1

Introduction

1.1 Regression and Classification

Quantifying the relationship between a response variable of interest and measurements taken on a set of possibly related observations is one of the most fundamental problems in statistics: a problem which conventionally is split into two distinct topics, regression and classification. Both focus on approximating the relationship between a set of input variables, or predictors, and an output variable, or response. In short, regression involves the case when the response variable is continuous, whereas classification is used when the response is categorical.

Figure 1.1 displays a dataset for which regression is appropriate. The data are taken from the Great Barrier Reef dataset described in detail in Poiner *et al.* (1997). The interest lies in determining the relationship between the longitude and the weight of fauna, given in terms of a score, captured at that longitude. By convention, we plot the response variable (score) on the y-axis with the predictor (longitude) on the x-axis. The aim of the regression analysis is to provide a good approximation to the true functional relationship between these two variables. The simplest such approximation widely used in data analyses assumes that the relationship is a straight line such that

$$\text{Score} = \beta_0 + \beta_1 \times \text{Longitude},$$

where β_0 and β_1 give the intercept and slope of the line. The line with β_0 and β_1 set so that they best fit the data, in least squares terms, is shown in Figure 1.1. However, this straight line seems to oversimplify the true relationship between the variables, and completely fails to model the particularly rapid drop in score for values of longitude greater than 143.3.

Figure 1.2 displays a dataset for which classification is appropriate. The dataset was originally described by Spyers-Ashby (1996) and was collected to determine how arm tremor measurements on an individual were related to the presence or absence of Parkinson's disease. The two axes represent measurements of two different types of arm tremor. We label these predictors X_1 and X_2 and write the set of all predictors as $\boldsymbol{X} = (X_1, X_2)$. Each sampled data point is represent by a cross if that particular patient had Parkinson's disease and a circle otherwise. The problem of interest is to split up the two-dimensional space in which the observed predictor values lie into regions, where the patients whose predictors lie in each one have a particular probability of

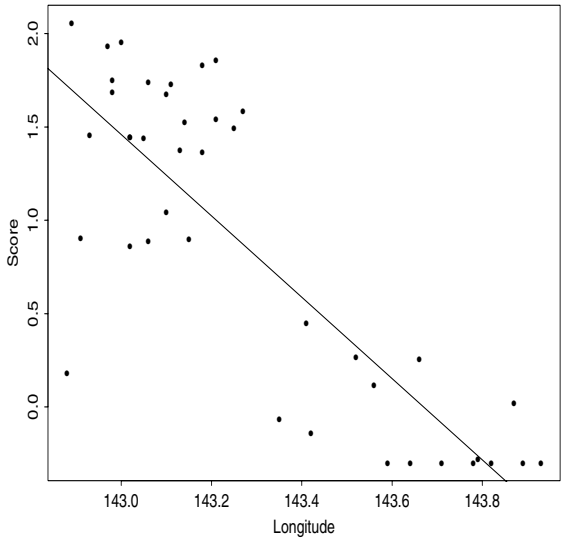

Figure 1.1 The Great Barrier Reef dataset together with the straight line which best fits the data in terms of sum of squared residuals.

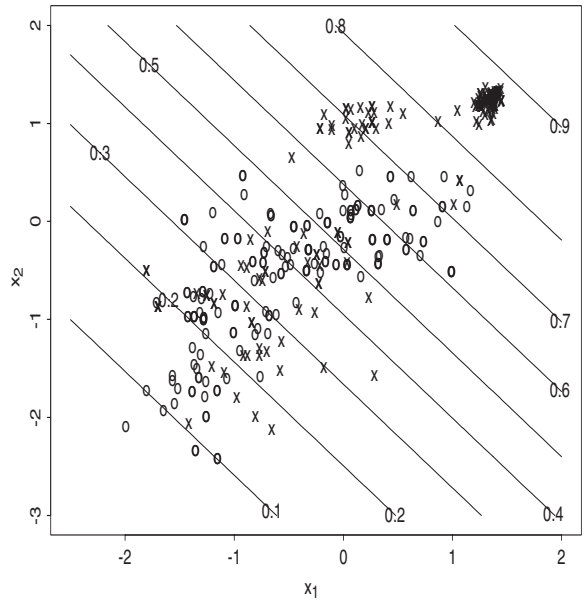

Figure 1.2 The arm tremor dataset with the crosses marking the location of individuals with Parkinson's disease and the circles representing those without it. The lines are the contours of probability found by fitting a straight line to $\text{logit}(p(y = 1 \mid \boldsymbol{x}))$.

having Parkinson's disease. One way to do this is to estimate a surface that captures how the probability of the presence of Parkinson's disease is related to $\boldsymbol{X}$.

A standard way to solve this classification problem is to use another model based on a straight line. That is, if $\boldsymbol{x} = (x_1, x_2)$ represents a general vector of measurements of the predictors, we take

$$\text{logit}\{p(y = 1 \mid \boldsymbol{x})\} = \log\left\{\frac{p(y = 1 \mid \boldsymbol{x})}{p(y = 0 \mid \boldsymbol{x})}\right\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

where the response y, equals one if the patient has Parkinson's disease, zero otherwise, and β_0, β_1 and β_2 are coefficients that define the probability surface. Using this underlying model, known as logistic regression, we find that the probability surface fitted is the simple two-dimensional plane represented by the contours in Figure 1.2. However, it appears that this plane does not fit the data adequately, especially failing to capture the near certainty of Parkinson's disease for those patients for which X_2 was greater than 0.5.

In both regression and classification problems the underlying focus is on determining some form of functional relationship between the predictors and responses. Despite their widespread use and popularity, we see from the two examples given above that linear models (i.e. those based on straight lines, or planes) will typically be too restrictive to accurately capture the actual underlying relationship. We need classes of more general, nonlinear models that are flexible enough to capture the complexity of the data.

The search for appropriate models for classification and regression was once the exclusive domain of the statistician but now it has been taken up by a wide range of other researchers, especially those in the more modern, computer-based research fields. Although linear methods still have uses in modern statistics, it has become apparent that more sophisticated approaches are required to model accurately a wide class of datasets. This need was recognised some time ago by various pioneering, far-sighted researchers (see, for example, Halpern 1973; Rosenblatt 1956; Whittaker 1923) who suggested models that would have advantages over the conventional linear model but, at the time, had no practical possibility of being implemented routinely using the computing methods and technology of their time. However, with the substantial increase in currently available computational power these, so-called nonlinear, methods have gained popularity.

The other consequence of readily available computing power is the ease with which data can now be stored and collected. A knowledge of data analysis, and how to extract information from data, is therefore increasingly seen as important. This has implications in a wide variety of areas, from the social sciences (responses to questionnaires in a marketing campaign), medical sciences (classification of EEG traces, detection of relevant gene sequences), financial markets (understanding the volatility of a stock, assigning optimal portfolios) and commerce (predicting the value a client might be to a company over their lifetime, determining a person who is a poor credit risk). It is now commonplace for researchers from electrical engineering, computer science and bioinformatics, amongst others, to be involved with problems of classification and

regression, which have also become known by a bewildering collection of new names such as pattern recognition, machine learning, intelligent data analysis, knowledge discovery, data mining and artificial intelligence. All of which essentially describe an approach to either regression of classification.

1.2 Bayesian Nonlinear Methods

Before any data analysis takes place, the true functional form between the input and output variables is almost always unknown (this is why we are trying to determine it!). We may have some clues about the relationship, or we may have some ideas about what we think our approximation should look like, but apart from this we may know very little.

Among the obvious questions which may influence how we proceed are the following.

1. What type of approximating functions should we use, or do we have available?
2. How do we know when we have found the 'best' approximation to the truth?
3. How can we incorporate any quantitative or qualitative knowledge we have about the relationship?

We shall address these questions in turn.

1.2.1 Approximating functions

There are a huge number of ways to approximate the truth and no one single specific approach can be uniformly better than any other in terms of predictive ability (see, for example, the discussion in Friedman (1993)). Commonly used approximating functions include linear and generalised linear models, smoothing splines, neural networks, Fourier bases, wavelets, decision trees and kernel smoothers. All of these provide explicit models for the relationship between the responses and predictors. Making inference about the relationship between y and $\boldsymbol{x}$ by defining such models is the focus of this book. We shall introduce a range of such models, describing their strengths and weaknesses, and demonstrate how use them to provide probabilistically sound inferences.

1.2.2 The 'best' model

A model is necessarily an approximation to the truth. In any real data analysis situation no single model will completely capture the true relationship between the inputs and outputs. We may think of the 'best' model among a set of alternatives as the one which most closely captures the true relationship for the particular purpose we have in mind. For example, if our aim is to predict well as judged by minimising the squared error between our predictions and the actual outcome, we shall see that a

'supermodel', corresponding to an average over all the specific models we might have considered, is the 'best'. This process of averaging over models means that the resulting approximation is drawn from a wider, more flexible, class of functions than is provided by any single specific model. We shall learn how to construct algorithms based on Markov chain Monte Carlo methods that identify good models and, where required, combine them in a sensible way.

1.2.3 Bayesian methods

We should clearly be able to improve the quality of the models we develop by incorporating whatever *a priori* qualitative or quantitative knowledge we have available. For example, knowledge of the degree of smoothness of a regression relationship might lead us to favour classes of models which provide smoothness in the second derivative (e.g. smoothing splines). Such approaches naturally lead us to Bayesian methods. These allow us to assign prior distributions to the parameters in the model which capture known qualitative and quantitative features, and then to update these priors in the light of the data, yielding a posterior distribution via Bayes' Theorem

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}.$$

The ability to include prior information in the model is not only an attractive pragmatic feature of the Bayesian approach, it is theoretically vital for guaranteeing coherent inferences. Our aim in this book is to provide the reader with an insight into the field of Bayesian nonlinear modelling, detailing the models and demonstrating the wealth of information we can glean from them through determination of appropriate posterior distributions. Further, Bayesian methods perform well in relation to other approaches and numerous papers (which we shall reference as appropriate in the later chapters) demonstrate their empirical effectiveness.

To whet the appetite, Figures 1.3 and 1.4 display the outcomes of Bayesian nonlinear modelling approaches to the problems considered earlier in Figures 1.1 and 1.2. Here we have used standard Bayesian models to estimate the true relationship between the response and predictors. We shall consider the form and implementation of such models later on in the book. For now, we content ourselves with noting how these nonlinear models capture the observed features in the data dramatically better than the linear models described earlier.

1.3 Outline of the Book

The next chapter underpins the understanding of the book as a whole. It begins by giving an outline of why we choose to adopt a Bayesian approach to predictive inference and then goes on to demonstrate how we can implement such an approach. This is described with reference to the regression problem which dominates the first few chapters. In particular, Chapter 2 provides familiarity with analytic posterior inference for the Bayesian linear model (Lindley and Smith 1972) and with simulation algorithms.

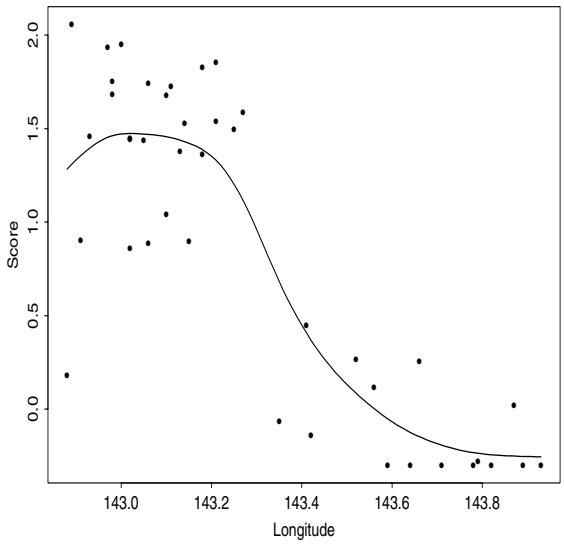

Figure 1.3 The Great Barrier Reef dataset together with the posterior mean Bayesian spline fit.

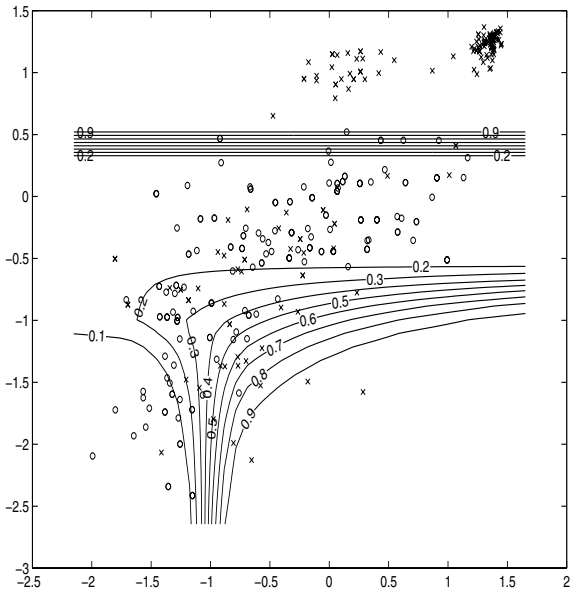

Figure 1.4 The arm tremor dataset together with the posterior mean Bayesian MARS fit to $\text{logit}(p(y = 1 \mid \boldsymbol{x}))$.

Chapter 3 goes through, in detail, the process of fitting simple Bayesian models, both analytically and via posterior simulation. We use the curve fitting problem to demonstrate simple applications of the methods with a view to aiding the understanding of readers new to the field.

Chapter 4 then introduces many models for fitting regression surfaces to data. It makes a distinction between models that are more suited to problems when there are many predictors and those that work well in fewer dimensions.

Bayesian generalised linear models form the focus of Chapter 5. Here, we show how to make posterior inference in these more difficult problems when analytic simplification is not available and simulation methods for high-dimensional posteriors are required. Generalised linear models provide a powerful toolkit for modelling in a wide variety of situations when the errors in the observations are not additive and normally distributed.

Chapter 6 covers Bayesian extensions to tree modelling. These are mainly used for the classification problem, although we include discussion of regression trees.

Chapter 7 extends the general method to partition models. These are similar to trees but allow for more flexible partitioning of the predictor space into regions where the data are assumed homogeneous.

Chapter 8 goes on to describe Bayesian nearest-neighbour modelling. Again, this method uses local structure in the data to lead us to a good way of making predictions.

Finally, Chapter 9 covers multiple response modelling when we wish to make inference about a vector of response variables given the predictors. This generalises the single response case, which is studied extensively throughout the rest of the book.