data contained in file D. First plot observations                # Plot points
      plot (D[,1], D[,2])
   Loess.model <- loess(D[,2]~D[,1], span=1, degree=2)            # Fit loess model
# Calculate predicted curve for Loess model
   x.limits <- seq(min(D[,1]), max(D[,1]), length=50              # Set range of x
   P.Loess <- predict.loess(Loess.model, x.limits, se.fit=T)      # Prediction
   lines(x.limits, D.INT$fit)                                     # Plot loess prediction
   Cubic.spline <- smooth.spline(D[,1], D[,2])                    # Fit cubic spline model
   lines(Cubic.spline)                                            # Plot cubic spline curve
```

An important parameter in evolutionary and ecological studies is the rate of increase of a population, denoted by the letter r. In an age-structured population, the value of r can be estimated from the Euler equation

$$1 = \sum_{x=0}^{\infty} e^{-rx} l_x m_x \tag{1.2}$$

where x is age, l_x is the probability of survival to age x and m_x is the number of female births at age x. Given vectors of survival and reproduction, the above equation can be solved numerically and hence r calculated. But having an estimate of a parameter is generally not very useful without also an estimate of

the variation about the estimate, such as the 95% confidence interval. There are two computer-intensive solutions to this problem, the jackknife (Chapter 3) and the bootstrap (Chapter 4). The jackknife involves the sequential deletion of a single observation from the data set (a single animal in this case) giving n (= number of original observations) data sets of $n-1$ observations whereas the bootstrap consists of generating many data sets by random selection (with replacement) from the original data set. For each data set, the value of r is calculated; from this set of values, each technique is able to extract both an estimate of r and an estimate of the desired confidence interval.

Perhaps one of the most important computer-intensive methods is that of hypothesis testing using randomization, discussed in Chapter 5. This method can replace the standard tests, such as the χ^2 contingency test, when the assumptions of the test are not met. The basic idea of randomization testing is to randomly assign the observations to the "treatment" groups and calculate the test statistic: this process is repeated many (typically thousands) times and the probability under the null hypothesis of "no difference" estimated by the proportion of times the test statistic from the randomized data sets exceeded the test statistic from the observed data set. To illustrate the process, I shall relate an investigation into genetic variation among populations of shad, a commercially important fish species.

To investigate geographic variation among populations of shad, data on mitochondrial DNA variation were collected from 244 fish distributed over 14 rivers. This sample size represented, for the time, a very significant output of effort. Ten mitochondrial haplotypes were identified with 62% being of a single type. The result was that almost all cells had less than 5 data points (of the 140 cells, 66% had expected values less than 1.0 and only 9% had expected values greater than 5). Following Cochran's rules for the χ^2 test, it was necessary to combine cells. This meant combining the genotypes into two classes, the most common one and all others. The calculated χ^2 for the combined data set was 22.96, which just exceeded the critical value (22.36) at the 5% level. The estimated value of χ^2 for the uncombined data was 236.5, which was highly significant ($P < 0.001$) based on the χ^2 with 117 degrees of freedom. However, because of the very low frequencies within many cells, this result was suspect. Rather than combining cells and thus losing information, we (Roff and Bentzen 1989) used randomization (Chapter 5) to test if the observed χ^2 value was significantly larger than the expected value under the null hypothesis of homogeneity among the rivers. This analysis showed that the probability of obtaining a χ^2 value as large or larger than that observed for the ungrouped data was less than one in a thousand. Thus, rather than being merely marginally significant the variation among rivers was highly significant.

Most of the methods described in this book follow the frequentist school in asking "What is the probability of observing the set of n data $x_1, x_2, \ldots, x_n$ given the set of k parameters $\theta_1, \theta_2, \ldots, \theta_k$?" In Chapter 7 this position is reversed by the Bayesian perspective in which the question is asked "Given the set of n data $x_1, x_2, \ldots, x_n$, what is the probability of the set of k parameters $\theta_1, \theta_2, \ldots, \theta_k$?" This "reversal" of perspective is particularly important when management decisions are required. For example, suppose we wish to analyze the effect of a harvesting strategy on population growth: in this case the question we wish to ask is "Given some observed harvest, say x, what is the probability that the population rate of increase, say θ, is less than 1 (i.e., the population is declining)?" If this probability is high then it may be necessary to reduce the harvest rate. In Bayesian analysis, the primary focus is frequently on the probability statement about the parameter value. It can, however, also be used, as in the case of the James–Stein estimator, to improve on estimates. Bayesian analysis generally requires a computer-intensive approach to estimate the posterior distribution.

Why S-PLUS?

There are now numerous computer packages available for the statistical analysis of data, making available an array of techniques hitherto not possible except in some very particular circumstances. Many packages have some computer-intensive methods available, but most lack flexibility and hence are limited in use. Of the common packages, SAS and S-PLUS possess the breadth of programming capabilities necessary to do the analyses described in this book. I chose S-PLUS for three reasons. First, the language is structurally similar to programming languages with which the reader may already be familiar (e.g., BASIC and FORTRAN. It differs from these two in being object oriented). In writing the coding, I have attempted to keep a structure that could be transported to another language: this has meant in some cases making more use of looping than might be necessary in S-PLUS. While this increases the run time, I believe that it makes the coding more readable, an advantage that outweighs the minor increase in computing time. The second reason for selecting S-PLUS is that there is a version in the public domain, known as R. To quote the web site (http://www.r-project.org/), "R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R." The programs written in this book will, with few exceptions, run under R. The user interface is definitely better in

S-PLUS than R. My third reason for selecting S-PLUS is that students, at present, can obtain a free version for a limited period at http://elms03.e-academy.com/splus/.

Further reading

Although S-PLUS has a fairly steep learning curve there are several excellent text books available, my recommendations being:

Spector, P. (1994). *An Introduction to S and S-PLUS*. Belmont, California: Duxbury Press.

Krause, A. and Olson, M. (2002). *The Basics of S-PLUS*. New York: Springer.

Crawley, M. J. (2002). *Statistical Computing: An Introduction to Data Analysis using S-PLUS*. UK: Wiley and Sons.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. New York: Springer.

An overview of the language with respect to the programs used in this book is presented in the appendices.

2

Maximum likelihood

Introduction

Suppose that we have a model with a single parameter, θ, that predicts the outcome of an event that has some numerical value y. Further, suppose we have two choices for the parameter value, say θ_1 and θ_2, where θ_1 predicts that the numerical value of y will occur with a probability p_1 and θ_2 predicts that the numerical value of y w'ill occur with a probability p_2. Which of the two choices of θ is the better estimate of the true value of θ ? It seems reasonable to suppose that the parameter value that gave the highest probability of actually observing what was observed would be the one that is also closer to the true value of θ. For example, if p_1 equals 0.9 and p_2 equals 0.1, then we would select θ_1 over θ_2, because the model with θ_2 predicts that one is unlikely to observe y, whereas the model with θ_1 predicts that one is quite likely to observe y. We can extend this idea to many values of θ by writing our predictive model as a function of the parameter values, $\varphi(\theta_i)=p_i$, where i designates particular values of θ. More generally, we can dispense with the subscript and write $\varphi(\theta)=p$, thereby allowing θ to take on any value. By the **principle of maximum likelihood** we select the value of θ that has the highest associated probability, p.

The important element of maximum likelihood estimation (often contracted to MLE) is that there is a definable probability function that can be used to generate the **likelihood** of the observed event. The most frequently used probability functions are the **normal distribution** and the **binomial distribution**.

There are three areas to be considered:

(1) **Point estimation**. Given some statistical model with k unknown parameters $\theta_1,\theta_2,\ldots,\theta_k$ how do we use MLE to obtain estimates of these parameters, denoted as $\hat{\theta}_1,\hat{\theta}_2,\ldots,\hat{\theta}_k$?

(2) **Interval estimation**. Having the set of estimates $\hat{\theta}_1,\hat{\theta}_2,\ldots,\hat{\theta}_k$ is only marginally useful, because we have no idea whether the estimates are

likely to be close to or far from the true values. In conjunction with point estimation we must, therefore, also estimate a confidence region for the estimates, typically 95%.

(3) **Hypothesis testing**. In many instances, we are interested in testing hypotheses about the parameter values: for example, given two data sets we could test the hypothesis that they have a common mean. Maximum likelihood provides a mechanism to both compare different parameter values and to compare different statistical models.

Point estimation

Why the mean?

The underlying distribution of much of statistical estimation is the normal distribution (Figure 2.1). Under this distribution, the probability of observing a value, say x, is given by

$$\varphi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{2.1}$$

where $\varphi(x)$ is called the **probability density function of x**. This function is symmetrical and characterized by two parameters μ and σ. Anyone who has had a first course in statistics will recognize these two as the "mean" and the "standard deviation," respectively. The mean is a measure of central tendency, and the standard deviation a measure of spread of the distribution (Figure 2.1). We typically estimate the parameter μ as the **arithmetic average**

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.2}$$

where n is the number of observations and x_i is the ith observation. The "hat" over μ indicates that this is an estimate of the true value of μ: this is a general symbol for the estimate of a parameter, but in the case of the average, we frequently use the symbol $\bar{x}$.

There are actually three measures of central tendency, the arithmetic average, the **mode** (the most commonly occurring value), and the **median** (the value that divides the sample into two equal portions). Why should we use the arithmetic average as the estimate of μ? The use of the arithmetic average as the preferred estimate of μ can be justified by the fact that it is the maximum likelihood estimate of μ. Suppose we have a sample of n observations