CHAPTER 1

Introduction

1.1 Background

In this book we consider a class of statistical models that is a natural generalization of classical linear models. *Generalized linear models* include as special cases, linear regression and analysis-of-variance models, logit and probit models for quantal responses, log-linear models and multinomial response models for counts and some commonly used models for survival data. It is shown that the above models share a number of properties, such as linearity, that can be exploited to good effect, and that there is a common method for computing parameter estimates. These common properties enable us to study generalized linear models as a single class, rather than as an unrelated collection of special topics.

Classical linear models and least squares began with the work of Gauss and Legendre (Stigler, 1981, 1986) who applied the method to astronomical data. Their data were usually measurements of continuous quantities such as the positions and magnitudes of the heavenly bodies and, at least in the astronomical investigations, the variability in the observations was largely the effect of measurement error. The Normal, or Gaussian, distribution was viewed as a mathematical construct developed to describe the properties of such errors; later in the nineteenth century the same distribution was used to describe the variation between individuals in a biological population in respect of a character such as height, an application quite different in kind from its use for describing measurement error, and leading to the numerous biological applications of linear models.

Gauss introduced the Normal distribution of errors as a device for describing variability, but he showed that many of the important properties of least-squares estimates depend not on Normality but on the assumptions of constant variance and indepen-

dence. A closely related property applies to all generalized linear models. In other words, although we make reference at various points to standard distributions such as the Normal, binomial, Poisson, exponential or gamma, the second-order properties of the parameter estimates are insensitive to the assumed distributional form: the second-order properties depend mainly on the assumed variance-to-mean relationship and on uncorrelatedness or independence. This is fortunate because, in applications, one can rarely be confident that all aspects of the assumed distributional form are correct.

Another strand in the history of statistics is the development of methods for dealing with discrete events rather than with continuously varying quantities. The enumeration of probabilities for various configurations in games of cards and dice was a matter of keen interest for gamblers in the eighteenth century. From their pioneering work grew methods for dealing with data in the form of counts of events. In the context of rare events, the basic distribution is that named after Poisson. This distribution has been applied to diverse kinds of events: a famous example concerns unfortunate soldiers kicked to death by Prussian horses (Bortkewitsch, 1898). The annual number of such incidents during the period 1875–1894 was observed to be consistent with the Poisson distribution having mean about 0.7 per corps per year. There is, however, some variation in this figure between corps and between years. Routine laboratory applications of the Poisson model include the monitoring of radioactive tracers by emission counts, counts of infective organisms as measured by the number of events observed on a slide under a microscope, and so on.

Closely related to the Poisson model are models for the analysis of counted data in the form of proportions or ratios of counts. The Bernoulli distribution is often suitable for modelling the presence or absence of disease in a patient, and the derived binomial distribution may be suitable as a model for the number of diseased patients in a fixed pool of patients under study. In medical and pharmaceutical trials it is usually required to study not primarily the incidence of a particular disease, but how the incidence is affected by factors such as age, social class, housing conditions, exposure to pollutants, and any treatment procedures under study. Generalized linear models permit us to study patterns of systematic variation in much the same way as ordinary linear models are used

to study the joint effects of treatments and covariates.

Some continuous measurements encountered in practice have non-Normal error distributions, and the class of generalized linear models includes distributions useful for the analysis of such data. The simplest examples are perhaps the exponential and gamma distributions, which are often useful for modelling positive data that have positively skewed distributions, such as occur in studies of survival times.

Before looking in more detail at the history of individual instances of generalized linear models, we make some general comments about statistical models and the part they play in the analysis of data, whether experimental or observational.

1.1.1 *The problem of looking at data*

Suppose we have a number of measurements or counts, together with some associated structural or contextual information, such as the order in which the data were collected, which measuring instruments were used, and other differences in the conditions under which the individual measurements were made. To interpret such data, we search for a pattern, for example that one measuring instrument has produced consistently higher readings than another. Such systematic effects are likely to be blurred by other variation of a more haphazard nature. The latter variation is usually described in statistical terms, no attempt being made to model or to predict the actual haphazard contribution to each observation.

Statistical models contain both elements, which we will call *systematic effects* and *random effects*. The value of a model is that often it suggests a simple summary of the data in terms of the major systematic effects together with a summary of the nature and magnitude of the unexplained or random variation. Such a reduction is certainly helpful, for the human mind, while it may be able to encompass say 10 numbers easily enough, finds 100 much more difficult, and will be quite defeated by 1000 unless some reducing process takes place.

Thus the problem of looking intelligently at data demands the formulation of patterns that are thought capable of describing succinctly not only the systematic variation in the data under study, but also for describing patterns in similar data that might

be collected by another investigator at another time and in another place.

1.1.2 *Theory as pattern*

We shall consider theories as generating patterns of numbers, which in some sense can replace the data, and can themselves be described in terms of a small number of quantities. These quantities are called *parameters*. By giving the parameters different values, specific patterns can be generated. Thus the very simple model

$$y = \alpha + \beta x,$$

connecting two quantities y and x via the parameter pair (α, β), defines a straight-line relationship between y and x. Suppose now that there is some causal relationship between x and y in which x is under control and affects y, and that y can be measured (ideally) without error. Then if we give x the values

$$x_1, x_2, \ldots, x_n,$$

y takes the values

$$\alpha + \beta x_1, \ \alpha + \beta x_2, \ldots, \alpha + \beta x_n$$

for the assigned values α and β. Clearly, if we know α and β we can reconstruct the values of y exactly from those of x, so that given $x_1, \ldots, x_n$, the pair (α, β) is an exact summary of $y_1, \ldots, y_n$ and we can move between the data and the parameters in either direction.

In practice, of course, we never measure the ys exactly, so that the relationship between y and x is only approximately linear. Despite this lack of exactness, we can still choose values of α and β, a and b say, that in some suitable sense best describe the now approximately linear relation between y and x. The quantities $a + bx_1$, $a + bx_2$, ..., $a + bx_n$, which we denote by $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ or $\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_n$, are the *theoretical* or *fitted values* generated by the model and the data. They do not reproduce the original data values $y_1, \ldots, y_n$ exactly. The pattern that they represent approximates the data values and can be summarized by the pair (a, b).

1.1.3 *Model fitting*

The fitting of a simple linear relationship between the ys and the xs requires us to choose from the set of all possible pairs of parameter values a particular pair (a, b) that makes the patterned set $\hat{y}_1, \ldots, \hat{y}_n$ closest to the observed data. In order to make this statement precise we need a measure of 'closeness' or, alternatively, of distance or discrepancy between the observed ys and the fitted $\hat{y}$s. Examples of such discrepancy functions include the L_1-norm

$$S_1(y, \hat{y}) = \sum |y_i - \hat{y}_i|$$

and the L_∞-norm

$$S_\infty(y, \hat{y}) = \max_i |y_i - \hat{y}_i|.$$

Classical least squares, however, chooses the more convenient L_2-norm or sum of squared deviations

$$S_2(y, \hat{y}) = \sum (y_i - \hat{y}_i)^2$$

as the measure of discrepancy. These discrepancy formulae have two implications. First, the straightforward summation of individual deviations, either $|y_i - \hat{y}_i|$ or $(y_i - \hat{y}_i)^2$, each depending on only one observation, implies that the observations are all made on the same physical scale and suggests that the observations are independent, or at least that they are in some sense exchangeable, so justifying an even-handed treatment of the components. Second, the use of arithmetic differences $y_i - \hat{y}_i$ implies that a given deviation carries the same weight irrespective of the value of $\hat{y}$. In statistical terminology, the appropriateness of L_p-norms as measures of discrepancy depends on stochastic independence and also on the assumption that the variance of each observation is independent of its mean value. Such assumptions, while common and often reasonable in practice, are by no means universally applicable.

The discrepancy functions just described can be justified in purely statistical terms. For instance, the classical least squares criterion arises if we regard the x-values as fixed or non-stochastic and the y-values are assumed to have the Normal, or Gaussian, distribution with mean μ, in which

$$\text{frequency of } y \text{ given } \mu \propto \exp\{-(y - \mu)^2/(2\sigma^2)\}, \qquad (1.1)$$

where μ is linearly related to x through the coefficients α and β. The scale factor σ, which is the standard deviation of y, describes the 'width' of the errors when measured about the mean value. In older statistical texts, 0.67σ is sometimes called the *probable error* in y.

We can look at the function (1.1) in two ways. If we regard it as a function of y for fixed μ, the function specifies the probability density of the observations. On the other hand, for a given observation y, we may regard (1.1) as a function of μ giving the relative plausibility of different values of μ for the particular value of y observed. It was this second interpretation, known as the likelihood function, whose value was first stressed by R.A. Fisher. We notice that the quantity $-2l$, where l is the logarithm of the likelihood function for a sample of n independent values, is equal to

$$\frac{1}{\sigma^2} \sum (y_i - \mu_i)^2.$$

In other words, apart from the factor σ^2, here assumed known, $-2l$ is identical to the sum-of-squares criterion. As μ varies, $-2l$ takes its minimum value at $\mu = \bar{y}$, the arithmetic mean of the observations. For a more complicated model in which μ varies in a systematic way from observation to observation, we define the closest set $\hat{\mu}$ to be that whose values maximize the likelihood or, equivalently, minimize $-2l$. More generally, we can extend our interest beyond the single point that minimizes $-2l$, to the shape of the likelihood surface in the neighbourhood of the minimum. This shape tells us, in Fisher's terminology, how much information concerning the parameters there is in the data.

Appendix A gives a concise summary of the principal properties of likelihood functions.

Reverting to our example of a linear relationship, we can plot on a graph with axes α and β, the contours of equal discrepancy $-2l$ for the given data y. In this particular instance, $-2l$ is a quadratic function of (α, β) and hence the contours are ellipses, similar in shape and orientation, with the maximum-likelihood estimate (a, b) situated at their centre. The information in the data on the parameters (α, β) is given by the curvature matrix or Hessian matrix of the quadratic. If the axes of the ellipses are not aligned with the (α, β) axes, the estimates are said to be correlated. The information is greatest in the direction for which

the curvature is greatest (see Fig. 3.8). In certain circumstances, the form of the information surface can be determined before an experiment is carried out. In other words, the precision achievable by a given experiment can sometimes be determined in advance and such information can be used to compute the experimental resources needed to estimate parameters with a required accuracy. A similar analysis will also show the parameter combinations that are badly estimated by the data and this information is often valuable in choosing among possible experimental designs. Alas, such calculations are not made nearly often enough!

1.1.4 *What is a good model?*

To summarize, we aim in model fitting to replace our data $\mathbf{y}$ with a set of fitted values $\hat{\boldsymbol{\mu}}$ derived from a model. These fitted values are chosen to minimize some criterion such as the sum-of-squares discrepancy measure $\sum_i (y_i - \hat{\mu}_i)^2$.

At first sight it might seem as though a good model is one that fits the observed data very well, i.e. that makes $\hat{\boldsymbol{\mu}}$ very close to $\mathbf{y}$. However, by including a sufficient number of parameters in our model, we can make the fit as close as we please, and indeed by using as many parameters as observations we can make the fit perfect. In so doing, however, we have achieved no reduction in complexity – produced no simple theoretical pattern for the ragged data. Thus simplicity, represented by parsimony of parameters, is also a desirable feature of any model; we do not include parameters that we do not need. Not only does a parsimonious model enable the research worker or data analyst to think about his data, but one that is substantially correct gives better predictions than one that includes unnecessary extra parameters.

An important property of a model is its *scope*, i.e. the range of conditions over which it gives good predictions. Scope is hard to formalize, but easy to recognize, and intuitively it is clear that scope and parsimony are to some extent related. If a model is made to fit very closely to a particular set of data, it will not be able to encompass the inevitable changes that will be found necessary when another set of data relating to the same phenomenon is collected. Both scope and parsimony are related to *parameter invariance*, that is to parameter values that either do not change as some external condition changes or that change in a predictable way.

Modelling in science remains, partly at least, an art. Some principles do exist, however, to guide the modeller. A first, though at first sight, not a very helpful principle, is that all models are wrong; some, though, are more useful than others and we should seek those. At the same time we must recognize that eternal truth is not within our grasp. A second principle (which applies also to artists!) is not to fall in love with one model to the exclusion of alternatives. Data will often point with almost equal emphasis at several possible models and it is important that the statistician recognize and accept this. A third principle recommends thorough checks on the fit of a model to the data, for example by using residuals and other statistics derived from the fit to look for outlying observations and so on. Such diagnostic procedures are not yet fully formalized, and perhaps never will be. Some imagination or introspection is required here in order to determine the aspects of the model that are most important and most suspect. Box (1980) has attempted a formalization of the dual processes of model fitting and model criticism.

1.2 The origins of generalized linear models

1.2.1 *Terminology*

This section deals with the origin of generalized linear models, describing various special cases that are now included in the class in approximately their chronological order of development. First we need to establish some terminology: data will be represented by a *data matrix*, a two-dimensional array in which the rows are indexed by experimental or survey units. In this context, units are the physical items on which observations are made, for example plots in an agricultural field trial, patients in a medical survey or clinical trial, quadrats in an ecological study and so on. The columns of the data matrix are the *variates* such as measurements or yields, treatments, varieties, plot characteristics, patient's age, weight, sex and so on. Some of the variates are regarded as responses or dependent variates, whose values are believed to be affected by the explanatory variables or covariates. The latter are unfortunately sometimes called independent variates. Tukey (1962) uses the terms *response* and *stimulus* to make this important distinction. Covariates may be quantitative or qualitative. Quantitative

variates take on numerical values: qualitative variates take on non-numerical values or *levels* from a finite set of values or labels. We shall refer to qualitative covariates as *factors*: such covariates include classification variables such as blocks, that serve to group the experimental units, and treatment indicators that may in principle be assigned by the experimenter to any of the experimental units. Dependent variables may be continuous, or discrete (in the form of counts), or they may take the form of factors, where the response is one of a finite set of possible values or classes. For examples of the latter type of response, see Chapter 5.

1.2.2 *Classical linear models*

In matrix notation the set of observations is denoted by a column vector of observations $\mathbf{y} = \{y_1, \ldots, y_n\}^T$. The set of covariates or explanatory variables is arranged as an $n \times p$ matrix $\mathbf{X}$. Each row of $\mathbf{X}$ refers to a different unit or observation, and each column to a different covariate. Associated with each covariate is a coefficient or parameter, usually unknown. The set of parameters is a vector of dimension $\mathbf{p}$, usually denoted by $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_p\}^T$. For any given value of $\boldsymbol{\beta}$, we can define a vector of residuals

$$\mathbf{e}(\boldsymbol{\beta}) = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}.$$

In 1805 Legendre first proposed estimating the βs by minimizing $\mathbf{e}^T\mathbf{e} = \sum_i e_i^2$ over values of $\boldsymbol{\beta}$. [Note that both Legendre and Gauss defined the residuals with opposite sign to that in current use, i.e. by $\mathbf{X}\boldsymbol{\beta} - \mathbf{y}$.] In 1809, in a text on astronomy, Gauss introduced the Normal distribution with zero mean and constant variance for the errors. Later in his *Theoria Combinationis* in 1823, he abandoned the Normal distribution in favour of the weaker assumption of constancy of variance alone. He showed that the estimates of β obtained by minimizing the least-squares criterion have minimum variance among the class of unbiased estimates. The extension of this weaker assumption to generalized linear models was given by Wedderburn (1974) using the concept of quasi-likelihood. This extension is discussed in Chapter 9.

Most astronomical data analysed using least squares were of the observational kind, i.e. they arose from observing a system, such as the Solar System, without perturbing it or experimenting with

it. The development of the theory of experimental design gave a new stimulus to linear models and is very much associated with R.A. Fisher and his co-workers.

1.2.3 *R.A. Fisher and the design of experiments*

In 1919, Fisher began work at the agricultural research station at Rothamsted. Within 10 years, he had, among other achievements, laid the foundations of the design of experiments, a subject that was substantially developed by his successor, F. Yates, and others at Rothamsted. In particular, Fisher stressed the value of factorial experiments in which several experimental and classification factors are studied simultaneously instead of being varied one at a time. Thus, with two factors under study, each having two levels, the one-at-a-time design (a) in Fig. 1.1 was replaced with the factorial design (b). In the latter case, all combinations of the two factors are studied.

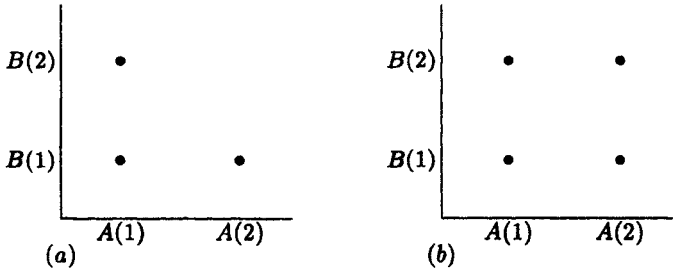

Fig. 1.1. (*a*) *Design for two factors, changing levels one at a time*; (*b*) *factorial design.*

The use of factorial designs increases the information per observation. Their analysis involves factorial models in which the yield or response is considered to be expressible as the sum of effects due to individual factors acting one at a time (main effects), effects due to pairs of factors above and beyond their separate contributions (two-factor interactions), and so on. Thus, the term 'factorial' refers to a particular class of design matrices or model matrices $\mathbf{X}$. In the case of factorial models, $\mathbf{X}$ is a matrix of zeros and ones only

and is sometimes called an *incidence matrix* for that particular design. Factorial models are often called analysis-of-variance models to be distinguished and contrasted with linear regression models for which the covariates are continuous and not restricted to the values zero and one. We shall use the terms factorial design and linear regression model as descriptors for different kinds of model matrices $\mathbf{X}$. However, we shall try not to make a major distinction, but rather to unify the ideas embodied in these two extremes. For instance, we shall include terms in which the slopes defined by regression coefficients are allowed to vary with the level of various indexing factors.

Fisher's influence on the development of generalized linear models extends well beyond models for factorial experiments and includes special models for the analysis of certain kinds of counts and proportions. We now consider some of the non-classical cases of generalized linear models that arose in the period 1922–1960.

1.2.4 *Dilution assay*

The original paper in this context is Fisher (1922), especially section 12.3. A solution containing an infective organism is progressively diluted and, at each dilution, a number of agar plates are 'streaked'. On some of these plates the infective organism produces a growth on the medium: the rest of the plates remain sterile. From the number of sterile plates observed at each dilution, an estimate of the concentration of infective organisms in the original solution is made.

Assuming for simplicity that dilutions are made in powers of two, the argument runs as follows. After x dilutions, the number of infective organisms, ρ_x, per unit volume is

$$\rho_x = \rho_0/2^x, \qquad x = 0, 1, \ldots$$

where ρ_0, which we wish to estimate, is the density of infective organisms in the original solution. Assuming that each agar plate is streaked using a volume, v, of solution, the expected number of organisms on any plate is $\rho_x v$ and, under suitable mixing conditions, the actual number of organisms follows the Poisson distribution with this parameter. Thus the probability that a plate is infected is just $\pi_x = 1 - \exp\{-\rho_x v\}$, the complement of the first

term in the Poisson series. It follows that at dilution x

$$\log\bigl(-\log(1-\pi_x)\bigr) = \log v + \log \rho_x = \log v + \log \rho_0 - x \log 2. \quad (1.2)$$

If at dilution x we have r infected plates out of m, the observed proportion of infected plates $y = r/m$ may be regarded as the realization of a random variable Y satisfying

$$E(Y \mid x) = \pi_x.$$

However, this time it is not the mean of Y that bears a linear relationship to x, but instead the transformation

$$\eta = \log\bigl(-\log(1-\pi_x)\bigr)$$

known as the complementary log log transformation. To make the linear relationship explicit, we write

$$\eta = \alpha + \beta x,$$

where $\alpha = \log v + \log \rho_0$ and $\beta = -\log 2$.

In this example, we have a slope β that is known a priori, an intercept α that bears a simple relationship to the quantity ρ_0 that we wish to estimate, and it is not the mean of Y that is linear in x but a known function of $E(Y)$, in this case the complementary log log function. For this dilution assay problem, Fisher showed how to apply maximum likelihood to obtain an estimator. He also used his concept of information to show that another estimator, based solely on the number of sterile plates over all dilutions, contained 87.7% of the information of the maximum-likelihood estimator. Nowadays, we can use a computer to calculate the maximum-likelihood estimate with minimal effort: alternative simpler estimators may still retain a certain appeal, but computational effort is no longer an important criterion for selection. The model just described is a particular instance of a generalized linear model. Fisher's estimation procedure is an early non-trivial application of maximum likelihood to a problem for which no closed-form solution exists.

1.2.5 *Probit analysis*

The technique known as probit analysis arose in connection with bioassay, and the modern method of analysis dates from Bliss (1935). In toxicology experiments, for example, test animals or insects are divided into sets, usually, but not necessarily of equal sizes. Each set of animals is subjected to a known level x of a toxin, or in other contexts, of a stimulant or dose. The dose varies from set to set but is assumed to be uniform within each set. For the jth set, the number y_j surviving out of the original m_j is recorded, together with the dose x_j administered. It is required to model the proportion surviving, π_x, at dose x as a function of x, which is usually measured in logarithmic units. The probit model is

$$\pi_x = \Phi(\alpha + \beta x), \tag{1.3}$$

where $\Phi(\cdot)$ is the cumulative Normal distribution function, and α and β are unknown parameters to be estimated. This model has the virtue that it respects the property that π_x is a probability and hence must lie between 0 and 1 for all values of x and for all parameter values. For this reason alone, it is not normally sensible to contemplate linear models for probabilities. Note also that if $\beta > 0$, the survival probability is monotonely increasing in the applied dose: otherwise, if $\beta < 0$, the survival probability is monotonely decreasing in the dose.

Because of the occurrence of $y_j = 0$ or $y_j = m_j$ at certain high or low doses, it is not feasible to take $\Phi^{-1}(y_j/m_j)$ as the response variable in order to make the model approximately linear in the parameters. Infinite values can be avoided by using a modified empirical transformation such as $\Phi^{-1}\{(y_j + \frac{1}{2})/(m_j + 1)\}$, but the choice of modification is to a large extent arbitrary.

Linearity in the parameters is an important aspect of the probit model (1.3). Note however, that the linearity does not occur directly in the expression for $E(Y)$ in terms of x nor in $E\{\Phi^{-1}(Y/m)\}$ as a function of x. The linearity in question arises in the expression for $\Phi^{-1}(\pi_x)$, the transformed theoretical proportion surviving at dose x. This is the same sense in which the model for the dilution assay (1.2), is linear, although the transformations required to achieve linearity are different in the two examples.

The probit model exhibits one further feature that distinguishes it from the usual Normal-theory model, namely that the variance of

the observed proportion surviving Y/m, is not constant but varies in a systematic fashion as a function of $\pi = E(Y/m)$. Specifically, under the usual binomial assumption, Y/m has variance $\pi(1-\pi)/m$, which has a maximum at $\pi = 0.5$. Generalized linear models accommodate unequal variances through the introduction of variance functions that may depend on the mean value through a known function of the mean.

1.2.6 *Logit models for proportions*

Dyke and Patterson (1952) published an analysis of some cross-classified survey data concerning the proportion of subjects who have a good knowledge of cancer. The recorded explanatory variables were exposures to various information sources, newspapers, radio, solid reading, lectures. All combinations of these explanatory variables occurred in the sample, though some combinations occurred much more frequently than others. A factorial model was postulated in which the logit or log odds of success, $\log\{\pi/(1-\pi)\}$ is expressed linearly as a combination of the four information sources and interactions among them. Success in this context is interpreted as synonymous with 'good knowledge of cancer'. Dyke and Patterson were successful in finding a suitable model of this kind, though the fitting, which was done manually, took several days. Similar computations done today take only a few seconds.

Dyke and Patterson's application of the linear logistic model was to survey data. Linear logistic models had earlier been used in the context of bioassay experiments (see, for example, Berkson, 1944, 1951).

1.2.7 *Log-linear models for counts*

The analysis of counted data has recently given rise to an extensive literature mainly based on the idea of a log–linear model. In such a model, the two components of the classical linear model are replaced in the following way:

	Classical linear model	*Log-linear model*
Systematic effects	additive	multiplicative
Nominal error distribution	Normal	Poisson

The Poisson distribution is the nominal distribution for counted data in much the same way that the Normal distribution is the bench-mark for continuous data. Such counts are assumed to take the values $0, 1, 2, \ldots$ without an upper limit. The Poisson distribution has only one adjustable parameter, namely the mean μ, which must be positive. Thus the mean alone determines the distribution entirely. By contrast, the Normal distribution has two adjustable parameters, namely the mean and variance, so that the mean alone does not determine the distribution completely.

Since the Poisson mean is required to be positive, an additive model for μ is normally considered to be unsatisfactory. All linear combinations $\eta = \sum \beta_j x_j$ become negative for certain parameter combinations and covariate combinations. Hence, although $\mu = \sum \beta_j x_j$ may be found to be adequate over the range of the data, it is often scientifically dubious and logically unsatisfactory for extrapolation. In the model with multiplicative effects, we set $\mu = \exp(\eta)$ and η rather than μ obeys the linear model. This construction ensures that μ remains positive for all η and hence positive for all parameter and covariate combinations.

The ideas taken from factorial design and regression models carry over directly to log-linear models except that the effects or parameters of interest are contrasts of log frequencies. For the purpose of explanation and exposition, such contrasts are usually best back-transformed to the original frequency scale and expressed as multiplicative effects.

It often happens with counted data that one of the classifying variables, rather than the counts themselves, is best regarded as the response. In this case, the aim usually is to model the way in which the remaining explanatory variables affect the relative proportions falling in the various categories of response. Normally, we would not aim to model the total numbers of respondents as a function of the response variables, but only the way in which these respondents are distributed across the k response categories. In this context, it is natural to consider modelling the errors by the multinomial distribution, which can be regarded as a set of k independent Poisson random variables subject to the constraint that their total is fixed. The relationship between Poisson log-linear models and multinomial response models is discussed further in section 6.4. It is possible, though not always desirable, to handle multinomial response models by using a suitably augmented log-

linear model.

Routine use of log-linear models has had a major impact on the analysis of counted data, particularly in the social sciences. Both log-linear and multinomial response models are special cases of generalized linear models and are discussed further in Chapters 4 to 6.

1.2.8 *Inverse polynomials*

Polynomials are widely used in biological and other work for expressing the shape of response curves, growth curves and so on. The most obvious advantage of polynomials is that they provide an infinite sequence of easily-fitted curves. The main disadvantage is that in most scientific work, the response is bounded, whereas polynomials, when extrapolated, become unbounded. Moreover, responses are often required to be positive, whereas polynomials are liable to become negative in certain ranges. In many applications, for example in the context of growth curves, it is common to find that the response approaches a plateau or asymptote as the stimulus increases. Polynomials do not have asymptotes and hence cannot be consistent with this known form of limiting behaviour.

Hyperbolic response curves of the form

$$x/y = \alpha + \beta x,$$

which do have asymptotes, have been used in a number of contexts such as the Michaelis–Menten equations of enzyme kinetics. The inverse polynomials introduced by Nelder (1966) extend this class of response curve to include inverse quadratic and higher-order inverse polynomial terms. More than one covariate can be included. Details are discussed in Chapter 8, which deals also with the case of continuous response variables in which the coefficient of variation rather than the variance is assumed constant over all observations.

1.2.9 *Survival data*

In the past 15 years or so, great interest has developed in models for survival in the context of clinical and surgical treatments. Similar problems, though with a rather different emphasis, occur in the analysis of failure times of manufactured components. In

medical experiments particularly, the data usually contain censored individuals. Such individuals are known to have survived up to a given time but their subsequent progress is not recorded either because the trial ends before the outcome is known or because the patient can no longer be contacted. In medical trials, such patients are said to be censored or 'lost to follow-up'. Aitkin and Clayton (1980) and Whitehead (1980) have shown how the analysis of censored survival data can be moulded into the framework of generalized linear models. This transformation is simplest to achieve when there are no time-dependent covariates: in more complicated cases, the computations are best done with the assistance of specially written computer programs.

1.3 Scope of the rest of the book

In Chapter 2, we outline the component processes in model fitting, describe the components of a generalized linear model, the definitions of goodness-of-fit of a model to data, a method for fitting generalized linear models and some asymptotic theory concerning the statistical properties of the parameter estimates. Chapter 3 deals with classical models for continuous data, in which the systematic effects are described by a linear model and the error variances are assumed constant and independent of the mean response. Many of the ideas introduced in this classical context carry over with little or no change to the whole class of generalized linear models. In particular, descriptive terms and model formulae that are used to specify design or model matrices are equally appropriate for all generalized linear models. The three subsequent chapters describe models that are relevant for data in the form of counts or proportions. Random variation in this context is often suitably described by the Poisson, binomial or multinomial distributions: systematic effects are assumed to be additive on a suitably chosen scale. The scale is chosen in such a way that the fitted frequencies are positive and the fitted proportions lie between 0 and 1. Where response categories are ordered, models are chosen that respect this order. Chapter 8 introduces generalized linear models for continuous data where, instead of assuming that the variance is constant, it is assumed instead that the coefficient of variation, σ/μ, is constant. In other words, the larger the mean response, the

greater the variability in the response. Examples are drawn from meteorology and the insurance industry.

A major extension of the applicability of generalized linear models was made by Wedderburn (1974) when he introduced the idea of quasi-likelihood. Wedderburn showed that often it is not necessary to make specific detailed assumptions regarding the random variation. Instead, many of the more useful properties of parameter estimates, derived initially from likelihood theory, can be justified on the grounds of weaker assumptions concerning independence and second moments alone. Specifically, it is necessary to know how the variance of each observation changes with its mean value but it is not necessary to specify the distribution in its entirety. Models based on quasi-likelihood are introduced informally, where appropriate, in earlier chapters, while in Chapter 9, a more systematic account is given.

Medical research is much concerned with the analysis of survival times of individual patients. Different patients have different histories and are assigned to one of several treatments. It is required to know how the survival time is affected by the treatment given, making such allowance as may be required for the differing histories of the various patients. There is a close connection between the analysis of survival times and the analysis of, say, 5-year survival rates. The latter problem falls under the rubric of discrete data or binary data. Such connections are exploited in Chapter 13 in order to handle survival times in the context of generalized linear models.

Frequently it happens that a model would fall into the linear category if one or two parameters that enter the model in a non-linear way were known a priori. Such models are sometimes said to be *conditionally linear*. A number of extensions to conditionally linear models are discussed in Chapter 11.

Chapter 10 discusses the simultaneous modelling of the mean and dispersion parameters as functions of the covariates, which are typically process settings in an industrial context.

Chapter 14 gives a brief introduction to problems in which there are several variance components, or dispersion components, associated with various sub-groups or populations. In this context it is usually unrealistic to assume that the observations are all independent.

1.4 Bibliographic notes

The historical development of linear models and least squares from Gauss and Legendre to Fisher has previously been sketched. For further historical information concerning the development of probability and statistics up to the beginning of the twentieth century, see the book by Stigler (1986).

The term 'generalized linear model' is due to Nelder and Wedderburn (1972), who showed how linearity could be exploited to unify apparently diverse statistical techniques.

For an elementary introduction to the subject, see the book by Dobson (1983).

1.5 Further results and exercises 1

1.1 Suppose that $Y_1, \ldots, Y_n$ are independent and satisfy the linear model

$$\mu_i = E(Y_i) = \sum_{j=1}^{p} x_{ij}\beta_j$$

for given covariates x_{ij} and unknown parameters $\boldsymbol{\beta}$. Show that if Y_i has the Laplace distribution or double exponential distribution

$$f_{Y_i}(y_i; \mu_i, \sigma) = \frac{1}{2\sigma} \exp\{-|y_i - \mu_i|/\sigma\}$$

then the maximum-likelihood estimate of $\boldsymbol{\beta}$ is obtained by minimizing the L_1-norm

$$S_1(y, \hat{y}) = \sum |y_i - \hat{y}_i|$$

over values of $\hat{y}$ satisfying the linear model.

1.2 In the notation of the previous exercise, show that if Y_i is uniformly distributed over the range $\mu_i \pm \sigma$, maximum-likelihood estimates are obtained by minimizing the L_∞-norm,

$$S_\infty(y, \hat{y}) = \max_i |y_i - \hat{y}_i|.$$

Show also that linearity of the model is irrelevant for the conclusions in both cases.

1.3 Justify the conclusion of the previous two exercises that the estimates of the regression parameters are unaffected by the value of σ in both cases. Show that the conclusion does not extend to either of the following distributions even though, in both cases, σ is a scale factor.

$$f_Y(y;\mu,\sigma) = \frac{\exp\{(y-\mu)/\sigma\}}{\sigma\{1+\exp\{(y-\mu)/\sigma\}\}^2}$$

$$f_Y(y;\mu,\sigma) = \frac{1}{\pi\sigma\{1+(y-\mu)^2/\sigma^2\}}$$

1.4 Find the maximum-likelihood estimate of σ for each model. Show that, for the models in Exercises 1.1 and 1.2, $\hat{\sigma}$ is a function of the minimized norm.

1.5 Suppose that X_1, X_2 are independent unit exponential random variables. Show that the distribution of $Y = \log(X_1/X_2)$ is

$$f_Y(y) = \frac{\exp(y)}{(1+\exp(y))^2}$$

for $-\infty < y < \infty$.

Find the distribution of Y if the Xs have the Weibull density

$$f_X(x) = \tau\rho(\rho x)^{\tau-1}\exp\{-(\rho x)^\tau\}, \qquad \rho,\tau,x > 0.$$

[Hint: first find the distribution of $(\rho X)^\tau$.]

1.6 The probable error, τ, of a random variable Y may be defined by

$$\text{pr}(|Y-M| \geq \tau) = 0.5,$$

where M is the median of Y. Find the probable errors of

1. the exponential distribution;
2. the double exponential distribution (Exercise 1.1);
3. the logistic distribution (Exercise 1.3);
4. the Cauchy distribution (Exercise 1.3);
5. the Normal distribution.

Discuss briefly the differences between the probable error and the inter-quartile range.

The historical definition of probable error appears to be vague. Some authors take M to be the mean; others take τ to be a multiple (0.67) of the standard deviation.