Applying Generalized Linear Models

James K. Lindsey

Springer

Springer Texts in Statistics

Advisors:
George Casella Stephen Fienberg Ingram Olkin

Springer
New York
Berlin
Heidelberg
Barcelona
Hong Kong
London
Milan
Paris
Singapore
Tokyo

Springer Texts in Statistics

Alfred: Elements of Statistics for the Life and Social Sciences
Berger: An Introduction to Probability and Stochastic Processes
Bilodeau and Brenner: Theory of Multivariate Statistics
Blom: Probability and Statistics: Theory and Applications
Brockwell and Davis: An Introduction to Times Series and Forecasting
Chow and Teicher: Probability Theory: Independence, Interchangeability, Martingales, Third Edition
Christensen: Plane Answers to Complex Questions: The Theory of Linear Models, Second Edition
Christensen: Linear Models for Multivariate, Time Series, and Spatial Data
Christensen: Log-Linear Models and Logistic Regression, Second Edition
Creighton: A First Course in Probability Models and Statistical Inference
Dean and Voss: Design and Analysis of Experiments
du Toit, Steyn, and Stumpf: Graphical Exploratory Data Analysis
Durrett: Essentials of Stochastic Processes
Edwards: Introduction to Graphical Modelling
Finkelstein and Levin: Statistics for Lawyers
Flury: A First Course in Multivariate Statistics
Jobson: Applied Multivariate Data Analysis, Volume I: Regression and Experimental Design
Jobson: Applied Multivariate Data Analysis, Volume II: Categorical and Multivariate Methods
Kalbfleisch: Probability and Statistical Inference, Volume I: Probability, Second Edition
Kalbfleisch: Probability and Statistical Inference, Volume II: Statistical Inference, Second Edition
Karr: Probability
Keyfitz: Applied Mathematical Demography, Second Edition
Kiefer: Introduction to Statistical Inference
Kokoska and Nevison: Statistical Tables and Formulae
Kulkarni: Modeling, Analysis, Design, and Control of Stochastic Systems
Lehmann: Elements of Large-Sample Theory
Lehmann: Testing Statistical Hypotheses, Second Edition
Lehmann and Casella: Theory of Point Estimation, Second Edition
Lindman: Analysis of Variance in Experimental Design
Lindsey: Applying Generalized Linear Models
Madansky: Prescriptions for Working Statisticians
McPherson: Statistics in Scientific Investigation: Its Basis, Application, and Interpretation
Mueller: Basic Principles of Structural Equation Modeling: An Introduction to LISREL and EQS

(continued after index)

James K. Lindsey

Applying Generalized Linear Models

With 35 Illustrations

Springer

James K. Lindsey
Department of Biostatistics
Limburgs Universitair Centrum
3590 Diepenbeek
Belgium

Library of Congress Cataloging-in-Publication Data
Lindsey, James K.
Applying generalized linear models / J.K. Lindsey
p. cm. — (Springer texts in statistics)
Includes bibliographical references (p. –) and index.
ISBN 0-387-98218-3 (hardcover : alk. paper)
1. Linear models (Statistics) I. Title. II. Series
QA279.L594 1997 97-6926
519.5'3—dc21

ISBN 0-387-98218-3 Springer-Verlag New York Berlin Heidelberg SPIN 10761357

Preface

Generalized linear models provide a unified approach to many of the most common statistical procedures used in applied statistics. They have applications in disciplines as widely varied as agriculture, demography, ecology, economics, education, engineering, environmental studies and pollution, geography, geology, history, medicine, political science, psychology, and sociology, all of which are represented in this text.

In the years since the term was first introduced by Nelder and Wedderburn in 1972, generalized linear models have slowly become well known and widely used. Nevertheless, introductory statistics textbooks, and courses, still most often concentrate on the normal linear model, just as they did in the 1950s, as if nothing had happened in statistics in between. For students who will only receive one statistics course in their career, this is especially disastrous, because they will have a very restricted view of the possible utility of statistics in their chosen field of work. The present text, being fairly advanced, is not meant to fill that gap; see, rather, Lindsey (1995a).

Thus, throughout much of the history of statistics, statistical modelling centred around this normal linear model. Books on this subject abound. More recently, log linear and logistic models for discrete, categorical data have become common under the impetus of applications in the social sciences and medicine. A third area, models for survival data, also became a growth industry, although not always so closely related to generalized linear models. In contrast, relatively few books on generalized linear models, as such, are available. Perhaps the explanation is that normal and discrete, as well as survival, data continue to be the major fields of application. Thus, many students, even in relatively advanced statistics courses, do not have

an overview whereby they can see that these three areas, linear normal, categorical, and survival models, have much in common. Filling this gap is one goal of this book.

The introduction of the idea of generalized linear models in the early 1970s had a major impact on the way applied statistics is carried out. In the beginning, their use was primarily restricted to fairly advanced statisticians because the only explanatory material and software available were addressed to them. Anyone who used the first versions of GLIM will never forget the manual which began with pages of statistical formulae, before actually showing what the program was meant to do or how to use it.

One had to wait up to twenty years for generalized linear modelling procedures to be made more widely available in computer packages such as Genstat, Lisp-Stat, R, S-Plus, or SAS. Ironically, this is at a time when such an approach is decidedly outdated, not in the sense that it is no longer useful, but in its limiting restrictions as compared to what statistical models are needed and possible with modern computing power. What are now required, and feasible, are nonlinear models with dependence structures among observations. However, a unified approach to such models is only slowly developing and the accompanying software has yet to be put forth. The reader will find some hints in the last chapter of this book.

One of the most important accomplishments of generalized linear models has been to promote the central role of the likelihood function in inference. Many statistical techniques are proposed in the journals every year without the user being able to judge which are really suitable for a given data set. Most *ad hoc* measures, such as mean squared error, distinctly favour the symmetry and constant variance of the normal distribution. However, statistical models, which by definition provide a means of calculating the probability of the observed data, *can* be directly compared and judged: a model is preferable, or more likely, if it makes the observed data more probable (Lindsey, 1996b). This direct likelihood inference approach will be used throughout, although some aspects of competing methods are outlined in an appendix.

A number of central themes run through the book:

- the vast majority of statistical problems can be formulated, in a unified way, as regression models;
- any statistical models, for the same data, can be compared (whether nested or not) directly through the likelihood function, perhaps, with the aid of some model selection criterion such as the AIC;
- almost all phenomena are dynamic (stochastic) processes and, with modern computing power, appropriate models should be constructed;
- many so called "semi-" and "nonparametric" models (although not nonparametric inference procedures) are ordinary (often saturated)

generalized linear models involving factor variables; for inferences, one must condition on the observed data, as with the likelihood function.

Several important and well-known books on generalized linear models are available (Aitkin *et al.*, 1989; McCullagh and Nelder, 1989; Dobson, 1990; Fahrmeir and Tutz, 1994); the present book is intended to be complementary to them.

For this text, the reader is assumed to have knowledge of basic statistical principles, whether from a Bayesian, frequentist, or direct likelihood point of view, being familiar at least with the analysis of the simpler normal linear models, regression and ANOVA. The last chapter requires a considerably higher level of sophistication than the others.

This is a book about statistical *modelling*, not statistical inference. The idea is to show the unity of many of the commonly used models. In such a text, space is not available to provide complete detailed coverage of each specific area, whether categorical data, survival, or classical linear models. The reader will not become an expert in time series or spatial analysis by reading this book! The intention is rather to provide a taste of these different areas, and of their unity. Some of the most important specialized books available in each of these fields are indicated at the end of each chapter.

For the examples, every effort has been made to provide as much background information as possible. However, because they come from such a wide variety of fields, it is not feasible in most cases to develop prior theoretical models to which confirmatory methods, such as testing, could be applied. Instead, analyses primarily concern exploratory inference involving model selection, as is typical of practice in most areas of applied statistics. In this way, the reader will be able to discover many direct comparisons of the application of the various members of the generalized linear model family.

Chapter 1 introduces the generalized linear model in some detail. The necessary background in inference procedures is relegated to Appendices A and B, which are oriented towards the unifying role of the likelihood function and include details on the appropriate diagnostics for model checking. Simple log linear and logistic models are used, in Chapter 2, to introduce the first major application of generalized linear models. These log linear models are shown, in turn, in Chapter 3, to encompass generalized linear models as a special case, so that we come full circle. More general regression techniques are developed, through applications to growth curves, in Chapter 4. In Chapter 5, some methods of handling dependent data are described through the application of conditional regression models to longitudinal data. Another major area of application of generalized linear models is to survival, and duration, data, covered in Chapters 6 and 7, followed by spatial models in Chapter 8. Normal linear models are briefly reviewed in Chapter 9, with special reference to model checking by comparing them to

nonlinear and non-normal models. (Experienced statisticians may consider this chapter to be simpler than the the others; in fact, this only reflects their greater familiarity with the subject.) Finally, the unifying methods of dynamic generalized linear models for dependent data are presented in Chapter 10, the most difficult in the text.

The two-dimensional plots were drawn with MultiPlot, for which I thank Alan Baxter, and the three-dimensional ones with Maple. I would also like to thank all of the contributors of data sets; they are individually cited with each table.

Students in the masters program in biostatistics at Limburgs University have provided many comments and suggestions throughout the years that I have taught this course there. Special thanks go to all the members of the Department of Statistics and Measurement Theory at Groningen University who created the environment for an enjoyable and profitable stay as Visiting Professor while I prepared the first draft of this text. Philippe Lambert, Patrick Lindsey, and four referees provided useful comments that helped to improve the text.

Diepenbeek J.K.L.
December, 1996

Contents

This page intentionally left blank

1
Generalized Linear Modelling

1.1 Statistical Modelling

Models are abstract, simplified representations of reality, often used both in science and in technology. No one should believe that a model could be true, although much of theoretical statistical inference is based on just this assumption. Models may be deterministic or probabilistic. In the former case, outcomes are precisely defined, whereas, in the latter, they involve variability due to unknown random factors. Models with a probabilistic component are called statistical models.

The one most important class, that with which we are concerned, contains the generalized linear models. They are so called because they generalize the classical linear models based on the normal distribution. As we shall soon see, this generalization has two aspects: in addition to the linear regression part of the classical models, these models can involve a variety of distributions selected from a special family, exponential dispersion models, and they involve transformations of the mean, through what is called a "link function" (Section 1.4.3), linking the regression part to the mean of one of these distributions.

1.1.1 A Motivating Example

Altman (1991, p. 199) provides counts of T_4 cells/mm^3 in blood samples from 20 patients in remission from Hodgkin's disease and 20 other patients in remission from disseminated malignancies, as shown in Table 1.1. We

TABLE 1.1. T_4 cells/mm^3 in blood samples from 20 patients in remission from Hodgkin's disease and 20 patients in remission from disseminated malignancies (Altman, 1991, p. 199).

Hodgkin's Disease	Non-Hodgkin's Disease
396	375
568	375
1212	752
171	208
554	151
1104	116
257	736
435	192
295	315
397	1252
288	675
1004	700
431	440
795	771
1621	688
1378	426
902	410
958	979
1283	377
2415	503

wish to determine if there is a difference in cell counts between the two diseases. To do this, we should first define exactly what we mean by a difference. For example, are we simply looking for a difference in mean counts, or a difference in their variability, or even a difference in the overall form of the distributions of counts?

A simple naive approach to modelling the difference would be to look at the difference in estimated means and to make inferences using the estimated standard deviation. Such a procedure implicitly assumes a normal distribution. It implies that we are only interested in differences of means and that we *assume* that the variability and normal distributional form are identical in the two groups. The resulting Student t value for no difference in means is 2.11.

Because these are counts, a more sophisticated method might be to assume a Poisson distribution of the counts within each group (see Chapter 2). Here, as we shall see later, it is more natural to use differences in logarithms of the means, so that we are looking at the difference between the means, themselves, through a ratio instead of by subtraction. However, this

TABLE 1.2. Comparison of models, based on various distributional assumptions, for no difference and difference between diseases, for the T_4 cell count data of Table 1.1.

	AIC			
Model	No difference	Difference	Difference in $-2\log(L)$	Estimate /s.e.
Normal	608.8	606.4	4.4	2.11
Log normal	590.1	588.6	3.5	1.88
Gamma	591.3	588.0	5.3	2.14
Inverse Gaussian	590.0	588.2	3.8	1.82
Poisson	11652.0	10294.0	1360.0	36.40
Negative binomial	589.2	586.0	5.2	2.36

model also carries the additional assumption that the variability will be different between the two groups if the mean is, because the variance of a Poisson distribution is equal to its mean. Now, the asymptotic Student t value for no difference in means, and hence in variances, is 36.40, quite different from the previous one.

Still a third approach would be to take logarithms before calculating the means and standard deviation in the first approach, thus, in fact, fitting a log normal model. In the Poisson model, we looked at the difference in log mean, whereas now we have the difference in mean logarithms. Here, it is much more difficult to transform back to a direct statement about the difference between the means themselves. As well, although the variance of the log count is assumed to be the same in the two groups, that of the count itself will not be identical. This procedure gives a Student t value of 1.88, yielding a still different conclusion.

A statistician only equipped with classical inference techniques has little means of judging which of these models best fits the data. For example, study of residual plots helps little here because none of the models (except the Poisson) show obvious discrepancies. With the direct likelihood approach used in this book, we can consider the Akaike (1973) information criterion (AIC) for which small values are to be preferred (see Section A.1.4). Here, it can be applied to these models, as well as some other members of the generalized linear model family.

The results for this problem are presented in Table 1.2. We see, as might be expected with such large counts, that the Poisson model fits very poorly. The other count model, that allows for overdispersion (Section 2.3), the negative binomial (the only one that is not a generalized linear model), fits best, whereas the gamma is second. By the AIC criterion, a difference between the two diseases is indicated for all distributions.

Consider now what would happen if we apply a significance test at the 5% level. This might either be a log likelihood ratio test based on the difference

in minus two log likelihood, as given in the second last column of Table 1.2, or a Wald test based on the ratio of the estimate to the standard error, in the last column of the table. Here, the conclusions about group difference vary depending on which distribution we choose. Which test is correct? Fundamentally, only one can be: that which we hypothesized *before* obtaining the data (if we did). If, by whatever means, we choose a model, based on the data, and then "test" for a difference between the two groups, the P-value has no meaning because it does not take into account the uncertainty in the model choice.

After this digression, let us finally draw our conclusions from our model selection procedure. The choice of the negative binomial distribution indicates heterogeneity among the patients with a group: the mean cell counts are not the same for all patients. The estimated difference in log mean for our best fitting model, the negative binomial, is -0.455 with standard error, 0.193, indicating lower counts for non-Hodgkin's disease patients. The ratio of means is then estimated to be $\exp(-0.455) = 0.634$.

Thus, we see that the conclusions drawn from a set of data depend very much on the assumptions made. Standard naive methods can be very misleading. The modelling and inference approach to be presented here provides a reasonably wide set of possible assumptions, as we see from this example, assumptions that can be compared and checked with the data.

1.1.2 History

The developments leading to the general overview of statistical modelling, known as generalized linear models, extend over more than a century. This history can be traced very briefly as follows (adapted from McCullagh and Nelder, 1989, pp. 8–17):

- multiple linear regression — a normal distribution with the identity link (Legendre, Gauss: early nineteenth century);
- analysis of variance (ANOVA) designed experiments — a normal distribution with the identity link (Fisher: 1920s → 1935);
- likelihood function — a general approach to inference about any statistical model (Fisher, 1922);
- dilution assays — a binomial distribution with the complementary log log link (Fisher, 1922);
- exponential family — a class of distributions with sufficient statistics for the parameters (Fisher, 1934);
- probit analysis — a binomial distribution with the probit link (Bliss, 1935);

- logit for proportions — a binomial distribution with the logit link (Berkson, 1944; Dyke and Patterson, 1952);
- item analysis — a Bernoulli distribution with the logit link (Rasch, 1960);
- log linear models for counts — a Poisson distribution with the log link (Birch, 1963);
- regression models for survival data — an exponential distribution with the reciprocal or the log link (Feigl and Zelen, 1965; Zippin and Armitage, 1966; Glasser, 1967);
- inverse polynomials — a gamma distribution with the reciprocal link (Nelder, 1966).

Thus, it had been known since the time of Fisher (1934) that many of the commonly used distributions were members of one family, which he called the *exponential family*. By the end of the 1960s, the time was ripe for a synthesis of these various models (Lindsey, 1971). In 1972, Nelder and Wedderburn went the step further in unifying the theory of statistical modelling and, in particular, regression models, publishing their article on *generalized linear models* (GLM). They showed

- how many of the most common linear regression models of classical statistics, listed above, were in fact members of one family and could be treated in the same way,
- that the maximum likelihood estimates for all of these models could be obtained using the same algorithm, *iterated weighted least squares* (IWLS, see Section A.1.2).

Both elements were equally important in the subsequent history of this approach. Thus, all of the models listed in the history above have a distribution in the *exponential dispersion family* (Jørgensen, 1987), a generalization of the exponential family, with some transformation of the mean, the link function, being related linearly to the explanatory variables.

Shortly thereafter, the first version of an interactive statistical computer package called GLIM (Generalized Linear Interactive Modelling) appeared, allowing statisticians easily to fit the whole range of models. GLIM produces very minimal output, and, in particular, only differences of log likelihoods, what its developers called deviances, for inference. Thus, GLIM

- displaced the monopoly of models based on the normal distribution by making analysis of a larger class of appropriate models possible by any statistician,
- had a major impact on the growing recognition of the likelihood function as central to all statistical inference,

- allowed experimental development of many new models and uses for which it was never originally imagined.

However, one should now realize the major constraints of this approach, a technology of the 1970s:

1. the linear component is retained;
2. distributions are restricted to the exponential dispersion family;
3. responses must be independent.

Modern computer power can allow us to overcome these constraints, although appropriate software is slow in appearing.

1.1.3 Data Generating Mechanisms and Models

In statistical modelling, we are interested in discovering what we can learn about systematic patterns from empirical data containing a random component. We suppose that some complex *data generating mechanism* has produced the observations and wish to describe it by some simpler, but still realistic, *model* that highlights the specific aspects of interest. Thus, by definition, models are never "true" in any sense.

Generally, in a model, we distinguish between systematic and random variability, where the former describes the patterns of the phenomenon in which we are particularly interested. Thus, the distinction between the two depends on the particular questions being asked. Random variability can be described by a probability distribution, perhaps multivariate, whereas the systematic part generally involves a regression model, most often, but not necessarily (Lindsey, 1974b), a function of the mean parameter. We shall explore these two aspects in more detail in the next two subsections.

1.1.4 Distributions

Random Component

In the very simplest cases, we observe some *response variable* on a number of independent units under conditions that we assume homogeneous in all aspects of interest. Due to some stochastic data generating mechanism that we imagine might have produced these responses, certain ones will appear more frequently than others. Our model, then, is some *probability distribution*, hopefully corresponding in pertinent ways to this mechanism, and one that we expect might represent adequately the frequencies with which the various possible responses are observed.

The hypothesized data generating mechanism, and the corresponding candidate statistical models to describe it, are scientific or technical constructs.

The latter are used to gain insight into the process under study, but are generally vast simplifications of reality. In a more descriptive context, we are just smoothing the random irregularities in the data, in this way attempting to detect patterns in them.

A probability distribution will usually have one or more unknown parameters that can be estimated from the data, allowing it to be fitted to them. Most often, one parameter will represent the average response, or some transformation of it. This determines the *location* of the distribution on the axis of the responses. If there are other parameters, they will describe, in various ways, the *variability* or *dispersion* of the responses. They determine the *shape* of the distribution, although the mean parameter will usually also play an important role in this, the form almost always changing with the size of the mean.

Types of Response Variables

Responses may generally be classified into three broad types:

1. measurements that can take any real value, positive or negative;
2. measurements that can take only positive values;
3. records of the frequency of occurrence of one or more kinds of events.

Let us consider them in turn.

Continuous Responses

The first type of response is well known, because elementary statistics courses concentrate on the simpler normal theory models: simple linear regression and analysis of variance (ANOVA). However, such responses are probably the rarest of the three types actually encountered in practice. Response variables that have positive probability for negative values are rather difficult to find, making such models generally unrealistic, except as rough approximations. Thus, such introductory courses are missing the mark. Nevertheless, such models are attractive to mathematicians because they have certain nice mathematical properties. But, for this very reason, the characteristics of these models are unrepresentative and quite misleading when one tries to generalize to other models, even in the same family.

Positive Responses

When responses are measurements, they most often can only take positive values (length, area, volume, weight, time, and so on). The distribution of the responses will most often be skewed, especially if many of these values tend to be relatively close to zero.

One type of positive response of special interest is the measurement of duration time to some event: survival, illness, repair, unemployment, and

so on. Because the length of time during which observations can be made is usually limited, an additional problem may present itself here: the response time may not be completely observed — it may be censored if the event has not yet occurred — we only know that it is at least as long as the observation time.

Events

Many responses are simple records of the occurrence of events. We are often interested in the *intensity* with which the events occur on each unit. If only one type of event is being recorded, the data will often take the form of *counts*: the number of times the event has occurred to a given unit (usual at least implicitly within some fixed interval of time). If more than one type of response event is possible, we have categorical data, with one category corresponding to each event type. If several such events are being recorded on each unit, we may still have counts, but now as many types on each unit as there are categories (some may be zero counts).

The categories may simply be nominal, or they may be ordered in some way. If only one event is recorded on each unit, similar events may be aggregated across units to form *frequencies* in a *contingency table*. When explanatory variables distinguish among several events on the same unit, the situation becomes even more complex.

Duration time responses are very closely connected to event responses, because times are measured between events. Thus, as we shall see, many of the models for these two types of responses are closely related.

1.1.5 Regression Models

Most situations where statistical modelling is required are more complex than can be described simply by a probability distribution, as just outlined. Circumstances are not homogeneous; instead, we are interested in how the responses change under different conditions. The latter may be described by *explanatory variables*. The model must have a systematic component.

Most often, for mathematical convenience rather than modelling realism, only certain simplifying assumptions are envisaged:

- responses are independent of each other;
- the mean response changes with the conditions, but the functional *shape* of the distribution remains fundamentally unchanged;
- the mean response, or some transformation of it, changes in some *linear* way as the conditions change.

Thus, as in the introductory example, we find ourselves in some sort of general linear regression situation. We would like to be able to choose from

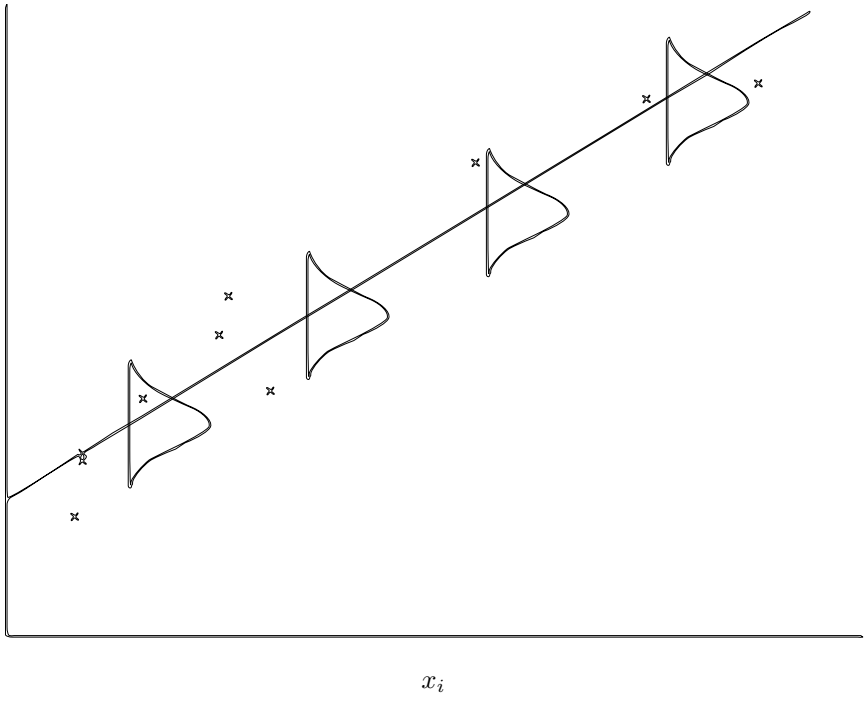

FIGURE 1.1. A simple linear regression. (The vertical axis gives both the observed y_i and its mean, μ_i.)

among the available probability distributions that which is most appropriate, instead of being forced to rely only on the classical normal distribution.

Consider a simple linear regression plot, as shown in Figure 1.1. The normal distribution, of constant shape because the variance is assumed constant, is being displaced to follow the straight regression line as the explanatory variable changes.

1.2 Exponential Dispersion Models

As mentioned above, generalized linear models are restricted to members of one particular family of distributions that has nice statistical properties. In fact, this restriction arises for purely technical reasons: the numerical algorithm, iterated weighted least squares (IWLS; see Section A.1.2) used for estimation, only works within this family. With modern computing power, this limitation could easily be lifted; however, no such software, for a wider family of regression models, is currently being distributed. We shall now look more closely at this family.

1.2.1 Exponential Family

Suppose that we have a set of independent random response variables, Z_i $(i = 1, \ldots, n)$ and that the probability (density) function can be written in the form

$$\begin{aligned} f(z_i; \xi_i) &= r(z_i)s(\xi_i)\exp[t(z_i)u(\xi_i)] \\ &= \exp[t(z_i)u(\xi_i) + v(z_i) + w(\xi_i)] \end{aligned}$$

with ξ_i a *location parameter* indicating the position where the distribution lies within the range of possible response values. Any distribution that can be written in this way is a member of the (one-parameter) exponential family. Notice the duality of the observed value, z_i, of the random variable and the parameter, ξ_i. (I use the standard notation whereby a capital letter signifies a random variable and a small letter its observed value.)

The *canonical form* for the random variable, the parameter, and the family is obtained by letting $y = t(z)$ and $\theta = u(\xi)$. If these are one-to-one transformations, they simplify, but do not fundamentally change, the model which now becomes

$$f(y_i; \theta_i) = \exp[y_i\theta_i - b(\theta_i) + c(y_i)]$$

where $b(\theta_i)$ is the normalizing constant of the distribution. Now, Y_i $(i = 1, \ldots, n)$ is a set of independent random variables with means, say μ_i, so that we might, classically, write $y_i = \mu_i + \varepsilon_i$.

Examples

Although it is not obvious at first sight, two of the most common discrete distributions are included in this family.

1. Poisson distribution

$$\begin{aligned} f(y_i; \mu_i) &= \frac{\mu_i^{y_i}\mathrm{e}^{-\mu_i}}{y_i!} \\ &= \exp[y_i\log(\mu_i) - \mu_i - \log(y_i!)] \end{aligned}$$

where $\theta_i = \log(\mu_i)$, $b(\theta_i) = \exp[\theta_i]$, and $c(y_i) = -\log(y_i!)$.

2. Binomial distribution

$$\begin{aligned} f(y_i; \pi_i) &= \binom{n_i}{y_i}\pi_i^{y_i}(1-\pi_i)^{n_i - y_i} \\ &= \exp\left\{y_i\log\left[\frac{\pi_i}{1-\pi_i}\right] + n_i\log(1-\pi_i) + \log\binom{n_i}{y_i}\right\} \end{aligned}$$

where $\theta_i = \log\left[\frac{\pi_i}{1-\pi_i}\right]$, $b(\theta_i) = n_i\log(1 + \exp[\theta_i])$, and $c(y_i) = \log\binom{n_i}{y_i}$. □

As we shall soon see, $b(\theta)$ is a very important function, its derivatives yielding the mean and the variance function.

1.2.2 Exponential Dispersion Family

The exponential family can be generalized by including a (constant) *scale parameter*, say ϕ, in the distribution, such that

$$f(y_i;\theta_i,\phi) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi)\right] \tag{1.1}$$

where θ_i is still the canonical form of the location parameter, some function of the mean, μ_i.

Examples

Two common continuous distributions are members of this family.

1. Normal distribution

$$\begin{aligned} f(y_i;\mu_i,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{e}^{\frac{-(y_i-\mu_i)^2}{2\sigma^2}} \\ &= \exp\left\{\left[y_i\mu_i - \frac{\mu_i^2}{2}\right]\frac{1}{\sigma^2} - \frac{y_i^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right\} \end{aligned}$$

where $\theta_i = \mu_i$, $b(\theta_i) = \theta_i^2/2$, $a_i(\phi) = \sigma^2$, and $c(y_i,\phi) = -[y_i^2/\phi + \log(2\pi\phi)]/2$.

2. Gamma distribution

$$\begin{aligned} f(y_i;\mu_i,\nu) &= \left(\frac{\nu}{\mu_i}\right)^{\nu}\frac{y_i^{\nu-1}\mathrm{e}^{-\frac{\nu y_i}{\mu_i}}}{\Gamma(\nu)} \\ &= \exp\{[-y_i/\mu_i - \log(\mu_i)]\nu + (\nu-1)\log(y_i) \\ &\qquad +\nu\log(\nu) - \log[\Gamma(\nu)]\} \end{aligned}$$

where $\theta_i = -1/\mu_i$, $b(\theta_i) = -\log(-\theta_i)$, $a_i(\phi) = 1/\nu$, and $c(y_i,\phi) = (\nu-1)\log(y_i) + \nu\log(\nu) - \log[\Gamma(\nu)]$. □

Notice that the examples given above for the exponential family are also members of the exponential dispersion family, with $a_i(\phi) = 1$. With ϕ known, this family can be taken to be a special case of the one-parameter exponential family; y_i is then the sufficient statistic for θ_i in both families.

In general, only the *densities* of continuous distributions are members of these families. As we can see in Appendix A, working with them implies that continuous variables are measured to infinite precision. However, the probability of observing any such point value is zero. Fortunately, such an approximation is often reasonable for location parameters when the sample size is small (although it performs increasingly poorly as sample size increases).

1.2.3 Mean and Variance

For members of the exponential and exponential dispersion families, a special relationship exists between the mean and the variance: the latter is

a precisely defined and unique function of the former for each member (Tweedie, 1947).

The relationship can be shown in the following way. For any likelihood function, $\mathrm{L}(\theta_i, \phi; y_i) = f(y_i; \theta_i, \phi)$, for one observation, the first derivative of its logarithm,

$$\mathrm{U}_i = \frac{\partial \log[\mathrm{L}(\theta_i, \phi; y_i)]}{\partial \theta_i}$$

is called the score function. (When this function, for a complete set of observations, is set to zero, the solution of the resulting equations, called the score equations, yields the maximum likelihood estimates.) From standard inference theory, it can be shown that

$$\mathrm{E}[\mathrm{U}_i] = 0 \tag{1.2}$$

and

$$\begin{aligned} \mathrm{var}[\mathrm{U}_i] &= \mathrm{E}[\mathrm{U}_i^2] \\ &= \mathrm{E}\left[-\frac{\partial \mathrm{U}_i}{\partial \theta_i}\right] \end{aligned} \tag{1.3}$$

under mild regularity conditions that hold for these families.

From Equation (1.1), for the exponential dispersion family,

$$\log[\mathrm{L}(\theta_i, \phi; y_i)] = \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)$$

Then, for θ_i,

$$\mathrm{U}_i = \frac{y_i - \frac{\partial b(\theta_i)}{\partial \theta_i}}{a_i(\phi)} \tag{1.4}$$

so that

$$\begin{aligned} \mathrm{E}[Y_i] &= \frac{\partial b(\theta_i)}{\partial \theta_i} \\ &= \mu_i \end{aligned} \tag{1.5}$$

from Equation (1.2), and

$$\mathrm{U}_i' = -\frac{\frac{\partial^2 b(\theta_i)}{\partial \theta_i^2}}{a_i(\phi)}$$

from Equation (1.4). This yields

$$\begin{aligned} \mathrm{var}[\mathrm{U}_i] &= \frac{\mathrm{var}[Y_i]}{a_i^2(\phi)} \\ &= \frac{\frac{\partial^2 b(\theta_i)}{\partial \theta_i^2}}{a_i(\phi)} \end{aligned}$$

from Equations (1.3), (1.4), and (1.5), so that

$$\text{var}[Y_i] = \frac{\partial^2 b(\theta_i)}{\partial \theta_i^2} a_i(\phi)$$

Usually, we can simplify by taking

$$a_i(\phi) = \frac{\phi}{w_i}$$

where w_i are known *prior weights*. Then, if we let $\partial^2 b(\theta_i)/\partial\theta_i^2 = \tau_i^2$, which we shall call the *variance function*, a function of μ_i (or θ_i) only, we have

$$\begin{aligned} \text{var}[Y_i] &= a_i(\phi)\tau_i^2 \\ &= \frac{\phi\tau_i^2}{w_i} \end{aligned}$$

a product of the dispersion parameter and a function of the mean only. Here, θ_i is the parameter of interest, whereas ϕ is usually a nuisance parameter. For these families of distributions, $b(\theta_i)$ and the variance function each uniquely distinguishes among the members.

Examples

Distribution	Variance function
Poisson	$\mu = \mathrm{e}^{\theta}$
Binomial	$n\pi(1-\pi) = n\mathrm{e}^{\theta}/(1+\mathrm{e}^{\theta})^2$
Normal	1
Gamma	$\mu^2 = (-1/\theta)^2$
Inverse Gaussian	$\mu^3 = (-2/\theta)^{3/2}$

□

Notice how exceptional the normal distribution is, in that the variance function does not depend on the mean. This shows how it is possible to have the classical linear normal models with constant variance.

1.3 Linear Structure

We have noted that one simplifying assumption in a model is often that some function of the mean response varies in a linear way as conditions change: the linear regression model. With n independent units observed, this can be written as a *linear predictor*. In the simplest case, the canonical location parameter is equated to a linear function of other parameters, of the form

$$\theta_i(\mu_i) = \sum_j x_{ij}\beta_j$$

or

$$\boldsymbol{\theta}(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$$

where $\boldsymbol{\beta}$ is a vector of $p < n$ (usually) unknown parameters, the matrix $\mathbf{X}_{n\times p} = [\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T]^T$ is a set of known explanatory variables, the conditions, called the *design* or *model matrix*, and $\mathbf{X}\boldsymbol{\beta}$ is the *linear structure*. Here, θ_i is shown explicitly to be a function of the mean, something that was implicit in all that preceded.

For a qualitative or factor variable, x_{ij} will represent the presence or absence of a level of a factor and β_j the effect of that level; for a quantitative variable, x_{ij} is its value and β_j scales it to give its effect on the (transformed) mean.

This strictly linear model (in the parameters, but not necessarily the explanatory variables) can be further generalized by allowing other smooth functions of the mean, $\eta(\cdot)$:

$$\boldsymbol{\eta}(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$$

called the linear predictor. The model now has both linear and nonlinear components.

1.3.1 Possible Models

In the model selection process, a series of regression models will be under consideration. It is useful to introduce terminology to describe the various common possibilities that may be considered.

Complete, full, or saturated model The model has as many location parameters as observations, that is, n linearly independent parameters. Thus, it reproduces the data exactly but with no simplification, hence being of little use for interpretation.

Null model This model has one common mean value for all observations. It is simple but usually does not adequately represent the structure of the data.

Maximal model Here we have the largest, most complex model that we are actually prepared to consider.

Minimal model This model contains the minimal set of parameters that must be present; for example, fixed margins for a contingency table.

Current model This model lies between the maximal and minimal and is presently under investigation.

The saturated model describes the observed data exactly (in fact, if the distribution contains an unknown dispersion parameter, the latter will often not even be estimable), but, for this very reason, has little chance of being adequate in replications of the study. It does not highlight the pertinent features of the data. In contrast, a minimal model has a good chance of fitting as well (or poorly!) to a replicate of the study. However, the important features of the data are missed. Thus, some reasonable balance must be found between closeness of fit to the observed data and simplicity.

1.3.2 Notation for Model Formulae

For the expression of the linear component in models, it is often more convenient, and clearer, to be able to use terms exactly describing the variables involved, instead of the traditional Greek letters. It turns out that this has the added advantage that such expressions can be directly interpreted by computer software. In this section, let us then use the following convention for variables:

quantitative	variate	`X,Y,...`
qualitative	factor	`A,B,...`

Note that these are abstract representations; in concrete cases, we shall use the actual names of the variables involved, with no such restrictions on the letters.

Then, the Wilkinson and Rogers (1973) notation has

Variable type	Interpretation	Algebraic component	Model term
Quantitative	Slope	βx_i	`X`
Qualitative	Levels	α_i	`A`
Interaction		$(\alpha\beta)_{ij}$	`A·B`
Mixed	Changing slopes	$\beta_i x$	`B·X`

Notice how these model formulae refer to *variables*, not to parameters.

Operators

The actual model formula is set up by using a set of operators to indicate the relationships among the explanatory variables with respect to the (function of) the mean.

Combine terms	+	X+Y+A+Y·A
Add terms to previous model	+	+X·A
Remove terms from model	−	−Y
No change	·	
Interaction	·	A·B
Nested model	/	A/B/C
		(=A+A·B+A·B·C)
Factorial model	*	A*B
		(=A+B+A·B)
Constant term	1	X−1
		(line through origin)

With some software, certain operator symbols are modified. For example, in R and S-Plus, the colon (:) signifies interaction. These software also allow direct specification of the response variable before the linear structure: $\mathtt{Y} \sim \mathtt{X1} + \mathtt{X2}$. I shall use the notation in the original article, as shown in the table, throughout the text.

Example

$$\begin{aligned}
(\mathtt{A} * \mathtt{B}) \cdot (\mathtt{C} + \mathtt{D}) &= (\mathtt{A} + \mathtt{B} + \mathtt{A} \cdot \mathtt{B}) \cdot (\mathtt{C} + \mathtt{D}) \\
&= \mathtt{A} \cdot \mathtt{C} + \mathtt{A} \cdot \mathtt{D} + \mathtt{B} \cdot \mathtt{C} + \mathtt{B} \cdot \mathtt{D} + \mathtt{A} \cdot \mathtt{B} \cdot \mathtt{C} + \mathtt{A} \cdot \mathtt{B} \cdot \mathtt{D} \\
(\mathtt{A} * \mathtt{B})/\mathtt{C} &= \mathtt{A} + \mathtt{B} + \mathtt{A} \cdot \mathtt{B} + \mathtt{A} \cdot \mathtt{B} \cdot \mathtt{C} \\
\mathtt{A} * \mathtt{B} * \mathtt{C} - \mathtt{A} \cdot (\mathtt{B} * \mathtt{C}) &= \mathtt{A} + \mathtt{B} + \mathtt{C} + \mathtt{B} \cdot \mathtt{C} \\
\mathtt{A} * \mathtt{B} * \mathtt{C} - *\mathtt{B} \cdot \mathtt{C} &= \mathtt{A} + \mathtt{B} + \mathtt{C} + \mathtt{A} \cdot \mathtt{B} + \mathtt{A} \cdot \mathtt{C} \\
\mathtt{A} * \mathtt{B} * \mathtt{C} - /\mathtt{A} &= \mathtt{A} + \mathtt{B} + \mathtt{C} + \mathtt{B} \cdot \mathtt{C}
\end{aligned}$$

□

1.3.3 Aliasing

For various reasons, the design matrix, $\mathbf{X}_{n\times p}$, in a linear model may not be of full rank p. If the columns, $\mathbf{x}_1, \ldots, \mathbf{x}_j$, form a linearly dependent set, then some of the corresponding parameters $\beta_1, \ldots, \beta_j$ are aliased. In numerical calculations, we can use a generalized inverse of the matrix in order to obtain estimates.

Two types of alias are possible:

Intrinsic alias The specification of the linear structure contains redundancy whatever the observed values in the model matrix; for example, the mean plus parameters for all levels of a factor (the sum of the matrix columns for the factor effects equals the column for the mean).

Extrinsic alias An anomaly of the data makes the columns linearly dependent; for example, no observations are available for one level of a

factor (zero column) or there is collinearity among explanatory variables.

Let us consider, in more detail, intrinsic alias. Suppose that the rank of $\mathbf{X}$ is $r < p$, that is, that there are $p - r$ independent constraints on p estimates, $\hat{\boldsymbol{\beta}}$. Many solutions will exist, but this is statistically unimportant because $\hat{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{\mu}}$ will have the same estimated values for all possible values of $\hat{\boldsymbol{\beta}}$. Thus, these are simply different ways of expressing the same linear structure, the choice among them being made for ease of interpretation.

Example

Suppose that, in the regression model,

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

$x_3 = x_1 + x_2$, so that β_3 is redundant in explaining the structure of the data. Once information on β_1 and β_2 is removed from data, no further information on β_3 remains. Thus, one adequate model will be

$$\eta = \beta_1 x_1 + \beta_2 x_2$$

However,

$$\eta = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3$$

is also possible if

$$\begin{aligned} \alpha_1 &= 0 \\ \alpha_2 &= \beta_2 \\ \alpha_3 &= \beta_3 \end{aligned}$$

or

$$\begin{aligned} \alpha_1 &= (2\beta_1 - \beta_2)/3 \\ \alpha_2 &= (2\beta_2 - \beta_1)/3 \\ \alpha_3 &= -(\beta_1 + \beta_2)/3 \end{aligned}$$

□

The first parametrization in this example, with say $\alpha_1 = 0$, is called the *baseline constraint*, because all comparisons are being made with respect to the category having the zero value. The second, where $\alpha_1 + \alpha_2 + \alpha_3 = 0$, is known as the *usual* or *conventional constraint*. Constraints that make the parameters as meaningful as possible in the given context should be chosen.

1.4 Three Components of a GLM

Consider again the simple linear (least-squares) regression plot of Figure 1.1. This model has been written, classically, as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{where} \quad \varepsilon_i \sim \mathrm{N}(0, \sigma^2)$$

but is more clearly seen to be

$$\mu_i = \beta_0 + \beta_1 x_i$$

where μ_i is the mean of a normal distribution with constant variance, σ^2.

From this simple model, it is not necessarily obvious that three elements are in fact involved. We have already looked at two of them, the probability distribution and the linear structure, in some detail and have mentioned the third, the link function. Let us look at all three more closely.

1.4.1 Response Distribution or "Error Structure"

The Y_i $(i = 1, \ldots, n)$ are independent random variables with means, μ_i. They share the same distribution from the exponential dispersion family, with a constant scale parameter.

1.4.2 Linear Predictor

Suppose that we have a set of p (usually) unknown parameters, $\boldsymbol{\beta}$, and a set of known explanatory variables $\mathbf{X}_{n\times p} = [\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T]^T$, the design matrix, are such that

$$\eta = \mathbf{X}\boldsymbol{\beta}$$

where $\mathbf{X}\boldsymbol{\beta}$ is the linear structure. This describes how the location of the response distribution changes with the explanatory variables.

If a parameter has a known value, the corresponding term in the linear structure is called an *offset*. (This will be important for a number of models in Chapters 3 and 6.) Most software packages have special facilities to handle this.

1.4.3 Link Function

If $\theta_i = \eta_i$, our generalized linear model definition is complete. However, the further generalization to noncanonical transformations of the mean requires an additional component if the idea of a linear structure is to be retained.

The relationship between the mean of the ith observation and its linear predictor will be given by a *link function*, $g_i(\cdot)$:

$$\begin{aligned} \eta_i &= g_i(\mu_i) \\ &= \mathbf{x}_i^T \boldsymbol{\beta} \end{aligned}$$

This function must be monotonic and differentiable. Usually the same link function is used for all observations. Then, the canonical link function is that function which transforms the mean to a canonical location parameter of the exponential dispersion family member.

Example

Distribution	Canonical link function	
Poisson	Log	$\eta_i = \log(\mu_i)$
Binomial	Logit	$\eta_i = \log\left[\frac{\pi_i}{1-\pi_i}\right] = \log\left[\frac{\mu_i}{n_i-\mu_i}\right]$
Normal	Identity	$\eta_i = \mu_i$
Gamma	Reciprocal	$\eta_i = \frac{1}{\mu_i}$
Inverse Gaussian	Reciprocal^2	$\eta_i = \frac{1}{\mu_i^2}$

□

With the canonical link function, all unknown parameters of the linear structure have sufficient statistics if the response distribution is a member of the exponential dispersion family and the scale parameter is known. However, the link function is just an artifact to simplify the numerical methods of estimation when a model involves a linear part, that is, to allow the IWLS algorithm to work. For strictly nonlinear regression models, it loses its meaning (Lindsey, 1974b).

Consider now the example of a canonical linear regression for the binomial distribution, called logistic regression, as illustrated in Figure 1.2. We see how the form of the distribution changes as the explanatory variable changes, in contrast to models involving a normal distribution, illustrated in Figure 1.1.

Link functions can often be used to advantage to linearize seemingly nonlinear structures. Thus, for example, logistic and Gomperz growth curves become linear when respectively the logit and complementary log log links are used (Chapter 4).

Example

The Michaelis–Menten equation,

$$\mu_i = \frac{\beta_1 x_i}{1 + \beta_2 x_i}$$

is often used in biology because of its asymptotic properties. With a reciprocal link, it can be written

$$\frac{1}{\mu_i} = \alpha_1 + \frac{\alpha_2}{x_i}$$

where $\alpha_1 = \beta_2/\beta_1$ and $\alpha_2 = 1/\beta_1$. □

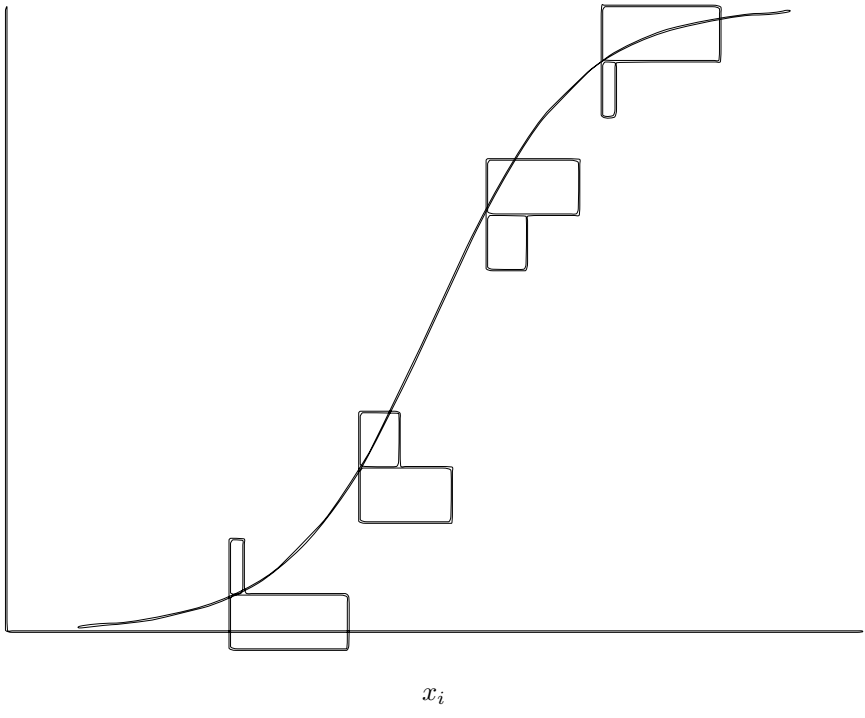

FIGURE 1.2. A simple linear logistic regression.

Thus, generalized linear models, as their name suggests, are restricted to having a linear structure for the explanatory variables. In addition, they are restricted to univariate, independent responses. Some ways of getting around these major constraints will be outlined in the next section and illustrated in some of the following chapters.

1.5 Possible Models

1.5.1 Standard Models

With GLM software, one can usually fit the following standard distributions, all members of the exponential dispersion family:

- Poisson
- binomial
- normal (also log normal)
- gamma (also log gamma, exponential, and Pareto)

- inverse Gaussian

and a series of link functions, only some of which are canonical,

- identity μ
- reciprocal $\frac{1}{\mu}$
- quadratic inverse $\frac{1}{\mu^2}$
- square root $\sqrt{\mu}$
- exponent $(\mu + c_1)^{c_2}$ (c_1 and c_2 known)
- log $\log(\mu)$
- logit $\log\left(\frac{\mu}{n-\mu}\right)$
- complementary log log $\log\left[-\log\left(\frac{\mu}{n}\right)\right]$
- probit $\Phi^{-1}\left(\frac{\mu}{n}\right)$

With some software, the user can define other models (distribution and/or link) if the distribution is a member of the exponential dispersion family.

1.5.2 Extensions

A number of tricks can be used with standard GLM software in order to fit certain models that are not generalized linear family.

Distributions Close to the Exponential Dispersion Family

If a distribution would be a member of the exponential dispersion family except for one (shape) parameter, an extra iteration loop can be used to obtain the maximum likelihood estimate of that parameter.

Example

The Weibull distribution,

$$f(y;\mu,\alpha) = \alpha\mu^{-\alpha}y^{\alpha-1}\mathrm{e}^{-(y/\mu)^{\alpha}}$$

with known shape parameter, α, is an exponential distribution (gamma with $\nu = 1$). If we take an initial value of the shape parameter, fit an exponential distribution with that value, and then estimate a new value, we can continue refitting until convergence. □

Parameters in the Link Function

Two possibilities are to plot likelihoods for various values of the unknown link parameter or to expand the link function in a Taylor series and include the first term as an extra covariate. In this latter case, we have to iterate to convergence.

Example

An exponent link with unknown power parameter ρ

$$\eta = \mu^{\rho}$$

can be estimated by including an extra term

$$-(\rho - \rho_0)\mu^{\rho_0} \log(\mu)$$

in the linear model. Change in likelihood will provide a measure of acceptability. □

Parameters in the Variance Function

In models from the exponential dispersion family, the likelihood equations for the linear structure can be solved without knowledge of the dispersion parameter (Section A.1.2). Some distributions have a parameter in the variance function that is not a dispersion parameter and, hence, cannot be estimated in the standard way. Usually, special methods are required for each case.

Example

Consider the negative binomial distribution with unknown power parameter, ζ, as will be given in Equation (2.4). If it were known and fixed, we would have a member of the exponential family. One approximate way in which this parameter can be estimated is by the method of moments, choosing a value that makes the Pearson chi-squared statistic equal to its expectation.

Another way, that I used in the motivating example and shall also use in Chapters 5 and 9, consists in trying a series of different values of the unknown parameter and choosing that with the smallest deviance. □

Nonlinear Structure

We can linearize a nonlinear parameter by a Taylor series approximation (Chapter 9), as for the link function.

Example

If $\beta h(x, \alpha)$ is the nonlinear term (for example, $\beta e^{\alpha x}$), then

$$h(x, \alpha) \doteq h(x, \alpha_0) + (\alpha - \alpha_0) \left[\frac{\partial h}{\partial \alpha}\right]_{\alpha=\alpha_0}$$

We can use two linear terms:

$$\beta h(x, \alpha_0) + \beta^* \left[\frac{\partial h}{\partial \alpha}\right]_{\alpha=\alpha_0}$$

where $\beta^* = \beta(\alpha - \alpha_0)$. At each iteration,

$$\alpha_{s+1} = \alpha_s + \frac{\beta^*}{\beta}$$

□

Survival Curves and Censored Observations

Many survival distributions can be shown to have a log likelihood that is essentially a Poisson distribution plus a constant term (an offset) not depending on the linear predictor (Section 6.3.2). A censored exponential distribution can be fitted with IWLS (no second iteration), whereas a number of others, including the Weibull, extreme value, and logistic distributions, require one simple extra iteration loop.

Composite Link Functions

The link function may vary with (subsets of) the observations. In many cases, this can be handled as for user-programmed link functions (and distributions). Examples include the proportional odds models for ordered variables in a contingency table and certain components of dispersion (of variance) in random effects and repeated measurements models.

1.6 Inference

Statistical software for generalized linear models generally produce deviance values (Section A.1.3) based on twice the differences of the log likelihood from that for a saturated model (that is, $-2\log[\mathrm{L}]$). However, as we have seen, the number of parameters in this saturated model depends on the number of observations, except in special cases; these models are a type of "semiparametric" model where the distribution is specified but the functional form of the systematic part, that is, the regression, is not. Hence, only differences in deviance, where this saturated term cancels out, may be

relevant. The one major exception is contingency tables where the saturated model has a fixed number of parameters, not increasing with the number of observations.

Thus, "semiparametric" and "nonparametric" models, that is, those where a functional form is not specified either for the systematic or for the stochastic part, are generally at least partially saturated models with a number of parameters that increases with the sample size. Most often, they involve a factor variable whose levels depend on the data observed. This creates no problem for direct likelihood inference where we condition on the observed data. Such saturated models often provide a point of comparison for the simpler parametric models.

In the examples in the following chapters, the AIC (Section A.1.4) is used for inference in the exploratory conditions of model selection. This is a simple penalization of the log likelihood function for complexity of the model, whereby some positive penalizing constant (traditionally unity) times the number of estimated parameters is subtracted from it. It only allows *comparison* of models; its absolute size is arbitrary, depending on what constants are left in the likelihood function and, thus, has no meaning.

For contingency tables, I shall use an AIC based on the usual deviance provided by the software. In all other cases, I base it on the complete minus two log likelihood, including all constants. The latter differs from the AIC produced by some of these packages by an additive constant, but has the important advantage that models based on different distributions can be directly compared. Because of the factor of minus two in these AICs, the penalty involves the subtraction of twice the number of estimated parameters. In all cases, a smaller AIC indicates a preferable model in terms of the data alone.

Generalized linear models provide us with a choice of distributions that frequentist inference, with its nesting requirements, does not easily allow us to compare. Direct likelihood inference overcomes this obstacle (Lindsey, 1974b, 1996b) and the AIC makes this possible even with different numbers of parameters estimated in the models to be compared.

In spite of some impressions, use of the AIC is not an automated process. The penalizing constant should be chosen, before collecting the data, to yield the desired complexity of models or smoothing of the data. However, for the usual sample sizes, unity (corresponding to minus two when the deviance is used) is often suitable. Obviously, if enough different models are tried, some will usually be found to fit well; the generalized linear model family, with its variety of distributions and link functions, already provides a sizable selection. However, a statistician will not blindly select that model with the smallest AIC; scientific judgment must also be weighed into the choice. Model selection is exploratory — hypothesis generation; the chosen model must then be tested, on new data, the confirmatory part of the statistical endeavour.

If the AIC is to be used for model selection, then likelihood intervals for parameters must also be based on this criterion for inferences to be compatible. Otherwise, contradictions will arise (Section A.1.4). Thus, with a penalizing constant of unity, the interval for one parameter will be $1/e = 0.368$ normed likelihood. This is considerably narrower than those classically used: for example a 5% asymptotic confidence interval, based on the chi-squared distribution, has a $\exp(-3.84/2) = 0.147$ normed likelihood. The AIC corresponding to the latter has a penalizing constant of 1.96, adding 3.84 times the number of estimated parameters, instead of 2 times, to the deviance. This will result in the selection of much simpler models if one parameter is checked at a time. (For example, in Section 6.3.4, the exponential would be chosen over the Weibull.)

For a further discussion of inference, see Appendix A.

Summary

For a more general introduction to statistical modelling, the reader might like to consult Chapter 1 of Lindsey (1996b) and Chapter 2 of Lindsey (1993).

Books on the exponential family are generally very technical; see, for example, Barndorff-Nielsen (1978) or Brown (1986). Chapter 2 of Lindsey (1996b) provides a condensed survey. Jørgensen (1987) introduced the exponential dispersion family.

After the original paper by Nelder and Wedderburn (1972) on generalized linear models, several books have been published, principally McCullagh and Nelder (1989), Dobson (1990), and Fahrmeir and Tutz (1994).

For much of their history, generalized linear models have owed their success to the computer software GLIM. This has resulted in a series of books on GLIM, including Healy (1988), Aitkin *et al.* (1989), Lindsey (1989, 1992), and Francis *et al.* (1993) and the conference proceedings of Gilchrist 1982), Gilchrist *et al.* (1985), Decarli *et al.* (1989), van der Heijden *et al.* (1992), Fahrmeir *et al.* (1992), and Seeber *et al.* (1995).

For other software, the reader is referred to the appropriate section of their manual.

For references to direct likelihood inference, see those listed at the end of Appendix A.

1.7 Exercises

1. (a) Figures 1.1 and 1.2 show respectively how the normal and binomial distributions change as the mean changes. Although informative, these graphics are, in some ways, fundamentally different. Discuss why.

(b) Construct a similar plot for the Poisson distribution. Is it more similar to the normal or to the binomial plot?

2. Choose some data set from a linear regression or analysis of variance course that you have had and suggest some more appropriate model for it than ones based on the normal distribution. Explain how the model may be useful in understanding the underlying data generating mechanism.

3. Why is intrinsic alias more characteristic of models for designed experiments whereas extrinsic aliases arises most often in observation studies such as sample surveys?

4. (a) Plot the likelihood function for the mean parameter of a Poisson distribution when the estimated mean is $\bar{y}_\bullet = 2.5$ for $n = 10$ observations. Give an appropriate likelihood interval about the mean.

   (b) Repeat for the same estimated mean when $n = 30$ and compare the results in the two cases.

   (c) What happens to the graphs and to the intervals if one works with the canonical parameter instead of the mean?

   (d) How do these results relate to Fisher information? To the use of standard errors as a measure of estimation precision?

2
Discrete Data

2.1 Log Linear Models

Traditionally, the study of statistics begins with models based on the normal distribution. This approach gives students a biased view of what is possible in statistics because, as we shall see, the most fundamental models are those for discrete data. As well, the latter are now by far the most commonly used in applied statistics. Thus, we begin our presentation of generalized linear regression modelling with the study of log linear models.

Log linear models and their special case for binary responses, logistic models, are designed for the modelling of frequency and count data, that is, those where the response variable involves discrete categories, as described in Section 1.1.4. Because they are based on the exponential family of distributions, they constitute a direct extension of traditional regression and analysis of variance. The latter models are based on the normal distribution (Chapter 9), whereas logistic and log linear models are based on the Poisson or multinomial distributions and their special cases, such as the binomial distribution. Thus, they are all members of the generalized linear model family.

Usually, although not necessarily, one models either the frequencies of occurrence of the various categories or the counts of events. Occasionally, as in some logistic regression models, the individual indicator variables of the categories are modelled. However, when both individual and grouped frequency data are available, they both give identical results. Thus, for the moment, we can concentrate, here, on grouped frequency data.

TABLE 2.1. A two-way table for change over time.

		Time 2	
		A	B
Time 1	A	45	13
	B	12	54

2.1.1 Simple Models

In order to provide a brief and simple introduction to logistic and log linear models, I have chosen concrete applications to modelling changes over time. However, the same principles that we shall study here also apply to cross-sectional data with a set of explanatory variables.

Observations over Time

Consider a simple two-way contingency table, Table 2.1, where some response variable with two possible values, A and B, was recorded at two points in time. A first characteristic that we may note is a relative stability over time, as indicated by the large frequencies on the diagonal. In other words, response at time 2 depends heavily on that at time 1, most often being the same.

As a simple model, we might consider that the responses at time 2 have a binomial distribution and that this distribution depends on what response was given at time 1. Thus, we might have the simple linear regression model

$$\log\left(\frac{\pi_{1|j}}{\pi_{2|j}}\right) = \beta_0 + \beta_1 x_j$$

where x_j is the response at time 1 and $\pi_{i|j}$ is the *conditional* probability of response i at time 2 given the observed value of x_j at time 1. Then, if $\beta_1 = 0$, this indicates independence, that is, that the second response does not depend on the first. In the Wilkinson and Rogers (1973) notation, the model can be written simply as the name of the variable:

```
TIME1
```

If the software also required specification of the response variable at the same time, this would become

```
TIME2 ~ TIME1
```

where `TIME2` represents a 2×2 matrix of the frequencies in the table, with columns corresponding to the two possible response values at the second time point.

This *logistic regression model*, with a logit link, the logarithm of the ratio of probabilities, is the direct analogue of classical (normal theory) linear

TABLE 2.2. A two-way table of clustered data.

		Right eye	
		A	B
Left eye	A	45	13
	B	12	54

regression. On the other hand, if x_j is coded $(-1, 1)$ or $(0, 1)$, we may rewrite this as

$$\log\left(\frac{\pi_{1|j}}{\pi_{2|j}}\right) = \mu + \alpha_j$$

where $\mu = \beta_0$, the direct analogue of an analysis of variance model, with the appropriate constraints. With suitable software, `TIME1` would simply be declared as a factor variable having two levels.

Example

The parameter estimates for Table 2.1 are $\hat{\beta}_0 = \hat{\mu} = 1.242$ and $\hat{\beta}_1 = \hat{\alpha}_1 = -2.746$, when x_j is coded $(0, 1)$, with an AIC of 4. (The deviance is zero and there are two parameters.) That with $\alpha_1 = \beta_1 = 0$, that is, independence, has AIC 48.8. Thus, in comparing the two models, the first, with dependence on the previous response, is much superior, as indicated by the smaller AIC. □

Clustered Observations

Let us momentarily leave data over time and consider, instead, the same table, now Table 2.2, as some data on the two eyes of people. We again have repeated observations on the same individuals, but here they may be considered as being made simultaneously rather than sequentially. Again, there will usually be a large number with similar responses, resulting from the dependence between the two similar eyes of each person.

Here, we would be more inclined to model the responses simultaneously as a multinomial distribution over the four response combinations, with *joint* probability parameters, π_{ij}. In that way, we can look at the association between them. Thus, we might use a log link such that

$$\log(\pi_{ij}) = \phi + \mu_i + \nu_j + \alpha_{ij} \tag{2.1}$$

With the appropriate constraints, this is again an analogue of classical analysis of variance. It is called a *log linear model*. If modelled by the Poisson representation (Section 2.1.2), it could be given in one of two equivalent ways:

`REYE * LEYE`

or

$$\texttt{REYE} + \texttt{LEYE} + \texttt{REYE} \cdot \texttt{LEYE}$$

With specification of the response variable, the latter becomes

$$\texttt{FREQ} \sim \texttt{REYE} + \texttt{LEYE} + \texttt{REYE} \cdot \texttt{LEYE}$$

where `FREQ` is a four-element vector containing the frequencies in the table. Notice that, in this representation of the multinomial distribution, the "response variable", `FREQ`, is not really a variable of direct interest at all.

Example

Here, the parameter estimates for Table 2.1 are $\hat{\phi} = 2.565$, $\nu_1 = 1.424$, $\hat{\mu}_1 = 1.242$, and $\hat{\alpha}_{11} = -2.746$, with an AIC of 8. (Again, the deviance is zero, but here there are four parameters.) That with $\alpha_{11} = 0$ has AIC 52.8. (This is 4 larger than in the previous case because the model has two more parameters, but the difference in AIC is the same.) The conclusion is identical, that the independence model is much inferior to that with dependence. □

Log Linear and Logistic Models

The two models just described have a special relationship to each other. With the same constraints, the dependence parameter, $\boldsymbol{\alpha}$, is identical in the two cases because

$$\log\left(\frac{\pi_{1|1}\pi_{2|2}}{\pi_{1|2}\pi_{2|1}}\right) = \log\left(\frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}\right)$$

The normed profile likelihoods for $\boldsymbol{\alpha} = \mathbf{0}$ are also identical, although the AICs are not because of the different numbers of parameters explicitly estimated in the two models (differences in AIC are, however, the same). This is a general result: in cases where both are applicable, logistic and log linear models yield the same conclusions. The choice is a matter of convenience.

This is a very important property, because it means that such models can be used for *retrospective sampling*. Common examples of this include, in medicine, case-control studies, and, in the social sciences, mobility studies.

These results extend directly to larger tables, including higher dimensional tables. There, direct analogues of classical regression and ANOVA models are still applicable. Thus, complex models of dependence among categorical variables can be built up by means of multiple regression. Explanatory variables can be discrete or continuous (at least if the data are not aggregated in a contingency table).

2.1.2 Poisson Representation

With a log linear model, we may have more than two categories for the response variable(s), so that we require a multinomial, instead of a binomial,

distribution. This cannot generally be directly fitted by standard generalized linear modelling software. However, an important relationship exists between the multinomial and Poisson distributions that makes fitting such models possible.

Consider independent Poisson distributions with means μ_k and corresponding numbers of events n_k. Let us condition on the observed total number of events, $n_\bullet = \sum_k n_k$. From the properties of the Poisson distribution, this total will also have the same distribution, with mean $\mu_\bullet = \sum_k \mu_k$. Then, the conditional distribution will be

$$\frac{\prod \frac{\mathrm{e}^{-\mu_k}\mu_k^{n_k}}{n_k!}}{\frac{\mathrm{e}^{-\mu_\bullet}\mu_\bullet^{n_\bullet}}{n_\bullet!}} = \binom{n_\bullet}{n_1 \cdots n_K} \prod_{k=1}^{K} \left(\frac{\mu_k}{\mu_\bullet}\right)^{n_k}$$

a multinomial distribution with probabilities, $\pi_k = \mu_k/\mu_\bullet$. Thus, any multinomial distribution can be fitted as a product of independent Poisson distributions with the appropriate conditioning on the total number of events.

Specifically, this means that, when fitting such models, the product of all explanatory variables must be included in the minimal log linear model, in order to fix the appropriate marginal totals in the table:

$$\texttt{R1} + \texttt{R2} + \cdots + \texttt{E1} * \texttt{E2} * \cdots \tag{2.2}$$

where `Ri` represents a response variable and `Ej` an explanatory variable. This ensures that all responses have proper probability distributions (Lindsey, 1995b). Much of log linear modelling involves searching for simple structural models of relationships among responses (`Ri`) and of dependencies of responses on explanatory variables.

2.2 Models of Change

One of the most important uses of log linear models has been in sample survey data. A particularly interesting area of this field is *panel data*. There, the same survey questions are administered at two or more points in time to the same people. In this way, we can study changes over time.

For simplicity, let us restrict attention, for the moment, to the observation of responses at only two points in time, that is, to two-dimensional tables, as in the simple example above. However, the generalization to more complex cases is fairly direct.

Suppose that the response has I categories, called the *states*, so that we have a $I \times I$ table and are studying changes in state over time. Then, our dependence parameter, $\boldsymbol{\alpha}$, in Equation (2.1), will be a $I \times I$ matrix, but with only $(I-1) \times (I-1)$ independent values, because of the need for constraints. When $I > 2$, the idea is to reduce this number of parameters

by structuring the values in some informative way, that is, to be able to model the specific forms of dependence among successive responses.

The minimal model will be independence, that is, when $\pi_{ij} = \pi_{i\bullet}\pi_{\bullet j}$ or, equivalently, $\alpha_{ij} = 0\ \forall i, j$. The maximal model is the saturated or "non-parametric" one. The latter is often not especially useful. Most interesting models, in this context, are based on *Markov chains*: the current response simply is made to depend on the previous one. These are models describing the *transition probabilities* of changing from one state to another between two points in time.

2.2.1 Mover–Stayer Model

Because, in the example above, we have noticed that there is often a rather large number of individuals who will give the same response the two times, let us first see how to model this.

Suppose that we have a mixture of two subpopulations or latent groups, one of which is susceptible to change while the other is not. This is called a *mover–stayer* model. We know that individuals recorded off the main diagonal will all belong to the first subpopulation, the movers, because they have changed. However, the main diagonal frequencies are more complex because they will contain both the stayers and any movers who did not happen to change within the observation period (more exactly, who were in the same place on both observation dates).

For a simple model, let us assume that the locations of the movers at the two points in time are independent. If we ignore the mixture on the diagonal, we can model the rest of the table by *quasi-independence*, that is, independence in an incomplete table where the diagonal is missing. Then, with this independence assumption, we can obtain estimates of the number of movers on the diagonal and, hence, of the number of stayers.

Example

A 10% sample is available, from the published migration reports of the 1971 British census, of people migrating between 1966 and 1971 among four important centres of population in Britain, the Metropolitan Counties. The results are given in Table 2.3. All are British residents who were born in the New Commonwealth. Here, the numbers not moving, those on the diagonal, are very extreme.

For this table, the deviance for independence

```
MOVE66 + MOVE71
```

is 19,884 (AIC 19,898) with nine degrees of freedom (d.f.), a strong indication of dependence, whereas that for the mover–stayer model (quasi-independence), fitted in the same way but to the table without the main diagonal, is 4.4 (26.4) with five d.f., a remarkable improvement. The loss of

TABLE 2.3. Place of residence in Britain in 1966 and 1971. (Fingleton, 1984, p. 142)

	1971			
1966	CC	ULY	WM	GL
Central Clydesdale	118	12	7	23
Urban Lancs. & Yorks.	14	2127	86	130
West Midlands	8	69	2548	107
Greater London	12	110	88	7712

four d.f. results from eliminating the diagonal entries; this is equivalent to allowing a separate parameter for each of them. This is taken into account in the AIC. Because the deviance is zero, the AIC for the saturated model of a contingency table is just two times the number of entries in the table, here 32, so that the mover–stayer model is also to be preferred to the saturated model.

Notice that the dependence arises almost entirely from stayers being in the same place at the two time points. The numbers of movers on the diagonal are estimated to be only 1.6, 95.2, 60.3, and 154.6, respectively. Thus, most people in the table can have their 1971 place of residence exactly predicted by that of 1966: they will be in the same place. This is the dependence just detected above. □

The mover–stayer model allows a different probability of staying within each response state. The special case where all of these probabilities are equal is called the *loyalty* model. This can be fitted by a factor variable with one level for all diagonal entries and a second for off-diagonal ones, instead of eliminating the diagonal.

Note, however, that, for calculating conditional probabilities, such models are really illegitimate. The probability of being in a given state at the second time point depends on knowledge about whether or not one is a mover (or loyal), but this cannot be known until the state at the second time is available.

2.2.2 Symmetry

Because, in panel data, the same response variables are being recorded two (or more) times, we might expect some symmetry among them.

Complete Symmetry

Suppose that the probability of changing between any pair of categories is the same in both directions:

$$\pi_{i|j} = \pi_{j|i} \qquad \forall i, j \tag{2.3}$$

a model of *complete symmetry*. In terms of Markov chains, this is equivalent to the combination of two characteristics, *reversibility* and *equilibrium*, that we shall now study. In fact, we can separate the two.

Equilibrium

Consider, first, equilibrium: the marginal probabilities are the same at the two time points,

$$\pi_{i\bullet} = \pi_{\bullet i} \qquad \forall i$$

In other words, the (marginal) distribution of the states remains the same at the different time points, hence the name. In the usual analysis of contingency tables, this is called *marginal homogeneity*; however, it is not a log linear model and, hence, not a generalized linear model. It requires special programming to be fitted. In our example below, this model would imply that the proportion of votes received by each party remained the same in the two elections.

Reversibility

Reversibility, on the other hand, implies (more or less) equal transition probabilities both ways between pairs of response categories, within the constraints of the marginal probabilities being those values observed, that is, the latter are not necessarily symmetric. In terms of log linear models, this is called *quasi-symmetry*. It can be fitted by creating a special factor variable with pairs of identical values for symmetric positions on either side of the main diagonal. This variable is fitted along with the two marginal parameters, whereas the main diagonal is weighted out (as in the mover–stayer model).

Combining this with marginal homogeneity, by removing the marginal parameters, we obtain complete *symmetry* (about the main diagonal) in the table, as described by Equation (2.3).

Example

A panel of 1651 voters in the elections in Sweden were randomly drawn from the electorate and followed from 1964 to 1970. They were interviewed immediately after each election. The results for panel members who voted for one of the four major parties are given in Table 2.4 for the elections of 1968 and 1970. Those who abstained or who voted for a minor party are omitted. The parties have been arranged from left to right. With the relatively large diagonal values, we may expect that a mover–stayer model would probably provide a substantial improvement over independence. However, there also appears to be a "distance" effect, in that a defecting voter seems more likely to switch to a nearby party on the left–right scale. We shall now examine this latter phenomenon more closely. Let us see how symmetry models can help us in this.

TABLE 2.4. Sweden election votes in 1968 and 1970. (Upton, 1978, p. 128, from Sarlvik)

	1970				
1968	SD	C	P	Con	Total
SD	850	35	25	6	916
C	9	286	21	6	322
P	3	26	185	5	219
Con	3	26	27	138	194
Total	865	373	258	155	1651

SD - Social Democrat C - Centre
P - People's Con - Conservative

For these data, the equilibrium or marginal homogeneity model has a deviance of 65.2 (AIC 91.2) with three d.f., whereas the reversibility or quasi-symmetry model

$$\texttt{VOTE68} + \texttt{VOTE70} + \texttt{SYMMETRY}$$

has 2.5 (28.5) with three d.f., as compared to an AIC of 32 for the saturated model. The complete symmetry model

$$\texttt{SYMMETRY}$$

has deviance and degrees of freedom that are each the sum of those for the two models just given. The overall election results changed, but, given this, the transfers between parties were equal in both directions. They are highest between adjacent parties. As we can see in the table, between the two years there was a shift in the margins toward the two central parties.□

From now on in the examples, the AIC will be indicated in parentheses after the deviance.

2.2.3 Diagonal Symmetry

Up until now, the models have not taken into account any ordering or distance among the response states. Suppose, for example, that all changes in either direction between adjacent categories have equal probability. In the same way, all changes two states apart have the same probability, and so on. This is called a *symmetric minor diagonals* model. In terms of Markov chains, it is a *random walk* without *drift*, where, however, jumps of more than one unit are allowed. (A strict random walk, with only one unit jumps, would only have positive frequencies, and probabilities, on the first two minor diagonals, one on either side of the main diagonal.)

Suppose that we keep the model for constant probabilities between all pairs of states the same distance apart. But now let us allow them to be

different in opposing directions. We, then, have an *asymmetric minor diagonals* model. This is a random walk with drift, because the probability will be higher to shift in one direction on the scale than in the other.

These models may be fitted by constructing the appropriate factor variable with a constant value for all positions on the pertinent diagonal or pair of diagonals.

Example

For the Swedish election data, the symmetric minor diagonals model has a deviance of 35.9 (55.9) with six d.f., and the asymmetric, 29.6 (53.6) with four d.f. Neither is an acceptable simplification of the quasi-symmetry model. This implies that probabilities of change among parties equal distances apart are not the same. □

2.2.4 Long-term Dependence

If we have panel data over more than two points in time, we can study more long-term dependence. We simply use the series of previous responses as explanatory variables in the log linear or logistic regression model, with an additional dimension to the table for each point in time. For example, if the response at any given point in time only depends on the immediately preceding one, we have a *first-order* Markov chain, the model we have been using above, with two points in time. This hypothesis can easily be checked by seeing if the log linear regression parameters for dependence further back in time could be set to zero.

Another possibility is to suppose that the way that the response depends on previous responses is identical at all points in time, called *stationarity*. With a sufficiently long series, this can also easily be checked by setting up the appropriate log linear regression model. These models will be further discussed in the Chapter 4.

2.2.5 Explanatory Variables

Usually, we also want to study how the responses depend on other explanatory variables, such as sex, age, marital status, medical treatment, and so on. Each such variable creates an additional dimension to the table. However, the variables are still introduced into the log linear model in exactly the same way.

Thus, for one such variable, our original model would become

$$\log(\pi_{i|jk}) = \phi + \nu_i + \mu_j + \theta_k + \alpha_{ij} + \beta_1 x_{ik} + \beta_2 x_{jk}$$

or

$$\texttt{R1} + \texttt{R2} + \texttt{EXP} + \texttt{R1} \cdot \texttt{R2} + \texttt{R1} \cdot \texttt{EXPL} + \texttt{R2} \cdot \texttt{EXPL}$$

where R1 and R2 are the responses at the two time points and EXPL is the explanatory variable. Here, α_{ij} or R1 · R2 measures the dependence between the two successive response, as before, whereas β_1 or R1 · EXPL measures the dependence of the first response on the explanatory variable, and β_2 or R2 · EXP that of the second response. Higher order interactions might also be introduced, if necessary.

As the number of such explanatory variables increases, the size of the table grows dramatically. Often, it is preferable to model the individual responses directly. This has the added advantage that the exact values of continuous explanatory variables can be used, instead of categorizing them in a table. Such models will be discussed further in Chapter 5.

2.3 Overdispersion

The binomial and Poisson distributions are members of the one-parameter exponential family. As we saw in Section 1.2.3, the variance of the distribution is related to the mean in a fixed way, $\text{var}[Y_i] = \tau^2$ (that is, $n\pi[1-\pi]$ and μ, respectively). Often, for count data, that is, several events recorded on the same units, as with the cell counts on patients in the introductory example of Chapter 1, the observed variance is greater than this, because the events will be interdependent. This is known as *overdispersion*. It may arise in several ways.

- The subjects in a group may not be homogeneous, as for the cell counts, so that π or μ is, in fact, different for different members. This could be corrected by introducing supplementary variables into the model to explain these differences, variables that, however, will usually not be available. This phenomenon is known variously as proneness, ability, or frailty, depending on the field of application, when due to inherent characteristics, but may also arise from differing environments.

- Each subject may start with the same mean parameter, but this may evolve over the time in which the count of events is being made, perhaps depending on previous events. This can only be detected, and distinguished from the first type, by recording the timing over all events. It may be a birth, contagion, or learning process.

To distinguish between these two types of explanations, we require the timings of the events. Then, we can use the models that will be developed in Chapters 5 and 7.

2.3.1 Heterogeneity Factor

When only the total counts of events on a unit are available, the simplest correction for making certain types of approximate inferences is to introduce a *heterogeneity factor*, ϕ, into the variance:

$$\text{var}[Y_i] = \phi\tau^2$$

Then, for the overdispersed binomial distribution, we have $n\phi\pi(1-\pi)$ and, for the Poisson, $\phi\mu$. This is not a statistical model, properly speaking; it has no probabilistic basis. All that it does is provide a correction to the standard errors, something that is not too useful in the direct likelihood approach where it is known that standard errors only give a crude approximation to the parameter precision obtained from an asymmetric likelihood function.

As with the normal distribution, the mean deviance for some maximal model can be used to provide an estimate for this parameter. Such a correction to inferences is usually only satisfactory if the number of observations under each explanatory variable condition is roughly equal.

2.3.2 Random Effects

A more complex, but more satisfactory, solution through modelling is to assume that the mean parameter, that is assumed to be varying in an unknown way among subjects, has some random distribution. This is known as a random effects model, obtained as a *compound distribution*. This corresponds to modelling the first way in which overdispersion might arise, as described above. (We look at the second way in Chapter 5.)

As we can see in Appendix A, each member of the exponential dispersion family has a corresponding compounding distribution, known as its *conjugate*, that yields an analytically closed-form compound distribution. For an exponential family distribution

$$f(y;\theta) = \exp[y\theta - b(\theta) + c(y)]$$

the conjugate distribution for the random parameter is

$$p(\theta;\zeta,\gamma) = \exp\left[\zeta\theta - \gamma b(\theta) + s(\zeta,\gamma)\right]$$

where $s(\zeta,\gamma)$ is a term not involving θ. This conjugate is also a member of the exponential family. The resulting compound distribution, for n observations, is

$$f(y;\zeta,\gamma) = \exp[s(\zeta,\gamma) + c(y) - s(\zeta+y,\gamma+n)]$$

This is not generally a member of the exponential family.

Examples

1. For the binomial distribution, the conjugate is the beta distribution,

$$p(\pi;\zeta,\gamma) = \frac{\pi^{\zeta-1}(1-\pi)^{\gamma-1}}{B(\zeta,\gamma)}$$

where $B(\cdot)$ is the beta function, and the compound distribution is called beta-binomial:

$$f(y;\zeta,\gamma) = \binom{n}{y}\frac{B(\zeta+y,\gamma+n-y)}{B(\zeta,\gamma)}$$

2. For the Poisson distribution, the conjugate is the gamma distribution,

$$p(\mu;\zeta,\gamma) = \frac{\mu^{\zeta}\gamma^{\zeta+1}\mathrm{e}^{-\gamma\mu}}{\Gamma(\zeta+1)}$$

where $\Gamma(\cdot)$ is the gamma function, yielding a negative binomial distribution:

$$\begin{aligned} f(y;\zeta,\gamma) &= \frac{\Gamma(y+\zeta+1)}{\Gamma(\zeta+1)y!}\gamma^{\zeta+1}(\gamma+1)^{-(y+\zeta+1)} \\ &= \frac{\Gamma(y+\zeta+1)}{\Gamma(\zeta+1)y!}\left(\frac{\gamma}{\gamma+1}\right)^{\zeta+1}\left(\frac{1}{\gamma+1}\right)^{y} \end{aligned} \tag{2.4}$$

□

Another common possibility is to use a normal compounding distribution, but then the model cannot be written in closed form (unless the original distribution is normal), and numerical integration is necessary in order to fit it. Still another approach will be presented in Section 3.3.2.

2.3.3 Rasch Model

The mover–stayer model is a special case of a random effects model in which there are only two values of the random parameter(s), that is, two latent groups in the population. If a set of discrete responses is available for each individual, more elaborate groups can be devised. Thus, for example, with R binary responses, the n individuals might be distinguished by their $R+1$ different possible total numbers of positive responses. This is the Rasch model that has its origin in educational research.

In educational testing, items in a test often have a binary, true/false, response. Each subject replies to a series of questions making up the test. Responses will vary according to the ability of the subject and to the difficulty of each item. The latter may be assumed to have some latent, or unobserved, variable in common, so that taking the total number of positive responses makes sense. Rasch (1960) introduced a binary data model

whereby the probability of response y_{ik} of the subject i to item k is given by

$$\Pr(y_{ik}|\kappa_i) = \frac{e^{y_{ik}(\kappa_i - \upsilon_k)}}{1 + e^{\kappa_i - \upsilon_k}} \tag{2.5}$$

The data are represented by an $n \times R$ matrix of zeros and ones.

In order to allow for variation among individuals, Rasch proposed using a conditional likelihood approach, because conditioning on the marginal totals, $y_{i\bullet}$, eliminates the nuisance parameter, κ_i, from the likelihood function.

Subsequently, Tjur (1982) showed that the conditional model can be fitted as a log linear model. The margins for each item, $\mathtt{R}_k$, are fitted, as well as a factor variable for `TOTAL` score, with $R + 1$ possible values,

$$\mathtt{R}_1 + \cdots + \mathtt{R}_R + \mathtt{TOTAL}$$

This can also be thought of as a model for quasi-independence in a $2^R \times (R+1)$ table containing structural zeros, because each combination of item responses can only give one score. It is also a generalization of the quasi-symmetry model, because all units with the same total number of correct responses are treated symmetrically. Differences among groups can also be introduced into the model.

This model obviously has much wider application than simply to educational testing.

Example

A Danish Welfare Study looked at various personal hazards, as given in Table 2.5. Each question had five parts, yielding five binary responses. The question on work asked if the person was often exposed to noise, bad light, toxic substances, heat, and dust. That for psychic inconvenience asked if the person suffered from neuroses, sensitivity to noise, sensitivity to minor difficulties, troubling thoughts, and shyness, whereas that for physical inconveniences asked about diarrhea, back pain, colds, coughing, and headaches. (For work hazards, "yes, always" and "yes, sometimes" have been grouped together as the positive response.) Thus, there is obviously no relationship between, say, the first question of each of the three types.

Because there is no such connection among questions, we may expect that the mean number of positive responses will vary according to subject area. We can begin by ignoring differences among individuals and fit models for independence among questions. The simplest model is

$$\mathtt{Q1} + \mathtt{Q2} + \mathtt{Q3} + \mathtt{Q4} + \mathtt{Q5} + \mathtt{TYPE}$$

that does not allow for these differences among the three types. It has a deviance of 6803.3 with 88 d.f. (AIC 6819.3), clearly very bad. Introducing

TABLE 2.5. Answers to questions on work, psychic inconvenience, and physical inconvenience in the Danish Welfare Study. (Andersen, 1991, p. 483)

Answer	Work	Psychic	Physical
YYYYY	70	34	16
YYYYN	15	10	5
YYYNY	34	21	24
YYYNN	6	17	16
YYNYY	39	17	11
YYNYN	21	4	12
YYNNY	38	15	79
YYNNN	49	20	98
YNYYY	103	63	18
YNYYN	39	21	11
YNYNY	129	45	19
YNYNN	66	38	15
YNNYY	115	42	9
YNNYN	116	29	9
YNNNY	217	65	54
YNNNN	409	92	97
NYYYY	4	14	37
NYYYN	8	4	38
NYYNY	7	35	73
NYYNN	3	32	82
NYNYY	12	21	55
NYNYN	22	12	70
NYNNY	16	99	365
NYNNN	60	172	689
NNYYY	24	46	30
NNYYN	27	27	61
NNYNY	54	115	84
NNYNN	99	176	178
NNNYY	63	98	44
NNNYN	193	107	131
NNNNY	168	776	454
NNNNN	1499	2746	2267

the interaction between type and question,

$$(\texttt{Q1} + \texttt{Q2} + \texttt{Q3} + \texttt{Q4} + \texttt{Q5}) * \texttt{TYPE}$$

gives a deviance of 4015.2 with 78 d.f. (4051.2).

We can now try to take into account individual variation among respondents by classifying them using the only information we have about them, that is, according to the number of positive replies,

$$(\texttt{Q1} + \texttt{Q2} + \texttt{Q3} + \texttt{Q4} + \texttt{Q5}) * \texttt{TYPE} + \texttt{SCORE}$$

This model has a deviance of 516.2 with 74 d.f. (560.2), showing that a very large variability among individuals is present. These differences among individuals may not act in the same way for all three types of questions, yielding the model

$$(\texttt{Q1} + \texttt{Q2} + \texttt{Q3} + \texttt{Q4} + \texttt{Q5} + \texttt{SCORE}) * \texttt{TYPE}$$

that has a deviance of 403.1 with 66 d.f. (463.1), again a considerable improvement.

However, although we have greatly improved the fit, this final model is still not satisfactory. (The saturated model has an AIC of 192.) Dependence among the replies to the five questions for an individual are not completely explained by heterogeneity of the participating respondents.

Because of the poor fit, it may be useful to examine the residual plots (Appendix B). Cook's distances are given in Figure 2.1. We discover one very large value, that corresponding to the frequency of 776, the second last in the column psychic. (Elimination of this value, a procedure not to be recommended without good reason, reduces the deviance by 28.5 for one d.f., with an AIC of 434.6.) The Q–Q plot is shown in Figure 2.2. It is far from the 45° straight line, confirming the global lack of fit of the model. □

Residual plots break down the global fit of a model, as measured by its deviance, into its component parts. If the fit is good, none of the components can generally be exceptionally large. Thus, such plots are often a useful procedure for examining the details of the fit of an unsatisfactory model. For reasons of space, they will not be discussed in the examples of the following chapters.

Summary

Standard log linear and logistic models, with their canonical link functions, have proven their worth, based on their ability to decompose probabilities in a multiplicative way. Nothing has been said in this chapter about link functions other than the canonical ones; for example, probit or complementary log log.

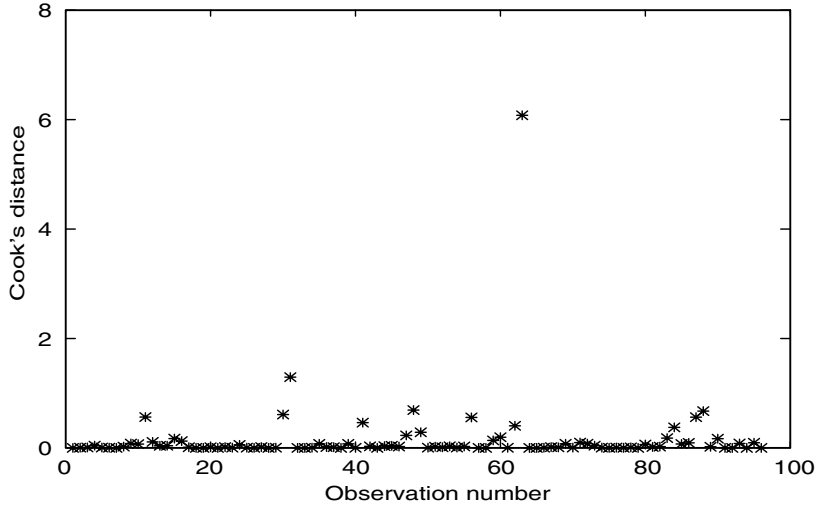

FIGURE 2.1. Plot of Cook's distances for the final model for the Danish Welfare Study of Table 2.5.

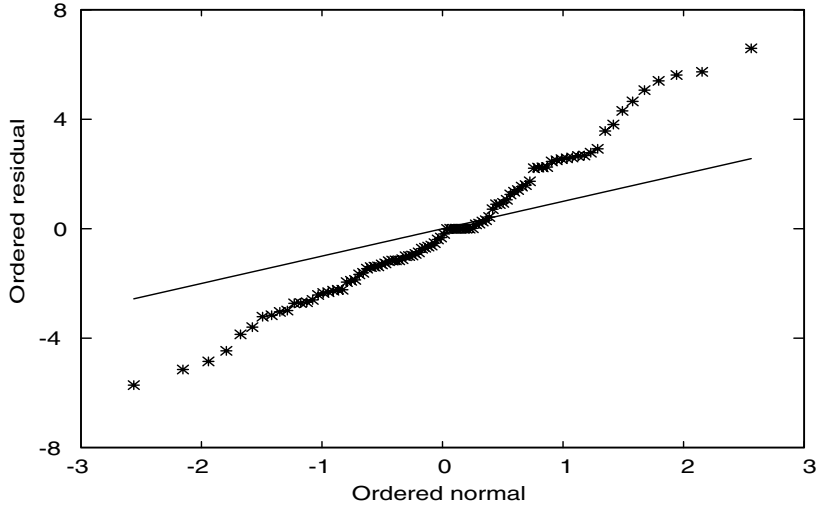

FIGURE 2.2. Q–Q plot for the final model for the Danish Welfare Study of Table 2.5.

Thus, categorical data models have become one of the most important areas of statistical modelling. In applied statistics, they have clearly displaced the monopoly of the normal linear models of earlier years. The general methods are well known and cannot be reproduced in detail here. A number of special cases will be covered in subsequent chapters. Many specialized books are now available showing various aspects of categorical data analysis. The reader may like to consult Agresti (1990), Collett (1991), Lindsey (1995b), and Morgan (1992).

2.4 Exercises

1. The following table provides voting changes between the British elections of 1964 and 1966 (Upton, 1978, p.111, from Butler and Stokes) for a panel of voters.

	1964			
1966	Conservative	Liberal	Labour	Abstention
Conservative	157	4	17	9
Liberal	16	159	13	9
Labour	11	9	51	1
Abstention	18	12	11	15

The panel members are those staying in the same constituency over the three-year period. They had only these three choices of parties for which they could vote. Explore possible models for these data. We are particularly interested in party loyalty and distance of moves among parties.

2. A classical study of British intergenerational social mobility (Bishop *et al.*, 1975, p. 100, from Glass) gave following results:

	Father				
Son	A	B	C	D	E
A	50	45	8	18	8
B	28	174	84	154	55
C	11	78	110	223	96
D	14	150	185	714	447
E	3	42	72	320	411

(The categories are A: Professional, high administrative, B: Managerial, executive, high supervisory, C: Low inspectional, supervisory, D: Routine nonmanual, skilled manual, and E: Semiskilled and unskilled manual.) Develop a model for the dependence of the son's profession on the father's. Of special interest will be how far along

the profession scale a family can move in one generation. Is there a drift in one direction or the other?

3. Employees aged 30–39 in Royal Ordinance factories from 1943 to 1946 had eye tests for unaided distance vision (Stuart, 1953, from the Association of Optical Practitioners):

Right eye grade	Left eye grade				
	4	3	2	1	Total
	Female				
4	1520	266	124	66	1976
3	234	1512	432	78	2256
2	117	362	1772	205	2456
1	36	82	179	492	789
Total	1907	2222	2507	841	7477
	Male				
4	821	112	85	35	1053
3	116	494	145	27	782
2	72	151	583	87	893
1	43	34	106	331	514
Total	1052	791	919	480	3242

Study the relationship between the two eyes. For example, do both eyes of a person generally have equally good vision? Is a left eye being poorer than the right as probable as the reverse? Do any such results that you find hold for the two sexes?

4. In a study of the social system of adolescents, all students in ten high schools in the United States of America were given several questions about the "leading crowd". Among other things, they were asked if they considered themselves to be members. One attitude question also asked was if being a member obliges one sometimes to go against his or her principles. Responses were recorded at the beginning and end of the school year, with the results for the boys given below (Coleman, 1964, p. 171):

Attitude 1	Member 1	Member 2	Attitude 2	
			Favourable	Unfavourable
Favourable	Yes	Yes	458	140
Unfavourable			171	182
Favourable	No		184	75
Unfavourable			85	97
Favourable	Yes	No	110	49
Unfavourable			56	87
Favourable	No		531	281
Unfavourable			338	554

How are membership and attitude interrelated? How do responses at the end of the year depend on the earlier ones?

5. The Six Cities study looked at the longitudinal effects of air pollution on health. Part of the data concern children in Steubenville, Ohio, USA, who were followed each year from ages seven through ten (Fitzmaurice and Laird, 1993). Each year, the wheezing status of the child was recorded. Whether the mother smoked or not when the child was seven is also available (although this could have evolved over time).

Age				Smoking	
7	8	9	10	No	Yes
No	No	No	No	237	118
			Yes	10	6
		Yes	No	15	8
			Yes	4	2
	Yes	No	No	16	11
			Yes	2	1
		Yes	No	7	6
			Yes	3	4
Yes	No	No	No	24	7
			Yes	3	3
		Yes	No	3	3
			Yes	2	1
	Yes	No	No	6	4
			Yes	2	2
		Yes	No	5	4
			Yes	11	7

Ignoring for the moment the fact that these are longitudinal data, determine if there is heterogeneity among the respondents. Does the child's wheezing depend on the smoking status of the mother?

6. Responses to three questions on abortion in surveys conducted over three years (Haberman, 1979, p. 482) are given below.

	Year		
Response	1972	1973	1974
YYY	334	428	413
YYN	34	29	29
YNY	12	13	16
YNN	15	17	18
NYY	53	42	60
NYN	63	53	57
NNY	43	31	37
NNN	501	453	430

The questions were: Should a pregnant woman be able to obtain a legal abortion (1) if she is married and does not want more children; (2) if the family has very low income and cannot afford any more children; (3) if she is not married and does not want to marry the man? The respondents were white Christians questioned by the General Social Survey in the United States of America. This is not a panel, so that the subjects involved are different each year. Interest centres on what changes in attitude to abortion are occurring over the years? Use a Rasch model to control for variability among respondents in the surveys when you study this question.

This page intentionally left blank

3

Fitting and Comparing Probability Distributions

3.1 Fitting Distributions

We have now seen how to analyze events by means of statistical models. Next, we are going to discover that *all* statistical observations can be analyzed as events. Even the record of a theoretically continuous variable, such as the weight of a person at a given point in time, is an event, and a discrete one at that, because any instrument can only measure to a finite precision.

Thus, in Chapter 2, we saw how a multinomial distribution can be fitted as a product of Poisson distributions if we condition on the total number of observations. A multinomial distribution is the most general distribution for independent observations, in the sense that it makes the least assumptions about the form. Some would call it a "nonparametric" model because it follows exactly the empirical data without imposing a smooth functional form. We can, however, go further and add structure to the distribution, making it parametric. We can impose a functional form defining a relationship among the probabilities (Lindsey, 1974a, 1995b; Lindsey and Mersch, 1992).

3.1.1 Poisson Regression Models

A Poisson regression model allows the frequencies of events, as represented in a contingency table, to depend on one or more variables, provided that the events are independent given these variables. Thus, in its general canonical

TABLE 3.1. People recalling one stressful event in the preceding 18 months. (Haberman, 1978, p. 3)

Month	1	2	3	4	5	6	7	8	9
Respondents	15	11	14	17	5	11	10	4	8
Month	10	11	12	13	14	15	16	17	18
Respondents	10	7	9	11	3	6	1	1	4

form, it is

$$\log(\mu_i) = \sum_j \beta_j x_{ij}$$

Conditional on these variables, the events are assumed independently to follow a Poisson distribution. This is a quite general result that turns out to be applicable to most observable phenomena for which statistical models are used. It depends on two critical facts:

- theoretically continuous variables can only be observed to finite precision, and
- a joint distribution for ordered dependent observations can be decomposed into a product of independent conditional distributions; see Equation (5.1).

Thus, one major exception to such applications is spatial data (Chapter 8), because they cannot generally be ordered.

The simplest example of a contingency table is one-dimensional; it can be represented by a histogram, where the categories are the values of the response variable. The multinomial model applied to such data is just a "nonparametric" model, a saturated model with one parameter for each category, the only constraint being that the sum of the probabilities for all categories be unity. However, if the values of the variable have more structure than just being nominal, we can make the probabilities of the categories, estimated by using the frequencies in the histogram, depend in some way on these values. In order to see how to proceed, it will be useful first to look at an example.

Example

Consider a Poisson regression model applied to a study on the relationship between life stresses and illnesses. One randomly chosen member of each randomly chosen household in a sample from Oakland, California, USA, was interviewed. In a list of 41 events, respondents were asked to note which had occurred within the last 18 months. The results given in Table 3.1 are for those recalling only one such stressful event. The canonical Poisson

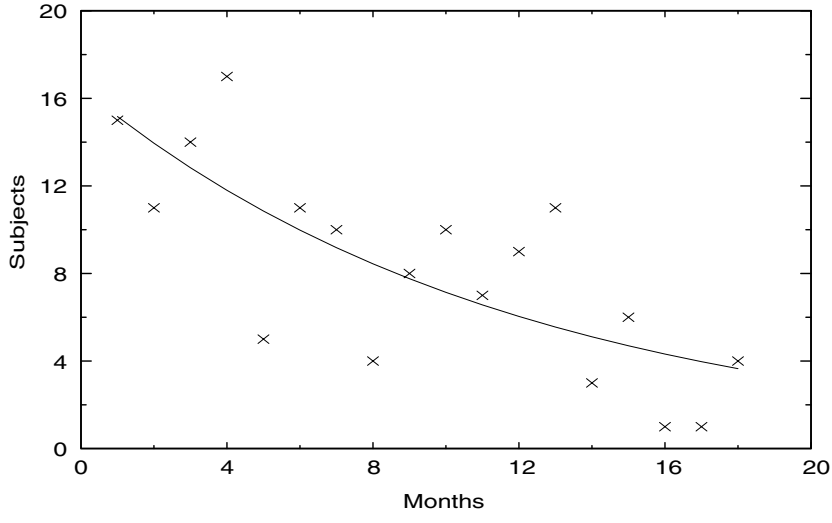

FIGURE 3.1. Data on recall of events over 18 months, with the fitted Poisson regression model.

regression model for these data would be

$$\log(\mu_i) = \beta_0 + \beta_1 y_i \tag{3.1}$$

where μ_i is the average number of respondents giving month i and $y_i = i$ is the month or simply

```
MONTHS
```

The maximum likelihood estimates are $\hat{\beta}_0 = 2.803$ and $\hat{\beta}_1 = -0.0838$. The model can be rewritten

$$\begin{aligned} \tilde{\mu}_i &= \mathrm{e}^{\hat{\beta}_0 + \hat{\beta}_1 y_i} \\ &= 16.494\mathrm{e}^{-0.0838 y_i} \end{aligned}$$

where the tilde indicates that $\tilde{\mu}_i$ is not the maximum likelihood estimate of the mean but is calculated from the estimates for the regression. This is a model of exponential decline (Chapter 4). The data and the estimated model for the mean (remember Figures 1.1 and 1.2) can be plotted as in Figure 3.1

Let us now look at these same data in another way. We take Y_i, the number of months to a recalled event, as our random variable, with an exponential distribution:

$$f(y_i; \phi) = \phi \mathrm{e}^{-\phi y_i} \tag{3.2}$$

We can already notice the similarity with the equation for the estimate of the mean just given. However, our data are grouped into one-month intervals, so that we have

$$\begin{aligned} \Pr\left(y_i - \frac{\Delta_i}{2} < Y_i \leq y_i + \frac{\Delta_i}{2}\right) &= \int_{y_i - \frac{\Delta_i}{2}}^{y_i + \frac{\Delta_i}{2}} \phi \mathrm{e}^{-\phi u_i} du_i \\ &\doteq \phi \mathrm{e}^{-\phi y_i} \Delta_i \end{aligned} \tag{3.3}$$

where Δ_i is one month. Let $\pi_i = \Pr(y_i - \Delta_i/2 < Y_i \leq y_i + \Delta_i/2)$, the multinomial probabilities for the table. Multiplying each side of Equation (3.3) by $n_\bullet$ and taking logarithms, we obtain the following relationship among these probabilities:

$$\begin{aligned} \log(n_\bullet \pi_i) &= \log(n_\bullet \phi \Delta_i) - \phi y_i \\ &= \beta_0 + \beta_1 y_i \end{aligned}$$

where $\beta_0 = \log(n_\bullet \phi \Delta_i)$ and $\beta_1 = -\phi$ in Equation (3.1). We thus discover that this is identical to our Poisson regression model above, with $\mu_i = n_\bullet \pi_i$. □

This surprisingly simple result can easily be generalized.

3.1.2 Exponential Family

Our exponential distribution is one of the simplest members of the *linear* exponential family (Section 1.2.1):

$$f(y_i; \boldsymbol{\theta}) = \exp[\boldsymbol{\theta}^T \mathbf{t}(y_i) - b(\boldsymbol{\theta}) + c(y_i)]$$

where $\mathbf{t}(y_i)$ is the vector of sufficient statistics for the canonical parameters, $\boldsymbol{\theta}$. However, for any empirically observable data, the probability function, $\pi_i = \Pr(y_i - \Delta_i/2 < Y_i \leq y_i + \Delta_i/2)$, should be used in the likelihood function, even if the variable is theoretically continuous. We, thus, apply the same approximation to the integral, as above, for all continuous variables.

If we, again, take logarithms, we may note that

$$\log(n_\bullet \pi_i) = \boldsymbol{\theta}^T \mathbf{t}(y_i) - b(\boldsymbol{\theta}) + c(y_i) + \log(n_\bullet \Delta_i)$$

This, not surprisingly, is linear in $\mathbf{t}(y_i)$ and in $\boldsymbol{\theta}$. Because $c(y_i) + \log(n_\bullet \Delta_i)$ contains no unknown parameters, it can be used as an offset, so that we have

$$\log(n_\bullet \pi_i) = \beta_0 + \boldsymbol{\beta}^T \mathbf{t}(y_i)$$

where $\boldsymbol{\beta} = \boldsymbol{\theta}$ and $\beta_0 = -b(\boldsymbol{\theta})$, the latter being the normalizing constant for the family. (In fact, $\log[n_\bullet]$ can also be absorbed into β_0, and this is often

simpler.) Thus, for frequency data, the usual likelihood function for any distribution from the linear exponential family can be exactly represented as a Poisson linear regression model. The only difference is that the normalizing constant is taken as an additional unknown parameter. Any other distribution can also be represented as a Poisson regression equation, but it will no longer be linear.

In the Poisson regression, the "explanatory" variables are the sufficient statistics of the appropriate member of the exponential family. The model terms for common distributions include

Distribution	Statistic	Offset
Uniform	—	
Geometric	y_i	
Exponential	y_i	
Poisson	y_i	$-\log(y_i!)$
Binomial	y_i	$\log \binom{n_i}{y_i}$
Normal	$y_i,\ y_i^2$	
Inverse Gaussian	$y_i,\ y_i^{-1}$	$-1.5\log(y_i)$
Gamma	$y_i,\ \log(y_i)$	
Pareto	$\log(y_i)$	
Log normal	$\log(y_i),\ \log^2(y_i)$	

Examples

1. For the Poisson distribution with mean, λ, we fit

$$\log(\mu_i) = \beta_0 + \beta_1 y_i$$

with offset, $\log(n_\bullet/y_i!)$, so that $\beta_0 = -\lambda$ and $\beta_1 = \log(\lambda)$. Notice that we are using a Poisson regression model to fit a Poisson distribution, where the latter has no explanatory variables (for the moment).

2. For the normal distribution, with mean, λ, and variance, σ^2, we fit

$$\log(\mu_i) = \beta_0 + \beta_1 y_i + \beta_2 y_i^2$$

so that $\beta_1 = \lambda/\sigma^2$ and $\beta_2 = -1/(2\sigma^2)$, the canonical parameters. □

Thus, a normal distribution can be fitted as a Poisson regression for relatively small samples, whereas a Poisson distribution only approaches normality asymptotically.

3.2 Setting Up the Model

3.2.1 Likelihood Function for Grouped Data

To recapitulate, for any (conditionally) independent set of events, we have the multinomial likelihood function

$$\mathrm{L}(p) \propto \prod \pi_i^{n_i}$$

where n_i is the frequency in category i and

$$\begin{aligned} \pi_i &= f(y_i;\theta), \qquad y \text{ discrete} \\ &= \int_{y_i-\frac{\Delta_i}{2}}^{y_i+\frac{\Delta_i}{2}} f(u_i;\theta)du_i, \qquad y \text{ continuous} \\ &\doteq f(y_i;\theta)\Delta_i \end{aligned}$$

Fixing $n_\bullet = \sum n_i$ is the condition required in order to have a proper probability distribution with $\sum \pi_i = 1$. The result is equivalent to a Poisson likelihood

$$\prod \frac{\mu_i^{n_i}\mathrm{e}^{\mu_i}}{n_i!}$$

where $\log(\mu_i) = \boldsymbol{\theta}^T\mathbf{t}(y_i) - b(\boldsymbol{\theta}) + c(y_i) + \log(n_\bullet\Delta_i)$ for the exponential family.

In log linear models for categorical data, conditioning on the total number of observations in the Poisson likelihood ensures that the total multinomial probability of all categories included in the model equals unity. This is accomplished by keeping the intercept in the model to fix the marginal total. However, we have seen that the intercept is $\beta_0 = -b(\theta)$, the normalizing constant of the exponential family (Section 1.2.1). Thus, this fitting procedure ensures that we have a proper probability distribution that sums to one over the observed categories.

In log linear models, we distinguish between *sampling* and *structural* zeros. Sampling zeros occur for categories that did not happen to be observed, but might have been, in the given data. Structural zeros are categories that are impossible for the given data. These two must be treated differently: categories with sampling zeros are included in a log linear model, whereas those for structural zeros are not. This has pertinence for our present approach.

For a probability distribution, the response variable usually has some range such as $y > 0$. This means that an infinite number of categories, $(y_i - \Delta_i/2, y_i + \Delta_i/2]$, is possible, but only a few have been observed. The others are sampling zeros and should be included in the model. However, most have extremely small probabilities and their exclusion does not affect the accuracy of estimation of the normalizing constant (the intercept), nor of the parameter estimates. Thus, a small number of sampling zeros can be

TABLE 3.2. Employment durations of two grades of Post Office staff. (Burridge, 1981)

	Grade			Grade	
Months	1	2	Months	1	2
1	22	30	13	0	1
2	18	28	14	0	0
3	19	31	15	0	0
4	13	14	16	1	1
5	5	10	17	1	1
6	6	6	18	1	0
7	3	5	19	3	2
8	2	2	20	1	0
9	2	3	21	1	3
10	1	0	22	0	1
11	0	0	23	0	1
12	1	1	24	0	0

added in the tails for enough categories to obtain any required precision of the parameter estimates.

3.2.2 Comparing Models

Once we begin working in the Poisson regression model context, there is no reason why we should restrict ourselves to "explanatory" variables that are the sufficient statistics for only one distribution. We saw that such statistics include y_i, y_i^2, $\log(y_i)$, $\log^2(y_i)$, and y_i^{-1}. Any number of these, and others, can be included in a Poisson regression model for such frequency data and standard model selection techniques used to determine which can be eliminated.

Thus, for example, to compare log normal and gamma distributions, the statistics y_i, $\log(y_i)$, and $\log^2(y_i)$ would be used. If y_i can be eliminated, we have a log normal distribution, whereas, if $\log^2(y_i)$ is unnecessary, we have a gamma distribution. This may also mean that some more complex combination proves necessary, not corresponding to any distribution commonly used.

Example

The employment durations of staff, aged 25 to 44, recruited to the British Post Office in the first quarter of 1973 and classified into two grades, are shown in Table 3.2. Losses were due to resignations, deaths, etc.

Burridge (1981) fits a gamma distribution to these data. The corresponding regression model, ignoring grade, with y_i and $\log(y_i)$ gives a deviance of 77.0 (AIC 85.0) on 44 d.f. If we also try a number other statistics, we find

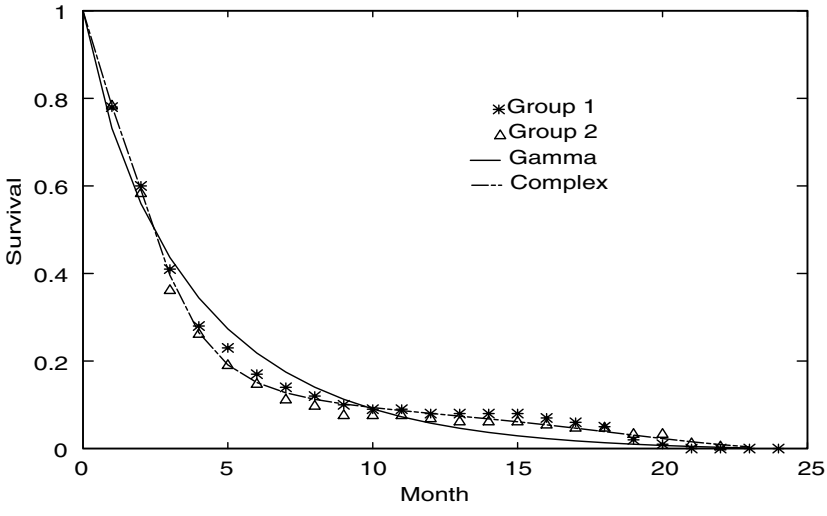

FIGURE 3.2. Fitted gamma and more complex survivor functions for the Post Office data of Table 3.2.

that a model with y_i, y_i^{-2}, $\log(y_i)$, $\log^2(y_i)$, and $\sqrt{y_i}$ has a deviance of 35.4 (49.4) with 41 d.f., although this is still not as good as the AIC for the saturated model (48). This model yields a survivor function, plotted in Figure 3.2, that, at first, drops more quickly than the gamma, but then stays at a higher level for long durations. As can be seen, this follows the observed values much more closely than does the gamma distribution. Examination of the residuals (not shown) reveals no systematic pattern. □

If categorical explanatory variables are available, we have a generalized linear model and the data will take the form of a multidimensional contingency table. Again, standard regression techniques can be applied. The same sufficient statistics for the response are used, but now in interaction with the explanatory variables, yielding the sufficient statistics for the new model. If the interaction terms are not necessary in the model, this implies that the corresponding explanatory variable can be eliminated: the response does not depend on it.

Example

For the Post Office data, there is no evidence of a difference in the distribution of employment durations between the two grades of employees. □

We also have a further generalization: the coefficient for a sufficient statistic may be zero in some categories of an explanatory variable and not in others. This means that we have different distributions for different categories of the explanatory variables. Thus, in one model, we can fit several

distributional forms. This might arise, for example, in a mortality study, where test and control subpopulations have different hazard functions.

3.3 Special Cases

3.3.1 Truncated Distributions

So far, we have only considered sampling zeros in our probability distribution. Suppose, now, that we treat some zero categories as impossible, although they would otherwise have reasonably large probabilities in the model. Then, these are structural zeros and must be left out of the data. In this way, we are fitting a *truncated* distribution. With this regression approach, this can be done even more easily than for complete distributions (because sampling zeros do not need to be added, at least in the direction of truncation). All of the techniques described above can still be used.

Example

Data were obtained from a postal survey, as given in Table 3.3. Obviously, houses with zero occupants could not reply to the postal questionnaire, so that a truncated distribution, without the zero category, is appropriate. For these counts of numbers of people in houses, the truncated Poisson distribution

$$\Pr(y_i;\lambda) = \frac{\mathrm{e}^{-\lambda}\lambda^{y_i}}{(1-\mathrm{e}^{-\lambda})y_i!} \tag{3.4}$$

might be a suitable hypothesis. Notice that this is still a member of the exponential family.

We can extend the vector of frequencies, say to length ten, by including zeros. Then, the number of occupants is fitted with the Poisson offset given above, including $\log(n_\bullet)$. The resulting deviance is 12.6 (AIC 16.6) with eight d.f., although there are now four sampling zeros in the table (the AIC for the saturated model is 20). The parameter estimates are $\hat{\beta}_0 = 6.63$ and

TABLE 3.3. Numbers of occupants in houses replying to a postal survey. (Lindsey and Mersch, 1992)

Occupants	Houses
1	436
2	133
3	19
4	2
5	1
6	0
7	1

$\hat{\beta}_1 = -0.55$. From the latter parameter, we have $\lambda = \exp(\beta_1)$, whereas, from the former, we have $\lambda = \log(e^{-\beta_0} + 1)$. Both calculations yield $\hat{\lambda} = 0.577$ for this truncated Poisson distribution. □

3.3.2 Overdispersion

In Section 2.3, we briefly looked at the problem of overdispersion in categorical data. Several models have been proposed that are difficult to fit by the usual methods, but that can be easily handled by the methods of this chapter. We shall look particularly at overdispersed binomial data, where the variance is larger than that expected for this distribution: $n\pi(1-\pi)$.

Altham (1978) proposes two extensions of the binomial distribution, an additive and a multiplicative one. The latter, which is a member of the exponential family, interests us here:

$$f(y;\pi,\psi) = c(\pi,\psi)\binom{n}{y}\pi^y(1-\pi)^{n-y}\psi^{y(n-y)}$$

with sufficient statistics, y and $y(n-y)$, and offset, $\log\binom{n}{y}$, whereas the binomial distribution only has y.

Efron (1986) develops what he calls the double binomial distribution, also a member of the exponential family:

$$f(y;\pi,\psi) = c(\pi,\psi)\binom{n}{y}\frac{n^{n\psi}}{n^n}\frac{y^y(n-y)^{n-y}}{y^{y\psi}(n-y)^{(n-y)\psi}}\pi^{y\psi}(1-\pi)^{(n-y)\psi}$$

with sufficient statistics, y and $-y\log(y) - (n-y)\log(n-y)$, and offset, $\log\binom{n}{y}$. In the two cases, setting $\psi = 1$ yields a binomial distribution. However, the normalizing constants, $c(\cdot)$ (which are different) are intractable, so that the models are difficult to fit by standard methods.

Example

Consider the numbers of male and female children in families of a fixed size. The data in Table 3.4 are for the first 12 children in 6115 families of size 13, obtained from hospital records in the nineteenth century in Saxony. The deviance for the binomial distribution is 97.0 (AIC 101.0) with 11 d.f., indicating a poor fit as compared with the saturated model that has an AIC of 24. The multiplicative generalization of the binomial has 14.5 (20.5) with ten d.f., whereas the double binomial has 13.1 (19.1) with the same degrees of freedom. The fitted proportions for the three models are plotted in Figure 3.3. □

TABLE 3.4. Numbers of male children among the first 12 children in 6115 families of size 13. (Sokal and Rohlf, 1969, p. 80, from Geissler)

Males	Families
0	3
1	24
2	104
3	286
4	670
5	1033
6	1343
7	1112
8	829
9	478
10	181
11	45
12	7

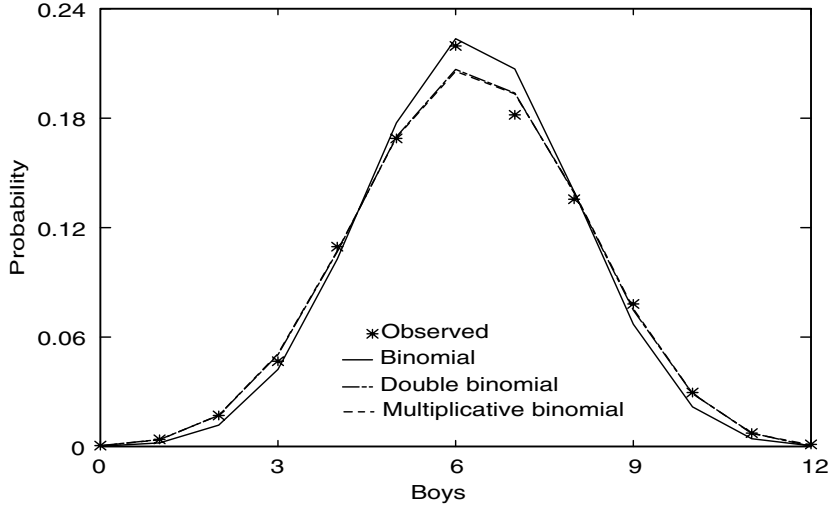

FIGURE 3.3. Fitted binomial, multiplicative binomial, and double binomial models for the sibship data of Table 3.4.

3.3.3 Mixture Distributions

One special application of truncated distributions is of particular importance. Suppose that one category of the response variable contains a mixture of two populations. For example, events may follow a Poisson distribution, but certain, unidentifiable, individuals may not be able to have events. The zero category will be a mixture of cases not happening to have an event and those not able to have one. If this category is structurally eliminated, for example, by weighting it out during the fit, the number of cases not happening to have an event can subsequently be estimated from the model. This is the same procedure that we used in Section 2.2.1 for the mixture on the diagonal in the mover–stayer model.

Example

In the early detection of cancer, the study of genetic lesions, as indicated, for example, by micronuclei counts, is very important (Lindsey and Laurent, 1996). Common factors causing such damage include chemical agents, such as ethylene oxide, and radiation. However, the samples obtained will involve mixtures of affected and unaffected cells. Thus, if the counts of micronuclei in the affected cells are assumed to follow a Poisson distribution, an estimate of the proportion of affected cells can be obtained from a Poisson regression model representation of a truncated Poisson distribution.

Table 3.5 gives data on micronucleus assays exposed to six levels of x-rays, with the age and sex of the ten subjects. Notice the large numbers of zero counts, especially at low levels of x-rays. Under radiation, most, if not all, cells should be affected, especially if, as here, the study is done *in vitro*.

We, first, fit a truncated Poisson model (with zero counts weighted out), assuming a mixture for the cells without micronuclei, in spite of the fact that we expect all cells to be exposed. The model with six discrete dose levels

$$\texttt{FACDOSE} * \texttt{COUNT} + \texttt{SUBJECT} * \texttt{FACDOSE}$$

has a deviance of 276.4 (AIC 465.4, counting each zero count to be estimated as a parameter) with 289 d.f. Adding age,

$$\texttt{FACDOSE} * \texttt{COUNT} + \texttt{AGE} * \texttt{COUNT} + \texttt{SUBJECT} * \texttt{FACDOSE}$$

decreases the deviance by only 0.22 (AIC, 467.1, one d.f.) whereas adding sex,

$$\texttt{FACDOSE} * \texttt{COUNT} + \texttt{SEX} * \texttt{COUNT} + \texttt{SUBJECT} * \texttt{FACDOSE}$$

decreases it by 0.12 (467.2, one d.f.). A linear effect in dose, replacing the factor variable,

$$\texttt{LINDOSE} * \texttt{COUNT} + \texttt{SUBJECT} * \texttt{FACDOSE}$$

TABLE 3.5. Micronucleus assays for peripheral blood lymphocytes obtained from ten people and exposed *in vitro* to six levels of x-rays (0, 0.5, 1, 2, 3, 4 Gy), the lines of the table for each subject. (Thierens *et al.*, 1991)

Number of micronuclei													
0	1	2	3	4	5	6	0	1	2	3	4	5	6
Male, age 24							Female, age 25						
990	10	0	0	0	0	0	987	13	0	0	0	0	0
960	37	3	0	0	0	0	958	39	1	2	0	0	0
919	69	12	0	0	0	0	895	94	11	0	0	0	0
774	187	34	5	0	0	0	794	175	30	1	0	0	0
614	285	85	12	2	2	0	590	281	107	19	2	1	0
429	345	166	47	12	1	0	424	330	169	53	18	6	0
Male, age 28							Female, age 30						
990	9	1	0	0	0	0	975	22	3	0	0	0	0
964	33	3	0	0	0	0	934	65	1	0	0	0	0
925	72	3	0	0	0	0	878	109	12	1	0	0	0
795	182	19	4	0	0	0	756	203	36	5	0	0	0
689	213	79	17	2	0	0	599	285	87	22	6	1	0
516	311	28	31	11	3	0	410	347	168	58	15	2	0
Male, age 42							Female, age 39						
981	17	1	0	1	0	0	985	15	0	0	0	0	0
931	64	4	1	0	0	0	955	42	3	0	0	0	0
878	110	10	2	0	0	0	901	86	11	2	0	0	0
761	206	25	8	0	0	0	775	174	44	7	0	0	0
563	283	120	27	5	2	0	560	308	109	18	4	0	1
456	281	154	71	34	4	0	428	321	163	68	16	3	1
Male, age 50							Female, age 44						
985	14	1	0	0	0	0	983	15	2	0	0	0	0
950	45	5	0	0	0	0	940	55	5	0	0	0	0
900	93	7	0	0	0	0	874	109	16	1	0	0	0
753	206	39	2	0	0	0	740	213	39	8	0	0	0
579	319	78	21	1	1	1	613	295	82	9	1	0	0
421	334	157	72	12	3	1	439	370	155	30	5	1	0
Male, age 54							Female, age 53						
976	21	3	0	0	0	0	971	28	1	0	0	0	0
936	61	3	0	0	0	0	939	53	7	1	0	0	0
895	94	11	0	0	0	0	894	91	14	1	0	0	0
760	207	32	1	0	0	0	759	198	35	5	3	0	0
583	302	97	12	6	0	0							
485	319	147	35	11	2	1	405	331	186	58	13	5	2

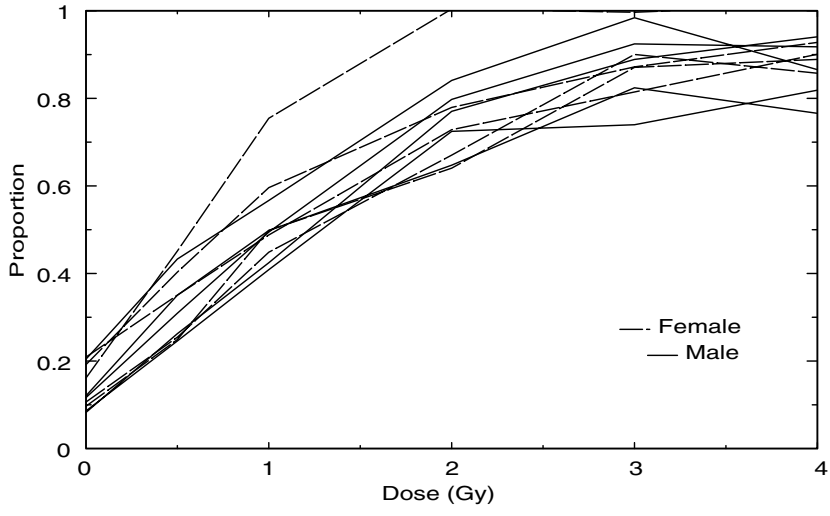

FIGURE 3.4. Variation in the estimated proportion of susceptible cells with radiation dose.

raises the deviance by 6.45 (463.8, four d.f.), whereas log dose

$$\texttt{LOGDOSE} * \texttt{COUNT} + \texttt{SUBJECT} * \texttt{FACDOSE}$$

raises it 45.15 (502.5, four d.f.). Replacing the sex and age variables by a ten-level factor variable distinguishing among the individuals

$$\texttt{LINDOSE} * \texttt{COUNT} + \texttt{SUBJECT} * \texttt{COUNT} + \texttt{SUBJECT} * \texttt{FACDOSE}$$

reduces the deviance (with linear dose) by 78.64 to 204.17 (AIC 403.2) with 284 d.f. Thus, there is more variability among the individuals than can be accounted for by sex and age alone.

If, instead, we fit the untruncated Poisson model to all of the data, we obtain a deviance of 1233.8 (1374.8) with 343 d.f. for this same model with individual differences and linear dose. Thus, the mixture model is very much better for these data. This places in question the idea that all cells are equally affected when exposed *in vitro* to radiation.

Because all cells should be exposed when radiation is applied *in vitro*, these results may indicate that some cells are not susceptible ever to produce micronuclei under these conditions. The estimated proportions of susceptible cells are plotted against dose in Figure 3.4. This proportion ranges from about 25% to 50% for low radiation levels and seems to reach a maximum of between 75% and 100% at the higher doses. □

TABLE 3.6. Numbers of lambs born to ewes in two years. (Plackett, 1965, from Tallis)

	1953		
1952	0	1	2
0	58	26	8
1	52	58	12
2	1	3	9

3.3.4 Multivariate Distributions

If there is more than one response variable, a multivariate distribution can easily be fitted by this method. For the distribution to be proper, the margins for all combinations of the explanatory variables must be included so that the model must be based on Equation (2.2). Interdependencies among the responses are introduced by including interactions among their sufficient statistics.

Example

The numbers of lambs born to sheep in two consecutive years were recorded, as in Table 3.6. We can fit a bivariate Poisson distribution to these data in a way very similar to that used above for the postal survey. With independence between the two years, the number of lambs for each year is just fitted,

$$\texttt{BIRTH52} + \texttt{BIRTH53}$$

with, as offset, minus the log factorial of each of these. The AIC is 74.7, as compared to 18 for the saturated model. If we add the interaction between these two variables to allow for dependence,

$$\texttt{BIRTH52} * \texttt{BIRTH53}$$

the AIC is 62.2, indicating dependence, but still a very bad fit. The lack of fit arises from the small numbers of twins, much less than predicted by a bivariate Poisson distribution. □

Summary

We have seen that, using Poisson regression models for categorical data, we can

- fit a large number of probability distributions, even multiparameter ones not usually considered;
- estimate the normalizing integration constant numerically, even if it is analytically complex;

- fit truncated distributions;
- fit multivariate distributions;
- fit certain simple mixture distributions;
- fit generalized linear models, even with different distributions for different categories of the explanatory variables;
- compare and select models quickly by standard regression techniques.

If nonlinear Poisson regression is available, then any distribution, and not just members of the exponential family, can be fitted.

However, one should not conclude that this procedure should be universally used for all data. Its drawback is that there is an additional parameter, the normalizing constant that is taken to be unknown. It can only be estimated if there are sufficient observations in the various categories of the response variable. In most situations, this constant is a known function of the other parameters, so that more traditional procedures are more efficient. These results do, however, have theoretical interest, showing the unity of all modelling.

For a more detailed presentation of these methods, and other examples, see Lindsey (1995b, pp. 125–149). We shall see in Chapter 7 that the same type of procedures can also be applied to proportional hazards models using a counting process approach.

3.4 Exercises

1. In a classic experiment, Rutherford and Geiger counted alpha particles from radioactive decay of polonium in 72-second intervals (Hand *et al.*, 1994, p. 223, from Rutherford and Geiger):

Scintillations	Frequency	Scintillations	Frequency
0	57	8	45
1	203	9	27
2	383	10	10
3	525	11	4
4	532	12	0
5	408	13	1
6	273	14	1
7	139	15	0

What distribution do you expect would fit these data? Does it?

2. In the effort to improve urban methods of transportation, studies have been made for quite some time in an attempt to estimate the number of passengers carried by each car in urban traffic. As shown in the table below, the numbers of occupants, including the driver, were recorded in passenger cars passing the intersection of Wilshire and Bundy Boulevards in Los Angeles, California, USA, on Tuesday, 24 March 1959, between 10:00 and 10:40 (Derman *et al.*, 1973, p. 278, from Haight):

Occupants	Cars
1	678
2	227
3	56
4	28
5	8
6+	14

Notice that the variable of interest, the number of occupants of a car, must take values greater than or equal to one. Few distributions have this characteristic. This indicates that you may need to consider a truncated distribution or to reconstruct a new variable before proceeding. What distribution might fit these data?

3. Consider another example of the numbers of male and female children in families of a fixed size, similar to that analyzed above. The data in the table below are for all children in 53,680 families of size 8 obtained from the same hospital records in the nineteenth century in Saxony (Fisher, 1958, p. 67, from Geissler):

Males	Families
0	215
1	1485
2	5331
3	10649
4	14959
5	11929
6	6678
7	2092
8	342

Again, one might think that a binomial distribution would be appropriate for these data. What assumptions about births are being made? Why might it be preferable to ignore the last birth, as in the example above? Determine if this model is acceptable and if not, try to find a better one. Because the sample size is so large (almost ten times

that for families of size 12), you may need to adjust the AIC before beginning, as described at the end of Section A.1.4.

4. The following table gives the days between coal-mining disasters in Great Britain, 1851–1962 (Lindsey, 1992, pp. 66–67, from Jarrett, 1979):

Days	Disasters
0–20	28
20–40	20
40–60	17
60–80	11
80–100	14
100–120	6
120–140	13
140–200	17
200–260	16
260–320	11
320–380	13
380–440	3
440–500	4
> 500	17

They concern explosions in mines involving more than 10 men killed, as originally recorded in the *Colliery Year Book and Coal Trades Directory* produced by the National Coal Board in London, UK. Can a distribution be found to describe these data adequately? Notice the fairly large number of very large times, called censored (Chapter 6), characteristic of this type of duration data. These are not truncated, because we know their minimum value (we never even know how many values are truncated). Can some way be devised to take censoring into account?

5. The following table shows the joint distribution of income and wealth (1000 DKr) in Denmark, 1974 (Andersen, 1991, p. 350):

	Wealth				
Income	0	1–50	50–150	150–300	300+
0–40	292	126	22	5	4
40–60	216	120	21	7	3
60–80	172	133	40	7	7
80–110	177	120	54	7	4
110+	91	87	52	24	25

The wealth and income are fairly crudely grouped so that any results can only be approximate. Can you find a bivariate distribution that describes the dependencies between these two variables?

6. Lengths of marriage (years) before divorce were recorded in Liège, Belgium, in 1984 (Lindsey, 1992, pp. 14–15):

Years	Divorces	Years	Divorces
1	3	27	14
2	18	28	17
3	59	29	12
4	87	30	17
5	82	31	10
6	90	32	11
7	91	33	13
8	109	34	7
9	94	35	9
10	83	36	9
11	101	37	9
12	91	38	10
13	94	39	5
14	63	40	3
15	68	41	3
16	56	42	4
17	62	43	6
18	40	44	0
19	43	45	0
20	41	46	1
21	28	47	0
22	24	48	2
23	39	49	0
24	34	50	0
25	14	51	0
26	22	52	1

These data include all divorces in the city that year and are unusual for duration data in that they are retrospective, with all marriages ending at the same time (to the nearest year). Find a distribution to represent these data.

7. In the Type II Coronary Intervention study, patients with Type II hyperlipoproteinemia and coronary heart disease were assigned at random to a daily dose of 24 g of cholestyramine or to placebo. After five years, the numbers of vascular lesions were counted on each patient's angiogram (Barnhart and Sampson, 1995):

Number	Cholestyramine	Placebo
0	5	2
1	4	4
2	6	6
3	5	4
4	7	6
5	7	9
6	6	7
7	6	5
8	7	2
9	2	4
10	2	4
11	1	2
12	0	0
13	0	2
14	1	0

Find an appropriate distribution to determine if there is a difference between the two treatments. Suggest a better way in which the data might have been recorded (Chapter 7).

4
Growth Curves

Longitudinal data involve observations of responses over time that can be modelled as a stochastic process (Lindsey, 1992, 1993). They differ from most other types of data in that the dependence of present response on past history must be taken into account.

Among longitudinal data, growth curves have a number of special characteristics, only some of which they share with other series of observations over time:

- the growth profile will generally be a nonlinear function of time, often reaching an asymptote;
- by definition, growth is not stationary; occasionally, the increments, or innovations, may be;
- random variability will generally increase with size;
- the successive responses are measured on the same subject so that they will generally not be independent;
- different individuals may have different growth rates or profiles, either inherently or due to environmental effects.

In a first step, in this chapter, we shall ignore the possibility of dependence, that is, the last two points, and see how we can use standard regression techniques to handle some of the other characteristics of such curves.

4.1 Exponential Growth Curves

4.1.1 Continuous Response

One of the simplest curves for growth has the nonlinear exponential functional form:

$$y = \alpha e^{\beta t}$$

This will only be realistic in the early stages of growth; nothing can continue growing exponentially forever. On the other hand, with $\beta < 0$, it may be a reasonable model of exponential decline (Section 3.1.1). In addition, as it stands, it is only a deterministic relationship between response and time.

A stochastic element (the "error structure") can be introduced in at least two ways. The log response may have some distribution in the exponential family, such as the normal or gamma, yielding a log normal or log gamma distribution with the identity link and linear predictor,

$$\mu_{\log(y)} = \log(\alpha) + \beta t \tag{4.1}$$

This is the type of model, with a normal distribution, favoured by econometricians, those who seem to believe that almost any response is normal if measured on the log scale!

Another possibility is to use the untransformed response in a normal or gamma distribution with a log link such that the predictor is

$$\mu_y = \alpha e^{\beta t}$$

or

$$\log(\mu_y) = \log(\alpha) + \beta t \tag{4.2}$$

Both of these models are easily fitted as generalized linear models; the resulting curves can differ rather significantly, both in their fit to observed data and in the predictions they yield.

In the first model, the curve goes through the geometrical mean of the data, whereas, in the second, it goes through the arithmetic mean. Note that the variability modelled by the two models is also quite different. For example, for the first model, with a normal distribution, the variance of the log response is constant, implying that the variance of the response is increasing with size. In the second, again with a normal distribution, the variance of the response is constant. If Equation (4.2) is used with a gamma distribution, the ratio of standard deviation to mean, the coefficient of variation, is assumed constant. Other distributions will carry still other assumptions about how the variance is changing with mean response, that is, over time.

TABLE 4.1. Gross domestic fixed capital formation in the United Kingdom, 1948–1967 (read across rows). (Oliver, 1970)

1422	1577	1700	1889	2106	2359	2552	2829	3103	3381
3492	3736	4120	4619	4731	4906	5860	6331	6686	7145

Example

Let us look at a series on one subject, a country. In economics, much attention is given to comparing growth rates among different countries. The gross domestic fixed capital formation plays a key role in economic growth. Models often take it to be growing exponentially, but an important question, before making comparisons, is how to associate a random component with this deterministic statement. Estimates for only one country, the United Kingdom from 1948 to 1967, are shown in Table 4.1 at current prices. Thus, the growth is due both to investment and to rising prices.

A log normal model, using Equation (4.1), is estimated as

$$\mu_{\log(y)} = 7.23 + 0.084t$$

taking $t = 0$ in 1947, so that $\hat{\alpha} = 1376.09$ and $\hat{\beta} = 0.084$. The AIC is 254.3. The normal model, using Equation (4.2), yields

$$\log(\mu_y) = 7.26 + 0.081t$$

with $\hat{\alpha} = 1425.10$ and $\hat{\beta} = 0.081$, not very different. The AIC is 256.9, somewhat worse.

The respective predictions for 1968 are 8047 and 7878; these are quite different. And, indeed, the value actually observed was 7734, indicating that the normal distribution with log link predicted better, at least for that year, although the AIC is poorer.

Another possibility is to fit a gamma distribution with a log link. This gives the equation

$$\log(\mu_y) = 7.23 + 0.084t$$

so that $\hat{\alpha} = 1377.46$ and $\hat{\beta} = 0.084$. The AIC is 254.3 and the prediction for 1968 is 8051, both similar to the log normal. Thus, there is indication of a skewed distribution. These curves are plotted in Figure 4.1 with the data points. The deviations from the fitted line seem to be indicating that some form of dependence among successive observations is present. □

4.1.2 Count Data

When we have count data, the canonical linear predictor for the Poisson distribution has the form of Equation (4.2), so that simple linear Poisson

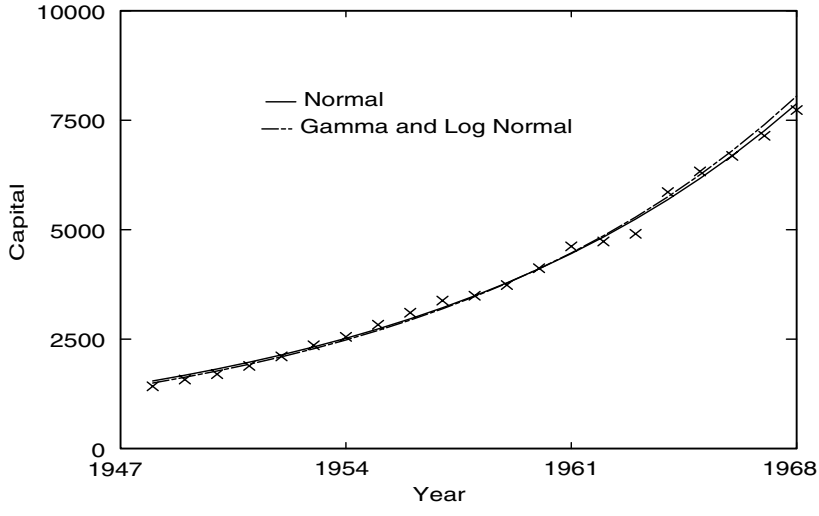

FIGURE 4.1. Fitted curves for the capital formation data in Table 4.1.

regression yields an exponential growth curve. We already studied an example of exponential decay, rather than growth, in Section 3.1.1.

Example

The acquired immune deficiency syndrome (AIDS) has had a substantial impact on the costs of health care for some time now. Thus, it is very important to be able to project the size of the epidemic accurately into the future. Here, we shall look at the increases in the numbers of AIDS cases diagnosed in the United States of America between 1982 and 1990, as shown in Table 4.2. One major problem with such data is the underreporting in the last quarters, because all cases have not yet reached the central recording office. Any projections must take this into account.

The simple exponential growth curve based on the Poisson distribution, using only the data to the middle of 1986 to avoid the problem of reporting delays, gives

$$\log(\mu_y) = 5.45 + 0.166t$$

so that $\hat{\alpha} = 233.1$ and $\hat{\beta} = 0.166$. This is plotted as the solid line in Figure 4.2. The AIC is 421.1. Hopefully, the model is unrealistic! □

4.2 Logistic Growth Curve

With the exponential growth curve, the response continues to increase indefinitely. For many phenomena, this is not reasonable, except over short

TABLE 4.2. AIDS cases reported, by quarter, as diagnosed in the United States of America, 1982–1990. (Hay and Wolak, 1994)

	Quarter			
1982	185	201	293	381
1983	536	705	769	851
1984	1148	1372	1573	1746
1985	2157	2578	2997	3107
1986	3775	4263	4692	4935
1987	5947	6409	6756	6920
1988	7560	7677	7674	7625
1989	8109	8224	7818	6935
1990	5922	329		

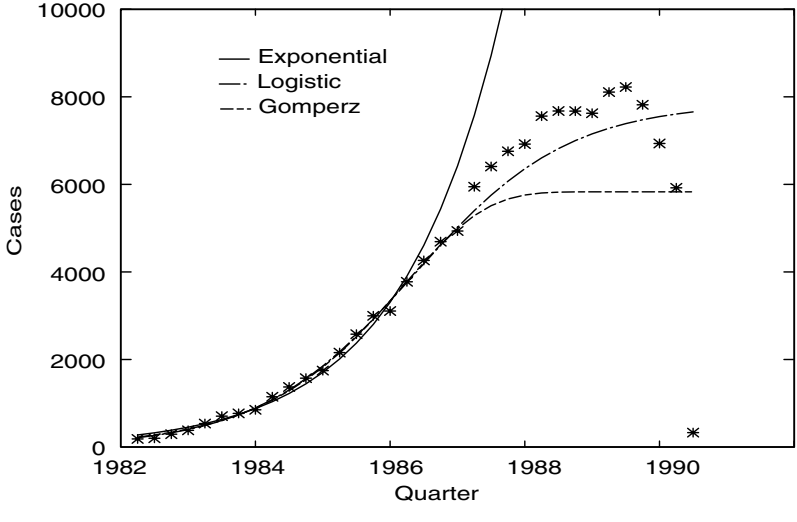

FIGURE 4.2. Fitted curves for the AIDS cases in Table 4.2.

periods. Another possibility is the symmetric S-shaped logistic curve that has the deterministic form

$$y = K\frac{\alpha e^{\beta t}}{1 + \alpha e^{\beta t}}$$

where K is the asymptotic maximum value of the response.

We can transform this to a linear structure by using a logit link:

$$\log\left(\frac{y}{K - y}\right) = \log(\alpha) + \beta t$$

This could easily be fitted as a generalized linear model, if K were known. This is an example of a link function containing an unknown parameter (Section 1.5.2). Notice that such a model does not allow y to become larger than K, even randomly.

When the asymptote is unknown, the model can be fitted for a number of values of $K > \max(y)$, in some search pattern, and that with the maximum likelihood chosen. Note that the deviance produced by most generalized linear models software will not work here; the complete likelihood is necessary because K is contained in the combinatorial not included in the usual deviance.

Example

For the AIDS data, the asymptote is estimated as 7860 cases per quarter using only the data until the middle of 1986. The curve is plotted in Figure 4.2. The AIC is 260.3, a major improvement on the exponential curve. However, this model is also obviously unreasonable, because some of the following partly-recorded quarters already have more cases than predicted at the asymptote. This problem might be overcome by using more of the data in the fit.

The normed profile likelihood for the asymptote is plotted in Figure 4.3 using an AIC-based likelihood interval, that is, a normed likelihood of 1/e. The upper limit of reasonable values of the asymptote lies above the so far observed values. □

One big disadvantage of the logistic curve is that it is symmetric so that the lower bend must be the same as the upper one. In many biological phenomena, this would not be expected to be the case.

4.3 Gomperz Growth Curve

A second commonly used growth curve with an asymptote is the Gomperz curve

$$y = K\left(1 - e^{-\alpha e^{\beta t}}\right)$$

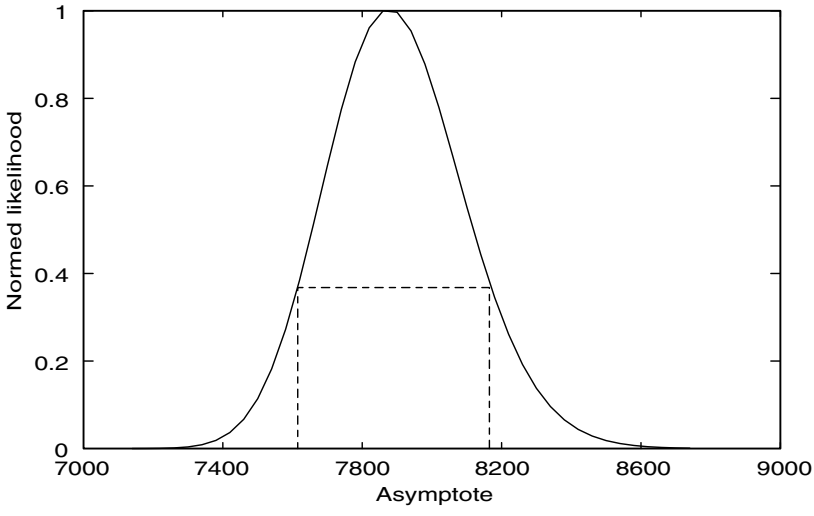

FIGURE 4.3. Normed profile likelihood for the logistic growth curve, with an AIC-based likelihood interval.

where, again, K is the unknown asymptotic maximum value. And again, we can obtain a linear structure, this time for a complementary log log link:

$$\log\left[-\log\left(\frac{K-y}{K}\right)\right] = \log(\alpha) + \beta t$$

We can use the same iterative procedure as before.

Example

For the AIDS data, this asymptote is estimated as 5830 cases per quarter, again using data until the middle of 1986. The curve is plotted in Figure 4.2. This curve, with an AIC of 287.0, is even more unreasonable than the logistic.

The normed profile likelihood for the asymptote is plotted in Figure 4.4. The likelihood interval is even narrower than that for the logistic growth curve, excluding all reasonable values. □

For the data in this example, all of the curves, although very different, follow the observations rather closely over the time period used in fitting the model. We see that it is difficult or impossible to predict the upper part of such a curve only from information on the lower part. We can also see that prediction intervals can be very misleading if the model is not appropriate. Below, we shall look at some more complex models, taking into account the distribution of reporting delays for these AIDS data.

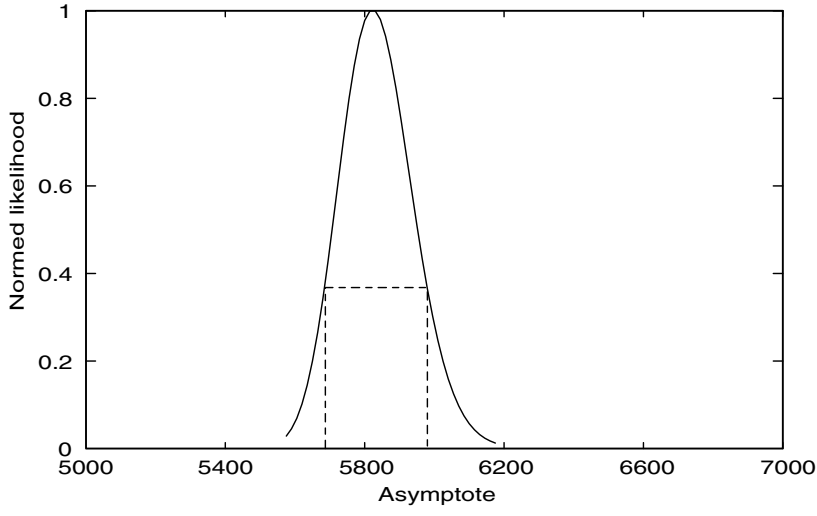

FIGURE 4.4. Normed profile likelihood for the logistic growth curve with an AIC-based likelihood interval.

4.4 More Complex Models

The exponential, logistic, and Gomperz models, in addition to polynomials (that rarely are justified), are those most frequently used for growth curves. However, more complex functions of time can also be used; an example is given in Chapter 10. Obviously, if other supplementary information is available, this should also be incorporated in the model. In most cases, this will involve explanatory variables, but, in some cases, multivariate processes will be required.

A simple function for a bivariate Poisson process can be written

$$\lambda(t, u) = \lambda_{Dt}\lambda_{Ru} \tag{4.3}$$

where, in our example, t will be the diagnosis time with mean incidence, λ_{Dt}, and u the reporting delay, with mean rate, λ_{Ru}, both growth functions of unspecified form. In this model, the two processes are assumed to be independent, so that the reporting delay distribution is the same at all points in time; that is, it is stationary. More complex models can be built up from this, usually by adding the appropriate interaction terms.

Example

For the above data on the growth of the AIDS epidemic, the problem of reporting delays was mentioned. We shall now incorporate this information in the model. In this way, we shall be able to use all of the observations, not just the earlier, complete ones. One problem that was not discussed above

is the large number of observations, 132,170 in the complete data set. This will mean that the usual inference procedures, including the standard AIC, will point to very complex models. In spite of this, I shall continue in the usual way, but discuss the consequences further below.

Data on incidence of AIDS with reporting delays take the form of a rectangular contingency table with observations in one triangular corner missing (see Exercise 4.7 for an example). The two dimensions of such a table are the diagnosis period and the reporting delay. The margin for diagnosis period (as in Table 4.2) gives the total incidence over time, but with the most recent values too small because of the missing triangle of values not yet reported. Hay and Wolak (1994) give diagnoses and reporting delays in the United States of America by quarter, except for an additional zero delay category, yielding a 34×17 table with a triangle of 120 missing cells (not reproduced here, although one margin was given in Table 4.2).

We are now simultaneously interested in how AIDS cases are occurring in the population over time (as above) and how these cases, occurring at a given time point, actually subsequently arrive at the central office. The latter process may be evolving over time, for example, if reporting improves or if the increasing number of cases swamps available facilities. Then our model will be a bivariate process, as described above.

The bivariate Poisson process of Equation (4.3), where we do not specify how the number of cases is growing, just corresponds to a log linear model for independence that can be fitted as

$$\texttt{FACDELAY} + \texttt{FACQUARTER}$$

where `FACDELAY` and `FACQUARTER` are appropriate factor variables. When there are missing cells, we have a quasi-independence model where the triangle of missing data is weighted out (to obtain predictions of incidence below, it must not be simply left out of the model). The fitted values for the diagnosis-time margin in this "nonparametric" (quasi-) stationary model are plotted as the solid line in Figure 4.5. We discover a predicted leveling off of AIDS cases for 1990. For this model, we obtain a deviance of 5381.6 (AIC 5481.6) with 392 d.f., indicating substantial nonstationarity.

A completely "nonparametric" nonstationary model, that is, the saturated model (`FACDELAY*FACQUARTER`), will not provide estimates for the missing triangle of values. Some assumption must be made about the evolution of reporting delays over time, the strongest being stationarity, that is, no change, that we have just used. One possible simple nonstationary model is the following interaction model:

$$\texttt{FACDELAY} + \texttt{FACQUARTER} + \texttt{LINDELAY} \cdot \texttt{FACQUARTER} + \texttt{LINQUART} \cdot \texttt{FACDELAY}$$

where `LINDELAY` and `LINQUART` are linear, instead of factor, variables. With such grouped data for `LINDELAY`, as described above, we use centres of three-month quarterly periods, but with an arbitrary 0.1 value for the zero

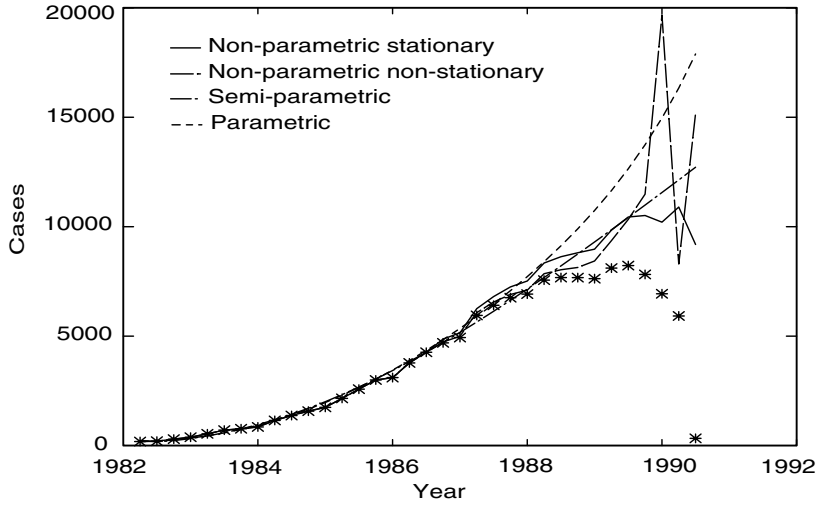

FIGURE 4.5. Estimated AIDS incidence in the U.S.A. taking into account reporting delays.

reporting delay category (to allow for logarithms below). Because we are interested in rates or intensities per unit time (months), we use an offset of log(3) for all delay periods, in all models, except for an arbitrarily log(0.2) for the zero reporting delay. For these data, the deviance decreases by 2670.0 (AIC 2905.6) on 47 d.f. with respect to the (quasi-) stationary model, a strong indication of nonstationarity. For comparison with subsequent AICs, those for these two models are given in line (1) of Table 4.3.

This nonstationary model has been plotted as the dashed line in Figure 4.5. It yields completely unstable predictions for the last quarters. Notice that this and the previous model follow the diagnosed cases exactly for the period until 1989 (24 quarters), where there were no missing cases due to reporting delays. □

A simple parametric bivariate model could have an exponential growth in one dimension and a Weibull distribution for the other (Chapters 6 and 7). This can be written

$$\lambda(t,u) = \alpha e^{\beta_2 u} t^{\beta_1}$$

This can be fitted as a log linear model using a linear time variable for the first dimension and the logarithm of the time for the second, instead of the factor variables above. Again, this is a (quasi-) stationary model. Notice that we do not include factor variables to fix the marginal totals of the contingency table at their observed values.

We can generalize this model by taking other transformations of time. Again, nonstationarity can be introduced by means of interactions between

TABLE 4.3. Deviances for a series of models fitted to the reporting delay data of Hay and Wolak (1994).

Model	Stationary		Nonstationary	
	d.f.	AIC	d.f.	AIC
(1) "Nonparametric"	392	5481.6	345	2905.6
(2) "Semiparametric"	422	6083.4	390	3385.5
(3) Parametric	435	6295.3	426	4020.8

the (transformed) time variables for the two time dimensions. We might consider time, its reciprocal, and its logarithm. Such a bivariate (quasi-) stationary model would be written

$$\lambda(t,u) = \alpha \mathrm{e}^{\beta_1 t + \beta_2/t} t^{\beta_3} \mathrm{e}^{\beta_4 u + \beta_5/u} u^{\beta_6}$$

Example

If we apply this model to the AIDS reporting delays, we use

$$\begin{aligned} &\texttt{LINQUART} + \texttt{RECQUART} + \texttt{LOGQUART} \\ &\quad + \texttt{LINDELAY} + \texttt{RECDELAY} + \texttt{LOGDELAY} \end{aligned}$$

The deviance for this model is 899.7 larger than that for the corresponding stationary "nonparametric" model, on 43 d.f., indicating a considerably poorer model. If we add all nine interaction terms to yield a nonstationary model, we obtain a deviance 2302.5 on nine d.f. smaller than the previous one, but 1277.2 larger than the corresponding nonstationary "nonparametric" model above, however with 81 fewer parameters. Again, the AICs for these two parametric models are summarized in line (3) of Table 4.3.

The fitted values for the margin for diagnoses for this nonstationary parametric model have been plotted, as the dotted line, in Figure 4.5. Notice how the curve no longer follows the random fluctuations of the completed diagnosis counts up until 1987. This accounts for much of the lack of fit of this model. Not surprisingly, the predictions no longer indicate a levelling off.

Much of the apparent lack of fit of the parametric models may be due to the form of the intensity function for delays that is high for short delays, but then descends rapidly for longer delays. This can be checked by fitting a parametric intensity for diagnosis and a "nonparametric" one for the delay. For (quasi-) stationarity, we find a reasonable model to be

$$\texttt{LINQUART} + \texttt{RECQUART} + \texttt{LOGQUART} + \texttt{FACDELAY}$$

where FACDELAY is again a factor variable, and for nonstationarity

$$\begin{aligned} &\texttt{RECQUART} + \texttt{LOGQUART} + \texttt{FACDELAY} \\ &\quad + \texttt{FACDELAY} \cdot (\texttt{LINQUART} + \texttt{LOGQUART}) \end{aligned}$$

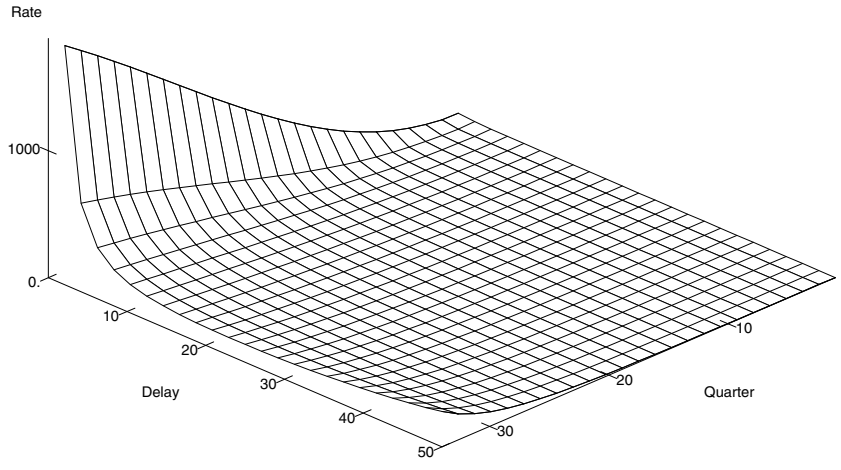

FIGURE 4.6. The estimated rates of AIDS reporting in the United States of America, from the nonstationary parametric model.

with AICs as given in line (2) of Table 4.3. The fitted values for this nonstationary "semiparametric" model give substantially lower prediction than those for our nonstationary parametric model, as can be seen in Figure 4.5. The difference in deviance between them is 707.3 with 36 d.f. These are both quite different from the very unstable nonstationary "nonparametric" model, although the latter has a deviance 569.9 on 45 d.f. smaller than the "semiparametric" model.

Other nonstationary models, parametric for diagnoses, give fairly similar results for the bivariate intensity. By playing with various transformations of time, a somewhat better parametric model can be found, with predictions very close to those for the "semiparametric" model. We discover that returns with short delays are rapidly increasing with time, but those for longer delays are growing even more rapidly. This can be seen from the plot in Figure 4.6. In contrast, a stationary model, such as that used by Hay and Wolak (1994), has reporting rates increasing at all delays in the same way. Thus, the missing triangle will contain estimates that are too low, explaining the leveling off indicated by such a model in 1990.

It is interesting to compare these results with those for England, as seen in Figure 4.7, where delays are becoming shorter so that the stationary model, that ignores this, predicts a rise in AIDS cases. See also Lindsey (1996a) and Exercise 4.7 below. □

Note that we have not attempted to construct a likelihood interval for the estimates, as we did above for the marginal data without reporting

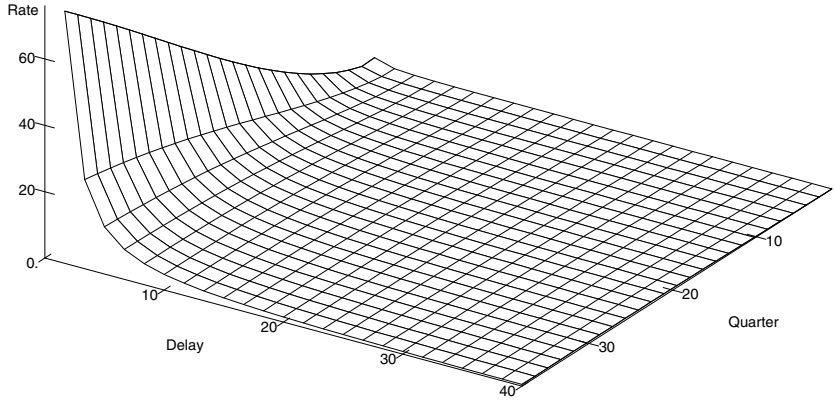

FIGURE 4.7. The estimated rates of AIDS reporting in the United Kingdom, from the nonstationary parametric model.

delays. This would be an essential further step that needs to be carried out before the projections from these models are useful. However, they depend critically on the model, as we saw above. Indeed, the intervals published in the papers cited above for the two sets of data, using stationary models in both cases, do not even cover the point projection from our better fitting nonstationary models!

In this example, the AIC for the saturated model is 916. We saw in Table 4.3 that none of the models fitted comes close to this value. This is due to the large number of observations in this data set. For such models to be selected, a factor considerably greater than two times the number of estimated parameters would have to be added to the deviance. In other words, to compensate for the extremely large number of observations, the smoothing factor, a, of Section A.1.4 would have to be smaller.

Summary

In this chapter, we have been primarily concerned with the nonlinear form of the regression curve for growth data. So far, we have ignored the dependence within the series of values due to observations coming from the same individual(s). We look at this in the next chapter. Combining nonlinear regression models with dependence among observations is only slowly becoming feasible (Chapter 10) in the generalized linear model context. The books on longitudinal data cited in the next chapter are useful references.

4.5 Exercises

1. The area (unreported units) of a bean plant leaf grown at constant temperature was recorded for 36 consecutive days, with measurements starting at day 17 (Scallon, 1985) (read across rows):

0.745	1.047	1.695	2.428	3.664	4.022	5.447
6.993	8.221	8.829	10.080	12.971	14.808	17.341
19.705	22.597	24.537	25.869	27.816	29.595	30.451
30.817	32.472	32.999	33.555	34.682	34.682	35.041
35.356	35.919	36.058	36.454	36.849	37.200	37.200
37.200						

   Fit an appropriate growth curve to these data.

2. The following table gives the heights (cm) of two sunflowers measured at weekly intervals (Sandland and McGilchrist, 1979, from Doyle):

4	2
4.5	3
7	5
10	9
16	13
23	19
32	31
47	51
77	78
130	91
148	93
153	96
154	91
154	93

   Find a growth curve to describe these data. Is the difference between the two plants due to the smaller initial size of the second or to the rate of growth?

3. Weights (g) of two pregnant Afghan pikas were recorded over 14 equally spaced periods from conception to parturition (Sandland and McGilchrist, 1979, from Puget and Gouarderes):

251	258
254	263
267	269
267	266
274	282
286	289
298	295
295	308
307	338
318	350
341	359
342	382
367	390
370	400

Find a growth curve to describe these data. Can you detect any difference between the two animals?

4. In the study of glucose turnover in human beings, 26 volunteers were injected with deuterium-labelled glucose and the deuterium enrichment (atom% excess) measured at various time points. The results for subject 15 are given in the following table (Royston and Thompson, 1995, from Edwards):

Time (min)	Glucose (atom% excess)
2	1.49
3	1.39
4	1.35
5	1.22
6	1.22
8	1.14
10	1.05
12.5	0.98
15	0.92
20	0.88
30	0.79
40	0.65
50	0.65
60	0.59
75	0.48
90	0.38
120	0.40
150	0.32
180	0.31

It is expected that the resulting curve will be asymptotically zero after infinite time. Try to find an appropriate model.

5. In progressive exercise tests to exhaustion, human beings require increasing amounts of oxygen to support the increased metabolic rate. At a certain point, this demand for oxygen increases very rapidly. The point just below this abrupt change is known as the anaerobic threshold. In a kinesiology experiment, a subject performed an exercise task at such a gradually increasing level. The oxygen uptake and the expired ventilation (in unreported units) were recorded (Bennett, 1988, from Hughson):

Oxygen uptake	Expired ventilation	Oxygen uptake	Expired ventilation
574	21.9	2577	46.3
592	18.6	2766	55.8
664	18.6	2812	54.5
667	19.1	2893	63.5
718	19.2	2957	60.3
770	16.9	3052	64.8
927	18.3	3151	69.2
947	17.2	3161	74.7
1020	19.0	3266	72.9
1096	19.0	3386	80.4
1277	18.6	3452	83.0
1323	22.8	3521	86.0
1330	24.6	3543	88.9
1599	24.9	3676	96.8
1639	29.2	3741	89.1
1787	32.0	3844	100.9
1790	27.9	3878	103.0
1794	31.0	4002	113.4
1874	30.7	4114	111.4
2049	35.4	4152	119.9
2132	36.1	4252	127.2
2160	39.1	4290	126.4
2292	42.6	4331	135.5
2312	39.9	4332	138.9
2475	46.2	4390	143.7
2489	50.9	4393	144.8
2490	46.5		

What form of relationship is there between these two variables? Can the threshold level be detected?

6. AIDS cases were reported, by quarter, as diagnosed in the United Kingdom, 1982–1992 (Healy and Tillett, 1988; de Angelis and Gilks, 1994) are given in the following table:

1982	3	1	3	4
1983	3	2	12	12
1984	14	15	30	39
1985	47	40	63	65
1986	82	120	109	120
1987	134	141	153	173
1988	174	211	224	205
1989	224	219	253	233
1990	281	245	260	285
1991	271	263	306	258
1992	310	318	273	133

Try various growth curves for these data.

7. The values in the previous exercise are the marginal totals from the table below and on the next page (de Angelis and Gilks, 1994) that shows reporting delays from 1983 to 1992:

	Delay period (quarters)														
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14+
83	2	6	0	1	1	0	0	1	0	0	0	0	0	0	1
	2	7	1	1	1	0	0	0	0	0	0	0	0	0	0
84	4	4	0	1	0	2	0	0	0	0	2	1	0	0	0
	0	10	0	1	1	0	0	0	1	1	1	0	0	0	0
	6	17	3	1	1	0	0	0	0	0	0	1	0	0	1
	5	22	1	5	2	1	0	2	1	0	0	0	0	0	0
85	4	23	4	5	2	1	3	0	1	2	0	0	0	0	2
	11	11	6	1	1	5	0	1	1	1	1	0	0	0	1
	9	22	6	2	4	3	3	4	7	1	2	0	0	0	0
	2	28	8	8	5	2	2	4	3	0	1	1	0	0	1
86	5	26	14	6	9	2	5	5	5	1	2	0	0	0	2
	7	49	17	11	4	7	5	7	3	1	2	2	0	1	4
	13	37	21	9	3	5	7	3	1	3	1	0	0	0	6
	12	53	16	21	2	7	0	7	0	0	0	0	0	1	1
87	21	44	29	11	6	4	2	2	1	0	2	0	2	2	8
	17	74	13	13	3	5	3	1	2	2	0	0	0	3	5
	36	58	23	14	7	4	1	2	1	3	0	0	0	3	1
	28	74	23	11	8	3	3	6	2	5	4	1	1	1	3
88	31	80	16	9	3	2	8	3	1	4	6	2	1	2	6
	26	99	27	9	8	11	3	4	6	3	5	5	1	1	3
	31	95	35	13	18	4	6	4	4	3	3	2	0	3	3
	36	77	20	26	11	3	8	4	8	7	1	0	0	2	2

	Delay period (quarters)														
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14+
89	32	92	32	10	12	19	12	4	3	2	0	2	2	0	2
	15	92	14	27	22	21	12	5	3	0	3	3	0	1	1
	34	104	29	31	18	8	6	7	3	8	0	2	1	2	–
	38	101	34	18	9	15	6	1	2	2	2	3	2	–	–
90	31	124	47	24	11	15	8	6	5	3	3	4	–	–	–
	32	132	36	10	9	7	6	4	4	5	0	–	–	–	–
	49	107	51	17	15	8	9	2	1	1	–	–	–	–	–
	44	153	41	16	11	6	5	7	2	–	–	–	–	–	–
91	41	137	29	33	7	11	6	4	–	–	–	–	–	–	–
	56	124	39	14	12	7	10	–	–	–	–	–	–	–	–
	53	175	35	17	13	11	–	–	–	–	–	–	–	–	–
	63	135	24	23	12	–	–	–	–	–	–	–	–	–	–
92	71	161	48	25	–	–	–	–	–	–	–	–	–	–	–
	95	178	39	–	–	–	–	–	–	–	–	–	–	–	–
	76	181	–	–	–	–	–	–	–	–	–	–	–	–	–
	67	–	–	–	–	–	–	–	–	–	–	–	–	–	–

Develop a model for dependence between the AIDS cases over the years and the reporting delay. Show that the latter is decreasing with time, as in Figure 4.7.

5
Time Series

When a series of responses is being observed over time on the same subject, one may expect to find some dependence among them. Thus, responses closer together in time may usually be expected to be more closely related. In the simplest models, as in Chapter 4, this dependence may be ignored, but this is most useful as a null hypothesis to which more complex models can be compared. Now, we shall look at this dependence, but use simpler systematic components than in the previous chapter.

Because of the dependence among successive responses, a multivariate distribution will be required in order to model them. The generalized linear model family is basically univariate; for multivariate data, tricks are often possible. (We have already seen some in previous chapters.) Here, we can use the fact that any multivariate distribution for ordered responses can be decomposed into univariate distributions as follows:

$$\begin{aligned}\Pr(y_1, y_2, y_3, \ldots | \mathbf{X}) &= \Pr(y_1|\mathbf{x}_1)\Pr(y_2|y_1, \mathbf{x}_1, \mathbf{x}_2) \\ &\quad \times \Pr(y_3|y_1, , y_2, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)\cdots \qquad (5.1)\end{aligned}$$

where $\mathbf{x}_t$ are time-varying explanatory variables. Thus, many such multivariate structures can be fitted as generalized linear models, at least if the dependence structure is linear. In this chapter, we shall consider only discrete time processes; for continuous time, see Chapter 10.

5.1 Poisson Processes

5.1.1 Point Processes

Suppose that we have a series of identical events randomly spaced over time, called a *point process*. Here, we shall be interested in the frequency with which such events are occurring, but we might also consider the time between successive events. These are actually just two different aspects of the same thing. When these events are occurring to an individual subject, they are a simple case of what is known as an event or life history (Chapter 7). Alternatively, we can consider the accumulating number of events, $N(t)$, called the *counting process* (for obvious reasons).

5.1.2 Homogeneous Processes

Suppose that the numbers of events in nonoverlapping time intervals are independent. Then, the simplest model assumes that these counts of events, Y, in the nonoverlapping intervals, Δt, have a Poisson distribution with constant mean, μ, per unit time, a *Poisson process*,

$$f(y) = \frac{(\mu\Delta t)^y \mathrm{e}^{-\mu\Delta t}}{y!}$$

where Δt is the interval of observation. μ is called the *rate* or *intensity* of the process. Because of the additive property of this distribution, the size of the time intervals, Δt, is unimportant although, in more complex models, the larger it is, the less precisely will the mean or intensity be estimated.

Because the basis of this model is the Poisson distribution, such processes for subjects under different conditions can easily be constructed as log linear models in a way similar to what we did in Chapter 3. More complex processes, that introduce some dependence in time, can be developed by introducing time-varying explanatory variables.

5.1.3 Nonhomogeneous Processes

In a *nonhomogeneous* Poisson process, the rate is varying over time. For example, suppose that the rate depends on the time, t, since the previous event in one of the following ways:

$$\log(\mu_t) = \nu t + \sum_i \beta_i x_i$$

an *extreme value process*, or

$$\log(\mu_t) = \nu \log(t) + \sum_i \beta_i x_i$$

a *Weibull process*; both are simple log linear models. The rate may also depend on the total time since the process began, giving rise to a *trend*.

TABLE 5.1. Suicides in the United States of America, 1968–1970. (Haberman, 1978, pp. 44 and 51)

Year	Jan.	Feb.	Mar.	Apr.	May	June
1968	1720	1712	1924	1882	1870	1680
1969	1831	1609	1973	1944	2003	1774
1970	1867	1789	1944	2094	2097	1981
	July	Aug.	Sept.	Oct.	Nov.	Dec.
1968	1868	1801	1756	1760	1666	1733
1969	1811	1873	1862	1897	1866	1921
1970	1887	2024	1928	2032	1978	1859

Example

Consider the number of suicides each month in the United States of America for 1968, 1969, and 1970, from the National Center for Health Statistics, as given in Table 5.1. These are highly aggregated data.

Suppose, first, that we only have available the data for 1968 and wish to construct a model for them. We readily see that they are varying considerably over time, but the appropriate form of the dependence is not obvious. One possibility is to use a four-level factor variable to have a different suicide rate for each season. This gives a deviance of 12.6 with eight d.f. (AIC 20.6), whereas the saturated model, with a different intensity each month, has an AIC of 24. However, spring (March, April, May) has a considerably higher rate than all of the other seasons; thus, we can group the latter together, for a model with a deviance of 14.4 on ten d.f. (18.4).

A second possible approach is to fit harmonics, that is, sines and cosines (for example, $\sin[t\pi/p]$ where p is the period) to describe the variation over the year. First and second harmonic models have deviances of 18.7 with nine d.f. (24.7) and 7.5 with seven d.f. (17.5), respectively. These curves are plotted in Figure 5.1.

Now we can apply our estimated models for 1968 to the three years simultaneously. Suicides are increasing, so we use a linear trend over the three years, along with the respective parameter values estimated from the models for 1968 to predict suicides in the two following years. The deviances for the two seasonal models are 74.40 and 74.81, whereas those for the harmonic models are 117.79 and 102.10, respectively. The harmonic models have overcompensated for the exceptionally low rate in June 1968 that does not recur in the following years.

Of course, once the data for these years are available, we would go on to use them to improve the model, obtaining better information about the choice between the two types of model and better parameter estimates. □

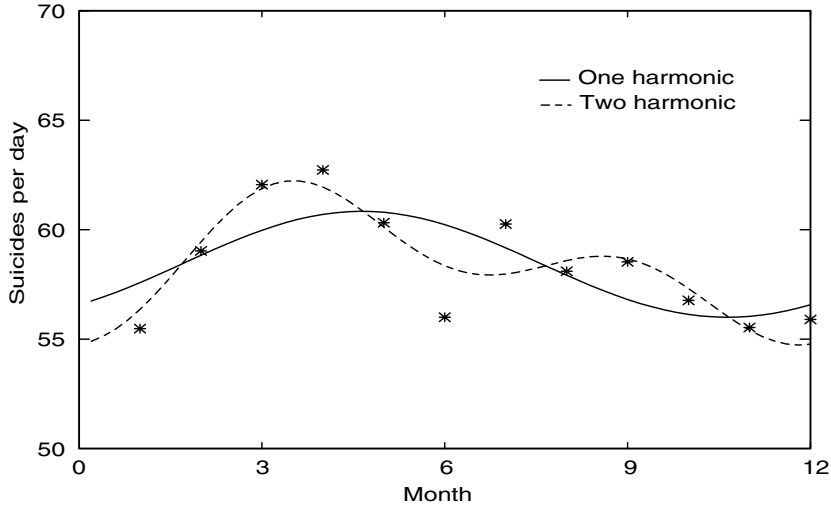

FIGURE 5.1. First and second harmonic models for the suicide data of Table 5.1.

5.1.4 Birth Processes

One important case of a nonhomogeneous Poisson process is the birth, learning, or contagion process, mentioned in Section 2.3, where the intensity depends on the number of previous events. In a pure birth process, we have the log linear model

$$\log(\mu_t) = \log(N_t) + \sum_i \beta_i x_i$$

where N_t is the total number of events up until time t.

This can easily be generalized in many ways. For example, $\log(N_t)$ can be multiplied by an unknown parameter to be estimated. A Weibull birth process would be

$$\log(\mu_t) = \log(N_t) + \nu \log(t) + \sum_i \beta_i x_i$$

Such a "birth" variable can also be introduced into other models, such as logistic regression.

This is one way of taking time dependence among responses into account, rather than just letting the intensity change over time. It is closely related to the Rasch model (Section 2.3.3). In a longitudinal context, the latter has the number of (previous) events as a factor variable, so that dependence on the number of previous events is nonlinear. This might be called a "nonparametric" birth model.

Example

A panel study was conducted to follow the interurban moves of people in the United States of America, as shown in Table 5.2. A model where moving or not in the fifth period depends only on age and ownership status (and their interaction),

$$\texttt{M5} * \texttt{AGE} * \texttt{OWNER} + \texttt{M1} * \texttt{M2} * \texttt{M3} * \texttt{M4} * \texttt{AGE} * \texttt{OWNER}$$

has an AIC of 370.0 (as compared to 256 for the saturated model).

When dependence on the number of previous moves (0–4, not the logarithm, which fits more poorly) is added to the model,

$$\texttt{M5} * \texttt{AGE} * \texttt{OWNER} + \texttt{M5} \cdot \texttt{NUMBER} + \texttt{M1} * \texttt{M2} * \texttt{M3} * \texttt{M4} * \texttt{AGE} * \texttt{OWNER}$$

the AIC is reduced to 344.5. When this number is used as a factor variable, the Rasch model,

$$\texttt{M5} * \texttt{AGE} * \texttt{OWNER} + \texttt{M5} \cdot \texttt{FNUMBER} + \texttt{M1} * \texttt{M2} * \texttt{M3} * \texttt{M4} * \texttt{AGE} * \texttt{OWNER}$$

it becomes 325.8. The probability of moving increases with the number of previous moves, but not in a linear fashion. This can also be interpreted as there being considerable heterogeneity in the population. A mover–stayer model, eliminating those who did not move in the first four periods (the fifth move cannot be included in a model to predict if there will be a move then!), could also be fitted, but it fits more poorly than the previous two. Although we have not found an adequate simple model, the main conclusion is that older owners are less likely to be moving. We shall come back to these data in the next section. □

5.2 Markov Processes

Another simple way to take the time dependence into account, as we saw in Chapter 2, is to condition on the *type* of previous response, taking it to be a given constant, once it has been observed. Thus, the value of the response, Y_t, will depend on that at time $t-1$, that is, on y_{t-1}. If $Y_t|y_{t-1}$ is independent of earlier responses, we have the (first-order) *Markov property.*

Thus, we can create a new explanatory variable, say $x_t = y_{t-1}$. Such a variable is called the *lagged* response. However, this means that we have no value for the first time period and must drop that response from the model, except as a fixed value upon which y_2 is conditioned.

TABLE 5.2. Interurban moves of a sample of people over five two-year periods in Milwaukee, Wisconsin, USA. M indicates that the person moved in the two-year period and S that he or she stayed in the same place during that period. (Crouchley *et al.*, 1982, from Clark *et al.*)

	Renters		Owners	
	Age			
Sequence	25–44	46–64	25–44	46–64
SSSSS	511	573	739	2385
SSSSM	222	125	308	222
SSSMS	146	103	294	232
SSSMM	89	30	87	17
SSMSS	90	77	317	343
SSMSM	43	24	51	22
SSMMS	27	16	62	19
SSMMM	28	6	38	5
SMSSS	52	65	250	250
SMSSM	17	20	48	14
SMSMS	26	19	60	25
SMSMM	8	4	10	3
SMMSS	8	9	54	21
SMMSM	11	3	18	1
SMMMS	10	3	21	1
SMMMM	4	1	8	2
MSSSS	41	29	134	229
MSSSM	16	15	23	10
MSSMS	19	13	36	25
MSSMM	2	4	1	0
MSMSS	11	10	69	24
MSMSM	11	2	15	3
MSMMS	1	9	13	2
MSMMM	2	2	2	0
MMSSS	7	5	40	18
MMSSM	4	2	9	2
MMSMS	8	1	15	3
MMSMM	1	0	5	0
MMMSS	8	1	22	7
MMMSM	3	2	7	2
MMMMS	5	0	9	2
MMMMM	6	3	5	0

5.2.1 Autoregression

The simple Markov model with the normal distribution is called (first-order) *autoregression* or an AR(1):

$$\mu_{t|t-1} = \rho y_{t-1} + \sum_i \beta_i x_{it}$$

where ρ is the *autoregression* coefficient and $\mu_{t|t-1}$ is the *conditional* mean response.

The unconditional, or marginal, mean in this model is somewhat more complicated:

$$\mu_t = \sum_{k=1}^{t} \rho^{t-k} \sum_i \beta_i x_{ik}$$

Thus, the current mean response depends on all of the previous values of the explanatory variables, in a geometrically decreasing fashion if $0 < \rho < 1$. This marginal model is fitted implicitly when the conditional model is used.

However, as it stands, this model makes no constraint on the value of ρ. Obviously, if $|\rho| > 1$, the situation rapidly becomes explosive. When $|\rho| < 1$, the series is said to be stationary and ρ is the *autocorrelation*. As its name suggests, ρ is then the correlation between consecutive responses. Responses t time units apart have a correlation of ρ^t. By adding lags further back in time, say p, we can check if the process is of higher than first order, an AR(p).

Thus, in this model, ρ plays two roles simultaneously. It describes the (decreasing) dependence of the marginal mean of all previous values of the explanatory variables. But it also gives the dependence among responses.

An apparently simpler model would have the marginal mean only dependent on the current explanatory variables:

$$\mu_t = \sum_i \beta_i x_{it}$$

However, this yields a conditional model of the form

$$\mu_{t|t-1} = \rho \left(y_{t-1} - \sum_i \beta_i x_{i,t-1} \right) + \sum_i \beta_i x_{it} \tag{5.2}$$

This may also be written

$$\mu_{t|t-1} - \sum_i \beta_i x_{it} = \rho \left(y_{t-1} - \sum_i \beta_i x_{i,t-1} \right)$$

The current expected (fitted value) residuals are being correlated with the previous observed residuals. Because the β_i appear twice in Equation (5.2),

TABLE 5.3. Canadian lynx trapped from 1821 to 1934 (read across rows). (Andrews and Herzberg, 1985, p. 14)

269	321	585	871	1475	2821	3928	5943	4950	2577
523	98	184	279	409	2285	2685	3409	1824	409
151	45	68	213	546	1033	2129	2536	957	361
377	225	360	731	1638	2725	2871	2119	684	299
236	245	552	1623	3311	6721	4254	687	255	473
358	784	1594	1676	2251	1426	756	299	201	229
469	736	2042	2811	4431	2511	389	73	39	49
59	188	377	1292	4031	3495	587	105	153	387
758	1307	3465	6991	6313	3794	1836	345	382	808
1388	2713	3800	3091	2985	3790	374	81	80	108
229	399	1132	2432	3574	2935	1537	529	485	662
1000	1590	2657	3396						

once as a product with ρ, this model has a nonlinear structure and is more difficult to fit.

In the latter model, ρ has exactly the same interpretation as in the former, in terms of autocorrelation, but it no longer also relates the previous explanatory variables to the marginal mean. Thus, we can either have a simple marginal model and a complex conditional model or vice versa. A further generalization introduces another set of regression parameters:

$$\mu_{t|t-1} = \rho y_{t-1} + \sum_i \alpha_i x_{i,t-1} + \sum_i \beta_i x_{it}$$

This is a linear model. Note, however, that, if none of the x_{it} are varying over time, these three models are identical.

Example

The counts of lynx trapped in the MacKenzie River District of Northwest Canada over 114 years, given in Table 5.3, are a classical time series that has been analyzed many times. These data have traditionally been modelled by a log normal distribution.

We do not have any explanatory variables, but we can look at various autoregression models. A quick check readily reveals that an AR(2),

$$\texttt{LAG1} + \texttt{LAG2}$$

is required but that an AR(3) is unnecessary. Models with various distributional assumptions and with lagged counts or lagged log counts might be considered. The AICs for some of these are given in the first panel of Table 5.4 for the AR(1) and AR(2) (we ignore the delay column for the moment).

TABLE 5.4. AICs for various autoregression models for the Canadian lynx data of Table 5.3.

		Dependence variable					
		Counts			Log counts		
Distribution	Link	Lag1	Lag2	Delay	Lag1	Lag2	Delay
Normal	Identity	1895	1848	1825	1900	1878	1873
	Log	1913	1893	1867	1889	1835	1824
Log normal	Identity	1818	1786	1772	1771	1684	1669
	Log	1824	1800	1787	1774	1684	1674
Gamma	Reciprocal	1843	1838	1828	1800	1769	1769
	Log	1814	1769	1761	1755	1679	1665
Log gamma	Reciprocal	1840	1821	1812	1788	1706	1701
	Log	1835	1807	1798	1782	1697	1686
Inverse	Rec. quad.	—	—	—	—	—	—
Gaussian	Log	1839	1814	1789	1809	1770	1749
Log inverse	Rec. quad.	1853	1840	1833	1804	1734	1733
Gaussian	Log	1843	1815	1808	1791	1710	1699
Poisson	Log	94851	73446	58555	65794	34651	30475
Negative binomial	Log	1846	1804	1794	1789	1728	1730

Here, the log normal depending on lagged log counts, with either identity or log link, fits best.

The number of lynx trapped should, however, depend on the previous birth rate. Lynx require two years to mature to mating age; that may explain the second-order model. But the dependence is positive with the previous year's count and negative with that two years before. Hence, the series seems to be measuring trapping pressure and not population size or density.

There may also by a threshold size of the population, over which births decrease (Tong, 1990, pp. 376–377). We can make trappings depend differently on previous values according to this threshold by introducing an interaction term:

$$(\texttt{LAG1} + \texttt{LAG2}) * \texttt{THRESHOLD}$$

I shall take `THRESHOLD` as a binary indicator, unity if the count two years before exceeded 1800 and zero otherwise, a value found after some experimentation. The AICs for this "delay" model are also given in the first panel of Table 5.4. The log normal distribution with an identity link is preferred. We find that the positive dependence for the previous year and negative for two years before are accentuated when the threshold is passed. The more lynx are trapped one year, the fewer two years hence. □

Random Walks

In autoregressive models, $\rho < 1$ if the situation is not rapidly to become explosive. Otherwise, the series is *nonstationary*. A case of special interest occurs when $\rho = 1$. Let us write Equation (5.2) in still another way:

$$\begin{aligned} \mu_{t|t-1} - \rho y_{t-1} &= \sum_i \beta_i x_{it} - \sum_i \rho \beta_i x_{i,t-1} \\ &= \sum_i \beta_i (x_{it} - \rho x_{i,t-1}) \end{aligned}$$

With $\rho = 1$, we are fitting successive differences, between the expected response and the previous one and between values of the explanatory variables. This is called first differencing. It implies that differences between successive responses will be stationary and, hence, is often used in an attempt to eliminate nonstationarity. The model, for the original response, not the differences, is known as a *random walk*; it is a generalized linear model because ρ is now known. Obviously, any explanatory variables that are not changing over time will drop out of the model when first differences are taken.

5.2.2 Other Distributions

Often, a log normal distribution, by taking logarithms of the responses, and their lagged values, is used in place of the normal, because responses are only positive and the distribution is skewed. An alternative rarely used, but that should be considered, is to take a log link, as we did for the exponential growth curve in Section 4.1. This can give a better prediction, because it fits a curve through the centre of the untransformed responses, in the least-squares sense, whereas the log normal distribution does not. However, the direct interpretation of ρ, as explained above, is lost.

If some other exponential family distribution is used, such as the gamma or inverse Gaussian, the identity link will no longer be canonical. Again, the interpretation of ρ is lost.

If a link other than the identity is used, it may often be useful to transform the lagged responses in the same way:

$$g(\mu_t) = \rho g(y_{t-1}) + \sum_i \beta_i x_{it}$$

although this is a rather peculiar model, as compared to transforming the response before taking the mean, which was what was done for the log normal distribution. As we have seen, transforming the mean through a link function is not at all the same as transforming the responses.

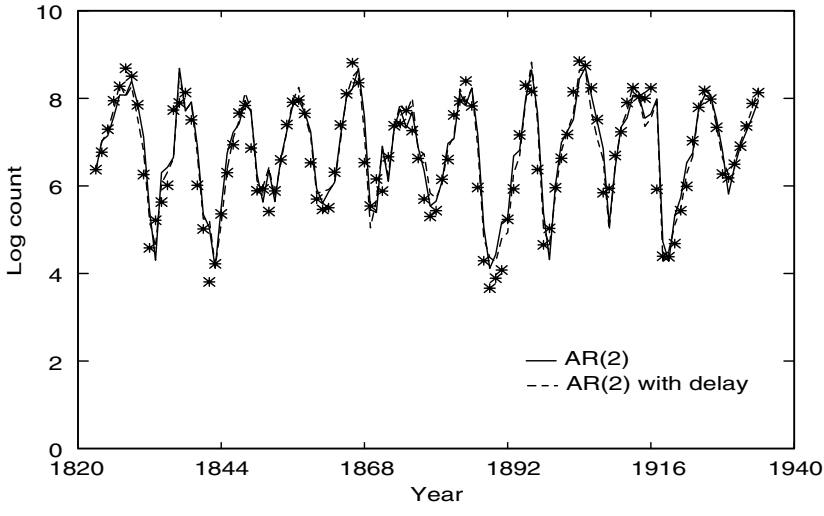

FIGURE 5.2. Gamma AR(2) models for the lynx data of Table 5.3.

Example

We continue with our example of lynx trappings. Among possible distributions, we see from Table 5.4 that the gamma distribution with a log link fits better than the log normal. Although these are count data, the Poisson model fits extremely poorly and, in fact, would require a higher-order autoregression. On the other hand, the negative binomial, fitted by plotting a normed profile likelihood of values of the power parameter (Section 1.5.2), is competitive, indicating the presence of substantial overdispersion.

For the gamma distribution with a log link, the estimated delay model is

$$\log(\mu_t) = 1.47 + 1.23x_{t-1} - 0.40x_{t-2}$$

when $x_{t-2} \leq 1800$ and

$$\log(\mu_t) = 4.98 + 1.46x_{t-1} - 1.11x_{t-2}$$

when $x_{t-2} > 1800$. The models for the gamma distributions, with and without threshold, are plotted in Figure 5.2. We see that the extremely large and small values are not well followed by the models.

In many respects, this has been an academic model fitting exercise, as is often the case with this data set. There does not appear to be any theoretical reason why such count data would follow a gamma distribution. We shall never be able to disaggregate the events to know the timings of individual trappings, and this probably would not make much sense anyway because they occurred over a vast geographical area. Thus, we should perhaps look for more reasonable models in the direction of compound distributions other

than the negative binomial. However, we may have isolated one interesting biological aspect of the phenomenon, the impact of trapping pressure two years before. □

Autoregression

In Equation (5.2), we saw how the current mean could be made to depend on the previous (fitted value) residual. When the distribution used in the model is not normal, several types of residuals are available (Section B.2). Thus, we can define an autoregression by

$$\mu_{t|t-1} = \rho\hat{\varepsilon}_{t-1} + \sum_i \beta_i x_{it} \tag{5.3}$$

where $\hat{\varepsilon}_{t-1}$ is an estimated residual from the previous time period. Because this residual is a function of the regression coefficients, $\boldsymbol{\beta}$, the model is nonlinear and requires extra iterations as compared to standard IWLS for generalized linear models. (The likelihood function is very complex.) However, estimation of this model can easily be implemented as an additional loop in standard generalized linear modelling software. Here, we shall use the deviance residuals.

Example

Beveridge (1936) gives the average rates paid to agricultural labourers for threshing and winnowing one rased quarter each of wheat, barley, and oats in each decade from 1250 to 1459. These are payments for performing the manual labour of a given task, not daily wages. He obtained them from the rolls of eight Winchester Bishopric Manors (Downton, Ecchinswel, Overton, Meon, Witney, Wargrave, Wycombe, Farnham) in the south of England. As well, he gives the average price of wheat for all of England, both as shown in Table 5.5. All values are in money of the time (pence or shillings), irrespective of changes in the currency: in fact, silver content was reduced five times (1300, 1344, 1346, 1351, and 1412) during this period.

Interestingly, at first sight, the Black Death of the middle of the fourteenth century seems to have had little effect. Beveridge (1936) states that, for the first 100 years of the series, the labourers' wages seemed to follow the price of wheat, but this is difficult to see for the rates plotted in Figure 5.3. He suggests that, during this period, such payments may only have been a substitute for customary service or allowances in kind, so that they were closely related to the cost of living. In the second part of the series, a true market in labour had begun to be established.

We shall consider models with the normal, gamma, and inverse Gaussian distributions, each with the identity and log links. Because we shall be looking at lagged values of the two variables, we weight out the first value. However, we should not expect strong relations with previous values because

TABLE 5.5. Rates (pence) for threshing and winnowing and wheat prices (shillings per quarter) on eight Winchester manors. (Beveridge, 1936)

Decade	Agricultural rate	Wheat price
1250–	3.30	4.95
1260–	3.37	4.52
1270–	3.45	6.23
1280–	3.62	5.00
1290–	3.57	6.39
1300–	3.85	5.68
1310–	4.05	7.91
1320–	4.62	6.79
1330–	4.92	5.17
1340–	5.03	4.79
1350–	5.18	6.96
1360–	6.10	7.98
1370–	7.00	6.67
1380–	7.22	5.17
1390–	7.23	5.45
1400–	7.31	6.39
1410–	7.35	5.84
1420–	7.34	5.54
1430–	7.30	7.34
1440–	7.33	4.86
1450–	7.25	6.01

these are ten-year averages. The plot of yearly rates published by Beveridge shows more discontinuity, especially from 1362 to 1370 (but he does not provide a table of these individual values). He interprets the irregularities in this decade as an (unsuccessful) attempt by William of Wykeham to reduce wages to their former level.

For a model with rates depending only on current wheat prices,

$$\texttt{RESIDUAL} + \texttt{WHEAT}$$

the normal distribution with an identity link, a classical autoregression model, fits best with an AIC of 249.7. Nevertheless, all of the models are fairly close, the worst being the gamma with log link that has 255.4. A normal model for dependence of rates on wheat prices, without the autoregression on the residuals, has an AIC of 265.3, showing the dependence over time. However, the need for regression on the residuals, here and below, probably arises because we have an inadequate set of explanatory variables. Adding dependence on wheat prices in the previous decade does not improve the model.

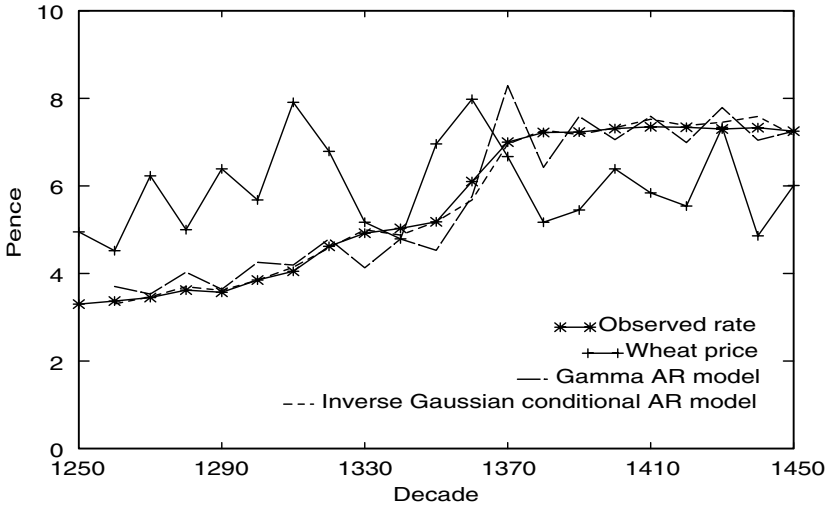

FIGURE 5.3. Plot of the labouring rates and wheat prices for the data of Table 5.5.

If we add a variable for the change due to the Black Death, at about 1370,

$$\texttt{RESIDUAL} + \texttt{WHEAT} + \texttt{WHEAT1} + \texttt{BDEATH}$$

the picture alters considerably. The gamma distribution with identity link now fits best (AIC 221.9 — a model with the break one decade earlier fits almost as well). In this model, the mean rate depends on present and lagged wheat prices, with a different intercept before and after this date. However, the dependence of the rate on wheat prices does not change at the break. In other words, there is no need for an interaction between the break indicator and these variables. The mean rate jumped up by an estimated three pence at this date, perhaps as an aftereffect of the Black Death. The dependence on the current and lagged wheat prices is almost identical, with coefficients of 0.25. This model is shown in Figure 5.3. We see that it does not follow the agricultural rate series very closely, but is irregular, like the wheat price series. Of course, the only information allowing prediction of changes in rate, besides the overall change in mean arising from the break variable, comes from these prices.

If we take into account the rate in the previous decade, we obtain a considerably better fit. Here, the previous wheat price is also necessary, but the break variable is not, so that there are three explanatory variables in the model (the current wheat price and the previous rate and wheat price),

$$\texttt{RESIDUAL} + \texttt{LAG1} + \texttt{WHEAT} + \texttt{WHEAT1}$$

that is, the same number as in the previous model (lagged rate replacing the break variable). The inverse Gaussian distribution with identity link

fits best, with an AIC of 169.2. (The normal model now has 171.5 and the gamma 169.8.) As might be expected, current rates depend most strongly on previous rates, but about four times as much on previous wheat price as on the current one. This conditional model is also plotted in Figure 5.3. It follows much more closely the agricultural rates, as would be expected.

In all of these models, we have not directly taken into account the time trend in the rate. We can do almost as well as the previous model with two simple regression lines on time, before and after the break due to the Black Death. In contrast to all of the previous models, here the autoregression on the residuals is not necessary:

$$\texttt{YEAR} * \texttt{BDEATH}$$

The best model is the normal distribution with a log link (AIC 174.3). The two regression models are

$$\begin{aligned} \log(\mu_t) &= 1.02 + 0.06103t, \qquad t \leq 12 \\ \log(\mu_t) &= 1.92 + 0.00342t, \qquad t > 12 \end{aligned}$$

This model can be only slightly improved by adding the logged rate, dependence on wheat prices not being necessary. □

This example shows how distributional assumptions can be very dependent on the type of dependence incorporated in the regression model.

5.2.3 Markov Chains

The above autoregressive approach can be applied directly to discrete data as well. This is what we were doing in Chapter 2. Suppose, first, that the subject can be in either one of two states at any given time. We have a binary response variable, the state, and can use logistic regression. Our model will be

$$\log\left(\frac{\pi_{t|t-1}}{1-\pi_{t|t-1}}\right) = \rho y_{t-1} + \sum_i \beta_i x_{it}$$

where y_t is a binary $(0, 1)$ variable. This is called a two-state, discrete time Markov chain.

The estimated conditional probabilities of staying in the same state or changing state at time t, given the state at the previous time, $t-1$, can be obtained by tabulation. We saw in Section 2.2.2 that these are known as the transition probabilities; they form a square matrix.

If there are more than two states, we must turn to log linear models. These are easier to represent with the Wilkinson and Rogers notation:

$$\mathtt{Y}_t * \mathtt{Y}_{t-1} + \mathtt{Y}_{t-1} * \mathtt{X}_1 * \cdots * \mathtt{X}_p$$

If the dependence between y_t and y_{t-1} is the same for all t, we have stationarity. This can easily be checked by comparing the appropriate log linear models, with an interaction between y_t and t.

Once again, by adding further lags, we can check the order of the Markov chain. We studied some special models related to Markov chains in Chapter 2. We thus see that, in the context of generalized linear models, autoregression and Markov chains are the same, but for a change of distributional assumptions.

Example

Let us consider again the mobility data of Table 5.2. With dependence on the previous status (mover or not), as well as ownership and age,

$$\texttt{M5} * \texttt{AGE} * \texttt{OWNER} + \texttt{M5} * \texttt{M4} + \texttt{M1} * \texttt{M2} * \texttt{M3} * \texttt{M4} * \texttt{AGE} * \texttt{OWNER}$$

the AIC is 342.5, not as good as the Rasch model. Dependence two periods back

$$\texttt{M5} * \texttt{AGE} * \texttt{OWNER} + \texttt{M5} * \texttt{M4} + \texttt{M5} * \texttt{M3} + \texttt{M1} * \texttt{M2} * \texttt{M3} * \texttt{M4} * \texttt{AGE} * \texttt{OWNER}$$

gives a further reduction to 304.4. Adding the Rasch model factor variable

$$\begin{aligned}\texttt{M5} * \texttt{AGE} * \texttt{OWNER} + \texttt{M5} \cdot \texttt{NUMBERF} + \texttt{M5} * \texttt{M4} + \texttt{M5} * \texttt{M3} \\ + \texttt{M1} * \texttt{M2} * \texttt{M3} * \texttt{M4} * \texttt{AGE} * \texttt{OWNER}\end{aligned}$$

reduces this to 284.5. Only by introducing most of the two-way interactions among the lagged variables and the other explanatory variables can the AIC be brought down to 241.3. The result is a complex model, including individual variability and time dependence, not easily interpretable. □

5.3 Repeated Measurements

Repeated measurements is a term most frequently used in medical statistics. It refers to studies where the same type of response is recorded several times on each of several subjects. The repetitions may occur more or less simultaneously, yielding clustered data, or extending over time, as longitudinal data. In the social sciences, the latter are often known as panel data, as we saw in Chapter 2.

The observations may be some response measured as a continuous variable, a series of events, categorized as some nominal or ordinal variable, or the durations between such events, event histories. If the observation of each subject extends over time, then time series methods, of the kinds just described, will need to be applied. However, responses may also occur more

or less simultaneously, as when the individual members of families are studied. Additional variables must be introduced into the models to account for differences among the subjects, such as different treatments.

Due to the repetition of observations, a second source of variability, besides that over time, will often be present. The set of observations on one subject will often be more similar than a random set across subjects. For example, some subjects, or the members of some families, will tend to have consistently higher responses than others. This can be called *intraclass dependence*.

One common approach to handling this is through a *random effects* model. The models are closely related to those for overdispersion presented in Section 2.3, including the Rasch model of Section 2.3.3. However, in general, they do not lead to generalized linear models and are often rather difficult to analyze.

Many of the examples in this chapter and others involve repeated measurements. No new examples will be provided here.

Summary

Books on time series abound, especially in the econometrics literature. For a simple introduction, readers might like to consult Diggle (1990). Those on statistical applications of more general stochastic processes are rarer; see Lindsey (1992). Recent books on longitudinal data and repeated measurements include Lindsey (1993), Diggle *et al.* (1994), Fahrmeir and Tutz (1994), and Hand and Crowder (1996).

5.4 Exercises

1. In Section 5.2, we looked at a classical data set modelled by autoregression techniques, the lynx data. Another such set involves the annual Wölfer sunspot numbers between 1770 and 1869 (Hand *et al.* 1994, pp. 85–86) (read across rows):

101	82	66	35	31	7	20	92	154	125
85	68	38	23	10	24	83	132	131	118
90	67	60	47	41	21	16	6	4	7
14	34	45	43	48	42	28	10	8	2
0	1	5	12	14	35	46	41	30	24
16	7	4	2	8	17	36	50	62	67
71	48	28	8	13	57	122	138	103	86
63	37	24	11	15	40	62	98	124	96
66	64	54	39	21	7	4	23	55	94
96	77	59	44	47	30	16	7	37	74

They measure the average number of sunspots on the sun each year. Can you find a Markov model of an appropriate order to describe these data adequately? A number of different suggestions have been made in the literature.

2. The following table gives the enrollment at Yale University, 1796–1975 (Anscombe, 1981, p. 130) (read across rows):

115	123	168	195	217	217	242	233	200
222	204	196	183	228	255	305	313	328
350	352	298	333	349	376	412	407	481
473	459	470	454	501	474	496	502	469
485	536	514	572	570	564	561	608	574
550	537	559	542	588	584	522	517	531
555	558	604	594	605	619	598	565	578
641	649	599	617	632	644	682	709	699
724	736	755	809	904	955	1031	1051	1021
1039	1022	1003	1037	1042	1096	1092	1086	1075
1134	1245	1365	1477	1645	1784	1969	2202	2350
2415	2615	2645	2674	2684	2542	2712	2816	3142
3138	3806	3605	3433	3450	3312	3282	3229	3288
3272	3310	3267	3262	2006	2554	3306	3820	3930
4534	4461	5155	5316	5626	5457	5788	6184	5914
5815	5631	5475	5362	5493	5483	5637	5747	5744
5694	5454	5036	5080	4056	3363	8733	8991	9017
8519	7745	7688	7567	7555	7369	7353	7664	7488
7665	7793	8129	8221	8404	8333	8614	8539	8654
8666	8665	9385	9214	9231	9219	9427	9661	9721

The primary irregularities in these data occur during the two World Wars. Develop an adequate Markov model for these count data. Among other possibilities, compare a normal model that uses differences in enrollment between years with a Poisson model that involves ratios of successive enrollment rates. Is there evidence of overdispersion?

3. Annual snowfall (inches) in Buffalo, New York, USA, was recorded from 1910 to 1972 (Parzen, 1979) (read across rows):

126.4	82.4	78.1	51.1	90.9	76.2	104.5	87.4	110.5
25.0	69.3	53.5	39.8	63.6	46.7	72.9	79.7	83.6
80.7	60.3	79.0	74.4	49.6	54.7	71.8	49.1	103.9
51.6	82.4	83.6	77.8	79.3	89.6	85.5	58.0	120.7
110.5	65.4	39.9	40.1	88.7	71.4	83.0	55.9	89.9
84.8	105.2	113.7	124.7	114.5	115.6	102.4	101.4	89.8
71.5	70.9	98.3	55.5	66.1	78.4	120.5	97.0	110.0

Find an appropriate model to describe these time series data. Is there evidence of a trend or of a cyclical phenomenon?

4. Beveridge (1936) also gives the average daily wages (pence) of several other classes of labourers each decade from 1250 to 1459 on several Winchester manors in England. Those for carpenters and masons (both in Taunton manor) are shown below, as well as the rates for agricultural labourers and the price of wheat used in the example above.

Decade	Agricultural rate	Carpenter's wage	Mason's wage	Wheat price
1250–	3.30	3.01	2.91	4.95
1260–	3.37	3.08	2.95	4.52
1270–	3.45	3.00	3.23	6.23
1280–	3.62	3.04	3.11	5.00
1290–	3.57	3.05	3.30	6.39
1300–	3.85	3.14	2.93	5.68
1310–	4.05	3.12	3.13	7.91
1320–	4.62	3.03	3.27	6.79
1330–	4.92	2.91	3.10	5.17
1340–	5.03	2.94	2.89	4.79
1350–	5.18	3.47	3.80	6.96
1360–	6.10	3.96	4.13	7.98
1370–	7.00	4.02	4.04	6.67
1380–	7.22	3.98	4.00	5.17
1390–	7.23	4.01	4.00	5.45
1400–	7.31	4.06	4.29	6.39
1410–	7.35	4.08	4.30	5.84
1420–	7.34	4.11	4.31	5.54
1430–	7.30	4.51	4.75	7.34
1440–	7.33	5.13	5.15	4.86
1450–	7.25	4.27	5.26	6.01

Can models similar to those for agricultural labourers be developed for the other two types of workers? Does it make any difference that these are wages and not rates for piece work?

5. A number of women in the United States of America were followed over five years, from 1967 to 1971, in the University of Michigan Panel Study of Income Dynamics. The sample consisted of white women who were continuously married to the same husband over the five-year period. Having worked in the year is defined as having earned any money during the year. The sample paths of labour force participation are given in the following table (Heckman and Willis, 1977):

				1971	
1970	1969	1968	1967	Yes	No
Yes	Yes	Yes	Yes	426	38
No				16	47
Yes	No			11	2
No				12	28
Yes	Yes	No		21	7
No				0	9
Yes	No			8	3
No				5	43
Yes	Yes	Yes	No	73	11
No				7	17
Yes	No			9	3
No				5	24
Yes	Yes	No		54	16
No				6	28
Yes	No			36	24
No				35	559

Study how the most recent employment record of each woman depends on her previous history. Is there indication of heterogeneity among the women? Notice that here there are two types of stable behaviour that might be classified stayers.

6. The numbers of deaths by horse kicks in the Prussian army from 1875 to 1894 for 14 corps (Andrews and Herzberg, 1985, p. 18) are as follows:

Corps	Year
G	0 2 2 1 0 0 1 1 0 3 0 2 1 0 0 1 0 1 0 1
I	0 0 0 2 0 3 0 2 0 0 0 1 1 1 0 2 0 3 1 0
II	0 0 0 2 0 2 0 0 1 1 0 0 2 1 1 0 0 2 0 0
III	0 0 0 1 1 1 2 0 2 0 0 0 1 0 1 2 1 0 0 0
IV	0 1 0 1 1 1 1 0 0 0 0 1 0 0 0 0 1 1 0 0
V	0 0 0 0 2 1 0 0 1 0 0 1 0 1 1 1 1 1 1 0
VI	0 0 1 0 2 0 0 1 2 0 1 1 3 1 1 1 0 3 0 0
VII	1 0 1 0 0 0 1 0 1 1 0 0 2 0 0 2 1 0 2 0
VIII	1 0 0 0 1 0 0 1 0 0 0 0 1 0 0 0 1 1 0 1
IX	0 0 0 0 0 2 1 1 1 0 2 1 1 0 1 2 0 1 0 0
X	0 0 1 1 0 1 0 2 0 2 0 0 0 0 2 1 3 0 1 1
XI	0 0 0 0 2 4 0 1 3 0 1 1 1 1 2 1 3 1 3 1
XIV	1 1 2 1 1 3 0 4 0 1 0 3 2 1 0 2 1 1 0 0
XV	0 1 0 0 0 0 0 1 0 1 1 0 0 0 2 2 0 0 0 0

G indicates the guard corps. This and corps I, VI, and XI have different organizations than the others. Can you detect any trends with time? Are there systematic differences among the corps?

7. Reanalyze the data on children wheezing in Exercise 2.5, taking into account the longitudinal aspect of the data.

This page intentionally left blank

6
Survival Data

6.1 General Concepts

6.1.1 Skewed Distributions

A duration is the time until some event occurs. Thus, the response is a non-negative random variable. If the special case of a survival time is being observed, the event is considered to be absorbing, so that observation of that individual must stop when it occurs. We first consider this case, although most of the discussion applies directly to more general durations such as the times between repeated events, called event histories (Chapter 7). Usually, the distribution of durations will not be symmetric, but will have a form like that in Figure 6.1 (this happens to be a log normal distribution). This restricts the choice of possible distributions to be used. For example, a normal distribution would not be appropriate. Suitable distributions within the generalized linear model family include the log normal, gamma, and inverse Gaussian.

6.1.2 Censoring

Because individuals are to be observed over time, until the prescribed event, and because time is limited and costly, not all individuals may be followed until an event. Such data are called *censored*. Censored observations are incomplete, but they still contain important information. We know that the event did not occur before the end of the observation period.

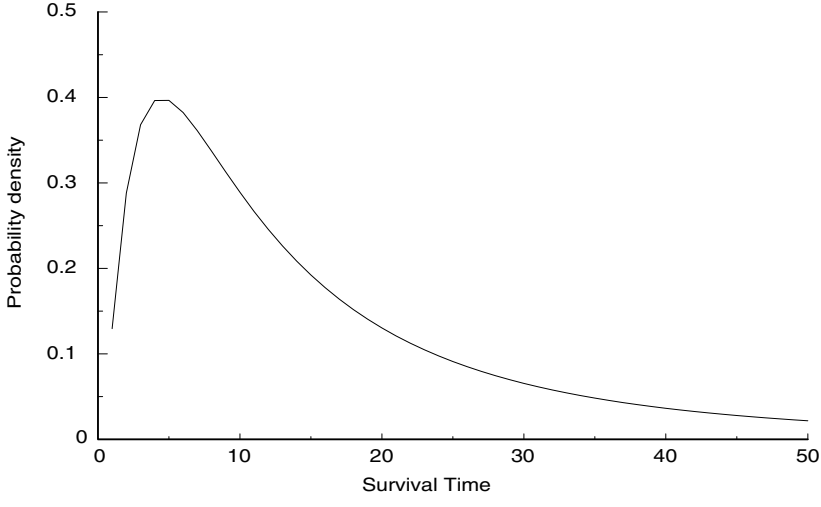

FIGURE 6.1. A typical density function for a survival curve.

Censoring can occur for a number of reasons. For example, the protocol for the study may specify observation over a fixed period of time or individual cases may disappear from the study for some reason.

Planned censoring may occur in two main ways:

- If recording of an event must stop after a fixed time interval, we have *Type I* or *time censoring*, most often used in medical studies.

- If the study must continue until complete information is available on a fixed number of cases, we have *Type II* or *failure censoring*, most common in industrial testing.

However, cases may drop out for reasons not connected with the study or beyond the control of the research worker. These may or may not be linked to the response or explanatory variables; for example, through side effects under a medical treatment. If they are related, the way in which censoring occurs cannot simply be ignored in the model. Thus, in complex cases, where censoring is not random, a model will need to be constructed for it, although generally very little information will be available in the data about such a model.

It is important to distinguish situations where censoring only depends on information already available at the censoring point from other possibilities because, for that case, such modelling may be possible. For example, the censoring indicator could be made to depend on the available explanatory variables in some form of regression model.

Even with ignorable causes of censoring, the analysis is further complicated because we cannot simply use the density function as it stands.

6.1.3 Probability Functions

If the probability density function is $f(t)$ and the cumulative probability function is $F(t)$, then the *survivor function* is

$$S(t) = 1 - F(t)$$

and the *hazard function*

$$\begin{aligned} h(t) &= \frac{f(t)}{S(t)} \\ &= -\frac{d \log S(t)}{dt} \end{aligned}$$

This is the rate or intensity of the point processes of the previous chapter. Then, we have

$$\begin{aligned} S(t) &= \exp\left[-\int_0^t h(u)du\right] \\ f(t) &= h(t)\exp\left[-\int_0^t h(u)du\right] \end{aligned}$$

where $\int h(u)du$ is called the *integrated hazard* or *intensity*.

Suppose that I_i is a code or indicator variable for censoring, with $I_i = 1$ if the observation i is completely observed and $I_i = 0$ if it is censored. Then, the probability for a sample of n individuals will be approximately (because the density assumes that one can actually observe in continuous time) proportional to

$$\prod [f(t_i)]^{I_i} [S(t_i)]^{1-I_i} \tag{6.1}$$

and a likelihood function can be derived from this. In most cases, this does not yield a generalized linear model.

6.2 "Nonparametric" Estimation

Before looking at specific parametric models for a set of data, it is often useful to explore the data by means of a "nonparametric" estimation procedure similar to those used in some of the previous chapters. As usual, such models are, in fact, highly parametrized, generally being saturated models. The most commonly used for survival data is the *Kaplan–Meier product limit estimate* (Kaplan and Meier, 1958).

TABLE 6.1. Remission times (weeks) from acute leukaemia under two treatments, with censored times indicated by asterisks. (Gehan, 1965, from Freireich *et al.*)

6-mercaptopurine
6 6 6 6* 7 9* 10 10* 11* 13 16 17* 19* 20* 22 23 25* 32* 32* 34* 35*
Placebo
1 1 2 2 3 4 4 5 5 8 8 8 8 11 11 12 12 15 17 22 23

If π_j is the probability of having an event at time t_j, conditional on not having an event until then, that is, on surviving to that time, the likelihood function is

$$\mathrm{L}(\boldsymbol{\pi}) = \prod_{j=1}^{k} \pi_j^{d_j} (1 - \pi_j)^{n_j - d_j}$$

where n_j is the number having survived and still under observation, and hence still known to be at risk just prior to t_j, called the *risk set*, d_j is the number having the event at time t_j, and π_j is the hazard or intensity at t_j. This is a special application of the binomial distribution, with maximum likelihood estimates, $\hat{\pi}_j = d_j/n_j$. Then, the product limit estimate of the survivor function is just the product of the estimated probabilities of not having the event at all time points up to the one of interest:

$$\hat{S}(t) = \prod_{j|t_j<t} \left(\frac{n_j - d_j}{n_j} \right)$$

a special application of Equation (5.1). This may be plotted in various ways (Lindsey, 1992, pp. 52–57) to explore what form of parametric model might fit the data. It provides a saturated model to which others can be compared.

Example

Table 6.1 gives a classical data set on the time maintained in remission for cases of acute leukaemia under two treatments. In this trial, conducted sequentially so that patients were entering the study over time, 6-mercaptopurine was compared to a placebo. The results in the table are from one year after the start of the study, with an upper limit of the observation time of about 35 weeks.

The Kaplan–Meier estimates of the survivor functions for these two groups are plotted in Figure 6.2. We see how the treatment group has longer estimated survival times than the placebo group. □

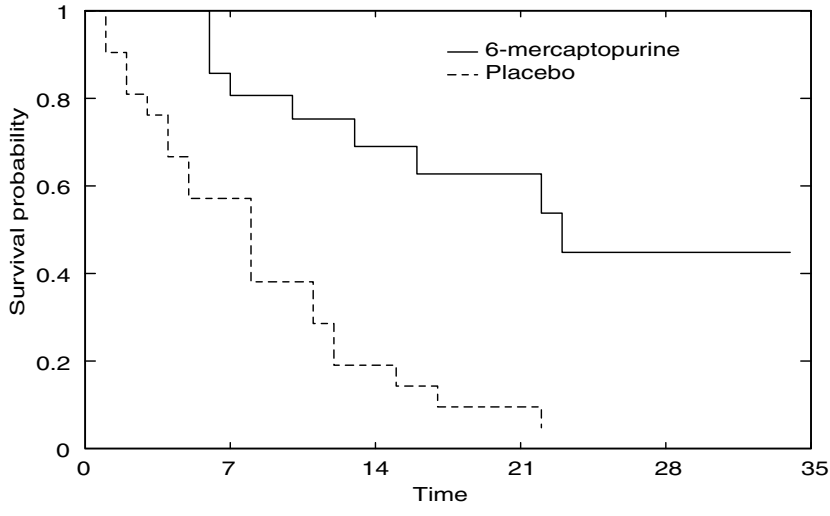

FIGURE 6.2. Kaplan–Meier curves for the survival data of Table 6.1.

6.3 Parametric Models

6.3.1 Proportional Hazards Models

Suppose now that the hazard function can be written in the form

$$h(t;\mathbf{x}) = h_0(t)\mathrm{e}^{\mathbf{x}^T\boldsymbol{\beta}} \tag{6.2}$$

where $\mathbf{x}$ is a vector of explanatory variables. This is called a *proportional hazards model*. Notice how h_0 depends only on time, t, and the other factor only on the explanatory variables, $\mathbf{x}$, so that hazard curves for different values of the explanatory variables will be proportional to h_0 at all time points, hence the model's name. If $h_0(t) = 1$, a constant, we have the hazard function for an exponential distribution, and if $h_0(t) = \alpha t^{\alpha-1}$, that for a Weibull distribution. If $h_0(t)$ is left unspecified, so that a factor variable in time must be used, we have the "semiparametric" *Cox model*.

Such models can simply be fitted as generalized linear models, based on the Poisson distribution, in at least two ways. We study one here for survival data and the other in the next chapter for event histories (as well as in Section 4.4). The second may be more computer intensive but has the advantage of easily allowing time-varying explanatory variables.

6.3.2 Poisson Representation

Aitkin and Clayton (1980) have demonstrated a useful relationship between proportional hazards models and the Poisson distribution that allows one

to fit censored data for proportional hazards as generalized linear models, although usually with one nonlinear parameter (Section 1.5.2).

With $h_0(\cdot)$ the *baseline hazard*, as above, we have

$$\begin{aligned} S(t_i) &= \exp\left[-H_0(t_i)\mathrm{e}^{\mathbf{x}_i^T\boldsymbol{\beta}}\right] \\ f(t_i) &= h(t_i)S(t_i) \\ &= h_0(t_i)\exp\left[\mathbf{x}_i^T\boldsymbol{\beta} - H_0(t_i)\mathrm{e}^{\mathbf{x}_i^T\boldsymbol{\beta}}\right] \end{aligned}$$

where

$$H_0(t) = \int_0^t h_0(u)du$$

is the integrated hazard. Then, from Equation (6.1), the likelihood function is

$$\begin{aligned} \mathrm{L}(\boldsymbol{\beta}) &= \prod_i \left[f(t_i)\right]^{I_i}\left[S(t_i)\right]^{1-I_i} \\ &= \prod_i \left[h_0(t_i)\exp(\mathbf{x}_i^T\boldsymbol{\beta})\right]^{I_i}\exp\left[-H_0(t_i)\mathrm{e}^{\mathbf{x}_i^T\boldsymbol{\beta}}\right] \\ &= \prod_i \left[\mu_i^{I_i}\mathrm{e}^{-\mu_i}\right]\left[\frac{h_0(t_i)}{H_0(t_i)}\right]^{I_i} \end{aligned}$$

with $\mu_i = H_0(t_i)\exp(\mathbf{x}_i^T\boldsymbol{\beta})$. Here, I_i can be interpreted as the cumulative number of events at time t_i for individual i, either zero or one.

In this equation, the first term is the likelihood function for the Poisson variables, I_i. The second term does not contain $\boldsymbol{\beta}$, but there may be other parameters, such as the α of the Weibull distribution, in $H_0(t)$. We can take the linear model to be

$$\log(\mu_i) = \log[H_0(t_i)] + \mathbf{x}_i^T\boldsymbol{\beta} \tag{6.3}$$

for fixed values of the parameters in $H_0(t)$ and use $\log[H_0(t_i)]$ as an offset. Iterations may then be performed on any unknown parameters in $H_0(t)$.

6.3.3 Exponential Distribution

For the exponential distribution, we have

$$\begin{aligned} H_0(t) &= t \\ h_0(t) &= 1 \end{aligned}$$

so that there are no extra parameters. Then, from Equation (3.2),

$$\begin{aligned} f(t_i) &= \phi_i\mathrm{e}^{-t_i\phi_i} \\ &= \exp\left[\mathbf{x}_i^T\boldsymbol{\beta} - t_i\mathrm{e}^{\mathbf{x}_i^T\boldsymbol{\beta}}\right] \end{aligned}$$

with $1/\phi_i = \mathrm{E}[T_i] = \exp[-\mathbf{x}_i^T\boldsymbol{\beta}]$ so that

$$\log(\mu_i) = \log(t_i) + \mathbf{x}_i^T\boldsymbol{\beta}$$

Here, ϕ_i is the constant (over time) intensity function under conditions x_i.

Example

For the leukaemia survival data of Table 6.1, each starred value will have a "response" value, I_i, of zero and the others a value of one. The offset is the logarithm of the time. Then, the exponential model without treatment differences has an AIC of 235.5, whereas that with such differences has 221.0. The constant difference in log risk is estimated to be $\hat{\beta} = 1.527$. □

6.3.4 Weibull Distribution

For the Weibull distribution, we have

$$\begin{aligned} H_0(t) &= t^\alpha \\ h_0(t) &= \alpha t^{\alpha-1} \end{aligned}$$

so that

$$\frac{h_0(t)}{H_0(t)} = \frac{\alpha}{t}$$

Then,

$$\begin{aligned} f(t_i) &= \alpha t_i^{\alpha-1}\phi_i \mathrm{e}^{-t_i^\alpha\phi_i} \\ &= \alpha t_i^{\alpha-1}\exp\left[\mathbf{x}_i^T\boldsymbol{\beta} - t_i\mathrm{e}^{\mathbf{x}_i^T\boldsymbol{\beta}}\right] \end{aligned} \tag{6.4}$$

with

$$\mathrm{E}[T_i] = \Gamma\left(1+\frac{1}{\alpha}\right)\mathrm{e}^{-\mathbf{x}_i^T\boldsymbol{\beta}/\alpha}$$

so that

$$\log(\mu_i) = \alpha\log(t_i) + \mathbf{x}_i^T\boldsymbol{\beta}$$

for a fixed value of α. On the other hand, the full log likelihood for unknown α is

$$\log(\mathrm{L}) = \sum_i I_i\log(\alpha) + \sum_i [I_i\log(\mu_i) - \mu_i]$$

and the score equation for α is

$$\frac{\partial\log(\mathrm{L})}{\partial\alpha} = \frac{\sum_i I_i}{\alpha} + \sum_i (I_i - \mu_i)\log(t_i)$$

Setting this equal to zero and solving, we obtain

$$\hat{\alpha} = \frac{\sum_i I_i}{\sum_i (\hat{\mu}_i - I_i) \log(t_i)}$$

This can be used to update the value of α in successive iterations.

Example

For the leukaemia survival data of Table 6.1, the Weibull model without treatment differences has an AIC of 236.5 and $\hat{\alpha} = 1.139$, whereas that with such differences has 218.9 with $\hat{\alpha} = 1.364$. The latter AIC is smaller than for the exponential model, indicating that $\alpha > 1$. The difference in log risk is estimated to be 1.726. □

6.4 "Semiparametric" Models

6.4.1 Piecewise Exponential Distribution

In the Weibull model, the hazard function changes continuously over time, whereas in the exponential model, it is constant at all time points. As we saw in the Section 5.1.3, the former represents a nonhomogeneous Poisson process. If the hazard function can be assumed to be constant in given intervals of time (perhaps simply because we have no information about what is happening in the interval), but to jump to a new level at the moment of changing interval, we have a *piecewise exponential* model. If the points where the hazard changes are known, this model can easily be fitted, using a factor variable to indicate the changes in level of the hazard, in place of the continuous function, $\log[H_0(t_i)]$, in Equation (6.3).

6.4.2 Cox Model

In the models of the previous section, the baseline hazard, $h_0(\cdot)$, was given a specific form. If this form is left undefined (that is, "nonparametric") before observing the data, the only information that we have is at the time points where an event occurs. Thus, the maximum likelihood estimates will have the hazard function changing at these points — its form is unknown in between and, hence, can be assumed to be constant. This yields the "semiparametric" *Cox model* that is, thus, equivalent to a piecewise exponential model, where the change points occur at the events. This is one of several possible ways in which it can be fitted. It requires special programming, available in many software packages.

Example

For the leukaemia survival data of Table 6.1, the Cox model without treatment differences has an AIC of 255.8, whereas that with such differences has 242.4. Although these are data originally used by Cox (1972) to illustrate use of this model, these values are both larger than and, hence, inferior to those obtained for the exponential and Weibull models above. This is due to the large number of parameters that must be estimated in this "semi-parametric" model. The difference in log risk is now estimated to be 1.521, very similar to that for the exponential model. □

Summary

Techniques for analyzing survival data are widely known and well documented. Because they are not generalized linear models, at least in the approaches described in this chapter, we shall not consider them further here. Many books on survival analysis are available; standard texts include those by Kalbfleisch and Prentice (1980), Lawless (1982), Cox and Oakes (1984), Fleming and Harrington (1991), and Andersen *et al.* (1993).

6.5 Exercises

1. Survival times (weeks) were recorded for patients with acute myelogeneous leukaemia, along with white blood cell counts (wbc) in thousands and AG-factors (Feigl and Zelen, 1965):

Time	wbc	Time	wbc	Time	wbc	Time	wbc
Positive AG-factor							
65	2.3	156	0.8	100	4.3	134	2.6
108	10.5	121	10.0	4	17.0	39	5.4
56	9.4	26	32.0	22	35.0	1	100.0
5	52.0	65	100.0	56	4.4	16	6.0
143	7.0	1	100.0				
Negative AG-factor							
65	3.0	17	4.0	7	1.5	16	9.0
3	10.0	4	19.0	2	27.0	3	28.0
4	26.0	3	21.0	30	79.0	4	100.0
22	5.3	8	31.0	43	100.0		

Patients were classified into the two groups according to morphological characteristics of their white blood cells: AG positive had Auer rods and/or granulature of the leukaemia cells in the bone marrow at diagnosis, whereas AG negative did not. No patients had palpable enlargement of the liver or spleen at diagnosis. The four largest white

blood cell counts were actually greater than 100,000. Notice that none of the survival times are censored. Do either of these variables help to predict survival time? Try transformations of the white blood cell counts.

2. A total of 90 patients suffering from gastric cancer were randomly assigned to two groups. One group was treated with chemotherapy and radiation, whereas the other only received chemotherapy, giving the following survival times in days (Gamerman, 1991, from Stablein *et al.*) (asterisks indicate censoring):

Chemotherapy + radiation				Chemotherapy			
17	167	315	1174*	1	356	524	977
42	170	401	1214	63	358	535	1245
44	183	445	1232*	105	380	562	1271
48	185	464	1366	125	383	569	1420
60	193	484	1455*	182	383	675	1460*
72	195	528	1585*	216	388	676	1516*
74	197	542	1622*	250	394	748	1551
95	208	567	1626*	262	408	778	1690*
103	234	577	1736*	301	460	786	1694
108	235	580		301	489	797	
122	254	795		342	499	955	
144	307	855		354	523	968	

Is there any evidence that radiation lengthens survival times?

3. The Eastern Cooperative Oncology Group in the United States of America conducted a study of lymphocytic nonHodgkin's lymphoma. Patients were judged either asymptomatic or symptomatic at the start of treatment, where symptoms included weight loss, fever, and night sweats. Survival times (weeks) of patients were classified by these symptoms (Dinse, 1982) (asterisks indicate censoring):

Asymptomatic			Symptomatic
50	257	349*	49
58	262	354*	58
96	292	359	75
139	294	360*	110
152	300*	365*	112
159	301	378*	132
189	306*	381*	151
225	329*	388*	276
239	342*		281
242	346*		362*

Patients with missing symptoms are not included. Do the symptoms provide us with a means of predicting differences in the survival time?

4. Fifty female black ducks, *Anas rubripes*, from two locations in New Jersey, USA, were captured by the U.S. Fish and Wildlife Service over a four-week period from 8 November to 14 December, 1983. The ducks were then fitted with radio emitters and released at the end of the year. Of these, 31 were born in the year (age 0) and 19 the previous year (age 1). Body weight (g) and wing length (mm) were recorded. Usually, these are used to calculate a condition index, the ratio of weight to wing length. The status of each bird was recorded every day until 15 February 1984 by means of roof-mounted antennae on trucks, strut-mounted antennae on fixed-wing airplanes, and hand-held antennae on foot and by boat. The recorded survival times were (Pollock *et al.*, 1989) (asterisks indicate censoring):

Age	Weight	Wing	Time	Age	Weight	Wing	Time
1	1160	277	2	0	1040	255	44
0	1140	266	6*	0	1130	268	49*
1	1260	280	6*	1	1320	285	54*
0	1160	264	7	0	1180	259	56*
1	1080	267	13	0	1070	267	56*
0	1120	262	14*	1	1260	269	57*
1	1140	277	16*	0	1270	276	57*
1	1200	283	16	0	1080	260	58*
1	1100	264	17*	1	1110	270	63*
1	1420	270	17	0	1150	271	63*
1	1120	272	20*	0	1030	265	63*
1	1110	271	21	0	1160	275	63*
0	1070	268	22	0	1180	263	63*
0	940	252	26	0	1050	271	63*
0	1240	271	26	1	1280	281	63*
0	1120	265	27	0	1050	275	63*
1	1340	275	28*	0	1160	266	63*
0	1010	272	29	0	1150	263	63*
0	1040	270	32	1	1270	270	63*
1	1250	276	32*	1	1370	275	63*
0	1200	276	34	1	1220	265	63*
0	1280	270	34	0	1220	268	63*
0	1250	272	37	0	1140	262	63*
0	1090	275	40	0	1140	270	63*
1	1050	275	41	0	1120	274	63*

What variables, including constructed ones, influence survival?

This page intentionally left blank

7
Event Histories

An event history is observed when, in contrast to survival data, events are not absorbing but repeating, so that a series of events, and the corresponding durations between them, can be recorded for each individual.

Many simple event histories can be handled in the generalized linear model context. Some of these were covered in Chapter 4. If the intervals between events are independently and identically distributed, we have a *renewal process* that can be fitted in the same way as ordinary survival models. This is generally only realistic in engineering settings, such as the study of times between breakdowns of machines or the replacement of burned out light bulbs. If the distribution of the intervals only depends on what has happened before an interval begins, the time series methods of Chapter 5 can be applied, by conditioning on the appropriate information.

However, if there are variables that are changing within the intervals, that is, time-varying explanatory variables, the probability distribution can no longer easily be modelled directly. Instead, one must work with the intensity function, as for the nonhomogeneous Poisson processes of Chapter 5.

In even more complex situations, an event signals a change of state of the subject, so that there will be a number of different intensity functions, one for each possible change of state. This is known as a *semiMarkov* or *Markov renewal process*.

7.1 Event Histories and Survival Distributions

An *event history* follows an individual over time, recording the times of occurrence of events. As we have seen, survival data are a special case, where the first event is absorbing so that the process stops. If the successive intervals in an event history are independent, they may simply be modelled as survival distributions. Then, we have a renewal process, so that things start over "as new" at each event. Usually, this will not be the case, because we are interested in modelling the evolution of each individual over time.

As we have seen in Chapter 6, survival data can be modelled equivalently by the probability density, the survival function, or the intensity function. Common survival distributions include the exponential, gamma, Weibull, extreme value, log normal, inverse Gaussian, and log logistic. Two important families of models are the proportional hazards, discussed in Chapter 6, with

$$h(t;\alpha,\beta) = h_0(t;\alpha)g^{-1}(\beta)$$

where $g(\cdot)$ is a link function, usually the log link, giving Equation (6.2), and the *accelerated lifetime* models with

$$h(t;\beta) = h_0(t\,\mathrm{e}^{-\beta})\mathrm{e}^{-\beta}$$

Both of these model the intensity, instead of the probability density, although they cover some of the densities mentioned above. The most famous example is the Cox proportional hazards model. In the present context, this is often called a multiplicative intensities model.

When the probability density is modelled directly, all conditions describing the subjects must be assumed constant within the complete duration until an event occurs, because that total duration is the response variable. Thus, even in the case of survival data, time-varying explanatory variables cannot easily be modelled by the density function. The solution is to allow the intensity function to vary over time and to model it directly: a non-homogeneous Poisson process. This may, however, render certain effects of interest difficult to interpret. For example, what does a difference in treatment effect in a randomized clinical trial really mean if time-varying variables are changing in different ways in the treatment groups after randomization?

In fact, except for the exponential distribution, the intensity function does change, in any case, over time, but only as a strict function of time since the beginning of the period. Thus, for example, for the Weibull distribution, the intensity function can be written

$$h(t;\alpha,\beta) = t^{\alpha-1}g^{-1}(\beta)$$

a member of both families mentioned above. However, we now would like to introduce observed explanatory variables that may change over time, even in between events.

7.2 Counting Processes

To go further, it is fruitful to consider the *counting process* approach, already mentioned in Chapter 5. As its name suggests, a counting process is a random variable over time, $N(t)$, that counts the number of events that have occurred up to t. We then study changes in this variable, $dN(t)$. To do this, let the corresponding intensity of the process be $h(t|\mathcal{F}_{t-};\boldsymbol{\beta})$ such that

$$h(t|\mathcal{F}_{t-};\boldsymbol{\beta})dt = \Pr[dN(t) = 1|\mathcal{F}_{t-}]$$

where $\boldsymbol{\beta}$ is a vector of unknown parameters and $\mathcal{F}_{t-}$ represents any relevant aspects of the history of the process up to, but not including t, called the *filtration*.

Then, the log likelihood function for observation over the interval $(0, T]$ can be shown to be

$$\log[\mathrm{L}(\boldsymbol{\beta})] \quad = \quad \int_0^T \log[h(t|\mathcal{F}_{t-};\boldsymbol{\beta})]dN(t) - \int_0^T h(t|\mathcal{F}_{t-};\boldsymbol{\beta})I(t)dt$$

where $I(t)$ is an indicator variable, with value one if the process is under observation at time t and zero otherwise.

Now, in any empirical situation, the process will only be observed at discrete time intervals, such as once an hour, once a day, or once a week. Suppose that these are sufficiently small so that generally at most one event occurs in any interval, although there will be a finite nonzero probability of more than one.

With M intervals of observation, not all necessarily the same size, the log likelihood becomes

$$\log[\mathrm{L}(\boldsymbol{\beta})] \quad = \quad \sum_{t=1}^{M} \log[h(t|\mathcal{F}_{t-};\boldsymbol{\beta})]\Delta N(t) - \sum_{t=1}^{M} h(t|\mathcal{F}_{t-};\boldsymbol{\beta})I(t)\Delta_t$$

where Δ_t is the width of the tth observation interval and $\Delta N(t)$ is the change in the count during that interval, generally with possible values zero and one. This is just the log likelihood of a Poisson distribution for response, $\Delta N(t)$, with mean, $h(t|\mathcal{F}_{t-};\boldsymbol{\beta})\Delta_t$. Conditional on the filtration, it is the likelihood for a (local) Poisson process (Lindsey, 1995c). The structure that we shall place on this likelihood will determine what stochastic process we are modelling.

7.3 Modelling Event Histories

Consider now, in more detail, the event history of one individual. For simplicity of notation, take all observation intervals to be equal. Suppose that

we have a stochastic process where the waiting times between events, measured in these units, are the random variables, $\{Y_k; k = 1, \ldots, K\}$, none being censored for the moment. Then, the number of observation intervals will be $M = \sum Y_k$ and the series $\{\Delta N(t); t = 1, \ldots, M\}$ will contain K ones, because that many events occurred, and $M - K$ zeros, the number of observation points where no event occurred. For example, the observed waiting times $\{5, 3, 2, 4\}$ between events would give the series $\{0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 1\}$, a point process (Section 5.1.1).

If we fit a Poisson regression model with constant mean to such an observed point process, we obtain the maximum likelihood estimate of the (constant) intensity,

$$h(t|\mathcal{F}_{t-}; \lambda) = \lambda$$

of a homogeneous Poisson process, that is, with exponentially distributed waiting times (Sections 5.1.2 and 6.3.3).

Note that we are fitting a Poisson model to what appear to be binary data. If, instead, we fit the binomial logistic model that has been more usual for discrete time counting processes, we exclude the finite probability of more than one event occurring to an individual in each discrete observation interval. This would be acceptable for survival data, where events are absorbing, but it is not for most event history data.

7.3.1 Censoring

Suppose now that some of the observations are censored, under the usual assumptions about such censoring, that is, that it only depends on the information available up until that time (Section 6.1.2). Then, the last value in the observation period will be a zero instead of a one. In the example, if the third time between events, of two units, were, in fact, censored, the one at position ten in the series would be a zero, and the same log linear model would be fitted.

7.3.2 Time Dependence

Next we can assume that the intensity varies as a function of time since the previous event, say z_t. The vector $\mathbf{z}$ starts at one and increases by one at each observation interval up until and including the time of an event, after which it is reinitialized to one. For our example without censoring, it would be $\{1, 2, 3, 4, 5, 1, 2, 3, 1, 2, 1, 2, 3, 4\}$.

Thus, if

$$h(t|\mathcal{F}_{t-}; \boldsymbol{\beta}) = \mathrm{e}^{\beta_0} z_t^{\beta_1}$$

we have a Weibull intensity. This is a Poisson regression model in $\log(z_t)$. For the Weibull model, the maximum likelihood estimate of the usual ex-

ponent of the Weibull distribution in Equation (6.4) will be given by $\hat{\alpha} = \hat{\beta}_1 + 1$.

In the same way, any proportional hazards model, with log link, can be fitted as a Poisson regression model. This includes the Cox model that is fitted by using z_t as a factor variable. In other words, it has a different intensity for each observed waiting time, a piecewise exponential distribution.

Thus, the well-known continuous time Nelson–Aalen (Poisson model) and Kaplan–Meier (binomial model, discussed in Section 6.2) estimates can be obtained by this procedure, because they are identical to discrete time estimates with mass points at the event times.

Time-varying explanatory variables that are predictable processes are easily included in the model. Instead of an explanatory variable vector being constant over all observation intervals for an individual, it is allowed to vary. With "continuous" monitoring, it may change in every observation interval.

For our example with 14 observation periods, suppose that a variable changes three times, taking the successive values $\{1, 3, 2\}$ in the time periods $\{5, 5, 4\}$. Then, the appropriate variable would be $\{1, 1, 1, 1, 1, 3, 3, 3, 3, 3, 2, 2, 2, 2\}$. This is introduced into the Poisson regression model in the usual way.

Examples

1. We can, first, try applying this method to the leukaemia survival data of Table 6.1. The total vector length is 541, corresponding to the total number of weeks of observation for all patients. For the exponential model, the results are exactly those obtained in Chapter 6. For the Weibull distribution, the AICs are 237.5 and 221.5, respectively without and with treatment effect. The latter is estimated as 1.662 with $\hat{\alpha} = 1.268$. These results are perhaps more accurate than the previous ones, because they do not assume that times are measured exactly. Here, there is little indication that $\alpha \neq 1$ because the AIC for the exponential distribution with treatment effect was 221.0. Finally, for the Cox model, the AICs are 256.1 and 242.9, whereas the treatment effect is 1.509. These are very close to those obtained previously.

2. A baboon troop was studied in the Amboseli reserve in Kenya for a period of a year. In the dynamics of social behaviour, birth–immigration and death–emigration are of interest. Thus, times of births, deaths, emigration, and immigration were recorded, as in Table 7.1. Here, we have four different events that may depend on the size of the troop and on the time since a previous event of the same type or since any previous event. There are ten births, 12 deaths, three immigrations, and three emigrations, for a total of 28 events in 373 days of observation. The event (response) vector contains four parts, one for each type of event, for a total length of 1492.

The basic model with the same intensity for all events has an AIC of 284.8, whereas that with a factor variable, `TYPE`, allowing for four different

TABLE 7.1. Times of four types of events in a baboon troop (Andersen *et al.*, 1993, p. 41, from Altmann and Altmann)

Days	Troop size	Number	Event
41	40	1	Birth
5	41	1	Birth
22	42	1	Birth
2	43	1	Death
17	42	1	Death
26	41	2	Immigration
55	43	1	Birth
35	44	1	Immigration
20	45	1	Emigration
5	44	1	Death
6	43	1	Emigration
32	42	1	Death
4	41	2	Death
22	39	1	Death
10	38	2	Birth
7	40	1	Death
4	39	1	Birth
17	40	1	Death
11	39	1	Emigration
3	38	1	Birth
4	39	1	Death
8	38	1	Death
2	37	1	Death
5	36	1	Birth
10	37	1	Birth

intensities, reflecting the relative numbers of events, has 280.9. Letting each intensity depend on the troop size

$$\texttt{NUMBER} * \texttt{TYPE}$$

reduces the AIC to 276.7. Dependence on log time from any previous event

$$(\texttt{NUMBER} + \texttt{LOGTIME}) * \texttt{TYPE}$$

further lowers it to 275.6. Dependence on log times since a previous birth, death, and immigration, but where dependence on births is the same for all types of events, yields the model,

$$(\texttt{NUMBER} + \texttt{LOGTIME} + \texttt{LOGDTIME} + \texttt{LOGITIME}) * \texttt{TYPE} + \texttt{LOGBTIME}$$

where `LOGTIME, LOGDTIME, LOGITIME,` and `LOGBTIME` denote respectively the logarithms of total time since any previous event, time since a death,

since an immigration, and since a birth. This gives a final AIC of 254.7 with 21 parameters (there are only 28 events).

The size of the troop most positively influences emigration but negatively for the births. Total log time since any previous event positively influences migration and negatively affects births and deaths. All events except births are negatively influenced by log time since a death, whereas only emigration is negatively affected by log time since immigration. Finally, log time since a birth positively affects the risk of all events. □

With the small sample size in this example, even the standard AIC has led us to a model that is very complex. We may suspect that additional explanatory variables, and more observations, would be necessary to develop a satisfactory model, but they would surely be difficult to obtain in such a context.

7.4 Generalizations

The approach outlined in this chapter is an extension of the standard modelling of Markov chains in discrete time to Markov processes. It also directly applies to semiMarkov processes or Markov renewal processes; these allow a change of state at each event.

The linear regression model for intensities of Aalen (1989) can be fitted by replacing the log link by the identity link, although there is a clear risk of obtaining negative intensities. Doubly stochastic processes can be handled by adding random effects to the log linear model (Section 2.3.2). Certain parametric models, other than the multiplicative intensities models, can be analyzed by adding a special iteration step to the log linear model algorithm.

The "exact" estimates of the continuous time model can be approached, to any degree of numerical precision, by reducing the size of the observation intervals. However, this is largely illusory, because more precision is being imputed to the estimates than is available in the empirically observable data.

This approach has the inconvenience that the variable vectors may be rather large if a number of individuals are observed over considerable periods of time. However, because the risk set is calculated automatically, the handicap in total computing time required will not be nearly as great as it might appear.

This general approach of disaggregating data to the level of individual observation points can and should be generally used for event histories. Aggregating event counts over longer periods carries the assumption that the intensity within that period is constant, something that may be contradicted by the unaggregated data. (Of course, "long" is always relative to the intensity of events occurring.) The Poisson or binomial distribution

may be chosen depending on whether or not more than one event is possible at each observation time point.

7.4.1 Geometric Process

Baskerville *et al.* (1984) present a novel form of clinical trial for evaluating therapeutic differences. The general principle is that the best available treatment should be provided to each patient so that, when there are alternatives, the appropriate treatment must be determined for each patient. Suppose that a trial will have a fixed length determined by v regularly scheduled visits to a clinic, but variable treatment lengths. Thus, each patient is assigned the series of treatments under study in a sequence determined in the standard way for crossover trials. However, as the trial proceeds, double blinded, each patient remains on the assigned first treatment in the sequence between all visits until either clinical deterioration or adverse effects of the treatment indicate that a change is necessary. Then, at the following visit, the patient is switched to the next treatment in the assigned sequence, and so on. Thus, certain patients may remain on the first treatment for all visits in the trial if that treatment proves effective or may run out of new treatments before the v visits and stop the trial if no treatment proves effective.

A model based on the geometric distribution can be used to describe the number of continuations of the same treatment on such a trial (Lindsey and Jones, 1996). If π_j is the probability of changing treatment when on treatment j and N_{ij} is the random variable representing the number of continuations for patient i under treatment j, then the model for that treatment is

$$\Pr(N_{ij} = n_{ij}) = \begin{cases} \pi_j(1-\pi_j)^{n_{ij}} & \text{if } n_{ij} < u \\ (1-\pi_j)^u & \text{if } n_{ij} \geq u \end{cases}$$

where u is the number of remaining visits in the trial at the moment when patient i begins treatment j; for the first treatment, $u = v$, the total number of visits. Then, the complete model for individual i is the product of these probabilities for all treatments used by the patient. This is a simple duration or event history model in discrete time. In the usual way for crossover trials, we can allow π_j to depend on the history of the patient: carryover effect, period effect, and so on.

At any point in time, an individual on treatment j has probability π_j of requiring a change. If we ignore the history of the patient for the moment, this can be set up as a simple binary logistic regression with

$$\log\left(\frac{\pi_j}{1-\pi_j}\right) = \mu + \alpha_j$$

where the response vector is a series of zeros for visits where no change is made, and a one for each change of treatment. Here, j indicates the

treatment during the interval since the previous visit, not that to which a change can be made at the present visit. We can now begin to complicate our model by adding aspects specific to patients. As usual in crossover trials, the model should be stratified by individual patient or a random effect introduced. Here, we shall use the former.

A carryover effect can be introduced as a factor variable indicating the previous treatment. However, the interpretation of this is not as clear as in a classical crossover design, because the period during which each treatment is applied is not of constant length. For example, one might easily imagine that the carryover effect of the previous treatment diminishes with time on the present treatment; this can be incorporated in the model, as we shall do below.

The period effect is also not as clearly defined as in the classical case. Do periods correspond to visits or to treatment regimes? An alternative approach is perhaps more useful. The probability of change may depend in some way on the time, that is, the number of visits that the patient has been on the present treatment. We can look at both this time and its logarithm in the logistic model. The probability may also be a function of the number of previous changes of treatment. These two variables together may account for period effects and provide a more informative description of the treatment change mechanism.

In summary, we have a number of variables, describing each patient's position in the trial, that can be introduced into the logistic regression model:

- individual;
- treatment in the previous interval leading up to the visit;
- previous treatment (before the last change);
- (a function of) time, that is, number of visits since the last treatment change;
- (a function of) the number of previous treatment changes.

Once the data are set up for these variables, the analysis is straightforward.

Example

Baskerville *et al.* (1984) provide data on clinical decisions in a trial of three bronchodilators (A, B, C) in the treatment of chronic asthma involving 36 patients, as shown in Table 7.2. Nine weekly visits were programmed. Only three patients remained on the same treatment throughout the trial, two on A and one on B; six others completed the treatment sequence (that is, ran out of new treatments) and stopped early. An example of the way the data are set up for the model is given in Table 7.3 for patient 3.

TABLE 7.2. Data for preferences from a crossover trial. (Baskerville *et al.*, 1984)

Sequence		Sequence	
1	AABBBBBBB	4	BBBBBBBCA
	AAAAAAAAA		BBBBBBBCC
	AAAAABBBC		BCCCCCCCC
	ABBBBCCCC		BBBBBBCCA
	AAAAAABBC		BBCCCCCCA
	AAAAAAAAA		BCCCAAA
2	AAACBBBBB	5	CCABBBBBB
	AACCCCCCC		CAB
	ACCBBBBBB		CCAAAAABB
	AAAAAACCC		CAABB
	AAAAAAACC		CCCCCCABB
	ACB		CCAABBBBB
3	BBBBBBBBB	6	CBBBBBBBB
	BACCCCCCC		CBBBBBBBB
	BBBBBBAAA		CCCBBBBBB
	BBBBAC		CCCCBBBAA
	BACCCCCCC		CCCCBBBAA
	BBAAAAACC		CBBBA

TABLE 7.3. Coded data for individual 3 with sequence, AAAAABBBC, from Table 7.2.

Change	Visit	Treatment	Previous treatment	Time on treatment	Number of changes
0	1	1	0	1	0
0	2	1	0	2	0
0	3	1	0	3	0
0	4	1	0	4	0
1	5	1	0	5	0
0	6	2	1	1	1
0	7	2	1	2	1
1	8	2	1	3	1
0	9	3	2	1	2

The basic model with only treatment and individual effects,

$$\texttt{TREATMENT} + \texttt{SUBJECT}$$

has a deviance of 250.00 (AIC 326.0) with 261 d.f., but this deviance has no absolute meaning because we are modelling binary data. The addition of the carryover effect from previous treatment

$$\texttt{TREATMENT} + \texttt{PREVIOUS} + \texttt{SUBJECT}$$

reduces the deviance by 17.97 (314.0) on three d.f. The inclusion of an interaction between treatment and carryover,

$$\texttt{TREATMENT} * \texttt{PREVIOUS} + \texttt{SUBJECT}$$

only further reduces it by 4.54 (315.5) on three d.f.

If we now take into account the period effect through the number of previous changes and the time since a change (that is, the number of visits between changes), we discover that carryover is no longer necessary. A model with only treatment, individual, number of changes, and time since a change

$$\texttt{TREATMENT} + \texttt{NUMBER} + \texttt{TIME} + \texttt{SUBJECT}$$

has a deviance of 215.32 (295.3) with 259 d.f., as compared to 232.03 (314.0) with 258 d.f. for the model described above with treatment, individual, and previous treatment. However, the model may be further simplified by combining treatments B and C, for a deviance of 215.84 (293.8) with 260 d.f. A log transformation of either the number of previous changes or the time since change does not improve the model. These results are summarized in the first three panels of Table 7.4.

The parameter estimates are -1.268 for treatments B and C as opposed to A, -1.656 for the number of previous changes, and 0.237 for the time since a change. The changes in deviance for removing these effects are respectively 6.08 (AIC 297.9), 33.63 (325.5), and 4.88 (296.7), each with one d.f. Thus, the probability of change is less when a patient is on treatment B and C than for A; it decreases as the number of changes increases; and it increases with time from the previous change. The carryover effect previously detected has apparently disappeared when some of the other aspects of the patients history are properly taken into account.

Above, we mentioned the theoretical possibility of a diminishing effect of carryover as time passes without treatment change. This can be modelled as an interaction between time and the previous treatment:

$$\texttt{TREATMENT} + \texttt{NUMBER} + \texttt{TIME} * \texttt{PREVIOUS} + \texttt{SUBJECT}$$

When this is added to our previously retained model, the deviance reduces to 199.09 (289.1) with 254 d.f., a useful improvement. Again, this can be

TABLE 7.4. Deviances, AICs, and degrees of freedom for various models applied to the preference data of Baskerville *et al.* (1984).

	AIC	d.f.
Treatment	326.0	261
	Carryover	
Treatment + Carryover	314.0	258
Treatment + Carryover + Interaction	315.5	255
	Period	
Treatment + Changes	298.3	260
Treatment + Changes + Time	295.3	259
Treatment (A/B+C) + Changes + Time	293.8	260
	Carryover + period	
Treatment (A/B+C) + Changes + Time + Time×Carryover	289.1	254
Treatment (A/B+C) + Changes + Time + Time×Carryover (A/B+C+0)	288.7	258

simplified by contrasting treatment A with the others, including no previous treatment (at the beginning of the trial). The deviance is increased by only 7.63 (288.7) with four d.f. These results are summarized in the last panel of Table 7.4. The corresponding parameter estimates for this final model are 4.133 for carryover and -0.758 for its interaction with time: if the previous treatment is A, the probability of change is increased, although this effect diminishes with time since the last change. Thus, we finally do detect the dependence of continuation on previous drug, although with B and C similar, rather than A and C, as Baskerville *et al.* suggested. Finally, the parameter estimate for treatment difference for B and C as opposed to A is now -2.344 with a change in deviance of 9.54 (296.2) on one d.f. if it is removed from the model. This is a larger estimate, with a larger change in deviance, than in the simpler model above. The same type of change is also observed for the effects of the number of previous events and the time since the last change. (Lindsey and Jones, 1996) □

7.4.2 Gamma Process

Proportional hazards models can easily be fitted by Poisson regression. Intensities based on other distributions are more difficult because the Poisson regression is no longer linear. In pioneering work, Hurley (1992) has applied this approach to nonhomogeneous gamma processes. Because the gamma distribution can be thought of as a sum of ν exponentially distributed times,

TABLE 7.5. AICs for various gamma process models fitted to the stream data of Tables 7.6 and 7.7.

Model	$\nu = 1$	$\nu = 2$	$\nu = 3$
Intercept	1341.5	1390.4	1493.0
One-year cycle	1322.2	1344.3	1420.0
Linear trend	1303.4	1300.1	1350.3
Six-month cycle	1299.7	1288.0	1329.8
Four-month cycle	1305.3	1288.4	1325.6

the intensity function can be written

$$h(t;\boldsymbol{\beta}) = h_0(t;\boldsymbol{\beta}) \Pr[Z(t) = \nu - 1 | Z(t) \leq \nu - 1] \tag{7.1}$$

where $Z(t)$ is a Poisson random variable with mean $\int_0^t h_0(u;\boldsymbol{\beta})du$ and h_0 is the baseline intensity. For $\nu = 1$, we have a Poisson process. Then, the baseline intensity can be used in a regression model. When the log link is used,

$$\log[h_0(t;\boldsymbol{\beta})] = \boldsymbol{\beta}^T \mathbf{x}(t) \tag{7.2}$$

this is called a modulated gamma process. Such a model is relatively difficult to fit because $\boldsymbol{\beta}$ appears both in the baseline intensity and in the conditional probability of Equation (7.1). For a fixed value of ν, an iterative procedure can be developed (Hurley, 1992).

Example

Two small unregulated streams in the Lake District of England, Great Eggleshope and Carl Becks, were monitored for bedload throughput in a study to investigate effects of human activities on salmonid fish (*Salmo trutta*). A bedload event is any discharge that transports coarse sand and granules, greater than 2 mm, into a specially constructed trap dug across each stream close to the catchment outlets. These events usually occur during heavy rainfall and may influence spawning success of the fish. The series for the two becks over a period of about five years are shown in Tables 7.6 and 7.7. Bedload events display a seasonal cycle with winter peaks and summer lows. A longer-term trend over time is also possible. Thus, five models were fitted with each of $\nu = 1$, 2, and 3, including, successively, constant intensity, one-year cycle, linear trend, six-month cycle, and four-month cycle. The corresponding AICs are given in Table 7.5. We see that a gamma process with $\nu = 2$, linear trend, and both year and six-month cycles is indicated. This has separate parameters for each stream. When a common model is fitted to the two, the AIC becomes 1273.0. Interestingly, the linear trend is negative (-0.00066). The intensity function over a year is plotted in Figure 7.1. □

TABLE 7.6. Times of bedloads in Carl Beck. (Hurley, 1992)

Time	Date	Magnitude	Time	Date	Magnitude
0	5 10 78	4.67	629	25 6 80	35.40
8	13 10 78	0.57	634	30 6 80	14.20
41	15 11 78	34.02	664	30 7 80	96.8
49	23 11 78	117.96	672	7 8 80	12.80
64	8 12 78	6.31	732	6 10 80	22.40
80	24 12 78	302.54	743	27 10 80	72.90
93	6 1 79	135.57	771	14 11 80	23.90
119	1 2 79	2.63	777	20 11 80	5.46
144	26 2 79	206.29	781	24 11 80	115.00
151	5 3 79	795.80	798	11 12 80	103.40
157	11 3 79	800.00	820	3 1 81	21.35
166	20 3 79	111.64	831	14 1 81	86.00
174	28 3 79	19.48	838	21 1 81	3.63
187	10 4 79	6.66	851	3 2 81	127.85
215	8 5 79	3.79	884	7 3 81	27.00
225	18 5 79	8.05	887	10 3 81	106.95
235	28 5 79	1.27	899	22 3 81	205.90
356	26 9 79	15.94	900	23 3 81	69.20
375	15 10 79	3.57	936	28 4 81	46.80
391	31 10 79	0.84	1087	26 9 81	141.60
394	3 11 79	6.75	1092	1 10 81	194.60
402	11 11 79	4.23	1099	8 10 81	10.30
407	16 11 79	0.49	1149	27 11 81	314.80
416	25 11 79	365.00	1185	3 1 82	1410.00
423	2 12 79	238.8	1245	3 3 82	35.5
424	3 12 79	21.20	1305	2 5 82	73.20
425	4 12 79	255.7	1360	26 6 82	78.90
429	8 12 79	21.50	1462	6 10 82	38.70
433	12 12 79	2.67	1508	21 11 82	344.30
436	15 12 79	2.39	1525	8 12 82	261.10
447	26 12 79	40.40	1553	6 1 83	94.50
456	4 1 80	30.40	1617	10 3 83	65.10
481	29 1 80	27.10	1661	23 4 83	32.50
491	8 2 80	96.00	1698	30 5 83	41.30
505	22 2 80	2.48	1829	8 10 83	6.10
523	11 3 80	1.85	1837	16 10 83	101.96
537	25 3 80	18.10	1891	9 12 83	22.20
541	29 3 80	2.61	1915	3 1 84	311.40
610	6 6 80	5.72	1932	20 1 84	115.50
618	14 6 80	17.20	1953	10 2 84	980.00
624	20 6 80	3.97	1999	20 3 84	22.30

TABLE 7.7. Times of bedloads in Great Eggleshope Beck. (Hurley, 1992)

Time	Date	Magnitude	Time	Date	Magnitude
0	8 12 78	1881.72	680	26 10 80	62.95
17	25 12 78	5438.17	710	17 11 80	9.27
81	27 2 79	217.49	713	20 11 80	9.41
85	3 3 79	5438.17	717	24 11 80	33.40
91	9 3 79	3641.50	725	2 12 80	9.88
107	25 3 79	715.05	734	11 12 80	24.70
114	1 4 79	351.88	737	14 12 80	550.10
124	11 4 79	2333.33	746	23 12 80	96.35
149	6 5 79	35.63	767	14 1 81	18.34
163	20 5 79	27.60	774	21 1 81	125.50
170	27 5 79	17.29	786	2 2 81	5440.00
292	26 9 79	7.39	820	7 3 81	629.78
300	4 10 79	15.85	837	24 3 81	593.19
309	13 10 79	61.76	872	28 4 81	709.94
321	25 10 79	6.06	1023	26 9 81	1590.05
330	3 11 79	30.14	1028	1 10 81	5438.17
338	11 11 79	29.60	1076	18 11 81	193.50
343	16 11 79	15.80	1081	23 11 81	5438.17
352	25 11 79	3461.71	1121	3 1 82	5438.17
359	2 12 79	159.30	1174	25 2 82	665.50
360	3 12 79	126.90	1241	2 5 82	33.00
361	4 12 79	1795.00	1302	2 7 82	168.70
369	12 12 79	105.40	1398	6 10 82	3401.40
374	17 12 79	30.50	1409	17 10 82	31.50
384	27 12 79	543.00	1435	12 11 82	119.10
392	4 1 80	32.90	1444	21 11 82	2590.00
418	30 1 80	151.60	1461	8 12 82	66.60
428	9 2 80	871.00	1472	19 12 82	2196.00
441	22 2 80	18.30	1489	6 1 83	692.80
459	11 3 80	9.98	1546	3 3 83	118.40
474	26 3 80	89.60	1597	23 4 83	548.20
480	1 4 80	14.00	1636	1 6 83	132.00
545	5 6 80	2276.00	1773	16 10 83	274.38
554	14 6 80	1970.00	1827	9 12 83	307.20
564	24 6 80	155.10	1842	24 12 83	2540.00
570	30 6 80	52.40	1861	13 1 84	1840.00
600	30 7 80	17.03	1883	4 2 84	4730.00
608	7 8 80	36.20	1935	26 3 84	277.40
630	29 8 80	10.20	2007	6 6 84	37.20

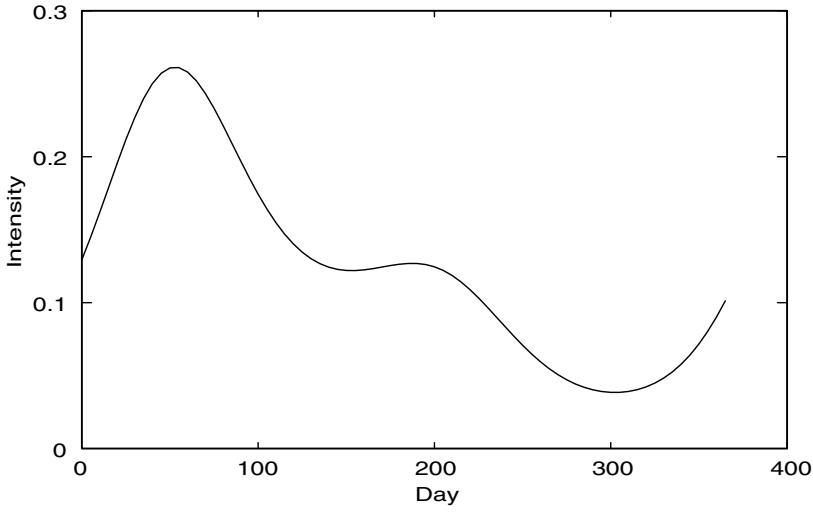

FIGURE 7.1. Variation in intensity of bed disturbances of the streams of Tables 7.6 and 7.7 over 365 days.

Summary

With the increasing possibilities of accumulating long series of data, event history analyses are growing in importance. Fortunately, analysis is fairly simple if somewhat computer intensive. For more details, see Blossfeld *et al.* (1989), Andersen *et al.* (1993), and Lindsey (1993, 1995b).

For crossover trials, see Jones and Kenward (1989) or Senn (1993).

7.5 Exercises

1. Repeat the exercises of the previous chapter using the methods of this chapter.

2. The following table gives the successive times (sec) at initiations of mating between flies (Aalen, 1978).

Ebony flies							
143	180	184	303	380	431	455	475
500	514	521	552	558	606	650	667
683	782	799	849	901	995	1131	1216
1591	1702	2212					
Oregon flies							
555	742	746	795	934	967	982	1043
1055	1067	1081	1296	1353	1361	1462	1731
1985	2051	2292	2335	2514	2570	2970	

Is there a difference over time in the intensity of mating between the two species?

3. In the Solomon–Wynne experiment (Kalbfleisch, 1985b, pp. 83–88), 30 dogs were taught to avoid an electrical shock. A dog was in a compartment with a floor through which a shock could be applied. The lights were turned out and a barrier raised; ten sec later, the shock occurred. Thus, the dog had ten sec, after the lights went out, to jump the barrier and avoid the shock. Each of the dogs was subjected to 25 such trials. Each line gives the results of the 25 successive trials for one dog, with a 1 indicating avoidance:

00101	01111	11111	11111	11111
00000	00100	00001	11111	11111
00000	11011	00110	10111	11111
01100	11110	10101	11111	11111
00000	00011	11111	11111	11111
00000	01111	00101	11111	11111
00000	10000	00111	11111	11111
00000	00110	01111	11111	11111
00000	10101	10100	01111	10110
00001	00110	10111	11111	11111
00000	00000	11111	10111	11111
00000	11111	00111	11111	11111
00011	01001	11111	11111	11111
00001	01101	11111	11111	11111
00010	11011	11111	11111	11111
00000	00111	11111	11111	11111
01010	00101	11101	11111	11111
00001	01011	11101	11111	11111
01000	01000	11111	11111	11111
00001	10101	10101	11111	11111
00011	11101	11111	11111	11111
00101	01111	11111	10011	11111
00000	00111	11111	11111	11111
00000	00011	10100	01101	11111
00000	01011	11010	11111	11111
00101	11011	01111	11111	11111
00001	01111	11111	11111	11111
00010	10111	01011	11111	11111
00001	10011	10101	01011	11111
00001	11111	01011	11111	11111

By conditioning on the numbers of previous events, determine if the dogs learn more from receiving or avoiding a shock.

4. The following table gives the June days with measurable precipitation (1) at Madison, Wisconsin, 1961–1971 (Klotz, 1973):

Year	
1961	10000 01101 01100 00010 01010 00000
1962	00110 00101 10000 01100 01000 00000
1963	00001 01110 00100 00010 00000 11000
1964	01000 00000 11011 01000 11000 00000
1965	10001 10000 00000 00001 01100 01000
1966	01100 11010 11001 00001 00000 11100
1967	00000 11011 11101 11010 00010 00110
1968	10000 00011 10011 00100 10111 11011
1969	11010 11000 11000 01100 00001 11010
1970	11000 00000 01000 11001 00000 10000
1971	10000 01000 10000 00111 01010 00000

Is there evidence of variation in periods without rainfall over the 11 years? Is there more chance of rain on a given day if it rained the day before?

5. The times (in days of service) at which valve seats were replaced on 41 diesel engines in a service fleet are shown in the following table (Lawless and Nadeau, 1995, from Nelson and Doganakov):

761	593
759	573 589
98 667	165 408 604 606
326 653 653 667	249 594
665	344 497 613
84 667	265 586 595
87 663	166 206 348 389
646 653	601
92 653	410 581 601
651	611
258 328 377 621 650	608
61 539 648	587
254 276 298 640 644	367 603
76 538 642	202 563 570 585
635 641	587
349 404 561 649	578
631	578
596	586
120 479 614	585
323 449 582	582
139 139 589	

The engines had 16 valves, but we do not know which ones were replaced. The last time for each engine is censored. We are interested in how the average number of valves replaced changes with age and whether the replacement rate increases with time. Develop an appropriate event history model for these data.

6. In the National Cooperative Gallstone Study (United States of America), an important point of interest was the safety of the drug chenodiol. A major concern was that, as the gallstones dissolve, they might pass into the biliary tree and lead to increased gallbladder symptoms. Progression of the disease is indicated by the occurrence of biliary tract pain, perhaps accompanied by other symptoms that might require surgical removal of the gallbladder, called cholecystectomy. Thus, in this study of the disease progression of floating gallstones, biliary pain may be experienced, and, if so, it may be followed by cholecystectomy. If the former does not occur, the latter will also be censored. A series of observations on 113 patients under two treatments, placebo and the drug chenodiol (Wei and Lachin, 1984), are presented on the next page.

Placebo				Treatment			
741*	741*	35	118	735*	735*	742*	742*
234	234	175	493	29	29	360*	360*
374	733*	481	733*	748*	748*	750	750*
184	491	738*	738*	671	671	360*	360*
735*	735*	744*	744*	147	147	360*	360*
740*	740*	380	761*	749	749	726*	726*
183	740*	106	735*	310*	310*	727*	727*
721*	721*	107	107	735*	735*	725*	725*
69	743*	49	49	757*	757*	725*	725*
61	62	727	727*	63	260	288	810*
742*	742*	733*	733*	101	744*	728*	728*
742*	742*	237	237	612	763*	730*	730*
700*	700*	237	730*	272	726*	360*	360*
27	59	363	727*	714*	714*	758*	758*
34	729*	35	733*	282	734*	600*	600*
28	497			615	615*	743*	743*
43	93			35	749*	743*	743*
92	357			728*	728*	733*	755*
98	742*			600*	600*	188	762*
163	163			612	730*	600*	600*
609	713*			735*	735*	613*	613*
736*	736*			32	32	341	341
736*	736*			600*	600*	96	770*
817*	817*			750*	750*	360*	360*
178	727			617	793*	743*	743*
806*	806*			829*	829*	721*	721*
790*	790*			360*	360*	726*	726*
280	737*			96	720*	363	582
728*	728*			355	355	324	324
908*	908*			733*	733*	518	518
728*	728*			189	360*	628	628
730*	730*			735*	735*	717*	717*
721*	721*			360*	360*		

[First column: time (days) to pain; second column: time (days) to cholecystectomy; with asterisks indicating censoring.] Although patients were randomly assigned to high dose (305), low dose (306), or placebo (305), the table only contains the results for high dose and placebo. Patients could be followed for as long as 28 months. Find an intensity model for these two events. Does treatment have an effect on either of them?

8
Spatial Data

Spatial data are similar to longitudinal data that evolve over time (Chapters 4, 5, and 7) in that there will generally be dependence among the responses. However, they are more complex because dependence can be on neighbours in all directions and not just through ordered unidimensional history. Thus, we shall not be able to develop a multivariate model that decomposes in a simple way, as in Equation (5.1). Generally, approximations have to be made.

8.1 Spatial Interaction

8.1.1 Directional Dependence

One of the simplest types of spatial data is the equivalent of a discrete time point process (Section 5.1.1): a record of presence or absence of some event at each point of a regular lattice. An important question that does not occur with time series data is whether dependence is directional. This can be studied by constructing a model whereby the probability of an event at each given point depends on the presence or absence of a neighbour in possibly different ways in each direction. This generally could involve a simple logistic model conditional on the nearest neighbours, a Markov process. If we are to construct a model for dependence on nearest neighbours, we lose the boundary cells because we do not have information on all of their neighbours (just as we lost the first observation of a time series). However, if all remaining observations are used, there will be a "circular" effect from

the illegitimate decomposition of the multivariate distribution. Each point would be used too many times because of the reciprocal dependencies. One possible remedy, with substantial loss of information, is to use only every other point (Besag, 1974).

One common multivariate model with dependence on neighbours in a lattice is the Ising model:

$$\Pr(\mathbf{y};\boldsymbol{\alpha},\boldsymbol{\beta}) = c(\boldsymbol{\alpha},\boldsymbol{\beta}) \exp\left(\sum_i y_i\alpha_i + \sum_i \sum_{j\neq i} \beta_{ij} y_i y_j\right) \tag{8.1}$$

where i indexes each position and j its neighbours and $c(\cdot)$ is the normalizing constant, generally an intractable function of the parameters. This distribution is a member of the exponential family. To be estimable, the model must have $\beta_{ij} = 0$ for j outside some neighbourhood of i.

For binary data, this is a rather peculiar model because the probability of all points with $y_i = 0$ is the same regardless of the configuration of their neighbours. Generally, the model is simplified by assuming stationarity, that is, that the parameters do not vary with i:

$$\Pr(\mathbf{y};\boldsymbol{\alpha},\boldsymbol{\beta}) = c(\boldsymbol{\alpha},\boldsymbol{\beta}) \exp\left(\sum_i y_i\alpha + \sum_i \sum_{j\neq i} \beta_j y_i y_j\right)$$

Usually, either four or eight neighbours are used so that $\boldsymbol{\beta}$ has this dimension. The original Ising model held β_j constant so that the probability just depends on the total number of neighbours and not on where they are located.

In spite of the complex normalizing constant, this model can be estimated by the techniques of Chapter 3 (see also Lindsey, 1995b, pp. 129–132). For example, the simplest model, for total number of neighbours, can be estimated from a cross-tabulation of an indicator of presence at each point by number of neighbours. Then, the Poisson regression has the model

$$\mathtt{C} + \mathtt{C} \cdot \mathtt{NUMBER}$$

where C is the binary indicator for each point. The one with four neighbours would require tabulation of the 2^5 table of all possible combinations of presence at a point with these neighbours, giving the Poisson regression,

$$\mathtt{C} + \mathtt{C} \cdot (\mathtt{N} + \mathtt{E} + \mathtt{S} + \mathtt{W})$$

where the four letters are binary indicators for a neighbour in the four directions. Notice how the main effects for number or for the directions are not included in the model. If they were, we would have a multivariate "distribution" for each point and its neighbours whereby each point would be included five times.

TABLE 8.1. Grid indicating presence of *Carex arenaria*. (Strauss, 1992, from Bartlett)

0	1	1	1	0	1	1	1	1	1	0	0	0	1	0	0	1	0	0	1	1	0	1	0
0	1	0	0	1	1	1	0	0	1	0	0	0	0	0	0	0	1	1	1	0	0	0	1
1	1	1	1	0	1	1	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0
0	0	1	1	1	1	1	0	1	0	1	0	0	0	0	0	1	1	0	0	0	0	0	0
0	1	1	0	1	1	1	1	1	0	0	0	0	1	1	1	0	1	1	0	0	0	0	0
0	1	0	0	0	1	0	1	0	1	1	1	1	1	0	0	0	0	1	1	0	0	0	0
0	1	0	1	1	0	1	0	1	0	0	0	1	0	0	1	1	0	0	1	0	0	0	0
0	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0
0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0
0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0
0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	1	1	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0
0	0	1	0	0	0	1	1	0	1	0	1	1	1	1	1	1	1	0	0	0	0	1	1
0	1	0	0	1	1	0	0	0	0	0	0	0	1	1	0	0	1	0	1	0	1	0	1
0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0	1	1	0	0	1
1	0	0	0	0	0	0	1	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	1	0	0	1	1	0	1	0	1
0	1	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0	1	1	1	1	1	0	1
1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	1	0	1
0	0	0	1	0	1	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1
1	0	0	0	0	1	1	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	1	0	0	0	0	1
0	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0

Example

Consider a 24×24 grid indicating the presence or absence of the plant, *Carex arenaria*, as shown in Table 8.1. The influence of neighbours can be summarized in a contingency table giving their numbers, as in Table 8.2. The plants obviously tend to have neighbours!

None of the Ising models described above fit well for the obvious reason, from Table 8.2, that the probability of zero is not constant for differing numbers of neighbours. □

Because of the intractable normalizing constant, this multivariate model has generally not been used. Instead, conditional models are applied, with the difficulties mentioned above. A simple conditional logistic (Ising) model has each lattice point depending on its four neighbours:

$$\mathtt{N} + \mathtt{E} + \mathtt{S} + \mathtt{W}$$

TABLE 8.2. Numbers of neighbours for the plant data of Table 8.1.

	Number of neighbours				
Presence	0	1	2	3	4
0	128	105	77	23	6
1	20	49	46	25	5

In terms of the Poisson regression above, this is

$$\mathtt{C} * (\mathtt{N} + \mathtt{E} + \mathtt{S} + \mathtt{W})$$

so that, in comparison to the multivariate Ising model above, it has the main effects for the neighbours (illegitimately) included. It does not correspond to the correct likelihood because of the circular effect mentioned above.

This approach, that has been called "pseudo-likelihood", may be misleading because of the misspecification of the multivariate distribution and, hence, of the likelihood function. The analysis should be repeated with every other point, once using only odd points and again only even points. Each of these analyses thus eliminates the dependency but loses information.

Example

For our plant data, using all of the points, this model has an AIC of 569.4. The north and south parameters, of similar size, are slightly large than the east and west ones, that are also of similar size. We can also add dependence on diagonal neighbours:

$$\mathtt{N} + \mathtt{E} + \mathtt{S} + \mathtt{W} + \mathtt{NE} + \mathtt{SE} + \mathtt{SW} + \mathtt{NW}$$

for an AIC of 566.2. Here, the north–west and south–east estimates are slightly smaller, although all are of similar size. Elimination of north–west reduces the AIC to 565.7, but no other parameter can be removed. The latter is most likely an artifact, because, if we remove two rows around the edge, its elimination is no longer possible. Adding dependence on points two steps away does not reduce the AIC.

We repeat the analysis only using odd or even points. The four and eight directional AICs are summarized in the first panel of Table 8.3. We see, indeed, that the two half-models are not independent: if they were, their deviances (not their AICs) would add to give that for the complete set of points.

The parameter values are considerably different for the three analyses. Consider, for example, those for the four main directions, as shown in Table 8.4. (These values remain similar when the four diagonal directions are introduced.) Evidently, there is some east–west relationship that is not detected when all of the data are used. The negative value indicates some form of repulsion. □

TABLE 8.3. AICs for spatial models for the plant data of Table 8.1.

Neighbours	All	Odd	Even
	Directions		
Four	569.4	276.3	281.4
Eight	566.2	279.0	277.1
	Total		
Four	563.6	281.6	285.8
Four + four	554.5	279.6	280.3
Eight	552.7	278.4	278.3

TABLE 8.4. Directional parameters for the plant data of Table 8.1.

Direction	All	Odd	Even
North	0.564	0.545	0.540
South	0.606	0.919	0.525
East	0.481	−0.311	1.200
West	0.475	1.169	−0.264

The logistic model looks at the presence or absence of the event. However, more than one event may have occurred at a point. If the counts are available, the Poisson log linear model can be used. Instead, suppose that we only have the binary indicator. The Poisson probability of no event is $\exp(-\mu)$, so that the probability of one or more events is $1 - \exp(-\mu)$. If μ in a Poisson regression model were to have a log link, this latter expression can be transformed, for binary data, as a complementary log log link.

Example

For our grid of plants, replacing the logit by the complementary log log link yields a slightly poorer model, with a marginally larger AIC. Apparently, presence of plants, and not their number at each point, is important in this plot. □

8.1.2 Clustering

Another important question is if events appear together on the lattice. This is closely related to birth processes for longitudinal data. Clustering can be studied in a number of ways. One traditional way has been to divide the area up in to a reasonably coarse grid and count the number of events in each square. If these counts have a Poisson distribution, it is an indication that they are distributed randomly. However, if the variance is much larger than the mean, it is an indication of clustering.

TABLE 8.5. Distribution of the plants in Table 8.1 divided into 3×3 squares.

Count	0	1	2	3	4	5	6	7	8	9
Frequency	8	9	17	7	10	6	7	0	0	0

Example

We can divide our lattice of plants up into a grid of 64 nonoverlapping 3×3 squares. The resulting frequencies of the different counts are given in Table 8.5. The Poisson distribution fitted to these data has $\hat{\mu} = 2.75$. The variance is estimated to be 3.41, considerably larger. The AIC is 20.5 for a Poisson distribution as compared to 20 for the saturated model, indicating a reasonable fit, although there is some possibility of nonrandomness. □

We can also look to see if the probability of an event at a given point depends primarily on the number of neighbours. This was the simplest Ising model described above. Positive values indicate clustering and negative values repulsion. Such a conditional logistic model, depending on the number of neighbours, is identical to the Poisson regression,

$$\texttt{C} + \texttt{FACNUMBER} + \texttt{C} \cdot \texttt{NUMBER}$$

where `FACNUMBER` is the appropriate factor variable for the total number of neighbours.

Example

In the study of the lattice of plants above, we noticed that the parameters for the various directions were of similar size, at least when all points are used. This is an indication that only the number of neighbouring plants is important.

A Poisson model with separate counts for east–west and for north–south,

$$\texttt{NNS} + \texttt{NEW}$$

has an AIC of 565.4 for all points, better than the Markov processes above. A model for a combined four-directional count has 563.6 (notice that this model could be fitted directly from Table 8.2 as a linear trend). If we add a parameter for the total of diagonal neighbours, we reach 554.5, whereas if we take one count for all neighbours in the eight directions, we have 552.7. Thus, when we use all points, we have a very strong indication of clustering without directional influence. However, when we perform the analysis separately for odd and even points, we see, in the lower panel of Table 8.3, that the earlier directional model is superior. Thus, clustering is not a sufficient explanation for these data.

The linearity of the dependence on numbers of neighbours can be checked by fitting a factor variable (replacing `NUMBER` by `FACNUMBER` in the interaction in the Poisson regression), a spatial Rasch model. There can be from

zero to four four-directional neighbours and from zero to eight in all directions. Introduction of such variables, for all points, raises the AIC, hence providing no indication of nonlinearity. □

8.1.3 One Cluster Centre

Unfortunately, such techniques as already presented only indicate the general presence of clustering. They do not show where the clusters are located, a more difficult problem. If only one point of clustering is thought to occur on the grid, as, for example, might occur with illness events around a pollution site, a response surface (Sections 9.1 and 9.3) can be added to the Ising model to determine the area of highest probability of an event. This involves making α_i in Equation (8.1) depend on the coordinates of each point. In the conditional logistic regression, this might give, in the simplest case,

$$\log\left(\frac{\pi_{hv}}{1-\pi_{hv}}\right) = \alpha_0 + \alpha_1 h + \alpha_2 v + \alpha_3 h^2 + \alpha_4 v^2 + \alpha_5 hv + \beta n_{hv}$$

where h and v are the coordinates and n_{hv} is the number of neighbours at each point. If not only the probability of an event but also the dependence on neighbours varies with position, β_{ij} can also be made to depend on the coordinates.

Example

For our grid of plants, the introduction of the response surface models raises the AIC. There is no evidence of a single maximum. □

If several centres of clustering are to be detected, more descriptive "nonparametric" techniques will generally be required.

8.1.4 Association

A common problem in spatial studies is to determine what different events are associated together. Here, we are not concerned with the distance to neighbours or with the number of close neighbours of the same kind, but with what different kinds of neighbouring events are found most often together. If the space is fairly uniform so that association is not changing, such data can be summarized as a square contingency table, with the size determined by the number of possible events. A number of the models in Section 2.2 may be of use here.

Example

Consider the species of trees that are associated together in a Michigan, USA, hardwood forest, as shown in Table 8.6. For each species, the number

TABLE 8.6. Association between tree species in a hardwood forest. (Hand *et al.* 1994, p. 19, from Digby and Kempton)

	Neighbour						
Species	Red oak	White oak	Black oak	Hickory	Maple	Other	Total
Red oak	104	59	14	95	64	10	346
White oak	62	138	20	117	95	16	448
Black oak	12	20	27	51	25	0	135
Hickory	105	108	48	355	71	16	703
Maple	74	70	21	79	242	28	514
Other	11	14	0	25	30	25	105

TABLE 8.7. Parameter estimates for the tree association data of Table 8.6.

		Estimate
Red oak	White oak	0.000
	Black oak	−1.538
	Hickory	0.503
	Maple	0.132
	Other	−1.751
White oak	Black oak	−1.107
	Hickory	0.620
	Maple	0.310
	Other	−1.395
Black oak	Hickory	−0.201
	Maple	−0.967
	Other	−13.80
Hickory	Maple	0.215
	Other	−1.082
Maple	Other	−0.735

of times that any given species is its nearest neighbour was counted. Notice that this table does not have a uniformly large diagonal, as in mobility studies. Only certain species, such as hickory and maple, seem to be clustered in more uniform groups.

If we apply the symmetry model of Section 2.2.2, we find that it fits very well. The AIC is 50.72 as compared to 72 for the saturated model. The parameter estimates for symmetry are given in Table 8.7. We see that the most frequent associations are red and white oaks each with hickory, whereas the least frequent are red and black oaks with other (the latter with none!), and red with black oak. However, recall that this model only looks at different neighbours, ignoring the main diagonal. □

8.2 Spatial Patterns

8.2.1 Response Contours

The search for patterns in spatial data is often complex. In most cases, there is no reason to expect a theoretical functional model, describing the shape of the pattern, to hold. Thus, "nonparametric" procedures, such as splines and kernel smoothing, will be appropriate to provide a descriptive plot that can be interpreted.

In simple cases, mainly if there is only one maximum or minimum, standard response surface methodology (Sections 9.1 and 9.3) can be applied, although it ignores the dependency among adjacent responses. This often simply means fitting a quadratic equation to the coordinates of the points, perhaps with transformations of the variables.

Example

Three sites, in Nevada, Texas, and Washington, USA, were proposed for high-level nuclear waste. The chosen site would eventually contain more than 68,000 canisters placed in holes, buried deep in the ground and surrounded by salt, about 30 ft apart over an area of about two square miles. The radioactive heat could cause tiny quantities of water in the salt to move toward the canisters until they were surrounded by water. The chemical reaction between salt and water could create hydrochloric acid that would corrode the canisters. Hence, the danger of leaks. Here, we consider the proposed site in Deaf Smith County, Texas.

Groundwater flow was measured in the region surrounding the proposed site to determine which way pollution might run if there was leakage. Piezometric head data were collected for the Wolfcamp Aquifer in West Texas and New Mexico by drilling 85 narrow pipes into the aquifer and letting the water find its own level in each pipe. Water level, above sea level (ft),

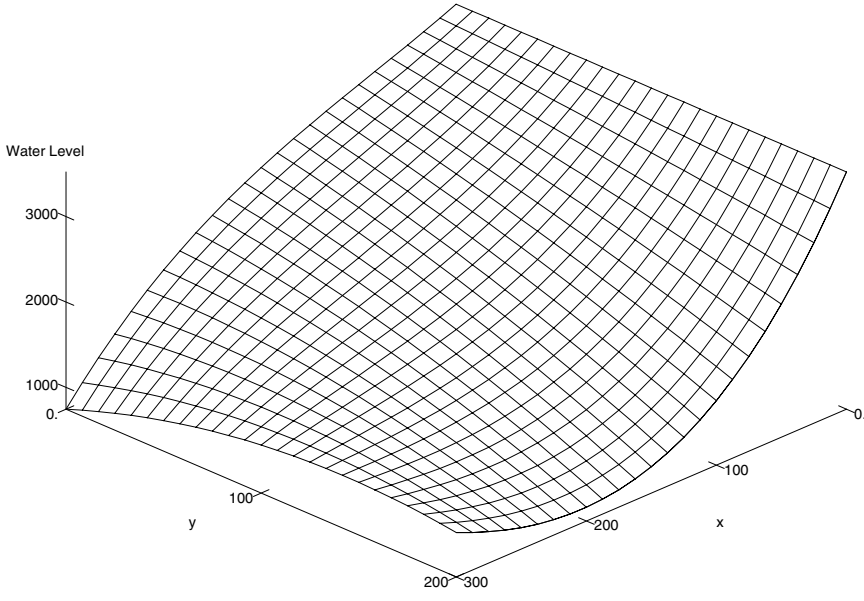

FIGURE 8.1. Linear response surface for the aquifer data in Table 8.8.

was measured (miles) at the 85 points, from an arbitrary origin, as shown in Table 8.8.

Because of the possibility of an irregular shape, it is perhaps best to start with at least a cubic surface. After elimination of some unnecessary terms, we obtain a model with a normal distribution and the equation

$$\begin{aligned}\mu_{hv} &= 3523 - 6.718h + 0.01526h^2 - 0.09097hv - 0.00007912h^3 \\ &\quad +0.0003995h^2v - 0.0001011hv^2\end{aligned}$$

(150 miles have been added to each h-value to make them all positive.) This model has an AIC of 1131 as compared to a full cubic model with 1135 and a full quadratic model with 1137. The h^3 term could also be eliminated without increasing the AIC, but, in this descriptive situation, it seems best to leave it in.

Neither transformation of the explanatory variables, nor change of distribution, nor change of link function improves the model. We see, in Figure 8.1, that the water level decreases sharply toward the higher values of the two coordinates. Although smoother, this figure has the same basic shape as that found by Cressie (1989) by much more complicated methods. It clearly answers the question about leakage. □

TABLE 8.8. Coordinates (miles), from an arbitrary origin, and water level (ft above sea level) around a proposed nuclear waste site. (Cressie, 1989)

h	v	Level	h	v	Level
42.78275	127.62282	1464	103.26625	20.34239	1591
−27.39691	90.78732	2553	−14.31073	31.26545	2540
−1.16289	84.89600	2158	−18.13447	30.18118	2352
−18.61823	76.45199	2455	−18.12151	29.53241	2528
96.46549	64.58058	1756	−9.88796	38.14483	2575
108.56243	82.92325	1702	−12.16336	39.11081	2468
88.36356	56.45348	1805	11.65754	18.73347	2646
90.04213	39.25820	1797	61.69122	32.49406	1739
93.17269	33.05852	1714	69.57896	33.80841	1674
97.61099	56.27887	1466	66.72205	33.93264	1868
90.62946	35.08169	1729	−36.65446	150.91456	1865
92.55262	41.75238	1638	−19.55102	137.78404	1777
99.48996	59.15785	1736	−21.29791	131.82542	1579
−24.06744	184.76636	1476	−22.36166	137.13680	1771
−26.06285	114.07479	2200	21.14719	139.26199	1408
56.27842	26.84826	1999	7.68461	126.83751	1527
73.03881	18.88140	1680	−8.33227	107.77691	2003
80.26679	12.61593	1806	56.70724	171.26443	1386
80.23009	14.61795	1682	59.00052	164.54863	1089
68.83845	107.77423	1306	68.96893	177.24820	1384
76.39921	95.99380	1722	70.90225	161.38136	1030
64.46148	110.39641	1437	73.00243	162.98960	1092
43.39657	53.61499	1828	59.66237	170.10544	1161
39.07769	61.99805	2118	61.87429	174.30178	1415
112.80450	45.54766	1725	63.70810	173.91453	1231
54.25899	147.81987	1606	5.62706	79.08730	2300
6.13202	48.32772	2648	18.24739	77.39191	2238
−3.80469	40.40450	2560	85.68824	139.81701	1038
−2.23054	29.91113	2544	105.07646	132.03181	1332
−2.36177	33.82002	2386	−101.64278	10.65106	3510
−2.18890	33.68207	2400	−145.23654	28.02333	3490
63.22428	79.49924	1757	−73.99313	87.97270	2594
−10.77860	175.11346	1402	−94.48182	86.62606	2650
−18.98889	171.91694	1364	−88.84983	76.70991	2533
−38.57884	158.52742	1735	−120.25898	80.76485	3571
83.14496	159.11558	1376	−86.02454	54.36334	2811
−21.80248	15.02551	2729	−72.79097	43.09215	2728
−23.56457	9.41441	2766	−100.17372	42.89881	3136
−20.11299	22.09269	2736	−78.83539	40.82141	2553
−16.62654	17.25621	2432	−83.69063	46.50482	2798
29.90748	175.12875	1024	−95.61661	35.82183	2691
100.91568	22.97808	1611	−87.55480	29.39267	2946
101.29544	22.96385	1548			

TABLE 8.9. Locations (km) of brain cancer cases in San Francisco, California, USA, (1977–1986). (Selvin, 1991, p. 123)

h	v	Distance	h	v	Distance
−1.136	0.718	1.007	−4.865	−1.154	3.169
−0.163	−3.429	4.067	−6.156	−1.854	4.636
−2.718	4.185	4.049	1.036	−0.013	3.043
−3.090	−0.532	1.313	0.474	−0.568	2.590
−4.103	0.111	2.105	−0.141	−5.267	5.774
−4.664	2.567	3.564	−2.741	4.196	4.064
−5.528	−1.899	4.105	−2.865	−2.249	2.597
−0.842	−0.363	1.288	−4.036	−1.093	2.412
2.330	−5.074	6.824	−4.513	3.042	3.794
−1.393	−2.411	2.681	−4.849	−1.142	3.149
−1.636	−5.274	5.486	−6.265	−0.282	4.292
−2.732	4.177	4.044	1.257	4.304	5.239
−3.055	−0.513	1.273	0.748	−1.574	3.271
−4.346	−3.017	3.982			

8.2.2 Distribution About a Point

The distribution of cases of illness about some point source, such as a site of pollution, is often of critical interest. This problem may be attacked in a number of ways. For example, one may check if the cases are uniformly distributed with distance from the source, or one may fit a response surface based on a Poisson distribution to the numbers of events in squares of a lattice around the source. One major problem is that population density may vary in the area. For Poisson regression, one can allow for this by using the logarithm of the population in each square as an offset, so that rates can be studied.

Example

A large microwave tower located near the centre of San Francisco, California, USA, was considered to be a possible source of exposure for brain cancer. In the period 1977 to 1986, 27 cases were found in white people under 21 years old, located as in Table 8.9. The tower is at (−2.0, 0.2). Unfortunately, we have no information about the distribution of population densities.

If we categorize the distances into 14 0.5-km groups. we obtain Table 8.10. The area covered in a 0.5-km ring increases with the square of the distance, so that we use the logarithm of this as an offset. (Thus, we are implicitly forced to assume that population density is uniform over this area.) When we fit a Poisson regression, the null model has an AIC of 53.9 and the saturated model 28, whereas the model linear in distance has 20.9.

TABLE 8.10. Counts of brain cancer cases at distances grouped by 0.5 km, from Table 8.9.

Distance	Count	Distance	Count
0.25	0	3.75	3
0.75	0	4.25	6
1.25	4	4.75	1
1.75	0	5.25	2
2.25	2	5.75	1
2.75	3	6.25	0
3.25	4	6.75	1

TABLE 8.11. Counts of brain cancer cases in one-km squares, from Table 8.9.

	h									
v	−6.5	−5.5	−4.5	−3.5	−2.5	−1.5	−0.5	0.5	1.5	2.5
−5.5	0	0	0	0	0	1	1	0	0	1
−4.5	0	0	0	0	0	0	0	0	0	0
−3.5	0	0	1	0	0	0	1	0	0	0
−2.5	0	0	0	0	1	1	0	0	0	0
−1.5	1	1	3	0	0	0	0	1	0	0
−0.5	1	0	0	2	0	0	1	1	1	0
0.5	0	0	1	0	0	1	0	0	0	0
1.5	0	0	0	0	0	0	0	0	0	0
2.5	0	0	1	0	0	0	0	0	0	0
3.5	0	0	1	0	0	0	0	0	0	0
4.5	0	0	0	0	3	0	0	0	1	0

The slope is estimated to be −0.714, indicating a decrease in cancer cases with distance from the tower.

If we divide the region into nonoverlapping one km squares, we obtain the results in Table 8.11. Fitting a quadratic Poisson response surface, even nonlinear, provides very little indication that the surface is not flat. This may be because of the low number of events or because the surface is too complex to be captured by such a simple model.

Thus, the first model, only taking into account distance but not direction, with its strong assumptions, detects a variation in cancer rates with distance from the point source. However, when the assumptions are weakened, by allowing differences with direction, the paucity of data prevent us from even detecting this. □

Summary

Spatial modelling is a difficult subject for which generalized linear models are not well suited, although they should form the basis for more sophist-

icated models. Most approaches are rather *ad hoc* and descriptive, using nonparametric methods. Texts include Cressie (1993), Ripley (1981, 1988), Stoyan *et al.* (1987), and Upton and Fingleton (1985, 1989).

8.3 Exercises

1. The ants, *Messor wasmanni*, in northern Greece build nests underground with a single opening at ground level. The following table gives the counts of nests in July, 1980, on an 8×8 grid, the northern part of which is open scrub land and the southern is a field (Harkness and Isham, 1983):

Scrub
1 0 1 1 1 0 0 1
0 1 0 1 0 1 1 1
2 1 1 0 2 0 0 0
1 1 1 0 1 0 1 2
Field
1 1 1 0 1 2 0 1
1 1 1 1 0 1 0 1
1 1 0 2 0 0 1 1
0 1 0 1 0 0 1 0

   The ants are unable to build nests in some parts of the scrub because of the bushes.

   This species of ants begin their journeys, to collect seeds for food, in columns following trails from which individuals branch off. The trails of different nests never cross. We are interested in any spatial relationship among the nests. Is there any indication of randomness, clustering, or repulsion?

2. On a smaller 4×8 grid, overlapping with that in the previous exercise, ant nests of the above species, *Messor wasmanni*, and another, *Catagliphis bicolor*, were enumerated (Harkness and Isham, 1983):

Messor	*Catagliphis*
2 1 1 0 3 0 2 0	0 1 2 0 1 1 0 0
0 1 1 2 1 1 2 1	0 0 1 0 1 0 0 0
1 2 0 2 0 1 0 1	0 1 1 1 0 1 1 0
0 2 1 2 1 0 0 1	0 0 0 1 0 0 0 2

   (Thus, these two subtables should be superimposed.) The latter leave their nests singly in all directions, often travelling 15 m or more. They

collect dead insects, mainly *Messor* ants. Large numbers of the latter are killed by a hunting spider, *Zodarium frenatum*, around the nest entrances.

Can a model be set up to describe the interactions between these two species on the same plot (the field)?

3. During World War II, German flying bombs were landing in London, England. It was important to learn if the bombs possessed a guidance system or were landing at random. South London was divided into 576 small subdivisions of equal area of about 0.25 km^2 and the number of hits in each recorded (Feller, 1957, p. 150):

Count	0	1	2	3	4	5
Frequency	229	211	93	35	7	1

Check whether or not the hits could be considered to be at random.

4. Cases of oral/pharyngeal cancer were recorded from 1978 to 1981 in Contra Costa County, California, USA, separately for males and females. The longitude and latitude of each point are given below (Selvin, 1991, p. 137):

Females		Males	
121.91	38.15	121.60	38.02
121.90	37.88	121.25	38.20
122.05	37.93	122.01	37.95
122.10	37.95	122.04	37.91
122.00	37.85	122.05	37.88
122.15	37.88	122.08	37.98
122.12	37.80	122.01	37.81
122.25	37.78	122.11	38.15
122.20	37.95	122.32	37.83
		122.31	37.84

One possible source of influence might be the oil refining industry that we can take to be located at (122.3, 38.1). Is there any indication of difference in clustering between the sexes? Is there a link with distance to the oil refinery?

5. A uniformity trial studies the yield of one variety of grain over a lattice of plots. The Plant Breeding Institute in Cambridge, England, carried out such a trial for spring barley in 1978. The yield (kg) in each plot, 5 ft wide and 14 ft long, is given in the following table (Kempton and Howes, 1981):

2.97	2.83	2.67	2.53	2.08	2.14	2.21
3.00	2.72	2.46	2.39	2.06	1.86	2.47
2.85	2.77	2.15	2.18	2.07	2.10	2.67
2.89	2.82	2.64	2.35	2.06	1.99	2.74
3.11	2.80	2.64	2.29	2.07	2.20	2.75
2.99	2.85	2.67	2.50	2.02	2.40	2.82
2.79	2.45	2.71	2.47	2.21	2.32	2.89
2.76	2.25	2.87	2.75	2.52	2.45	2.81
2.81	2.39	2.71	2.74	2.52	2.48	2.91
2.59	2.63	2.55	2.79	2.22	2.54	3.09
2.40	2.52	2.55	2.73	2.44	2.73	3.27
2.20	2.93	2.48	2.54	2.37	2.64	3.29
2.25	2.64	2.28	2.38	2.36	2.75	3.11
2.00	2.88	2.69	2.58	2.47	2.98	3.21
2.29	2.81	2.69	2.62	2.79	3.02	3.20
2.48	2.64	2.52	2.61	2.51	3.06	3.36
2.46	2.78	2.61	2.75	2.58	3.14	3.55
2.40	2.56	2.25	3.07	2.39	2.93	3.40
2.40	2.49	2.44	3.16	2.51	2.95	3.02
2.22	2.44	2.32	2.72	2.43	3.04	3.32
2.23	2.24	2.35	2.64	2.32	3.25	2.98
1.98	2.46	2.64	3.06	2.54	3.20	3.17
2.29	2.43	2.84	3.17	2.71	2.96	3.28
2.09	2.33	2.95	3.08	2.58	2.68	2.76
2.36	2.46	2.85	3.02	2.43	2.77	2.92
2.20	2.34	2.64	2.69	2.72	2.45	2.76
2.69	2.85	2.73	2.67	2.79	2.52	2.89
2.56	2.80	2.47	2.60	2.43	2.36	3.19

There appears to be considerable variability among the plots due to soil heterogeneity, perhaps caused by differences in water stress, depth, compaction, nutrients, or previous croppings. Can you construct a model to take into account the dependence among yields on neighbouring plots?

6. Contact sampling for vegetation in an area involves randomly selecting points in a habitat of interest. Then, the species covering the point and its closest neighbour of a different species are recorded. (Notice the difference with respect to Table 8.6.) Counts of nearest neighbours were recorded by this kind of sampling in a field in British Columbia, Canada, in the spring of 1980 (de Jong and Greig, 1985):

Point	Neighbour									
	1	2	3	4	5	6	7	8	9	10
1 *Agropyron repens*	—	4	43	0	8	1	23	24	2	12
2 *Agrostis alba*	1	—	6	3	3	0	11	7	1	3
3 *Dact. glomerata*	41	10	—	10	106	12	123	156	13	60
4 *Holcus lanatus*	0	1	9	—	2	3	10	8	0	2
5 *Lolium perenne*	10	1	83	2	—	4	22	32	6	7
6 *Phleum pratense*	0	0	18	0	2	—	9	12	1	2
7 *Poa compressa*	15	5	59	4	15	3	—	52	4	29
8 *Trifolium repens*	10	11	55	4	12	3	34	—	3	11
9 *Plant. lancelotata*	1	0	7	1	2	2	4	2	—	1
10 *Tar. officinale*	9	0	32	0	4	0	11	12	2	—

Are neighbouring relationships symmetrical within pairs of different species?

7. As shown below and on the next page, coal ash (%) was measured on the Roberta Mine Property in Greene County, Pennsylvania, USA (Cressie, 1986, from Gomez and Hazen). Grid spacings are 2500 ft^2.

	1	2	3	4	5	6	7	8
24								
23								
22							11.62	10.91
21					10.39	10.65	10.36	9.58
20				9.79	9.06	10.70	11.21	8.98
19				10.74	12.80	10.03	9.36	8.57
18				11.21	9.89	10.34	8.20	9.82
17			9.97	9.70	9.84	10.29	9.84	10.01
16	11.17	10.14	9.93	10.27	10.21	11.09	10.83	8.82
15	9.92	10.82	11.65	8.96	9.88	8.90	10.18	9.34
14	10.21	10.73	9.46	9.35	9.78	10.38	9.79	8.91
13			12.50	9.63	10.82	10.12	9.40	9.48
12		9.92	11.05	10.11	11.46	10.41	8.45	8.90
11		11.31	9.41	9.37	11.21	9.93	10.70	9.27
10		11.15	9.91	10.17	10.55	11.61	9.16	10.04
9			10.82	11.75	0.78	11.00	9.79	10.19
8		10.01	8.23	11.04	10.28	13.07	10.47	11.58
7			10.39	11.11	10.96	10.83	10.09	8.69
6			10.41	10.82	17.61	10.87		13.06
5			9.76	11.10	10.80	8.86	9.48	9.22
4			10.93	10.94	9.53	10.61	10.27	9.59
3				9.64	9.52	10.06	12.65	9.63
2				9.29	8.75	8.96	8.27	8.14
1				10.59	10.43	9.32		

	9	10	11	12	13	14	15	16
23		8.59	9.00	11.86	8.91	9.99		
22	8.76	8.89	9.10	7.62	9.65			
21	10.66	8.92	7.80	7.84	9.03	8.60		
20	9.27	8.19	7.88	7.61	8.20	8.77		
19	9.01	9.04	7.28	9.58	9.69	9.96	9.91	
18	10.06	8.58	8.89	8.64	7.04	8.81	7.95	
17	9.01	7.68	9.25	7.83	9.14		7.63	9.07
16	10.18	9.34	8.61					
15	10.56	9.06						
14	9.22	11.43						
13	10.99	9.92	7.85	8.21				
12	8.07	7.96	7.00	7.90				
11	9.28	10.13	8.61	8.78				
10	11.19	8.10	11.30					
9	9.15	8.15	9.20					
8	9.46	8.54	10.87					
7	11.17	9.39	9.56					
6	11.41	9.96	9.15					
5	9.61	8.20						
4	9.82	7.81						

Use a regression model in an attempt to find pattern in these data.

8. The following table shows iron ore (%Fe_2O_3) on a grid with 50 m by 50 m spacing of 50 m^2, where north is to the right (Cressie, 1986):

		47.1	45.6					
		50.5	46.9					
	50.4	51.4	53.5	56.2	54.3	55.1	57.8	60.8
	54.0	55.8	52.1	55.4	51.9	59.6	58.1	54.8
	51.3	47.5	53.5	54.3	52.0	53.2	57.8	58.9
	47.8	5109	55.7	50.0	56.7	58.4	56.1	59.1
	52.6	55.0	52.2	58.9	57.8	55.9	58.7	57.4
	50.5	48.0	47.2	56.5	50.3	59.6	58.7	59.1
	51.2	45.4	54.1	51.6	58.2	56.1	57.3	57.9
55.9	52.1	56.9	57.6	57.1	58.3	58.5	58.1	54.4
55.1	53.5	53.7	58.7	58.4	59.0	56.9	51.5	
53.1	56.7	57.5	58.6	55.7	58.9	56.3	54.2	
	51.5	55.1	59.1	59.6	59.7	56.6		
	56.2	53.5	56.1	56.6	52.9	54.3		
	56.8	47.5	56.7	56.9	52.8	52.5		
	59.7	57.6	55.0	59.0	57.1	54.3		
		60.3	56.4			51.3		

Use a regression model in an attempt to find pattern in these data. A north–south trend is suspected.

9
Normal Models

The classical normal linear models are especially simple mathematically, as compared to other members of the exponential dispersion family, for a number of reasons:

- the canonical link function for the normal distribution is the identity;
- the variance function does not depend on the mean;
- all cumulants after the second are zero;
- all dependence relationships in the multivariate normal distribution are contained in the (second-order) covariance or correlation matrix;
- in a multivariate normal distribution, the conditional distribution of one variable given the others is just the linear multiple regression model.

These simple features have made conceptualizing generalizations to other distributions more difficult and not always very obvious. Think, for example, of the choice of a transformed lagged response or of residuals in Chapter 5 for extending autoregression to other distributions than the normal.

Traditionally, in most situations, a normal linear model has simply been used for convenience. However, there is no longer any reason for this to be so with the wide availability of generalized linear modelling software. Instead, all of the assumptions should be checked in any data analysis. Here, we shall be particularly interested in looking at the distributional form, the link function, and linearity.

9.1 Linear Regression

One important application of classical linear regression is to response surfaces from designed experiments where some optimal combination of factors is being sought. Generally, two or three explanatory variables are varied over a range where the optimum is believed to occur. Then, in the simplest cases, a multiple regression, quadratic in these variables, is fitted to determine the maximum or minimum response.

Such techniques were originally developed for the control of industrial processes, such as chemical reactions, but very quickly came to be much more widely used. More complex models are often required in order to explore the whole shape of the surface in some region and not just to find the optimum. However, in most cases, no functional equation is available; thus, polynomials must still be used, interpreted as a Taylor series approximation to the function.

Example

In a classical experiment in fisheries biology (Fry and Hart, 1948), the effects of acclimation and test temperatures on the cruising speed of young goldfish, *Carassius auratus*, were studied. Speed measurements were carried out in a rotating annular chamber with a glass outer wall and metal inner wall. When the chamber was rotated, the fish reacted by swimming to maintain place with respect to the room. One rotation required the fish to swim about 2.5 ft.

A fish was subjected to a constant temperature for sufficiently long for it to become thermally adapted. It was then introduced into the chamber, generally at a different temperature, and allowed to settle down at a slow speed, after which the rotation rate was increased by gradual steps, lasting two minutes each, until the fish consistently fell behind. This usually lasted a total of about 20–25 min. Although three fish were used at each temperature combination, unfortunately, only the average is available. From the layout of the data in Table 9.1, we clearly see the area of temperature combinations in which the experimenters believed the maximum speed would be attained.

The simple quadratic normal regression model with response equation

$$\begin{aligned}\mu_i &= 21.84 + 0.9744x_{1i} + 4.493x_{2i} \\ &\quad -0.1842x_{1i}^2 - 0.2052x_{2i}^2 + 0.2803x_{1i}x_{2i}\end{aligned}$$

where x_{1i} is acclimation and x_{2i} test temperature, or

$$\texttt{LINTEST} + \texttt{LINACCLIM} + \texttt{SQUTEST} + \texttt{SQUACCLIM} + \texttt{LINTEST.LINACCLIM}$$

gives an AIC of 163.9. The optimum is at $x_1 = 23$ and $x_2 = 27$ where the mean speed is 93 ft/min, considerably lower than several of the observed values.

TABLE 9.1. Cruising speed (ft/min) of goldfish depending on acclimation and test temperatures (°C). (Fry and Hart, 1948; Lindsey *et al.* 1970)

Test	Acclimation temperature						
temperature	5	10	15	20	25	30	35
5	44		44				
10	58	57					
15	68		71		55		
20	68		79	100	79		30
25	41		77		96		58
30					100	98	70
35					68		76
38							84

The corresponding gamma distribution with identity link has an AIC of 160.6, suggesting that the distribution may be skewed or the variance may not be constant. The log normal distribution has an AIC of 161.1. Both give responses equations similar to that above. □

Notice that, for interpretability of the response surface equation, all terms in the multidimensional polynomial of a given order are required, although some may be eliminated if found to be unnecessary. Thus, if an interaction is included, the squared terms are as well.

9.2 Analysis of Variance

Traditional analysis of variance (ANOVA) is just a special case of multiple regression where indicator variables or orthogonal polynomials are used to describe the discrete levels of factor variables. Such indicator variables are created automatically by most generalized linear modelling software. As we saw in Section 1.3.3, there is generally no unique way of specifying such indicators. The software often provides a reasonable choice. Obviously, such factor variables can be mixed with continuous variables, in what was once known as analysis of covariance.

The term analysis of variance refers, not to the model, but to the method of determining which effects are statistically significant. Any given normal model is assumed to have a constant variance. However, for different models, the value of this constant variance will change, smaller variances corresponding to higher likelihoods. Thus, the variances of various models can be analyzed in this way, using the likelihood, in order to determine which is most suitable, hence the name.

Because a response surface experiment is based on factors, analysis of variance techniques can be used to check the linearity of the model, a kind

of "semiparametric" model. However, if the model is found to be nonlinear, factor variables do not yield a smoothed surface, like the quadratic model, from which an optimum can easily be determined.

Example

If we fit a full factorial model to the goldfish data, the variance will be estimated to be zero because we do not have available the replications at the response points. Instead, we can fit the two main effects as factors and add a linear interaction,

$$\texttt{FACTEST} + \texttt{FACACCLIM} + \texttt{LINTEST.LINACCLIM}$$

This gives an AIC of 161.8 for the normal model, 171.2 for the gamma, and 165.6 for the log normal, each with nine more parameters than in the quadratic regression. Only for the first do we have a hint of such nonlinearity.□

When the assumptions of the normal linear model are not fulfilled, one classical solution (Box and Cox, 1964) has been to transform the response variable in an attempt simultaneously to obtain normality, constant variance, and additive effects. This, in fact, generates a new family of distributions based on the normal distribution. By taking the log response, we have already used one member several times, the log normal distribution. With generalized linear models, it is possible to supplement this approach with other changes of distribution to alleviate normality constraints and constant variance, and with changes of link function that may facilitate additivity.

Example

A study was made for the International Wool Textile Organization on the behaviour of worsted yarn under cycles of repeated loading. Box and Cox (1964) analyze this $3 \times 3 \times 3$ factorial design to study the number of cycles to failure of the yarn, as shown in Table 9.2.

These authors find a log transformation of the response, that is, a log normal distribution, in an additive no-interaction model with logarithms of the explanatory variables

$$\texttt{LOGLENGTH} + \texttt{LOGAMPLITUDE} + \texttt{LOGLOAD}$$

Their model has an AIC of 332.6. The comparable model with additive factor variables, instead of logarithms,

$$\texttt{FACLENGTH} + \texttt{FACAMPLITUDE} + \texttt{FACLOAD}$$

has an AIC of 336.7, making the former preferable. However, the addition of the three first-order interactions of the factor variables to the latter model

$$\texttt{FACLENGTH} * \texttt{FACAMPLITUDE} + \texttt{FACLENGTH} * \texttt{FACLOAD} + \texttt{FACAMPLITUDE} * \texttt{FACLOAD}$$

reduces the AIC to 321.2.

TABLE 9.2. Cycles to failure of worsted wool. (Box and Cox, 1964, from Barella and Sust)

Length (mm)	Amplitude (mm)	Load (g)	Cycles
250	8	40	674
		45	370
		50	292
	9	40	338
		45	266
		50	210
	10	40	170
		45	118
		50	90
300	8	40	1414
		45	1198
		50	634
	9	40	1022
		45	620
		50	438
	10	40	442
		45	332
		50	220
350	8	40	3636
		45	3184
		50	2000
	9	40	1568
		45	1070
		50	566
	10	40	1140
		45	884
		50	360

TABLE 9.3. AICs for various models for the wool data of Table 9.2.

Distribution	Link	No interaction	Interaction
Normal	Identity	418.3	337.7
	Log	352.5	312.8
Log normal	Identity	336.7	321.2
Gamma	Reciprocal	377.3	307.3
	Log	336.5	321.4
Log gamma	Reciprocal	344.2	318.4
Inverse	Rec. quad.	—	358.7
Gaussian	Log	332.8	307.1
Log inverse Gaussian	Rec. quad	356.3	316.2
Poisson	Log	751.9	356.9
Negative binomial	Log	335.1	321.5

The AICs for a series of other models are summarized in Table 9.3. Both the gamma distribution with canonical link and the inverse Gaussian with a log link give comparable, and much improved, models. The latter has the advantage in that it can be slightly simplified by eliminating the amplitude by load interaction to give an AIC of 305.6. On the other hand, the Poisson and negative binomial models for counts are comparable with the log normal, even though overdispersion has been taken into account in the second of these.

Although Box and Cox argue that a normal model, additive in the logs of all variables, can have a physical interpretation, we see that it does not approach the goodness of fit of the nonnormal models. Indeed, even a model with the log inverse Gaussian distribution and identity link that is additive in the logs of all variables gives a slightly better fit, with an AIC of 331.1.□

9.3 Nonlinear Regression

9.3.1 Empirical Models

As we have seen in Section 9.1, a multiple regression with continuous explanatory variables fits a smooth response surface to the data. However, generally low-order polynomials are used, very much restricting the form of surface possible. On the other hand, factor variables simply follow the empirically observed mean values at each point in the response space, performing no smoothing at all. They can be used for checking the adequacy of a model, but, being "semiparametric", are of little use for interpreting

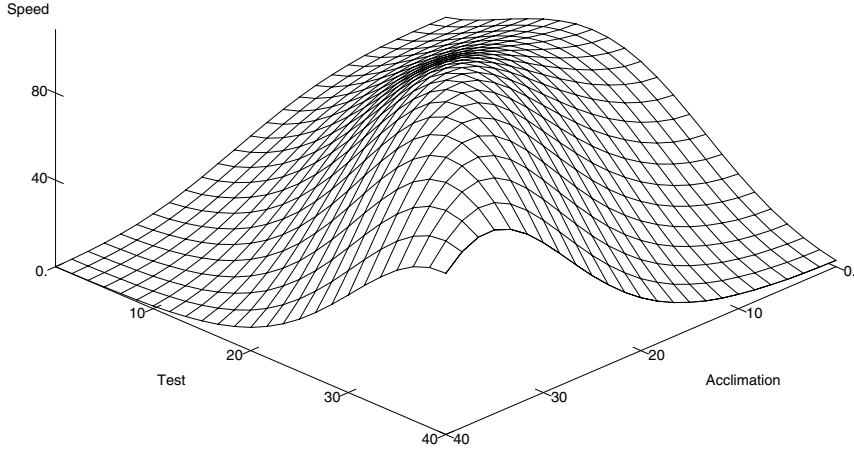

FIGURE 9.1. Nonlinear response surface for the goldfish data in Table 9.1.

the form of the surface. If no theory dictates what the model might be, one useful approach is to transform the explanatory variables (Box and Tidwell, 1962). Most commonly, a power transformation is estimated and some rounded value close to it chosen, as with the Box and Cox transformation of the response. This can be calculated by the linearization procedure described in Section 1.5.2 (see Lindsey, 1972, 1995b, pp. 234 and 265–266).

Example

For the goldfish data, with a normal distribution, the power transformed model has an AIC of 154.2, whereas the gamma with the identity link has 150.4 and the log normal 144.5. With a better response model, the need for a skewed distribution is confirmed; the log normal is now superior to the gamma. Thus, of all the models we have fitted, we choose the former. The response equation is

$$\begin{aligned}\mu_i &= 3.632 - 0.0004799x_{1i}^{1.64} + 0.02886x_{2i}^{1.29} - 0.00002636x_{1i}^{3.28} \\ &\quad -0.0004457x_{2i}^{2.57} + 0.0001637x_{1i}^{1.64}x_{2i}^{1.29}\end{aligned}$$

where μ_i here refers to average log speed. For these data, there appear to be no useful values to which the powers could be rounded so we keep the estimates as they are. The optimum is at about $x_1 = 26$ and $x_2 = 27$ where the mean speed is 100 ft/min. A three-dimensional plot of the fitted model is shown in Figure 9.1. □

9.3.2 Theoretical Models

As we saw in Section 1.4.3, the link function can be used to linearize certain nonlinear models. This is especially important when the response model has

TABLE 9.4. Measurements (sec) used to calibrate a Stormer viscometer. (Williams, 1959, p. 66)

Viscosity	Weight (g)		
(centipoise)	20	50	100
14.7	35.6	17.6	
27.5	54.3	24.3	
42.0	75.6	31.4	
75.7	121.2	47.2	24.6
89.7	150.8	58.3	30.0
146.6	229.0	85.6	41.7
158.3	270.0	101.1	50.3
161.1		92.2	45.1
298.3		187.2	89.0,86.5

a particular nonlinear functional form predicted by some scientific theory. We have already used this for growth curves in Chapter 4. Another more general example is inverse polynomials (Nelder, 1966).

Example

An experiment was performed to calibrate a Stormer rotational viscometer. This instrument measures the viscosity of a liquid as the time taken to complete a number of revolutions of its inner cylinder. The weight controlling the cylinder can also be varied. From theoretical considerations, the mean time depends on the viscosity (x_1) and the weight (x_2) in the following way:

$$\mu_i = \frac{\alpha_1 x_{1i}}{x_{2i} - \alpha_2}$$

The parameter α_2 can be interpreted as the weight required to overcome internal friction in the machine. Notice that the theory does not specify the distributional form of measurements around this mean. This equation may be linearized as

$$\frac{1}{\mu_i} = \beta_1 z_{1i} + \beta_2 z_{2i}$$

where $\beta_1 = 1/\alpha_1$, $\beta_2 = \alpha_2/\alpha_1$, $z_1 = x_2/x_1$, and $z_2 = -1/x_1$.

In calibration, liquids of known viscosity are used and times for 100 revolutions recorded under different weights. Those from one study are given in Table 9.4. When we fit a normal distribution with reciprocal link, the parameter estimates are $\hat{\beta}_1 = 0.0340$ and $\hat{\beta}_2 = 0.0755$, so that $\hat{\alpha}_1 = 29.40$ and $\hat{\alpha}_2 = 2.218$. This model has an AIC of 153.6. Setting $\beta_2 = 0$ is equivalent to $\alpha_2 = 0$, for a model with an AIC of 160.4.

The corresponding results for a gamma distribution with canonical link are $\hat{\beta}_1 = 0.0314$ and $\hat{\beta}_2 = 0.0905$, so that $\hat{\alpha}_1 = 31.85$ and $\hat{\alpha}_2 = 2.880$, with

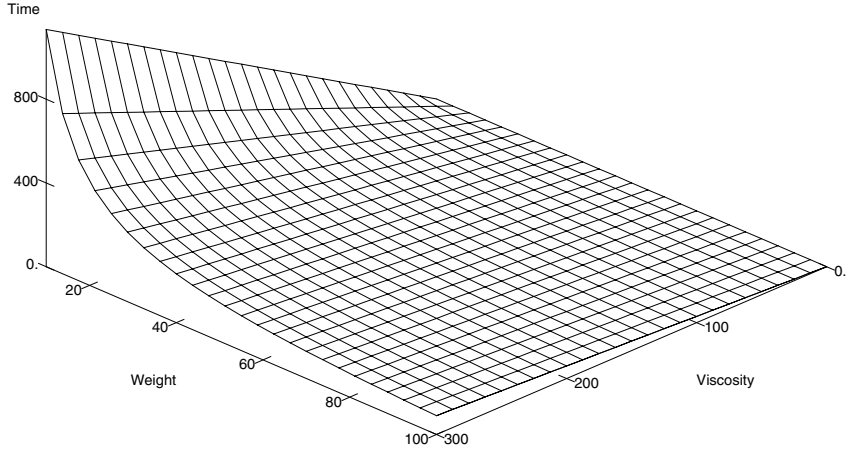

FIGURE 9.2. Nonlinear response surface for the viscometer data in Table 9.4.

an AIC of 185.2, considerably poorer. Setting $\alpha_2 = 0$ here yields an AIC of 185.0. This probably indicates that variability is almost completely due to symmetric measurement error about the theoretical equation for the mean.

The fitted response surface for the first model, with a normal distribution, has been plotted in Figure 9.2. □

Summary

Besides the basic material on normal linear models covered in any introductory statistics text, a vast number of more specialized books are available. To name just two, the reader may like to consider Searle (1971) and Jørgensen (1993), as well as the books on generalized linear models listed at the end of Chapter 1.

A number of good books on nonlinear models are available, including Bates and Watts (1988), Seber and Wild (1989), Ross (1990), and Huet *et al.* (1996).

9.4 Exercises

1. Above, we performed a reanalysis of the wool data that Box and Cox (1964) used to illustrate the use of transformations. In the following table, we have the other data set that they used:

	Treatment			
Poison	A	B	C	D
I	0.31	0.82	0.43	0.45
	0.45	1.10	0.45	0.71
	0.46	0.88	0.63	0.66
	0.43	0.72	0.76	0.62
II	0.36	0.92	0.44	0.56
	0.29	0.61	0.35	1.02
	0.40	0.49	0.31	0.71
	0.23	1.24	0.40	0.38
III	0.22	0.30	0.23	0.30
	0.21	0.37	0.25	0.36
	0.18	0.38	0.24	0.31
	0.23	0.29	0.22	0.33

These are survival times, in ten hour units, of animals in a 3×4 completely randomized factorial experiment with three types of poison and four treatments. The authors find a reciprocal transformation of the response (with a normal distribution). Check to see if this is a reasonable model.

2. The following table gives weight gains (g) in rats under six diets in a completely randomized experiment (Snedecor and Cochran, 1967, p. 347):

Protein					
High			Low		
Beef	Cereal	Pork	Beef	Cereal	Pork
73	98	94	90	107	49
102	74	79	76	95	82
118	56	96	90	97	73
104	111	98	64	80	86
81	95	102	86	98	81
107	88	102	51	74	97
100	82	108	72	74	106
87	77	91	90	67	70
117	86	120	95	89	61
111	92	105	78	58	82

Develop an appropriate model for this replicated factorial experiment. Take care to check for interaction effects.

3. Merino wether sheep are an important meat production animal in Australia. To assess energy maintenance requirements, weight (kg) and daily energy needs (mcal/sheep/day) were recorded on 64 grazing sheep (Wallach and Goffinet, 1987):

Weight	Energy	Weight	Energy	Weight	Energy
22.1	1.31	25.1	1.46	25.1	1.00
26.2	1.27	27.0	1.21	30.0	1.23
33.2	1.25	33.2	1.32	33.2	1.47
34.3	1.14	34.9	1.00	42.6	1.81
49.0	1.78	49.2	2.53	51.8	1.87
52.6	1.70	53.3	2.66	23.9	1.37
27.6	1.39	28.4	1.27	28.9	1.74
31.0	1.47	31.0	1.50	31.8	1.60
32.6	1.75	33.1	1.82	34.1	1.36
44.6	2.25	52.1	2.67	52.4	2.28
52.6	3.73	46.7	2.21	37.1	2.11
28.6	2.13	29.2	1.80	26.2	1.05
34.4	1.85	34.4	1.63	26.4	1.27
25.7	1.20	25.9	1.36	52.7	3.15
30.2	1.01	30.2	1.12	31.8	1.39
33.9	1.03	33.8	1.46	45.9	2.36
43.7	1.73	44.9	1.93	27.5	0.94
51.8	1.92	52.5	1.65	53.1	2.73
25.1	1.39	26.7	1.26	36.1	1.79
29.3	1.54	29.7	1.44	36.8	2.31
32.0	1.67	32.1	1.80		
34.2	1.59	44.4	2.33		

How successful are such energy requirements in predicting meat production?

4. A three-period crossover trial was performed on 12 subjects to determine gastric half-emptying time in minutes (Keene, 1995):

Treatment					
A		B		C	
Period	Response	Period	Response	Period	Response
1	84	2	62	3	58
3	87	2	108	1	38
3	85	1	85	2	96
1	82	3	46	2	61
2	83	1	70	3	46
2	110	3	110	1	66
3	215	2	86	1	42
1	50	3	46	2	34
2	92	3	50	1	80
1	70	2	61	3	55
3	97	1	40	2	78
2	95	1	147	3	57

Each line corresponds to the treatment and response of one subject. This is an analysis of variance type design for duration data. Find an appropriate model to determine if there are treatment effects.

5. Specimens from 50 species of timber were measured for modulus of rigidity, modulus of elasticity, and air-dried density (Brown and Maritz, 1982):

Rigidity	Elasticity	Density	Rigidity	Elasticity	Density
1000	99	25.3	1897	240	50.3
1112	173	28.2	1822	248	51.3
1033	188	28.6	2129	261	51.7
1087	133	29.1	2053	245	52.8
1069	146	30.7	1676	186	53.8
925	91	31.4	1621	188	53.9
1306	188	32.5	1990	252	54.9
1306	194	36.8	1764	222	55.1
1323	195	37.1	1909	244	55.2
1379	177	38.3	2086	274	55.3
1332	182	39.0	1916	276	56.9
1254	110	39.6	1889	254	57.3
1587	203	40.1	1870	238	58.3
1145	193	40.3	2036	264	58.6
1438	167	40.3	2570	189	58.7
1281	188	40.6	1474	223	59.5
1595	238	42.3	2116	245	60.8
1129	130	42.4	2054	272	61.3
1492	189	42.5	1994	264	61.5
1605	213	43.0	1746	196	63.2
1647	165	43.0	2604	268	63.3
1539	210	46.7	1767	205	68.1
1706	224	49.0	2649	346	68.9
1728	228	50.2	2159	246	68.9
1703	209	50.3	2078	237.5	70.8

Develop a model to describe how the modulus of rigidity depends on the other variables.

6. Dumbbell shapes of high-density polyethylene (Rigidex HO60–45P from BP Chemicals) were made using injection moulding, cut in two, and hot plate welded back together. The quality of the weld was measured as the ratio of the yield stress of the welded bar to the mean yield stress of unwelded bars, called the weld factor. Four variables were controlled: hot-plate temperature (°C), heating time (sec), welding time (sec), and pressure on the weld (bars), with results as follows (Buxton, 1991):

Hot-plate temperature	Heating time	Welding time	Pressure on weld	Weld factor
295	40	25	3.0	0.82
295	40	25	2.0	0.87
295	40	15	3.0	0.83
295	40	15	2.0	0.86
295	20	25	3.0	0.82
295	20	25	2.0	0.80
295	20	15	3.0	0.77
295	20	15	2.0	0.58
245	40	25	3.0	0.89
245	40	25	2.0	0.86
245	40	15	3.0	0.84
245	40	15	2.0	0.82
245	20	25	3.0	0.67
245	20	25	2.0	0.77
245	20	15	3.0	0.74
245	20	15	2.0	0.40
320	30	20	2.5	0.83
270	50	20	2.5	0.88
270	10	20	2.5	0.66
270	30	30	2.5	0.84
270	30	10	2.5	0.81
270	30	20	3.5	0.88
270	30	20	1.5	0.81
270	30	20	2.5	0.82
270	30	20	2.5	0.86
270	30	20	2.5	0.80
270	30	20	2.5	0.84
270	30	20	2.5	0.86
270	30	20	2.5	0.83
270	30	20	2.5	0.81
270	30	20	2.5	0.79
270	30	20	2.5	0.80
270	30	20	2.5	0.82
270	30	20	2.5	0.86

The goal of the experiment was to develop a suitable empirical model.

This page intentionally left blank

10
Dynamic Models

10.1 Dynamic Generalized Linear Models

An approach to longitudinal data, similar in some ways both to autoregression and to random effects, and in fact able to encompasses both, involves allowing the regression coefficients to be random, evolving over time according to a Markov process. This is called a *dynamic generalized linear model* and is usually estimated by a procedure called the Kalman filter. (Unfortunately, the usual software for generalized linear models generally cannot easily be adapted for this algorithm.) Although originally proposed as the dynamic linear model for normal data, it can be extended to other distributions.

10.1.1 Components of the Model

The regression model for the location parameter, now called the *observation* or *measurement equation*, will be written

$$g_i(\mu_{it}) \quad = \quad \boldsymbol{\lambda}_{it}^T \mathbf{x}_{it}$$

where $\boldsymbol{\lambda}_{it}$ is a random vector of coefficients, defining the *state* of individual i at time t, with a distribution conditional on the previous responses and on the explanatory variables, $\mathbf{x}_{it}$. In contrast to the random effects model, here coefficients can vary (over time) on the same individual, as well as across individuals. The state is simply the minimum past and present information necessary to predict a future response, the filtration of Chapter 7.

The state of the system is, then, taken to evolve over time according to a *state transition equation*

$$\mathrm{E}[\boldsymbol{\lambda}_{it}] = \mathbf{T}_{it}\boldsymbol{\lambda}_{i,t-1} \tag{10.1}$$

where $\mathbf{T}_{it}$ is a first-order Markovian *state transition matrix*. Note that $\boldsymbol{\lambda}_{it}$ may contain values before time t as well as present values. The distributions of Y_{it} and $\boldsymbol{\lambda}_{it}$ are assumed to be independent. Then, the multivariate probability is given by the recursive relationship

$$\Pr(y_1, \ldots, y_t) = \Pr(y_1)\Pr(y_2|y_1)\cdots\Pr(y_t|y_1, \ldots, y_{t-1}) \tag{10.2}$$

as in Equation (5.1).

10.1.2 Special Cases

The dynamic generalized linear model for an autoregression of order M has measurement and state equations that can be written

$$\begin{aligned} g_i(\mu_{it}) &= [1, 0, \ldots]\boldsymbol{\lambda}_{it} \\ \mathrm{E}\left[\begin{pmatrix} \lambda_{it} \\ \lambda_{i,t-1} \\ \vdots \\ \lambda_{i,t-M+1} \end{pmatrix}\right] &= \begin{pmatrix} \rho_{i1} & \cdots & \rho_{i,M-1} & \rho_{iM} \\ 1 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_{i,t-1} \\ \lambda_{i,t-2} \\ \vdots \\ \lambda_{i,t-M} \end{pmatrix} \end{aligned}$$

On the other hand, for a random effects model, the equations are

$$\begin{aligned} g_i(\mu_{it}) &= \mu + \lambda_{it} \\ \mathrm{E}[\lambda_{it}] &= 0 \end{aligned}$$

a special case of the model

$$g_i(\mu_i) = \lambda_i + \mathbf{x}_i^T\boldsymbol{\beta}$$

Simple models, such as these, can be combined in any desired way.

10.1.3 Filtering and Prediction

Filtering means estimating the current state, given responses up to the present. The Kalman filter is a sequential or recursive procedure, yielding new distributions at each time point. Using Bayes' formula (Section A.3.1), we have

$$p(\boldsymbol{\lambda}_{it}|\mathcal{F}_{it}) = \frac{\Pr(y_{it}|\boldsymbol{\lambda}_{it}, \mathcal{F}_{i,t-1})p(\boldsymbol{\lambda}_{it}|\mathcal{F}_{i,t-1})}{\Pr(y_{it}|\mathcal{F}_{i,t-1})} \tag{10.3}$$

where $\mathcal{F}_{it}$ denotes the history of responses for individual i up to and including time t, that is, the vector of responses $(y_{i1}, \ldots, y_{it})^T$, with all pertinent relationships among them. In Equation (10.3),

$$\Pr(y_{it}|\mathcal{F}_{i,t-1}) = \int_{-\infty}^{\infty} \Pr(y_{it}|\boldsymbol{\lambda}_{it}, \mathcal{F}_{i,t-1}) p(\boldsymbol{\lambda}_{it}|\mathcal{F}_{i,t-1}) d\boldsymbol{\lambda}_{it}$$

whereas $\Pr(y_{it}|\boldsymbol{\lambda}_{it}, \mathcal{F}_{i,t-1})$ is the usual distribution, if there were no random coefficients, defined by the observation equation, and $p(\boldsymbol{\lambda}_{it}|\mathcal{F}_{it})$ is called the *filtering* or *observation update*.

The *one-step-ahead prediction* or *time update*,

$$p(\boldsymbol{\lambda}_{it}|\mathcal{F}_{i,t-1}) = \int_{-\infty}^{\infty} p(\boldsymbol{\lambda}_{it}|\boldsymbol{\lambda}_{i,t-1}) p(\boldsymbol{\lambda}_{i,t-1}|\mathcal{F}_{i,t-1}) d\boldsymbol{\lambda}_{i,t-1} \tag{10.4}$$

is defined by the transition equation. Both of these integrals are usually complicated when the distributions are not normal and the link is not the identity. We shall be interested in the conditional distribution, $\Pr(y_{it}|\mathcal{F}_{i,t-1})$, to calculate the likelihood function.

Most often, such dynamic models are used in forecasting. However, here, they are particularly important as a unifying framework for many of the models of change over time, as well as a means of calculating the likelihood function and estimating the parameters in difficult cases (see, for example, de Jong, 1988). Thus, two advantages of this Kalman filter-type approach, even in the case of a model based on the normal distribution, are that it can be extended to handle unequally spaced time intervals and (randomly) missing observations and that it encompasses many useful models as special cases.

10.2 Normal Models

Dynamic models based on the normal distribution are generally called dynamic linear models. However, here we shall also be concerned with nonlinear regression models.

In general, we shall often wish to model responses that are unequally spaced in time, either because the study was planned that way or because some responses are randomly missing. When only heterogeneity across individuals is present, this creates no problem. However, time series methods will almost always be necessary and the problem becomes more complex. The usual autoregression models, such as the AR(1) (Section 5.2), cannot directly be applied, because the autocorrelation parameter, ρ, measures the *constant* stochastic dependence among *equally spaced* observations. Thus, it cannot account for the different degrees of stochastic dependence among successive responses that are not the same distance apart.

What we require is a continuous time autoregression (Priestley, 1981, pp. 156-174; Harvey, 1989, pp. 479-501). Such a continuous AR(1) process is defined by the linear differential equation

$$\mathrm{E}[dY_{it}] = -\kappa_i y_{it} dt$$

the mean of a normal distribution with variance ξdt. This gives an autocorrelation function

$$\rho_i(\Delta t) = e^{-\kappa_i |\Delta t|} \tag{10.5}$$

Thus, with some generalization, one possible autocorrelation function for this model is given by

$$\rho(u) = e^{-\kappa u^{\nu}}$$

In the same way as for the discrete time AR(1), a location model, that, if linear, might be represented by $\boldsymbol{\beta}_i^T \mathbf{z}_{it}$, will usually be included in the model to produce either a serial correlation or state dependence model.

In a general continuous time model such as this, there are a number of nonlinear parameters so that estimation is usually difficult. Direct maximization of the likelihood function is possible (Jennrich and Schluchter, 1986; Jones and Ackerson, 1990). This approach is especially useful when the numbers of responses on an individual is not too large.

A more interesting approach is to consider the continuous autoregression as a special case of a dynamic linear model (Jones, 1993) as described above. Estimation in this way gives the same results as by direct maximization of the likelihood function and is most suitable when the series of responses on an individual is fairly long.

10.2.1 Linear Models

The general observation or measurement equation (10.1) for responses following a normal distribution with an identity link can be written

$$\mathrm{E}[Y_{it}] - \boldsymbol{\beta}_i^T \mathbf{z}_{it} = \boldsymbol{\lambda}_{it}^T \mathbf{v}_{it} \tag{10.6}$$

with the state transition equation as in Equation (10.1).

Let us first look at the discrete time, state space representation of a first-order autoregression with a random effect. The measurement and state equations are

$$\mathrm{E}[Y_{it}] - \boldsymbol{\beta}_i^T \mathbf{z}_{it} = \lambda_{1it} + \lambda_{2it}$$

with conditional variance, say ψ^2, fixed initial conditions, and

$$\begin{aligned} \mathrm{E}[\lambda_{1it}] &= \rho_i \lambda_{1i,t-1} \\ \mathrm{E}[\lambda_{2it}] &= 0 \end{aligned}$$

with variances ξ and δ.

For a continuous AR(1), the measurement equation remains the same, whereas the state equation (Jones and Boadi-Boateng, 1991) for λ_{1it} is now assumed to be a continuous AR(1), so that

$$\mathrm{E}[d\lambda_{1it}] = -\kappa_i \lambda_{1it} dt$$

that is, the mean of a normal distribution with variance ξdt. The state equation for the random effect is now

$$\mathrm{E}[d\lambda_{2it}] = 0$$

with variance δdt. In this way, we have expressed our continuous time, serial correlation model with a random effect for heterogeneity as a dynamic linear model.

The Kalman filter of Equations (10.3) and (10.4) is applied in order to obtain the estimates. Because only the mean and variance are required to define the normal distribution, these equations can be written in a simple closed form, and the procedure is relatively straightforward. We move forward in time from the first observation, estimating the expected value of each successive response before going on to the next, building up the total multivariate probability as a product of conditional probabilities using Equation (10.2).

For simplicity, let us consider how this works with a discrete time series. The one-step-ahead prediction or time update for $\mathrm{E}[Y_{it}] - \boldsymbol{\beta}_i^T \mathbf{z}_{it}$ has mean

$$\hat{\boldsymbol{\lambda}}_{it|t-1} = \mathbf{T}_{it} \hat{\boldsymbol{\lambda}}_{i,t-1}$$

and covariance matrix

$$\mathbf{A}_{it|t-1} = \mathbf{T}_{it} \mathbf{A}_{i,t-1} \mathbf{T}_{it}^T + \hat{\boldsymbol{\Xi}}$$

where $\boldsymbol{\Xi}$ is a diagonal covariance matrix. In the random effects AR(1) model above, this would contain the covariance elements ξ and δ. $\mathbf{A}_{i,t-1}$ is the prior covariance of the estimation error

$$\mathbf{A}_{it} = \mathrm{E}[(\boldsymbol{\lambda}_{it} - \hat{\boldsymbol{\lambda}}_{it})(\boldsymbol{\lambda}_{it} - \hat{\boldsymbol{\lambda}}_{it})^T]$$

The filtering or observation update, using the next response, y_{it}, has posterior mean

$$\hat{\boldsymbol{\lambda}}_{it} = \hat{\boldsymbol{\lambda}}_{it|t-1} + \mathbf{A}_{it|t-1} \mathbf{v}_{it} (y_{it} - \hat{\boldsymbol{\lambda}}_{it|t-1}^T \mathbf{v}_{it} - \boldsymbol{\beta}_i^T \mathbf{z}_{it}) / c_{it}$$

and posterior covariance matrix

$$\mathbf{A}_{it} = \mathbf{A}_{it|t-1} - \mathbf{A}_{it|t-1} \mathbf{v}_{it} \mathbf{v}_{it}^T \mathbf{A}_{it|t-1} / c_{it}$$

where

$$c_{it} = \mathbf{v}_{it}^T \mathbf{A}_{it|t-1} \mathbf{v}_{it} + \psi^2 \tag{10.7}$$

As is usual for time series, the initial conditions must be chosen for $\boldsymbol{\lambda}_{i0}$ and for $\mathbf{A}_{i0}$.

To obtain the likelihood function, we can rewrite the observation equation (10.6) as

$$\mathrm{E}[Y_{it}] - \boldsymbol{\beta}_i^T \mathbf{z}_{it} = \hat{\boldsymbol{\lambda}}_{i,t-1}^T \mathbf{v}_{it} + (\boldsymbol{\lambda}_{it} - \hat{\boldsymbol{\lambda}}_{i,t-1})^T \mathbf{v}_{it}$$

from which the conditional distribution of Y_{it} given $\mathcal{F}_{i,t-1}$ has mean

$$\mathrm{E}[Y_{it}|\mathcal{F}_{i,t-1}] = \boldsymbol{\beta}_i^T \mathbf{z}_{it} + \hat{\boldsymbol{\lambda}}_{i,t-1}^T \mathbf{v}_{it}$$

with variance given by Equation (10.7). Then, the likelihood function for one individual is

$$\log(\mathrm{L}_i) = -\frac{1}{2} \sum_{t=1}^{R} \left[\log(2\pi c_{it}) + (y_{it} - \mathrm{E}[Y_{it}|\mathcal{F}_{i,t-1}])^2 / c_{it} \right]$$

For a discrete serial correlation AR(1) with the first observation stationary, $\mathrm{E}[Y_{i1}|\mathcal{F}_0] = \boldsymbol{\beta}_i^T \mathbf{z}_{i1}$ and $c_{i1} = \xi/(1-\rho^2)$, whereas $\mathrm{E}[Y_{it}|\mathcal{F}_{i,t-1}] = \boldsymbol{\beta}_i^T \mathbf{z}_{it} + \rho y_{i,t-1}$ and $c_{it} = \xi$ for $t > 1$. These results are readily generalized to continuous time.

Example

In the context of nonlinear growth curves, Heitjan (1991b) provides a pioneering analysis of data on the treatment of multiple sclerosis. In a randomized double-blinded clinical trial, patients received either (1) two placebos (P), (2) real azathioprine (AZ) and a placebo, or (3) real doses of azathioprine and methylprednisolone (MP). Blood samples were drawn from the 48 patients one or more times prior to therapy, at initiation, and in weeks 4, 8, and 12 plus every 12 weeks thereafter, and a measure of autoimmunity (AFCR) made. More details will be given below.

The data are plotted in Figure 10.1. The responses were measured at highly irregular periods in time. Except for the profiles of two or three patients with placebo, all three of the plots seem rather similar, although the placebo group does not approach as closely to the zero level of AFCR as do the two other groups. The responses are decreasing in time, except, perhaps, at the very end, and the average profiles appear to be nonlinear.

For the moment, we shall fit a number of linear and quadratic polynomial models, with and without treatment effects, and with a number of covariance structures, including random effects and/or AR(1). For the polynomials, time has been centred at the mean of 501.5 days. We follow Heitjan (1991b) in using a square root transformation on the AFCR responses. The AICs for these models are summarized in the upper panel of Table 10.1.

We see that the random effects models fit better than the AR(1) models. The quadratic location model with treatment differences and interaction

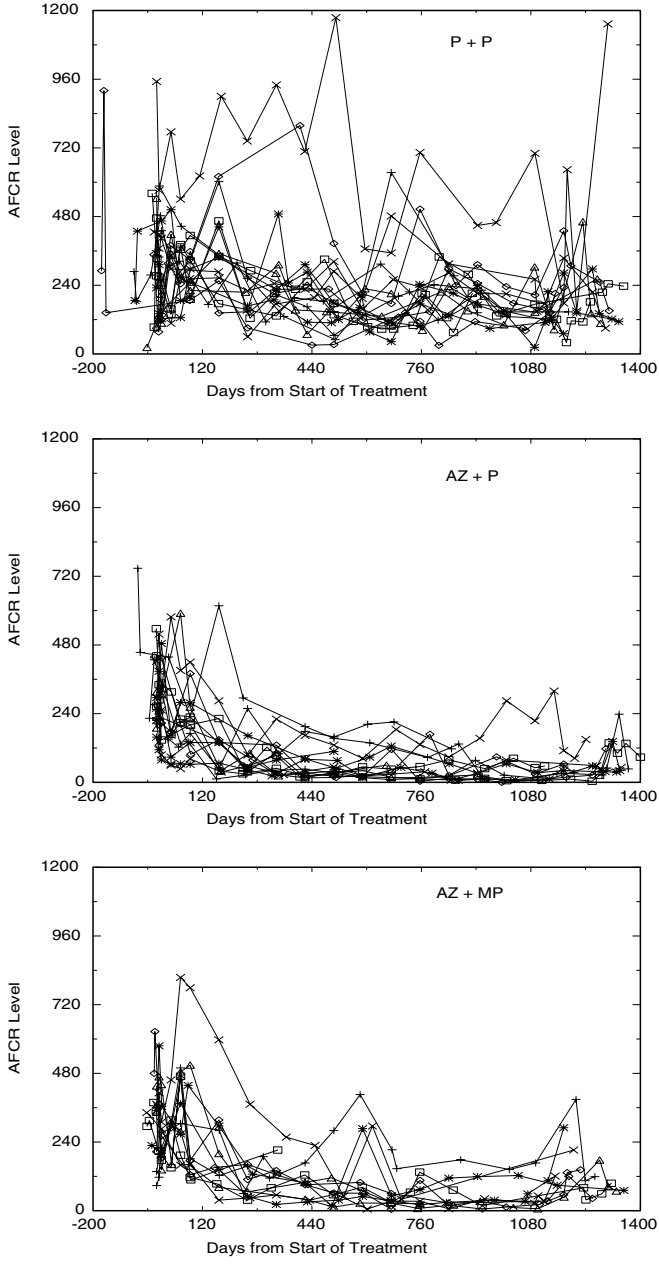

FIGURE 10.1. Plots of AFCR level against time in days for three treatments.

TABLE 10.1. AICs for models fitted to the multiple sclerosis data of Figure 10.1.

	Indep.	RE	AR	RE + AR
	Without dose			
	Polynomial curve			
Null	3516.2	3440.1	–	3175.4
Linear	3380.6	3105.7	3292.7	2998.4
+ Treatment	3335.7	3086.4	3183.3	3000.2
+ Interaction	3215.8	3062.6	3175.4	2996.7
Quadratic	3314.1	2985.8	3261.1	2984.8
+ Treatment	3147.3	2966.3	3118.8	2959.8
+ Interaction	3100.8	2898.0	3088.6	2898.6
	Logistic curve			
Null	3324.0	2996.6	3265.5	2983.8
Treatment	3120.5	2918.3	3104.7	2913.9
	With dose			
	Linear curve			
Null	3474.1	3218.4	3328.0	3176.4
Treatment	3316.9	3157.1	3243.4	3135.6
	Logistic curve			
Null	3444.8	3067.9	3323.5	3044.7
Treatment	3172.2	2913.1	3151.1	2908.8

fits much better than the others. For this model, the AR(1) is not necessary. This indicates considerable variability among the patients, but little stochastic dependence among successive responses of a given patient. Of the polynomial curves, this model, with a AIC of 2898.0 and 11 parameters, appears to be the best.

For our chosen model, the variance is $\hat{\xi} = 13.37$ and the additional variance component, and covariance, is $\hat{\delta} = 6.35$, giving a uniform intrapatient correlation of 0.32. The polynomial models are

$$\begin{aligned} \mathrm{E}[Y_{i1t}^{0.5}] &= 14.53 - 0.0038t + 0.0000039t^2 \\ \mathrm{E}[Y_{i2t}^{0.5}] &= 7.33 - 0.0096t + 0.000015t^2 \\ \mathrm{E}[Y_{i3t}^{0.5}] &= 8.84 - 0.0095t + 0.000014t^2 \end{aligned}$$

The corresponding estimated profiles are plotted in Figure 10.2. The slight rise at the end of the time period is plausible in light of the plots in Figure 10.1. The double placebo group is much worse than the other two. This may be due, in part, to the two or three control patients having high profiles. The two groups receiving real medicine are fairly close and show no significant difference. That the azathioprine and placebo group appears to be slightly better than the no placebo group may be due to one patient who had a higher profile in the latter group, as seen in Figure 10.1. □

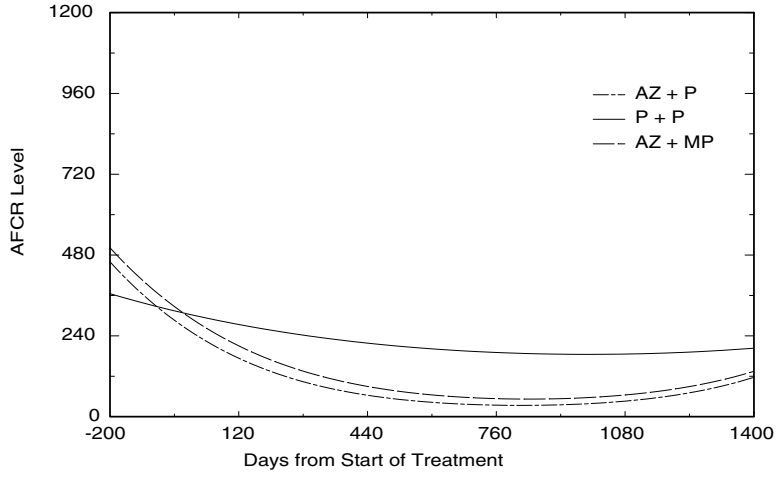

FIGURE 10.2. Response profiles of AFCR for the three treatments, using a quadratic polynomial model, from Figure 10.1.

10.2.2 Nonlinear Curves

Nelder (1961, 1962) has suggested a generalized form of the logistic growth curve, that Heitjan (1991a, 1991b) has further generalized by adding autocorrelation and random effects. Suppose that the change in mean response obeys the following differential equation:

$$\frac{d\mu_t}{dt} = \kappa_3 \mu_t [s(e^{\kappa_2}, \kappa_4) - s(\mu_t, \kappa_4)]$$

where

$$\begin{aligned} s(\mu, \kappa_4) &= \tfrac{\mu^{\kappa_4} - 1}{\kappa_4}, \quad \kappa_4 \neq 0 \\ &= \log(\mu), \quad \kappa_4 = 0 \end{aligned}$$

with initial condition $\mu_0 = \exp(\kappa_1)$ at t_0. The solution is

$$\begin{aligned} \mu_t &= e^{\kappa_2} \left[1 + \left(e^{(\kappa_2 - \kappa_1)\kappa_4} - 1\right) e^{-\kappa_3 (t - t_0) e^{\kappa_2 \kappa_4}}\right]^{-\frac{1}{\kappa_4}}, \quad \kappa_4 \neq 0 \\ &= e^{\kappa_2 + (\kappa_1 - \kappa_2) e^{-\kappa_3 (t - t_0)}}, \quad \kappa_4 = 0 \end{aligned} \tag{10.8}$$

Then, $\kappa_1 = \log(\mu_0)$ is the initial condition and $\kappa_2 = \lim_{t \to \infty} \log(\mu_t)$, the asymptote. The parameters κ_3 and κ_4 control the rate of growth. If $\kappa_3 < 0$ and $\kappa_2 > \kappa_1$, or $\kappa_3 > 0$ and $\kappa_2 < \kappa_1$, we have negative growth or decay. The parameter κ_4 determines the type of the curve, varying from the Mitscherlich ($\kappa_4 = -1$) through the Gomperz ($\kappa_4 = 0$), and the logistic ($\kappa_4 = 1$) to the exponential ($\kappa_4 \to \infty$ and $s[e^{\kappa_2}, \kappa_4] \to$ constant).

TABLE 10.2. Location parameter estimates for the generalized logistic model fitted to the multiple sclerosis data of Figure 10.1.

	P + P	AZ + P	AZ + MP
κ_{1j}	5.745	5.750	5.752
κ_{2j}	5.032	3.939	4.268
κ_{3j}	1.462	1.460	1.466
κ_{4j}	-1.563	-1.564	-1.561

Example

We can now apply this generalized logistic model to the multiple sclerosis data of Figure 10.1. We can expect this to be a negative growth curve because the AFCR level is decreasing over time. We fit models with identical generalized logistic curves for all three treatments and with completely different curves for each treatment. The covariance structure is kept the same for all treatments, as for the polynomial curve above. The AICs for the four possible covariance structures for each location model are given in the second panel of Table 10.1 above. The treatment differences are significant. However, now, the autocorrelation cannot be set to zero. The variance is $\hat{\xi} = 4.33$, the additional variance component, and part of the covariance, is $\hat{\delta} = 6.69$, and the autocorrelation $\hat{\rho} = 0.82$. However, this model, with four more parameters, has an AIC larger by 15.9 than the quadratic model without autocorrelation and, thus, is not as acceptable as that model.

These parameters may be compared with those for the quadratic model with autocorrelation: 13.40, 21.82, and 0.51. The logistic model has higher autocorrelation, but the intrapatient correlation is about the same: 0.62 as compared to 0.61 for the quadratic model with autocorrelation and 0.32 for that without. The logistic and quadratic models without autocorrelation have almost identical estimates for ξ and δ.

The parameters for the three treatments in our new model are given in Table 10.2. The estimated parameter values in the curves for the two treatment groups are more similar than that for the placebo group. The parameter for the asymptote, κ_{2j}, shows the main difference. The three location models are

$$\begin{aligned} g(\mathrm{E}[g^{-1}(Y_{i1t})]) &= e^{5.03}\left[1+\left(e^{1.16}-1\right)e^{-1.46(t-t_0)e^{-7.87}}\right]^{0.64} \\ g(\mathrm{E}[g^{-1}(Y_{i2t})]) &= e^{3.94}\left[1+\left(e^{2.83}-1\right)e^{-1.46(t-t_0)e^{-6.16}}\right]^{0.64} \\ g(\mathrm{E}[g^{-1}(Y_{i3t})]) &= e^{4.27}\left[1+\left(e^{2.32}-1\right)e^{-1.47(t-t_0)e^{-6.87}}\right]^{0.64} \end{aligned}$$

Here, the function $g(\cdot)$, both the link function and the inverse of the transformation of the data, is the square.

The curves are plotted in Figure 10.3. In contrast to the polynomial

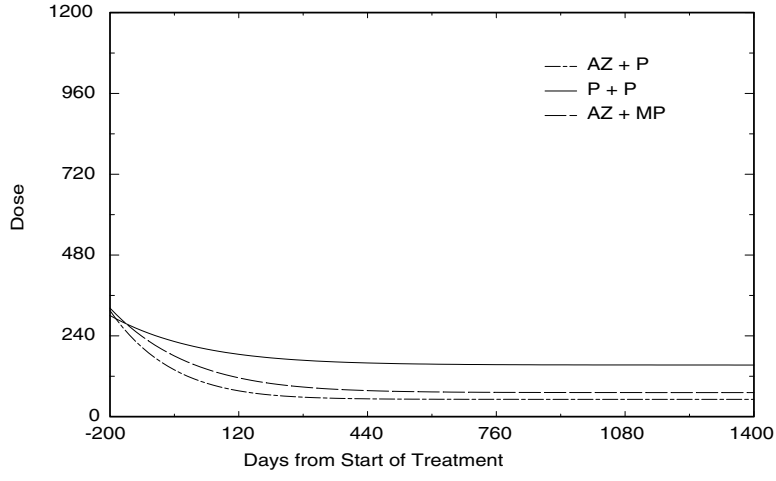

FIGURE 10.3. Response profiles of AFCR for the three treatments, using a generalized logistic model, from Figure 10.1.

curves in Figure 10.2, these cannot rise at the end of the period, but instead reach an asymptote. This may be part of the explanation for the poorer fit of this model. The autocorrelation parameter is attempting to compensate for this. □

Often, in growth studies, as elsewhere, time-varying covariates are available. One way in which to modify Equation (10.8) to incorporate them is to have them influence the asymptote, κ_2 (Heitjan, 1991b). Suppose, for treatment j, that this parameter is a function of time,

$$\kappa_{2jt} = \kappa_1 + \log\left(\frac{2}{1 + e^{\beta_j z_{jt}}}\right)$$

Then, if $z_{jt} = 0$, the asymptote is constant at the initial condition. If $z_{jt} = z_j$ is constant in time, the mean grows or decays from its initial condition, e^{κ_1}, to

$$\frac{2e^{\kappa_1}}{1 + e^{\beta_j z_j}}$$

a new equilibrium level. If $\beta_j = 0$, no growth or decay occurs for that treatment. If z_{jt} is a step function, the mean follows a piecewise generalized logistic curve. Then, if z_{jt} returns to zero, the mean goes back to the initial condition.

Example

For our multiple sclerosis study, the dose of the medication for each patient was varied in time as a function of the patient's condition. The changes in

dose occurred at irregular periods that did not correspond to the times when the response, AFCR, was measured. The profiles of dose changes are plotted in Figure 10.4. The doses were all zero up until time zero, at which point a dose of one unit was given to all patients. This was rapidly increased to a maximum value of around two, but, subsequently, often drastically reduced and increased, depending on the condition of the patient. Patients in the complete placebo group were matched with patients in the other groups and their doses followed that of the paired patient, so that neither the patient nor the physician would know who was on placebo.

The dose appears to be following a stochastic process that may be interdependent with that for AFCR, an internal time-dependent covariate in the terminology of Kalbfleisch and Prentice (1980, pp. 122–127), what is called an endogenous variable in the econometrics literature. However, here we follow Heitjan (1991b) in conditioning on the dose at a given time in modelling AFCR.

Our model has z_{ijt} as the strength of dose currently being administered, when the response is measured. Models with and without differences among groups were fitted using this model, with the four usual covariance structures. These give individual curves for each patient, instead of a mean curve for each group, because the dose profile varies among the patients. The various curves have the general form of those in Figure 10.3, but with visible wobble, following the dose of each patient.

The AICs are given in the bottom panel of Table 10.1. The model with differences in treatment has a consistently larger AIC than the corresponding model (quadratic or logistic) without the dose variable. However, the present model has fewer parameters than any of the latter models. For example, it has an AIC 10.2 larger than the corresponding quadratic model, with only three fewer parameters, and 10.8 larger than the quadratic without autocorrelation, with two fewer parameters.

Heitjan (1991b) also fits a linear model in dose for comparison. Here, the response only depends on time through the dose being administered. The AICs are given in the second last panel of Table 10.1. These are the poorest fitting models of all, although they have more parameters than many of the others. □

In this example, a sophisticated and theoretically appealing growth model has been found to fit less well than a simple polynomial. The reason appears to be that the response does not reach an asymptote. This should stimulate further theoretical developments to derive a more satisfactory biological model.

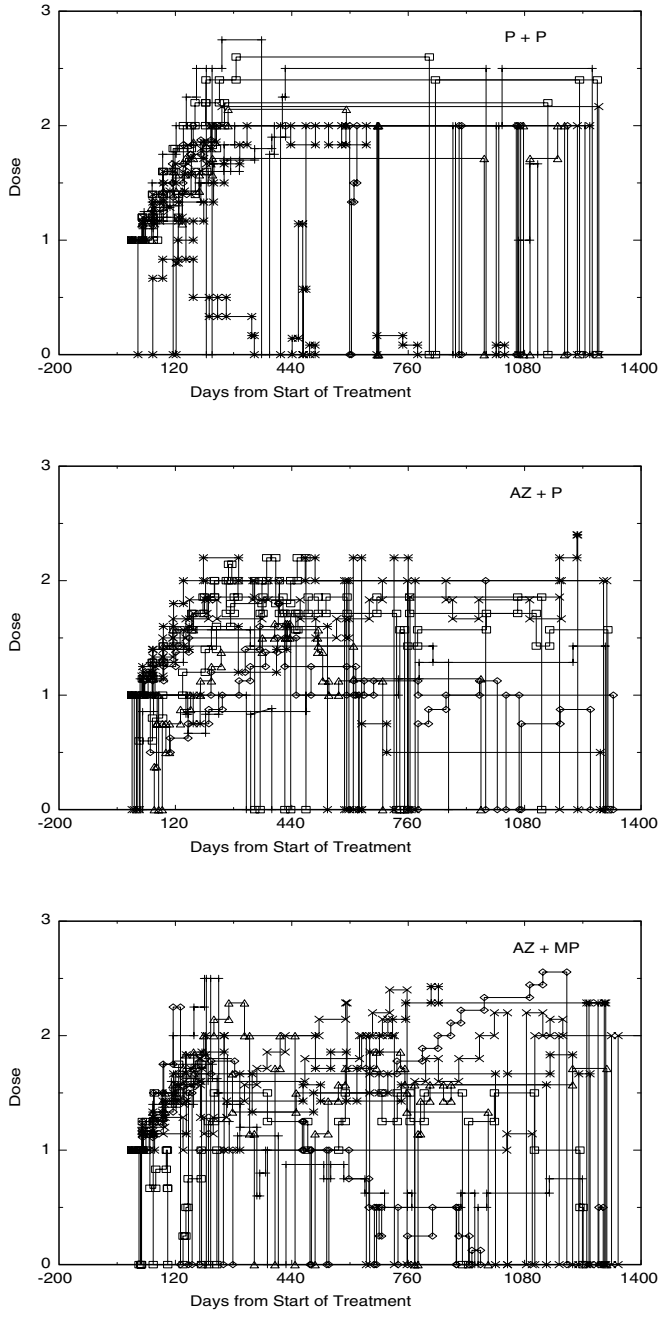

FIGURE 10.4. Plots of dose level against time in days for the three treatments.

10.3 Count Data

Dynamic generalized linear models are particularly useful when one needs to predict the evolution of several series of responses. However, they also provide a general means of estimating parameters for trend and seasonal effects.

In this section, we shall be specifically interested in models applicable to counts and categorical data. For the former, the Poisson and negative binomial processes are often suitable, and, for the latter, the binomial process. The conjugate distribution for the Poisson is the gamma, yielding a negative binomial likelihood, whereas that for the negative binomial is the beta distribution, giving a hypergeometric distribution. The conjugate for the binomial is also the beta distribution, yielding a beta-binomial distribution (Section 2.3.2). Here, we shall use an example of count data, so as to compare those two possible distributions for that type of data.

To estimate the parameters of the model, a Kalman filter procedure will be applied chronologically to the observations. Let us look, in more detail, at the steps for the gamma-Poisson process. The model has the mean parameter λ_{ij} from a Poisson distribution,

$$\Pr(y_{ij}|\lambda_{ij}) = \frac{\mathrm{e}^{-\lambda_{ij}}\lambda_{ij}^{y_{ij}}}{y_{ij}!}$$

taken to have a gamma distribution with density

$$p(\lambda_{ij}) = \frac{\upsilon_j^{-\kappa_j}\lambda_{ij}^{\kappa_j-1}\mathrm{e}^{-\frac{\lambda_{ij}}{\upsilon_j}}}{\Gamma(\kappa_j)}$$

over the individuals i in treatment j. Integration over λ_{ij} yields the marginal negative binomial distribution of Y_{ij},

$$\begin{aligned}\Pr(y_{ij}) &= \frac{\upsilon_j^{-\kappa_j}\int \mathrm{e}^{-\lambda_{ij}}\lambda_{ij}^{y_{ij}}\lambda_{ij}^{\kappa_j-1}\mathrm{e}^{-\frac{\lambda_{ij}}{\upsilon_j}}d\lambda_{ij}}{\Gamma(\kappa_j)y_{ij}!} \\ &= \frac{\Gamma(y_{ij}+\kappa_j)}{y_{ij}!\Gamma(\kappa_j)}\left(\frac{1}{1+\upsilon_j}\right)^{\kappa_j}\left(\frac{\upsilon_j}{1+\upsilon_j}\right)^{y_{ij}} \qquad (10.9)\end{aligned}$$

Here, the parameter estimates are updated dynamically. Thus, starting with some initial values, for one series, we have the prediction equations

$$\begin{aligned}\kappa_{t|t-1} &= \zeta\kappa_{t-1} \\ \frac{1}{\upsilon_{t|t-1}} &= \frac{\zeta}{\upsilon_{t-1}}\end{aligned}$$

and the updating equations

$$\begin{aligned} \kappa_t &= \kappa_{t|t-1} + y_t \\ \frac{1}{\upsilon_t} &= \frac{1}{\upsilon_{t|t-1}} + 1 \end{aligned} \tag{10.10}$$

where ζ is a discount factor between zero and one. Then, the deviance, derived from Equation (10.9), is given by

$$\begin{aligned} -2\sum_t \{&\log[\Gamma(y_t + \kappa_{t|t-1})] - \log[\Gamma(\kappa_{t|t-1})] - \kappa_{t|t-1}\log[\upsilon_{t|t-1}] \\ &-(\kappa_{t|t-1} + y_t)\log[1/\upsilon_{t|t-1} + 1]\} \end{aligned}$$

This must be minimized by some numerical procedure in order to estimate the regression parameters and, perhaps, the discount.

In the case of repeated measurements, we have several series. The most complex model involves applying the above procedure separately to each series, so that all parameters are different. For "parallel" series, υ_t and the regression coefficients are the same for all series, whereas κ_t is allowed to be different. The latter are updated, in Equation (10.10), using the respective values of y_t for each series. To fit a common model to all series, the mean response may be used in Equation (10.10). Similar procedures may be used for the models for other types of data mentioned above.

Example

The reported total numbers of deaths in the United Kingdom from bronchitis, emphysema, and asthma each month from 1974 to 1979, distinguished by sex, are presented in Table 10.3. Notice the particularly high values in the winter of 1975–76 that correspond to months 24 to 26.

We are interested in seeing if the number of deaths is changing over the years, a linear time trend. As well, we shall require a seasonal component, because the number of deaths varies regularly over the year. We shall use seasonal harmonics for the 12-month period. Thus, three models will be fitted: (1) separately to the data for each sex, (2) with the same trend and seasonal, but a different level for each sex, and (3) with all components the same for each sex. The resulting AICs for the gamma-Poisson (negative binomial) and beta-negative binomial (hypergeometric) are displayed in Table 10.4. We arbitrarily take a fixed discount of 0.7, although this could also have been estimated.

Although there is no clear saturated model, we take our most complex model as a baseline, giving it zero deviance, so that it can easily be compared with the others. We immediately see that all of the gamma-Poisson models are unacceptable as compared to the beta-negative binomial ones. A trend is not necessary and the seasonal components can be the same for

TABLE 10.3. Monthly numbers of deaths from bronchitis, emphysema, and asthma in the United Kingdom, 1974-1979. (Diggle, 1990, p. 238, from Appleton) (read across rows)

Males									
2134	1863	1877	1877	1492	1249	1280	1131	1209	1492
1621	1846	2103	2137	2153	1833	1403	1288	1186	1133
1053	1347	1545	2066	2020	2750	2283	1479	1189	1160
1113	970	999	1208	1467	2059	2240	1634	1722	1801
1246	1162	1087	1013	959	1179	1229	1655	2019	2284
1942	1423	1340	1187	1098	1004	970	1140	1110	1812
2263	1820	1846	1531	1215	1075	1056	975	940	1081
1294	1341								
Females									
901	689	827	677	522	406	441	393	387	582
578	666	830	752	785	664	467	438	421	412
343	440	531	771	767	1141	896	532	447	420
376	330	357	445	546	764	862	660	663	643
502	392	411	348	387	385	411	638	796	853
737	546	530	446	431	362	387	430	425	679
821	785	727	612	478	429	405	379	393	411
487	574								

TABLE 10.4. AICs and numbers of parameters for various models for the monthly deaths of Table 10.3.

Effect	Separate sexes		Different levels		Sexes together	
Gamma-Poisson						
Trend	7522.16	6	7520.18	5	29243.84	3
Seasonal	697.91	26	700.97	14	22426.61	13
Both	695.72	28	701.86	15	22427.55	14
Beta-negative binomial						
Trend	296.30	8	293.36	6	487.81	4
Seasonal	56.27	28	36.62	16	475.72	14
Both	60.00	30	38.62	17	477.72	15

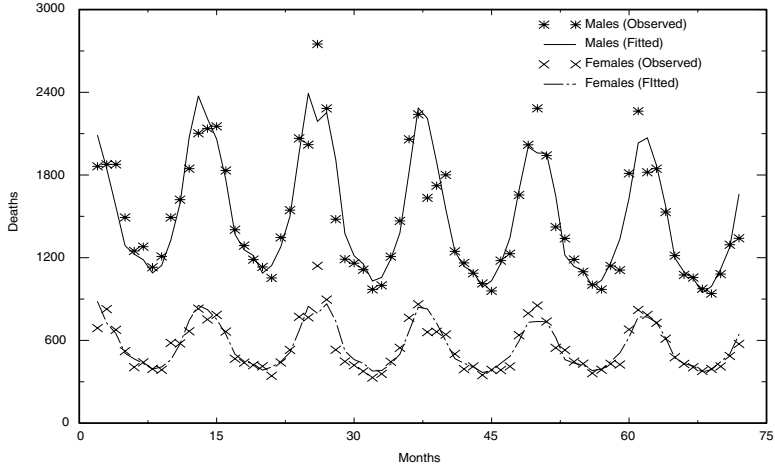

FIGURE 10.5. Monthly deaths from bronchitis, emphysema, and asthma, from Table 10.3, with the fitted dynamic generalized linear model.

deaths of both sexes. However, the level must be different for the two sexes. This model has a deviance of only 4.62 greater than the most complex one, but with 14 fewer parameters. The only further simplification is to reduce the number of harmonics. It is only possible to eliminate the two highest ones, with a further increase in deviance of 3.76. The fitted values of the final model are plotted in Figure 10.5, along with the observed numbers of deaths. The fitted lines follow the observed deaths fairly closely, with the exception of three high counts of male deaths and one of females. Male deaths are consistently higher than female, but with the same seasonal variation. There is no indication of a change in the number of deaths over the years. (Lindsey, 1993, pp. 206–209; Lindsey and Lambert, 1995) □

10.4 Positive Response Data

Dynamic generalized linear models can be applied to duration data, or at least to longitudinal data having positive response values that might follow a gamma, inverse Gaussian, or log normal distribution. Thus, the model is based on the density, not on the intensity function. Here, we shall use the gamma distribution, whose conjugate is also a form of gamma, allowing for frailty or heterogeneity among the individuals. The procedure for estimating the parameters is essentially the same as that described in Section 10.3, except for the change in distributions, and need not be repeated here.

TABLE 10.5. Plasma citrate concentrations (μmol/l) for ten subjects at 14 successive times during the day. (Andersen *et al.*, 1981, from Toftegaard Nielsen *et al.*)

93	109	114	121	101	109	112	107	97	117
89	132	121	124						
116	116	111	135	107	115	114	106	92	98
116	105	135	83						
125	166	180	137	142	114	119	121	95	105
152	154	102	110						
144	157	161	173	158	138	148	147	133	124
122	133	122	130						
105	134	128	119	136	126	125	125	103	91
98	112	133	124						
109	121	100	83	87	110	109	100	93	80
98	100	104	97						
89	109	107	95	101	96	88	83	85	91
95	109	116	86						
116	138	138	128	102	116	122	100	123	107
117	120	119	99						
151	165	156	149	136	142	121	128	130	126
154	148	138	127						
137	155	145	139	150	141	125	109	118	109
112	102	107	107						

Example

A study was conducted to find possible relationships between daily rhythms of plasma citrate and those of carbohydrate metabolites during feeding with a citrate-poor diet. Measurements of plasma citrate concentration (μmol/l) were made for ten subjects over 14 successive hourly observation points between eight in the morning and nine in the evening. The data are reproduced in Table 10.5. Meals were at eight in the morning, noon, and five in the afternoon.

Because interest centres on daily rhythms, a dynamic generalized linear model with harmonics may be appropriate. With short series, as in this example, fitting a different dynamic generalized linear model to each series is not reasonable; there would be too many parameters. Instead we fit "parallel" and identical series, with 12 harmonics for a half-day and a trend that might pick up a longer period. The resulting deviances are given in Table 10.6. The ten series are not identical, but have different levels, as already could be seen from Table 10.5. For example, subjects one and seven have consistently lower plasma citrate concentrations than the others. There is some indication of a trend. The harmonics can be reduced from 12 to 4 with

TABLE 10.6. AICs for several dynamic generalized linear models for the plasma citrate data of Table 10.5.

	Null	Trend	Harmonics	Both
Identical	1156.92	1155.57	1166.05	1164.96
"Parallel"	1103.68	1099.17	1099.87	1097.96

a reduction in AIC of 7.55. The observations are plotted, along with the fitted model, in Figure 10.6, arbitrarily separated into two plots to make them clearer. We see that the plasma citrate concentration is generally highest at about ten or eleven in the morning and lowest about four or five in the afternoon. There seems to be no relationship to the meal times of eight in the morning, noon, and five in the afternoon. □

10.5 Continuous Time Nonlinear Models

Let us now consider again the gamma-Poisson process of Section 10.3, with a time-dependent conditional mean, $\lambda_{ij}\exp(\beta_{ij}^T\mathbf{x}_{ij})$ and unequally spaced observation times $(t_{i1},\ldots,t_{iJ_i})$. We shall follow Lambert (1996). Here, $\log(\lambda_{ij})$ is the residual part of the mean at time t_{ij} for individual i that is not modelled by explanatory variables. If the data are autocorrelated, the values of two such residuals at time points close together on the same individual should be closely related. One way to allow for this is to give, at any time point, a gamma distribution to λ_{ij} with mean $\kappa_{i,j|j-1}$ and variance $\kappa_{i,j|j-1}/\upsilon_{i,j|j-1}$.

Once an observation is available, the prior distribution (that is, the posterior distribution at the previous observation time) has to be updated to account for the observed residual that it has "generated". In continuous time, the longer the time since the previous observation, the less weight is given to the prior in constructing the corresponding posterior distribution.

One way to allow residuals observed at near time points to be more closely related than those further back in time is to let the prior distribution $p(\lambda_{ij}|\mathcal{F}_{i,j|j-1})$ at t_{ij} have the same mode, but a larger Fisher information than the posterior distribution $p(\lambda_{i,j-1}|\mathcal{F}_{i,j-1})$ at time $t_{i,j-1}$. Thus, we may take

$$\begin{aligned}\kappa_{i,j|j-1} &= \kappa_{i,j-1}\\ \upsilon_{i,j|j-1} &= \rho(\Delta t_{ij})\upsilon_{i,j-1}\end{aligned}$$

where $\Delta t_{ij} = t_{ij} - t_{i,j-1}$ and $\rho(\cdot)$ is a monotonically decreasing function with values on $[0,1]$ such that $\rho(0) = 1$. Here, we shall use $\rho(\Delta t) = \mathrm{e}^{-\phi\Delta t}$, as in Equation (10.5).

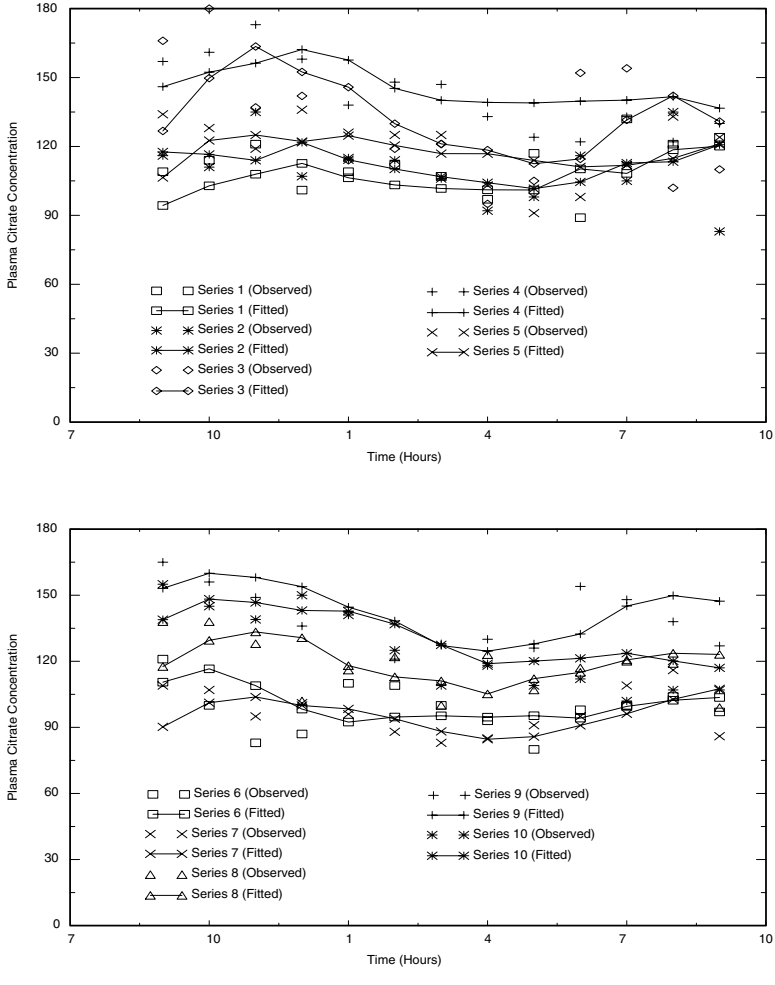

FIGURE 10.6. Hourly plasma citrate concentrations, from Table 10.5, with the fitted dynamic generalized linear model.

The distribution of λ_{ij} has to be updated, using Bayes' formula, once y_{ij} has been observed. The corresponding posterior will be gamma with

$$\begin{aligned}\kappa_{ij} &= \kappa_{i,j|j-1} + \frac{y_{ij} - \kappa_{i,j|j-1} e^{\beta_{ij}^T \mathbf{x}_{ij}}}{\upsilon_{i,j|j-1} + e^{\beta_{ij}^T \mathbf{x}_{ij}}} \\ \upsilon_{ij} &= \upsilon_{i,j|j-1} + e^{\beta_{ij}^T \mathbf{x}_{ij}}\end{aligned}$$

Because no past information on residuals is available at time t_{i0}, initial conditions have to be specified for starting the above iteration procedure. This can be done by using a vague prior for $\log(\lambda_{i1})$. One could, for example, take $\kappa_{i1|0} = 1$ and $\upsilon_{i1|0} = 0$.

Conditional on the first observation for each time series, the likelihood is a product over individuals of negative-binomial distributions:

$$\prod_i \prod_{j=2}^{n_i} \frac{\upsilon_{i,j|j-1}^{\kappa_{i,j|j-1}\upsilon_{i,j|j-1}}}{y_{ij}(\upsilon_{i,j|j-1} + e^{\beta_{ij}^T \mathbf{x}_{ij}})^{\kappa_{i,j|j-1}\upsilon_{i,j|j-1}+y_{ij}}} \frac{e^{\beta_{ij}^T \mathbf{x}_{ij} y_{ij}}}{B(\kappa_{i,j|j-1}\upsilon_{i,j|j-1}, y_{ij})}$$

Heterogeneity can, at least partially, be accounted for by assuming a possibly different evolution of the residuals for different individuals.

Additional flexibility can be allowed by introducing two further parameters:

1. The Poisson-gamma model can include the simple Poisson distribution as a special case by taking $\lambda_{ij}^{'}$, where

$$p(\lambda_{ij}^{'}|\mathcal{F}_{i,j|j-1}) \propto p(\lambda_{ij}|\mathcal{F}_{i,j|j-1})^{\delta}$$

   The gamma mixing distribution will reduce to a point if δ tends to zero and to a vague distribution if δ tends to infinity.

2. For robustness to extreme observations, another parameter, $0 \leq \alpha \leq 1$, can be added, such that the final equations are

$$\begin{aligned}\kappa_{ij}^{'} &= \kappa_{i,j|j-1}^{'} + \alpha \frac{y_{ij} - \kappa_{i,j|j-1}^{'} e^{\beta_{ij}^T \mathbf{x}_{ij}}}{\upsilon_{i,j|j-1}^{'}\delta + e^{\beta_{ij}^T \mathbf{x}_{ij}}} \\ \upsilon_{ij}^{'} &= \upsilon_{i,j|j-1}^{'} + \alpha \frac{e^{\beta_{ij}^T \mathbf{x}_{ij}}}{\delta}\end{aligned}$$

Example

Consider the growth of three closed colonies of *Paramecium aurelium* in a nutritive medium consisting of a suspension of the bacterium, *Bacillus pyocyaneous*, in salt solution, as shown in Table 10.7. At the beginning of each experiment, 20 *Paramecia* were placed in a tube containing 5 ml of the

TABLE 10.7. Sizes of three closed colonies of *Paramecium aurelium* (Diggle, 1990, p. 239, from Gause).

Day	Replicates		
0	2	2	2
2	17	15	11
3	29	36	37
4	39	62	67
5	63	84	134
6	185	156	226
7	258	234	306
8	267	348	376
9	392	370	485
10	510	480	530
11	570	520	650
12	650	575	605
13	560	400	580
14	575	545	660
15	650	560	460
16	550	480	650
17	480	510	575
18	520	650	525
19	500	500	550

medium at a constant 26° C. Each day, starting on the second, the tube was stirred, a sample of 0.5 ml taken, and the number of individuals counted. The remaining suspension was centrifuged, the medium drawn off, and the residue washed with bacteria-free salt solution to remove waste products. After a second centrifuging to remove this solution, fresh medium was added to make up the original volume.

One striking feature of this data set is the stabilization of the colony sizes after about ten days. Therefore the systematic part of the model should tend to an asymptote as time passes. The Nelder–Heitjan generalized logistic growth curve, discussed earlier, may be appropriate. On the other hand, the model considered by Diggle (1990, pp. 155), a quartic polynomial in time, does not allow for this.

We shall look at two families of models:

- ignoring the longitudinal aspect of the data set and simply assuming that the counts have a negative binomial distribution;
- taking the above gamma-Poisson dynamic model.

For each family of models, the two different systematic parts for the mean are fitted to the data.

TABLE 10.8. AICs and numbers of parameters for several models for the growth data of Table 10.7.

	Negative binomial		Gamma-Poisson	
Polynomial	565.7	6	562.2	8
Generalized logistic	566.5	5	562.8	7

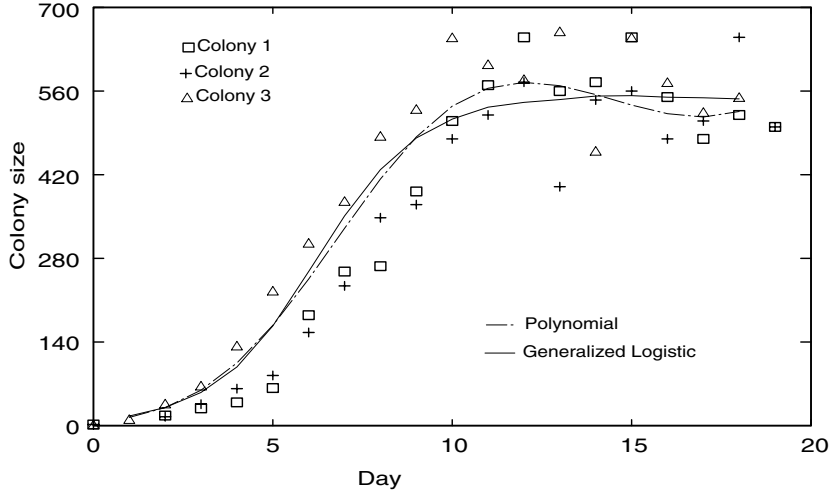

FIGURE 10.7. Plot of growth curves for three closed colonies of *Paramecium aurelium* from Table 10.7.

The AICs for the above four models are given in Table 10.8. We see that the gamma-Poisson model performs better than the independence negative binomial model. This indicates the need to model the serial association within individuals. Although the AIC provides no clear-cut choice for the systematic part, the gamma-Poisson model with a generalized logistic growth curve seems preferable for theoretical reasons as can be seen in Figure 10.7. It models the asymptotic behaviour of the colony development at the later times more appropriately.

The serial association parameters for the above models are respectively estimated to be $\hat{\phi} = 0.2245$, $\hat{\delta} = 0.000084$, $\hat{\alpha} = 0.05937$ and $\hat{\phi} = 0.1944$, $\hat{\delta} = 0.000077$, $\hat{\alpha} = 0.04837$ for the two best models. (Lambert, 1996) □

Summary

Procedures for dynamic generalized linear models are not yet well developed, other than in the normal case. There, they are very useful for fitting random effects and autoregression models when the observation times are

unequally spaced. In the more general case, the intractability of the integrals means that time-consuming numerical methods must be used or approximations made. Here, only the first two moments of the conjugate distribution were employed. However, the power of the procedure makes it one of the most promising avenues of research both in generalized linear models and in repeated measurements.

The development of dynamic generalized linear models is a fairly recent activity; see, especially, West *et al.* (1985), Kitagawa (1987), Fahrmeir (1989), Harvey (1989), West and Harrison (1989), and Harvey and Fernandes (1989). Other specialized books covering aspects of this topic include Jones (1993), Lindsey (1993), and Fahrmeir and Tutz (1994).

Appendix A
Inference

Most modern inference techniques have the likelihood function as a basis. Direct likelihood inference looks only at this, whereas frequentist and Bayesian inference make additional assumptions (Lindsey, 1996b). We shall only consider some aspects of the latter two approaches that are specific to generalized linear models, leaving the reader to consult a general inference book for more details.

A.1 Direct Likelihood Inference

A.1.1 Likelihood Function

In general, the probability of n *observed* response values, $\mathbf{y}^T = (y_1, \ldots, y_n)$, of a random variable, Y, is given by the joint probability distribution, $\Pr(y_{11} < Y_1 \leq y_{21}, \ldots, y_{1n} < Y_n \leq y_{2n})$. Notice that the observed value, y_i, of the random variable can only be known to some finite precision, that is, to lie within a unit of measurement, $y_{1i} < y_i \leq y_{2i}$, determined by the resolution of the instrument used. For independent observations, this probability becomes

$$\begin{aligned} &\Pr(y_{11} < Y_1 \leq y_{21}, \ldots, y_{1n} < Y_n \leq y_{2n}) \\ &\quad = \prod_{i=1}^{n} \Pr(y_{1i} < Y_i \leq y_{2i}) \end{aligned}$$

$$= \prod_{i=1}^{n} [F(Y_i \leq y_{2i}; \boldsymbol{\psi}) - F(Y_i \leq y_{1i}; \boldsymbol{\psi})]$$

where $\boldsymbol{\psi}$ is an unknown parameter vector. If observations are dependent, but occur serially, in time for example, the product of the appropriate conditional probabilities can be used, as in Equation (5.1).

In probability theory, the parameter values are assumed fixed and known, whereas the random variable is unknown so that probabilities of various possible outcomes can be calculated. However, for inference, because the random variable has been observed, the vector $\mathbf{y}$ is fixed and no longer variable.

Thus, the *likelihood function* is defined as the model taken as a function of the unknown parameter vector, $\boldsymbol{\psi}$, for the fixed given observed value of $\mathbf{y}$:

$$\mathrm{L}(\boldsymbol{\psi}; \mathbf{y}) = \Pr(y_{11} < Y_1 \leq y_{21}, \ldots, y_{1n} < Y_n \leq y_{2n}; \boldsymbol{\psi})$$

where y_i has unit of measurement, $\Delta_i = y_{2i} - y_{1i}$. In direct likelihood inference, *a model that makes the observed data more probable,* that is, best predicts them, *is said to be more likely* (to have generated those data).

Example

The likelihood function for the binomial distribution is

$$\mathrm{L}(\pi; y, n) = \binom{n}{y} \pi^y (1 - \pi)^{n-y}$$

Because we have a discrete distribution, the unit of measurement is one. □

For independent continuous response variables, the exact likelihood function is

$$\mathrm{L}(\boldsymbol{\psi}; \mathbf{y}) = \prod_{i=1}^{n} \int_{y_{1i}}^{y_{2i}} f(u_i; \boldsymbol{\psi}) du_i \tag{A.1}$$

This is based on the probability of empirically observing values from a continuous variable, thus providing the link between a theoretical model with such continuous variables and the discreteness of empirical data. Most often, the integral in this function is approximated by

$$\mathrm{L}(\boldsymbol{\psi}; \mathbf{y}) \doteq \prod_{i=1}^{n} f(y_i; \boldsymbol{\psi}) \Delta_i \tag{A.2}$$

where Δ_i is usually not a function of the parameters of interest.

Examples

1. An approximate likelihood for n observations from a normal distribution is

$$\mathrm{L}(\mu, \sigma^2; \mathbf{y}) \doteq \frac{\mathrm{e}^{-\frac{\Sigma(y_i-\mu)^2}{2\sigma^2}}}{(2\pi\sigma^2)^{\frac{n}{2}}} \prod \Delta_i$$

2. For censored survival times, if censoring occurs at the fixed time, t_c, the unit of measurement is (t_c, ∞) for events occurring after that time. Thus, Equation (A.2) will not provide a valid approximation and the exact factor in Equation (A.1), for example $\exp(-t_c/\mu)$ for an exponential distribution, must be used, as we do in Section 6.1.3. □

Except in very special cases, the likelihood function of Equation (A.1) will not be of the form of the exponential family (Section 1.2.1), even when the density is of this form. In particular, sufficient statistics will only exist, for continuous distributions, if measurements are assumed to be made with infinite precision, so that Equation (A.2) is applicable. This approximation depends not only on the Δ_i being reasonably small but also on the sample size being small. For fixed measurement precision, the quality of the approximate likelihood function in Equation (A.2) degrades rather rapidly as n increases, because it involves a growing product of small approximation errors.

A.1.2 Maximum Likelihood Estimate

A (vector of) parameter value(s) in a model function that makes the observed data most probable (given the model function) is called the *maximum likelihood estimate* (m.l.e). This need not be unique; several models may be equally likely for a given data set. Although this can create numerical problems for optimization routines, it is not a problem for drawing inferential conclusions.

For the classical normal linear model, the (approximate, in the sense above) likelihood surface is quadratic with respect to location parameters, meaning that the maximum likelihood estimates are unique. For other models, even in the generalized linear model family, it is not quadratic; there may be more than one maximum or a maximum at infinite parameter values. However, it is unique for the canonical link functions (sufficient statistics), as well as for probit and complementary log log links for the binomial distribution. There is no guarantee of uniqueness for the gamma distribution with an identity link.

Iterative Weighted Least Squares

Fitting the Model

When we use the method of maximum likelihood to estimate the linear parameters, β_j, by parametrization invariance, we also obtain estimates for the linear predictors, η_i, and the fitted values, μ_i.

Let

$$\mathbf{I} = \mathbf{X}^T\mathbf{V}^-\mathbf{X}$$

and

$$\mathbf{c} = \mathbf{X}^T\mathbf{V}^-\mathbf{z}$$

where

$$\begin{aligned} \mathbf{V} &= \text{diag}\left[\tau_i^2\left(\frac{d\eta_i}{d\mu_i}\right)^2\frac{\phi}{w_i}\right] \\ &= \text{diag}[v_i] \end{aligned} \tag{A.3}$$

say, with $\mathbf{V}^-$ its generalized inverse and

$$z_i = \eta_i + (y_i - \mu_i)\frac{d\eta_i}{d\mu_i}$$

The v_i^{-1} are called the *iterative weights*.

Then, we can solve

$$\mathbf{I}\hat{\boldsymbol{\beta}} = \mathbf{c}$$

iteratively. These are the *score equations* for weighted least squares with random variable $\mathbf{Z}$ having $\mathrm{E}[\mathbf{Z}] = \boldsymbol{\eta}$, $\mathrm{var}[\mathbf{Z}] = \mathbf{V}$, that is, with iterative weights, $\mathbf{v}^{-1}$. When we solve them, we obtain

$$\hat{\boldsymbol{\beta}} = \mathbf{I}^-\mathbf{c}$$

where $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimate of $\boldsymbol{\beta}$. We also have $\mathrm{var}[\hat{\boldsymbol{\beta}}] \doteq (\mathbf{X}^T\mathbf{V}^-\mathbf{X})^-$. Then, the estimates of $\boldsymbol{\eta}$ and $\boldsymbol{\mu}$ are $\mathbf{X}\hat{\boldsymbol{\beta}}$ and $g^{-1}(\mathbf{X}\hat{\boldsymbol{\beta}})$, respectively.

Proof:

For one observation from an exponential dispersion family, recall that

$$\begin{aligned} \log[\mathrm{L}(\theta_i, \phi; y_i)] &= \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \\ &= \frac{w_i[y_i\theta_i - b(\theta_i)]}{\phi} + c(y_i, \phi) \end{aligned}$$

$$\begin{aligned}\mu_i &= \frac{\partial b(\theta_i)}{\partial \theta_i} \\ \tau_i^2 &= \frac{\partial^2 b(\theta_i)}{\partial \theta_i^2} \\ &= \frac{d\mu_i}{d\theta_i}\end{aligned}$$

and that the score function for a linear regression coefficient is

$$\begin{aligned}\frac{\partial \log[\mathrm{L}(\theta_i, \phi; y_i)]}{\partial \beta_j} &= \frac{w_i(y_i - \mu_i)}{\phi} \frac{d\theta_i}{d\mu_i} \frac{d\mu_i}{d\eta_i} x_{ij} \\ &= \frac{w_i(y_i - \mu_i)}{\tau_i^2 \phi} \frac{d\mu_i}{d\eta_i} x_{ij}\end{aligned}$$

For independent observations, the (approximate) score equations are based on

$$\sum\nolimits_i \left[\frac{w_i(y_i - \mu_i)}{\tau_i^2 \phi} \frac{d\mu_i}{d\eta_i} x_{ij} \right] = \sum\nolimits_i \left[\frac{y_i - \mu_i}{v_i} \frac{d\eta_i}{d\mu_i} x_{ij} \right]$$

The Hessian is

$$\begin{aligned}\frac{\partial^2 \log(\mathrm{L})}{\partial \beta_j \partial \beta_k} &= \frac{\partial^2 \log(\mathrm{L})}{\partial \eta_i^2} x_{ij} x_{ik} \\ &= \left\{ \frac{\partial^2 \log(\mathrm{L})}{\partial \theta_i^2} \left(\frac{d\theta_i}{d\eta_i} \right)^2 + \frac{\partial \log(\mathrm{L})}{\partial \theta_i} \frac{d^2\theta_i}{d\eta_i^2} \right\} x_{ij} x_{ik} \\ &= \left\{ \left[-\tau_i^2 \left(\frac{d\theta_i}{d\mu_i} \right)^2 \left(\frac{d\mu_i}{d\eta_i} \right)^2 + (y_i - \mu_i) \frac{d^2\theta_i}{d\eta_i^2} \right] \frac{w_i}{\phi} \right\} x_{ij} x_{ik} \\ &= \left\{ \left[-\left(\frac{d\mu_i}{d\eta_i} \right)^2 \frac{1}{\tau_i^2} + (y_i - \mu_i) \frac{d^2\theta_i}{d\eta_i^2} \right] \frac{w_i}{\phi} \right\} x_{ij} x_{ik}\end{aligned}$$

Then, its expected value is

$$\begin{aligned}\mathrm{E}\left[\frac{\partial^2 \log(\mathrm{L})}{\partial \beta_j \partial \beta_k} \right] &= -\left(\frac{d\mu_i}{d\eta_i} \right)^2 \frac{w_i x_{ij} x_{ik}}{\tau_i^2 \phi} \\ &= -\frac{x_{ij} x_{ik}}{v_i}\end{aligned}$$

so that the *Fisher expected information* matrix is

$$\mathbf{I} = \mathbf{X}^T \mathbf{V}^- \mathbf{X}$$

given above.

The Newton–Raphson method, using expected second derivatives, called Fisher's scoring technique, has

$$\begin{aligned} \mathbf{I}(\hat{\boldsymbol{\beta}}_{s+1} - \hat{\boldsymbol{\beta}}_s) &= \mathbf{c}^* \\ \hat{\boldsymbol{\beta}}_{s+1} &= \hat{\boldsymbol{\beta}}_s + \mathbf{I}^- \mathbf{c}^* \end{aligned}$$

to pass from iterative step s to $s+1$, where $\mathbf{c}^* = \mathbf{X}^T\mathbf{V}^-(\mathbf{z}_s - \boldsymbol{\eta}_s)$ is a vector of length p. □

The conventional constraints (Section 1.3.3), although the most easily interpretable constraints in balanced cases like contingency tables, are not computationally simple and may not be helpful for unbalanced designs. Thus, unfortunately, they are not always provided by computer software for generalized linear models. Instead, a sweep operation can be used to evaluate the parameters sequentially and drop any that are aliased with previous ones (this is equivalent to setting them to zero). If we choose $\mathbf{I}^-$ to give such a solution, then $\text{var}(\hat{\boldsymbol{\beta}}) \doteq \mathbf{I}^-$, with rows and columns corresponding to aliased parameters set to zero. This also often could solve the problem of extrinsic alias, except for numerical approximation. Thus, with this method, we must watch out for one or more inflated standard errors of the parameters, indicating extrinsic alias.

A.1.3 Parameter Precision

Normed Likelihood

When only one functional form of statistical model is under consideration, it is often useful to compare all other models with that form to the most likely one using a special likelihood ratio, the *normed* or *relative likelihood function*:

$$\text{R}(\boldsymbol{\psi}; \mathbf{y}) = \frac{\text{L}(\boldsymbol{\psi}; \mathbf{y})}{\text{L}(\hat{\boldsymbol{\psi}}; \mathbf{y})}$$

This function can be directly used to obtain interval estimates of the parameter, that is, a set of plausible models. This is done by taking the set of all values of a (scalar) parameter with normed likelihood greater than some value, say a. However, the normed likelihood is only easily interpretable if the dimension of the parameter vector is one or two, in which case it can, for example, be plotted. We consider how to choose an appropriate value for a and how to handle vector parameters in the next subsection.

Example

Suppose that 3 heads are observed in 10 tosses of a coin. If we can assume the tosses to be independent and the probability of heads to be the same

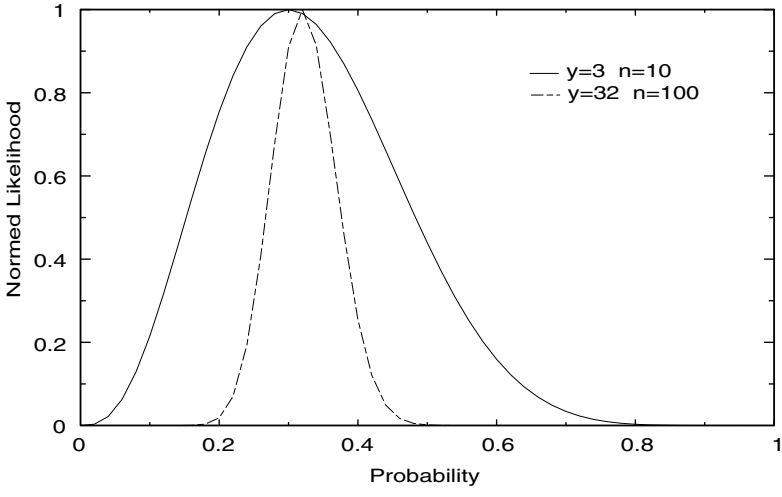

FIGURE A.1. Normed likelihood functions for two binomial experiments.

each time, then we can choose a binomial distribution as our model, yielding the normed likelihood,

$$\mathrm{R}(\pi) = \left(\frac{\pi}{0.3}\right)^3 \left(\frac{1-\pi}{0.7}\right)^7$$

We can plot this, as in Figure A.1. We see, for example, that any value of π in the interval $[0.10, 0.58]$ is at least one-fifth (that is, $a = 0.2$) as likely as $\hat{\pi} = 0.3$. Note that this interval is not symmetric about the m.l.e.

The same coin tossed 100 times yields, say, 32 heads. Now $\hat{\pi} = 0.32$ and the normed likelihood is

$$\mathrm{R}(\pi) = \left(\frac{\pi}{0.32}\right)^{32} \left(\frac{1-\pi}{0.68}\right)^{68}$$

This is the broken line in Figure A.1. The $a = 0.2$ likelihood interval is now $[0.23, 0.41]$. As might be expected, with more information available, our interval is narrower (and more symmetric). □

When the likelihood function has several local maxima, a disjoint *likelihood region* may be required. This will generally not be the case for generalized linear models.

Deviance

The *log likelihood function*

$$\mathrm{l}(\boldsymbol{\psi}; \mathbf{y}) = \log[\mathrm{L}(\boldsymbol{\psi}; \mathbf{y})]$$

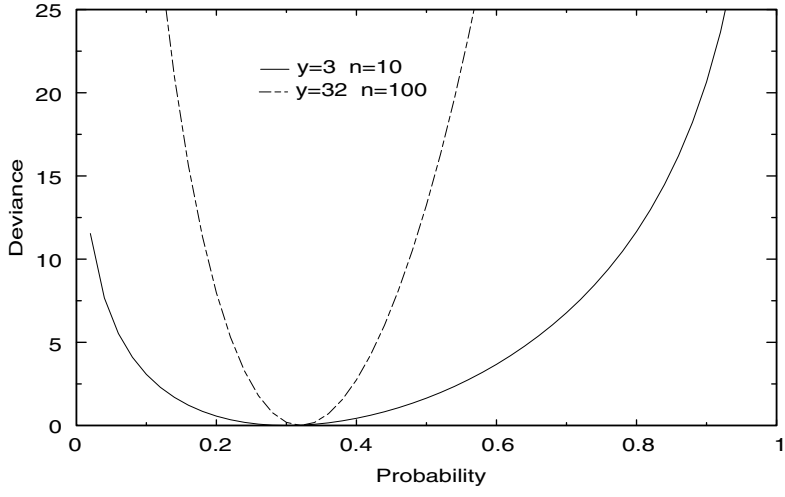

FIGURE A.2. Normed deviance functions for two binomial experiments.

is often convenient because it allows models (that can be decomposed as a product) to be compared additively instead of multiplicatively.

Often the log normed likelihood is used, but because it is always nonpositive, it can be multiplied by an arbitrary negative number. When this number is -2, it has come to be called the *deviance* (at least for one-parameter exponential family models):

$$\mathrm{D}(\boldsymbol{\psi}) = -2[\mathrm{l}(\boldsymbol{\psi};\mathbf{y}) - \mathrm{l}(\hat{\boldsymbol{\psi}};\mathbf{y})] \tag{A.4}$$

a nonnegative number, that will be used in Section A.2.1. The larger the deviance, the further the model under consideration is from the most likely model, in the set under study, given the observed data.

Example

The same coin experiments can be plotted as log normed likelihoods or deviances, as is seen in Figure A.2. □

Again, such plots allow the range of plausible values of the parameter, given the observed data, to be determined.

Profile Likelihood

When the parameter vector of a given model is of high dimension, one often wishes to be able to study one parameter component in isolation. Let us call this component i of the parameter vector, $\boldsymbol{\psi}$, the parameter of interest, λ, and the other, nuisance, parameters, $\boldsymbol{\phi}$, so that $\boldsymbol{\psi} = (\lambda, \boldsymbol{\phi}^T)^T$. Then, a (normed) *profile likelihood* for the parameter of interest is defined by

$$\mathrm{R}_p(\lambda;\mathbf{y}) \quad = \quad \max_{\psi|\lambda}\mathrm{R}(\boldsymbol{\psi};\mathbf{y})$$

$$= \quad \mathrm{R}(\hat{\boldsymbol{\phi}}_\lambda; \mathbf{y})$$

where $\hat{\boldsymbol{\phi}}_\lambda$ is the m.l.e. of $\boldsymbol{\phi}$ for the given fixed value of λ.

Example

For n independent observations generated from a normal distribution, the (approximate) likelihood is

$$\mathrm{L}(\mu, \sigma^2; \mathbf{y}) = (2\pi\sigma^2)^{-\frac{n}{2}} \mathrm{e}^{-\frac{\Sigma(y_i-\mu)^2}{2\sigma^2}}$$

where Δ_i is taken to be unity.

Suppose that the mean, μ, is of interest. For any given value of μ, $\hat{\sigma}^2_\mu = \sum(y_i - \mu)^2/n$ so that the normed profile likelihood is

$$\mathrm{R}_p(\mu) = \left[\frac{\sum(y_i - \hat{\mu})^2}{\sum(y_i - \mu)^2}\right]^{\frac{n}{2}}$$

This just involves a ratio of sums of squared deviations from the mean, where $\sum(y_i - \hat{\mu})^2$ is the smallest such sum for the observed data and the given model. Thus, inferences about the mean involve comparing estimates of the variance for the corresponding models (Section 9.2). □

A profile likelihood function for one parameter may be plotted in the same way as a one-dimensional normed likelihood.

Example

Suppose that we have a family of models indexed by two parameters, α and β, and observe data yielding a likelihood function such as that in Figure A.3, where three contours of constant likelihood are represented. The maximum likelihood estimate, $(\hat{\alpha}, \hat{\beta})$, lies at the diamond in the centre. In this figure, the two diagonal dashed lines trace the profile likelihoods for α and β. Each of these could be plotted as a two-dimensional curve like those in Figure A.1. □

Greater care must be taken in using this procedure than for the simpler normed likelihoods of one-dimensional models because it produces the profile from what may be a complex multidimensional surface. Thus, it can give a narrower range of likely values of the parameter of interest than is necessarily appropriate. This is because, at each point, it takes all other parameters to be fixed at their m.l.e. and does not take into account the varying width of the likelihood surface in those directions, that is, the varying amounts of information about these other parameters.

A.1.4 Model Selection

With a wide set of possible models available, as in the family of generalized linear models, model selection is very important. It often involves searching

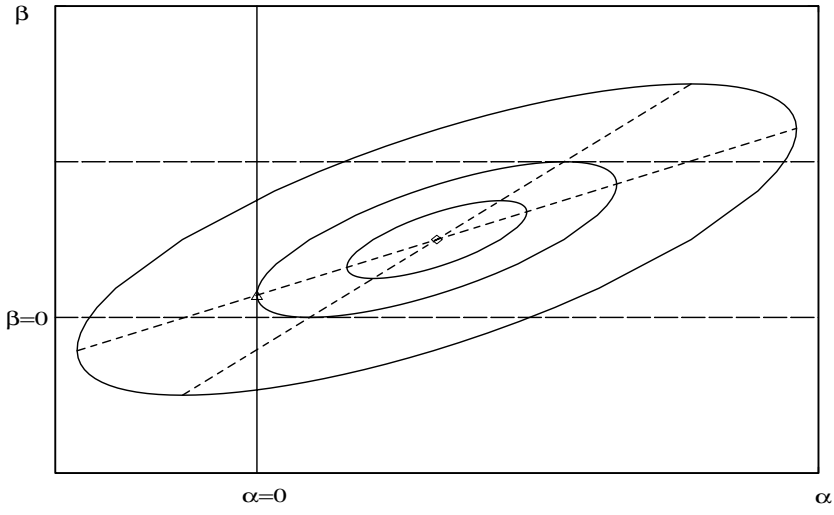

FIGURE A.3. Contours of constant normed likelihood for a two-parameter model, with submodels for $\alpha = 0$ and $\beta = 0$.

for the simplest reasonable model that will adequately describe the observed data. As we have seen, normed likelihoods and deviances can provide a measure of the distance of each model from the data, of *relative* goodness of fit. However, if enough parameters are included, the model can be brought to follow all of the irregularities of the observed data exactly and patterns of interest will not be isolated. Some smoothing is required.

The degrees of freedom provide a corresponding measure of complexity of a set of models; a model with a smaller number of estimated parameters generally smoothes the data more. The normed likelihood must be calibrated with the degrees of freedom to be interpretable. Thus, similar-sized normed likelihoods or deviances, for a series of different models, are not comparable if their degrees of freedom differ.

Two different types of model selection may be distinguished:

1. A complex model containing many parameters may be under consideration. A simpler submodel is to be selected by eliminating some of the parameters (or, conversely, some parameter may be added to a simple model).

2. Several distinct model functions, usually with different parameter sets, may be in competition.

Both situations require some means of calibrating normed likelihoods for models of different complexity to make them comparable.

Parameter Elimination

The clearest case of the procedures required for parameter elimination occurs when all parameters are independent of each other.

Example

Imagine that a series of p independent experiments, each yielding n observations, is generated from some distributions, each with one parameter, ψ_j. We wish to see if each parameter could be eliminated, say by setting it to zero. The likelihood function for all parameters being zero is the product of the individual functions. Suppose that we require the normed likelihood for an individual experiment, j, to be at least a for $\psi_j = 0$ to be plausible, that is, to be able to eliminate that one parameter. In this case of independent experiments, we would like to obtain the same final result whatever sequence we use to eliminate the parameters, one or more at a time, in a series of steps. Then, for all inferences to be *compatible* in this way, the combined likelihood for r ($\leq p$) experiments need only be greater than a^r for it to be plausible that this set of parameters are all zero, that is, to eliminate these r parameters. □

Now let us look more closely at how we can interpret a multidimensional likelihood function. For simplicity, suppose again that we have a family of models indexed by only two parameters, α and β, and observe data yielding a likelihood function such as that in Figure A.3. First, we shall consider all models within the outer contour, say 4%, as being reasonably plausible given the data. The model with $\alpha = \beta = 0$ lies within this region.

Next, let us, instead, check to see if either parameter individually could be zero. From our previous discussion, we believe that we could use a region of 20% for compatible inferences, because $0.2^2 = 0.04$. If this is the middle contour, we discover that models with $\alpha = 0$, the solid vertical line, and those with $\beta = 0$, the lower dashed horizontal line, each lie just on that contour. Thus, both sets of models, or at least one member of each of them, are plausible. We can decide to eliminate either α or β from our model. If we first eliminate α, the likelihood function for the remaining models of interest now lies on the solid vertical line through $\alpha = 0$ in Figure A.3.

By the decision to eliminate a parameter, we are excluding all models in the old 4% region defined by the outer contour, except those in the new subspace intersecting that region, the vertical line. In Figure A.3, when α is eliminated, all models within the outer contour are excluded, even although they are plausible, except for those on the line segment, $\alpha = 0$, contained within the contour. As well, our new most plausible model, with the constraint, $\alpha = 0$, will generally be considerably less plausible than the old most plausible model. The danger with this stepwise approach is that, at each step, we obtain a region with many plausible models, but then select only one subset as being of interest. The advantage is that we obtain

a simpler model. We have been obliged to decide what reduction in possible plausibility we are prepared to trade for a simpler model.

To continue, after eliminating only α, we now renormalize this likelihood, with $\alpha = 0$, to make inferences about β. To do this, we divide all likelihood values on this vertical line by the constrained maximum, at say $\hat{\beta}_\alpha$, with $\alpha = 0$, indicated by the triangle on that line. By this act of renormalization, we assume that only this family with $\alpha = 0$ now contains models of interest.

We would, then, like the new likelihood interval of plausible values to be that between the two points where the vertical line cuts the outer contour of 4%, our region for two parameters. With one parameter, we again take a 20% interval, the end points of which will lie on this contour of the two-dimensional normed likelihood surface, as required. Notice, however, that, if $\alpha = 0$ had lain within the 20% contour, the likelihood at $\hat{\beta}_\alpha$ would be greater than that on the 20% contour; a 20% interval for β, with α eliminated, will generally be narrower than the 4% contour. By eliminating unnecessary parameters, we can increase the precision of the remaining ones; see Altham (1984).

Suppose, now, instead that $\beta = 0$ is the upper horizontal dashed line in Figure A.3. Then, either α or β can be eliminated individually, but the second parameter will not subsequently be removed. Thus, two distinct models will result, depending on the order of the operations. Indeed, the point for both parameters being zero lies outside the 4% contour, so that they cannot simultaneously be removed. When parameters are not orthogonal, as indicated by the slope of the contours with respect to the axes, stepwise elimination can yield different results depending on the order followed. Of course, the likelihood surface will usually have a more complex shape than that in Figure A.3.

The only way in which successively renormalized likelihoods can provide compatible inferences is if all are calibrated as a^p, where p is the number of estimated parameters and a is some positive constant (chosen before beginning the study). If we start with a one-parameter region of 20%, we have $a = 0.2$ and the comparable two-parameter region will be 4%, and so on. The smaller the value chosen for a, the less plausible need be the simpler model with parameters eliminated. Thus, the smaller a is, the wider will be the precision interval about each parameter, giving more chance for it to include zero, so that the simpler will be the model chosen. Then, a can be called the *smoothing factor*, because simpler models smooth the data more.

Nonnested Models

As we see in Chapter 3, for independent observations, the most complex model that can be fitted to given data will be based on a multinomial distribution. These multinomial probabilities can often be structured in several different ways by the use of various competing model functions, perhaps of different complexity. The simplest case is that just studied, where success-

ive functions are obtained by adding or removing parameters. However, there is no reason that models need be nested in this way. Comparison of completely different functional forms for the multinomial probabilities may be of interest.

Our calibration of normed likelihood contours by a^p, according to the complexity of the models, does not require the models to be nested. It is applicable for all model selection procedures on a given set of data. Thus, for example, the calibrated deviance can be written

$$\mathrm{D}_c = \mathrm{D} - 2p \log(a) \tag{A.5}$$

where p is the number of estimated parameters and a the smoothing factor. Then, before beginning the data analysis (preferably before collecting the data), a can be chosen to obtain the required level of simplicity of the model selected. Equation (A.5) is based on what is sometimes called a *penalized likelihood*, because a provides a penalty against complex models (its logarithm is called the penalizing constant in Section 1.6). However, for it to be possible to compare models from different distributions, each likelihood or deviance must contain the complete probability distribution of the data, not just those factors involving the parameters (as is usually done).

Often, $a = 1/e$ is a reasonable choice; this is called the Akaike (1973) information criterion (AIC). It has been used throughout this book, although cases where it may be inappropriate are pointed out. Thus, an example is given in Section 4.4 where it is not really appropriate because the sample size is too large; there, a smaller value of a would be more suitable. The AIC was originally derived by asymptotic arguments that are inapplicable for direct likelihood inference, and even in contradiction with it. Here, the smoothing factor, $a = 1/e$, of the AIC is only appropriate if a reasonably small sample is available, not an asymptotically large one.

Other possibilities for the choice of a have also been proposed, including making it a function of the sample size, for example, $a = 1/\sqrt{n}$ or $a = 1/\sqrt{\log(n)}$, the former sometimes called the Bayesian information criterion (BIC). Although they will asymptotically select the "correct" model among two (or a finite number of) possible alternatives, this is not pertinent in the usual model selection conditions, where the set of possible models is not a limited and none is the "true" one. Here, these choices of a tend to select rather simple models as n increases, indicating a null model, with only one parameter, as n becomes very large. A subsidiary problem is the definition of n, itself. Is it the number of independent individuals observed or the number of distinct response values observed? In the former case, n would be unity for a single time series! But, in the latter case, for correlated responses, n does not convey the amount of information in the data.

Orthogonal Parameters

If the inclusion of parameter set A in the model produces the same reduction in deviance whether parameter set B is already in or not, sets A and B are said to be orthogonal, usually by the design of the experiment. If the sets are not orthogonal, the order of inclusion can be important for the conclusions. Parameters are generally only orthogonal, in this sense, for normal linear models.

A.1.5 Goodness of Fit

We can compare the likelihood of the current model (L_c) with that of the full or saturated model (L_f), as described in Section 1.3.1. The *scaled deviance* is often defined as

$$\mathrm{D}(c, f) = -2 \log \left(\frac{\mathrm{L}_c}{\mathrm{L}_f} \right)$$

in GLM terminology. However, this can be confusing, because in the general statistical literature, as in Equation (A.4) above, it is usually simply called the deviance.

For the exponential dispersion family,

$$\mathrm{D}(c, f) = 2 \sum\nolimits_i [y_i(\hat{\theta}_i - \tilde{\theta}_i) + b(\tilde{\theta}_i) - b(\hat{\theta}_i)]/a_i(\phi)$$

where the tilde ($\tilde{}$) indicates the current model and the hat ($\hat{}$) the saturated model. With $a_i(\phi) = \phi/w_i$, the deviance, in GLM terminology, is

$$\begin{aligned} \mathrm{D}_U(c, f) &= \phi \mathrm{D}(c, f) \\ &= 2 \sum_i w_i [y_i(\hat{\theta}_i - \tilde{\theta}_i) + b(\tilde{\theta}_i) - b(\hat{\theta}_i)] \end{aligned}$$

that can be calculated from the data. However, it is perhaps clearer to call this the *unscaled deviance*.

Example

	Deviance	Unscaled deviance
Poisson	$\sum \left[y_i \log \left(\frac{y_i}{\hat{\mu}_i} \right) - y_i + \hat{\mu}_i \right]$	Same
Binomial	$\sum \left[y_i \log \left(\frac{y_i}{\hat{\mu}_i} \right) + (n_i - y_i) \log \left(\frac{n_i - y_i}{n_i - \hat{\mu}_i} \right) \right]$	Same
Normal	$\frac{1}{\sigma^2} \sum (y_i - \hat{\mu}_i)^2$	$\sum (y_i - \hat{\mu}_i)^2$
Gamma	$2\nu \sum \left[\log \left(\frac{y_i}{\hat{\mu}_i} \right) + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} \right]$	$2 \sum \left[\log \left(\frac{y_i}{\hat{\mu}_i} \right) + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} \right]$

For the normal distribution with the identity link, the unscaled deviance gives the residual sum of squares for the current model. For the Poisson distribution with the log link, the deviance gives the log likelihood ratio statistic (G^2) for contingency tables. □

In a linear regression context, the mean parameter, μ, is allowed to vary with some explanatory variables. This regression model will contain a number of parameters, less than the number of independent observations, n. For a probability distribution with known scale parameter, say $f(y; \boldsymbol{\mu})$, or when there are replicated observations for at least some values of the explanatory variables, a special kind of a saturated model can be developed. This has a different μ_i for every distinct combination of the explanatory variables. One potential problem with this approach is that the number of parameters in such a saturated "semiparametric" model is not necessarily fixed but may increase with the number of observations. In certain cases, comparison to such a saturated model may be relatively uninformative about goodness of fit.

Example

For n Bernoulli observations y_i, with probabilities π_i, the deviance for goodness of fit of the model $\pi_i = \pi$, a constant, as compared to the saturated model is given by

$$\begin{aligned} \mathrm{D}(\pi) &= 2\sum_{i=1}^{n}\left[y_i \log\left(\frac{y_i}{\pi}\right) + (1-y_i)\log\left(\frac{1-y_i}{1-\pi}\right)\right] \\ &= 2n[\bar{y}_\bullet \log(\pi) + (1-\bar{y}_\bullet)\log(1-\pi)] \end{aligned}$$

because y_i is either zero or one. This contains only $\bar{y}_\bullet$, the m.l.e. of π, and not the individual observations, and so is, in fact, of little use for measuring goodness of fit. If information is available on the order in which the responses were generated, a suitable saturated model might be developed from that. □

The situation is not so simple when the probability distribution has more than one parameter. One will usually be a mean parameter and the regression model will be constructed in terms of it. In the simplest cases, the other parameter(s) are held constant for all values of the explanatory variables. If the mean is allowed to take a different value for each observation when the values of the explanatory variables are all different, there will be more parameters than observations and no information will remain in the data about the other parameter(s) in the probability distribution once the mean parameters are estimated. Although the model is saturated, it is of little use, unless there are replications of observations of the response for at least some values of the explanatory variables.

A.2 Frequentist Decision-making

The frequentist approach is primarily concerned with the long-run properties of statistical decision-making procedures — how they perform in repeated applications. It is based on probabilities obtained by integration over the sample space, thus depending on all possible outcomes and not just those actually observed. Thus, conclusions are critically dependent on the choice of the distribution of the responses in the model (explaining the attractiveness of nonparametric methods). The procedures are always implicitly testing this choice, as well as the explicit hypotheses under consideration.

This approach concentrates on testing if hypotheses are true and on the point estimation of parameters, based on the distributions of statistics calculable from the data. Parameter precision (confidence intervals) is based on testing. Asymptotic methods, for large sample sizes, are important and will be emphasized here.

In contrast to direct likelihood, this approach, and the next, must make the assumption that some model (function) is true in order to be able to draw conclusions.

A.2.1 Distribution of the Deviance Statistic

In order to calculate probabilities on the sample space for use in tests and confidence intervals, we require the distribution of available statistics, the most important of which is the deviance. Suppose that the parameter vector, $\boldsymbol{\beta}$, has a fixed length p. Then, asymptotically, for a known value of $\boldsymbol{\beta}$ (the hypothesis), the (scaled) deviance has a chi-squared distribution:

$$-2\{\log[\mathrm{L}(\boldsymbol{\beta};\mathbf{y})] - \log[\mathrm{L}(\hat{\boldsymbol{\beta}};\mathbf{y})]\} \sim \chi^2_p \quad \text{if } \mathrm{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$$

under mild regularity conditions.

Proof:

For the gradient or score function, $\mathbf{U}$, and the Hessian (or negative observed information), $\mathbf{H} = \partial\mathbf{U}/\partial\boldsymbol{\theta}$, of the log likelihood,

$$\begin{aligned} \mathrm{E}[\mathbf{U}] &= \mathbf{0} \\ \mathrm{E}[\mathbf{U}\mathbf{U}^T] &= \mathrm{E}[-\mathbf{H}] \\ &= \mathbf{I} \end{aligned}$$

the expected information matrix. Suppose that there exists a unique maximum of the log likelihood function at $\hat{\boldsymbol{\beta}}$ that is near the true value of $\boldsymbol{\beta}$.

The second-order Taylor series approximation to the log likelihood function is

$$\begin{aligned} \log[\mathrm{L}(\boldsymbol{\beta};\mathbf{y})] \doteq\ & \log[\mathrm{L}(\hat{\boldsymbol{\beta}};\mathbf{y})] + (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T\mathbf{U}(\hat{\boldsymbol{\beta}}) \\ & + \frac{1}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T\mathbf{H}(\hat{\boldsymbol{\beta}})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) \end{aligned}$$

Because, by definition, $\mathbf{U}(\hat{\boldsymbol{\beta}})=\mathbf{0}$, and replacing $\mathbf{H}(\hat{\boldsymbol{\beta}})$ by $-\mathbf{I}$,

$$-2\{\log[\mathrm{L}(\boldsymbol{\beta};\mathbf{y})] - \log[\mathrm{L}(\hat{\boldsymbol{\beta}};\mathbf{y})]\} \doteq (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T\mathbf{I}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})$$

Now, the first-order Taylor series approximation for the score is

$$\mathbf{U}(\boldsymbol{\beta}) \doteq \mathbf{U}(\hat{\boldsymbol{\beta}}) + \mathbf{H}(\hat{\boldsymbol{\beta}})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})$$

where $\mathbf{H}(\hat{\boldsymbol{\beta}})$ is evaluated at $\hat{\boldsymbol{\beta}}$. Because, asymptotically, $\mathbf{H}$ equals its expected value, for large samples,

$$\mathbf{U}(\boldsymbol{\beta}) \doteq \mathbf{U}(\hat{\boldsymbol{\beta}}) - \mathbf{I}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})$$

But, by definition, $\mathbf{U}(\hat{\boldsymbol{\beta}})=\mathbf{0}$, so

$$(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \doteq \mathbf{I}^{-1}\mathbf{U}$$

With $\mathbf{I}$ fixed,

$$\begin{aligned}\mathrm{E}[\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}] &\doteq \mathbf{I}^{-1}\mathrm{E}[\mathbf{U}]\\ &= \mathbf{0}\end{aligned}$$

as might be expected, and

$$\begin{aligned}\mathrm{E}[(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T] &\doteq \mathbf{I}^{-1}\mathrm{E}[\mathbf{U}\mathbf{U}^T]\mathbf{I}^{-1}\\ &= \mathbf{I}^{-1}\end{aligned}$$

Thus, for large samples,

$$\hat{\boldsymbol{\beta}}-\boldsymbol{\beta} \sim \mathrm{N}(\mathbf{0},\mathbf{I}^{-1})$$

and

$$(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T\mathbf{I}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) \sim \chi^2_p$$

Therefore,

$$-2\{\log[\mathrm{L}(\boldsymbol{\beta};\mathbf{y})] - \log[\mathrm{L}(\hat{\boldsymbol{\beta}};\mathbf{y})]\} \sim \chi^2_p$$

□

If $\mathbf{I}$ depends on $\boldsymbol{\beta}$, one often uses $\mathbf{I}(\hat{\boldsymbol{\beta}})$ or, even better, $-\mathbf{H}(\hat{\boldsymbol{\beta}})$, because it is not an average over the sample space.

Notice that this proof does not generally hold for saturated models because then $\boldsymbol{\beta}$ is not a vector of fixed length, but grows with sample size. However, consider a log linear model for a contingency table of fixed size. To compare the current to the saturated model, we have

$$\begin{aligned}\mathrm{D}(c,f) &= -2\log[\mathrm{L}_c(\hat{\boldsymbol{\beta}};\mathbf{y})/\mathrm{L}_f(\hat{\boldsymbol{\beta}};\mathbf{y})]\\ &= -2\{[\log\mathrm{L}_f(\boldsymbol{\beta};\mathbf{y}) - \log\mathrm{L}_f(\hat{\boldsymbol{\beta}};\mathbf{y})]\\ &\quad -[\log\mathrm{L}_c(\boldsymbol{\beta};\mathbf{y}) - \log\mathrm{L}_c(\hat{\boldsymbol{\beta}};\mathbf{y})]\\ &\quad -[\log\mathrm{L}_f(\boldsymbol{\beta};\mathbf{y}) - \log\mathrm{L}_c(\boldsymbol{\beta};\mathbf{y})]\}\end{aligned}$$

Under certain regularity conditions, the first term has a χ^2_n distribution (where n is here the number of cells in the table), the second, χ^2_p, and the third is a positive constant, near zero if the current model describes the data well. Thus

$$\mathrm{D}(c, f) \sim \chi^2_{n-p}$$

if the current model is good.

More generally, under mild regularity conditions, suppose that model 2 is nested in model 1, that is, that the parameter space under model 2 is a subspace of that of model 1 (and that the latter has a fixed number of parameters). If model 2 is correct, $\mathrm{D}(2,1) = -2\log(\mathrm{L}_2/\mathrm{L}_1)$ is distributed as χ^2 with $p_1 - p_2$ d.f., where p_i is the (fixed) number of independent parameters estimated under model i.

The distribution is exact for the normal distribution with the identity link and known variance, but approximate otherwise. However, generally, the deviance, $\mathrm{D}(2,1)$, is a function of ϕ. In the exponential dispersion family, ϕ is usually unknown. But the ratio of mean deviances, $\mathrm{D}(i,j)/(p_i - p_j)$, does not involve the scale parameter, so that it has approximately an F distribution. (Again, this is exact for the normal distribution with the identity link and variance now unknown.) The approximation, for nonnormal models, will generally be better for the difference between two deviances, expressing the effect of adding or removing a term, than for the deviance with respect to the saturated model. The latter would give the goodness of fit of the current model under the conditions explained above. Specifically, for binary data, as we have seen (Section A.1.5), this absolute deviance is uninformative because it only depends on the sufficient statistics.

A.2.2 Analysis of Deviance

Suppose that model 2 has parameters $\beta_1, \ldots, \beta_t, \beta_{t+1}, \ldots, \beta_p$, all linearly independent, whereas model 3 has only $\beta_1, \ldots, \beta_t$, with the models otherwise identical. Then, under model 3, $\mathrm{D}(3,2) \sim \chi^2_{p-t}$. We can test $\beta_{t+1} = \beta_{t+2} = \cdots = \beta_p = 0$ by comparing $\mathrm{D}(3,2)$ to a χ^2_{p-t} (or a ratio of them to F). Note also that $\mathrm{E}[\chi^2_{p-t}] = p - t$, so that large deviances, say more than twice the degrees of freedom, are suspect.

Suppose now model 1 to be saturated while model 2 has p independent parameters. As we have seen, $\mathrm{D}(2,1)$ can represent a test for lack of fit. If model 3, nested in model 2, has t ($< p$) independent parameters,

$$\begin{aligned} \mathrm{D}(3,1) - \mathrm{D}(2,1) &= -2\log(\mathrm{L}_3/\mathrm{L}_1) + 2\log(\mathrm{L}_2/\mathrm{L}_1) \\ &= -2\log(\mathrm{L}_3/\mathrm{L}_2) \\ &= \mathrm{D}(3,2) \end{aligned}$$

is distributed as χ^2_{p-t} when model 3 correct. Thus, for a sequence of k nested models, each with p_i independent parameters, form $\mathrm{D}(i,1)$ ($i = 1, \ldots, k$).

We, then, can create a table of differences of deviance analogous to sums of squares of an ANOVA table.

An alternative method for the elimination of individual parameters, including levels of factor variable, is to use the fact that $\hat{\beta}/\text{s.e.}(\hat{\beta})$ has an (approximate) Student t distribution with $n - t$ d.f. However, this should be used with great care for nonnormal models, where it can be very misleading. It is a useful quick rule of thumb that should be checked by looking at the corresponding change in deviance.

A.2.3 Estimation of the Scale Parameter

Much can be learned about generalized linear models without estimation of the scale parameter. However, it is often useful to have an estimate of this as well.

We know that $\mathrm{D}(c, f) \sim \chi^2_{n-p}$ for a model with p independent parameters, whereas $\mathrm{D}_U(c, f)$ can be calculated from the data. Then, under a reasonable model, we shall have $\mathrm{E}[\mathrm{D}(c, f)] = n - p$ and

$$\begin{aligned} \hat{\phi} &\doteq \frac{\mathrm{D}_U(c, f)}{\mathrm{E}[\mathrm{D}(c, f)]} \\ &= \frac{\mathrm{D}_U(c, f)}{n - p} \end{aligned}$$

This is not the maximum likelihood estimate (that divides by n), but a moment estimate and also the conditional maximum likelihood estimate.

A.3 Bayesian Decision-making

Often, when studying statistical models, prior beliefs about the unknown parameters are available. If these can be quantified as a *prior probability distribution*, they can be combined with the new information from the data just observed, contained in the likelihood function, to provide a new *posterior distribution*. This is known as the Bayesian approach.

In contrast to the frequentist approach that integrates over the sample space, this approach requires integration over the space of all possible models. These must *all* be specified before beginning the study, and a prior probability given to each. Any model with zero prior probability must have a zero posterior probability, so that this method excludes unexpected discoveries. Through the *likelihood principle*, it can provide conclusions only from the observed data, but only conditional on one of the models being true. The set of models (for example, defined by a model function) being used cannot. itself, be placed in question. This contrasts with the frequentist approach that cannot draw any conclusions without simultaneously questioning the validity of the whole set (that is, of the model function itself).

A.3.1 Bayes' Formula

Suppose that we have an accepted (true) model function containing the parameter, $\boldsymbol{\theta}$. Now suppose that this parameter is itself taken to be a random variable. We can imagine at least two ways in which this might occur. A population may be heterogeneous, so that some parameter, such as a mean, varies, in an unknown way not of direct interest, among subgroups (Section 2.3.2). Or, as here, we have some relative weights representing beliefs about the possible true value of $\boldsymbol{\theta}$ before making observations, that could be described by a probability distribution. Now consider Bayes' formula

$$p(\boldsymbol{\theta}|y) = \frac{f(y|\boldsymbol{\theta})p(\boldsymbol{\theta})}{f(y)} \tag{A.6}$$

Here, $p(\boldsymbol{\theta})$ is the prior distribution for $\boldsymbol{\theta}$ and $p(\boldsymbol{\theta}|y)$ its posterior distribution, whereas $f(y|\boldsymbol{\theta})$ is just the likelihood function. Note that distributions of observable quantities will be denoted, as usual, by $f(\cdot)$, and those for parameters by $p(\cdot)$.

Thus, posterior probability statements can simply be made about the parameter lying in any region of the parameter space. Because they involve integration, however, they may often be computationally complex.

A.3.2 Conjugate Distributions

One class of prior distributions is of particular interest in the context of generalized linear models. It would seem to be a desirable property that $p(\boldsymbol{\theta}|y)$ and $p(\boldsymbol{\theta})$ have the same functional form. Then, $p(\cdot)$ is called the *conjugate prior* for $f(y;\boldsymbol{\theta})$. It is said to be *closed under sampling*, because the distribution has the same form before and after observing the sample data. This implies that the marginal distribution of Y, in the denominator of Equation (A.6), called a *compound distribution*,

$$f(y) = \int_{\Theta} f(y|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$$

must have an analytically tractable form. In the two families that interest us most, conjugate distributions exist and can even be written down in general for each family.

Exponential Family

Recall, for an exponential family in canonical form, that

$$\log[f(y|\theta)] = \theta y - b(\theta) + c(y)$$

In order for the prior and posterior to have the same form within this family, it is easy to demonstrate that the conjugate prior must have the general form

$$\log[p(\theta;\boldsymbol{\zeta},\gamma)] = \theta\zeta - \gamma b(\theta) + s(\zeta,\gamma) \tag{A.7}$$

where ζ and γ are new parameters, with $\gamma > 0$ and $s(\zeta, \gamma)$ is not a function of θ. This is still an exponential family, but with two parameters and with canonical statistics, θ and $b(\theta)$. Thus, when we multiply the prior distribution and the likelihood function for n observations together in Bayes' formula, ζ and γ in the prior become $\zeta + y_\bullet$ and $\gamma + n$ in the posterior distribution. It is as if we had γ more observations than the n actually observed. Then, the marginal compound distribution is given by

$$\log[f(\mathbf{y}; \boldsymbol{\zeta}, \gamma)] = s(\zeta, \gamma) + c(\mathbf{y}) - s(\zeta + y_\bullet, \gamma + n)$$

Although the prior and posterior for θ are members of the exponential family, the compound distribution for Y generally will not be.

Example

The Poisson distribution, with $\theta = \log(\mu)$,

$$f(y|\mu) = \exp[y \log(\mu) - \mu - \log(y!)]$$

is a member of both the linear exponential and exponential dispersion families. The conjugate prior distribution for μ is

$$p(\mu; \zeta, \gamma) = \exp[\zeta \log(\mu) - \mu\gamma + s(\zeta, \gamma)]$$

that is a gamma distribution with

$$s(\zeta, \gamma) = (\zeta + 1) \log(\gamma) - \log[\Gamma(\zeta + 1)]$$

The resulting marginal compound distribution is

$$\begin{aligned} f(y; \zeta, \gamma) &= \exp[s(\zeta, \gamma) - \log(y!) - s(\zeta + y, \gamma + 1)] \\ &= \frac{\Gamma(\zeta + y + 1)\gamma^{\zeta+1}}{\Gamma(\zeta + 1)y!(\gamma + 1)^{\zeta+y+1}} \end{aligned}$$

that is a negative binomial distribution. The latter is commonly used when the fixed relationship between the mean and variance of the Poisson distribution does not hold, so that there is overdispersion of the random variable (see Section 2.3).

For n observations, the posterior distribution will be

$$\begin{aligned} p(\mu|\mathbf{y}; \zeta, \gamma) = \exp\{&(\zeta + y_\bullet) \log(\mu) - \mu(\gamma + n) \\ &+ (\zeta + y_\bullet + 1) \log(\gamma + n) - \log[\Gamma(\zeta + y_\bullet + 1)]\} \end{aligned}$$

again a gamma distribution. □

Exponential Dispersion Family

Consider now the exponential dispersion family,

$$\log[f(y|\theta;\phi)] = \frac{y\theta - b(\theta)}{\phi} + c(\phi, y)$$

Here, a conjugate prior, for θ alone, can still be written in the general form of Equation (A.7). Then, the marginal distribution is given by

$$\log[f(\mathbf{y};\zeta,\gamma)] = s(\zeta,\gamma) + c(\phi,\mathbf{y}) - s\left(\zeta + \frac{y_\bullet}{\phi}, \gamma + \frac{n}{\phi}\right)$$

that will generally not be an exponential dispersion model.

Example

Because the normal distribution, with $\theta = \mu$, is given by

$$f(y|\mu;\sigma^2) = \exp\left[\frac{y\mu - \frac{\mu^2}{2}}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right]$$

the conjugate prior distribution for μ is

$$p(\mu;\zeta,\gamma) = \exp\left[\zeta\mu - \frac{\gamma\mu^2}{2} + s(\zeta,\gamma)\right]$$

that is, itself, a normal distribution with mean ζ/γ, variance $1/\gamma$, and

$$s(\zeta,\gamma) = -\frac{\zeta^2}{2\gamma} - \frac{\log(2\pi/\gamma)}{2}$$

The resulting marginal compound distribution is also normal,

$$\begin{aligned} f(y;\zeta,\gamma,\sigma^2) &= \exp\left\{s[\zeta,\gamma] - \frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right. \\ &\qquad \left. -s\left[\zeta + \frac{y}{\sigma^2}, \gamma + \frac{1}{\sigma^2}\right]\right\} \\ &= \exp\left\{\frac{\zeta y}{\gamma\sigma^2+1} - \frac{\zeta^2}{2\gamma(\gamma\sigma^2+1)}\right. \\ &\qquad \left. -\frac{\gamma y^2}{2(\gamma\sigma^2+1)} - \frac{\log[2\pi(\gamma\sigma^2+1)/\gamma]}{2}\right\} \end{aligned}$$

with mean ζ/γ and variance $\sigma^2 + 1/\gamma$. For n observations, this becomes a multivariate normal distribution with mean and variance as given, and constant covariance $1/\gamma$. Models based on it are called random effects or variance components models.

For n observations, the posterior distribution will be

$$p(\mu|\mathbf{y};\zeta,\gamma,\sigma^2) = \exp\left\{\left(\zeta + \frac{y_\bullet}{\sigma^2}\right)\mu \right.$$
$$\left. -\frac{(\gamma + n/\sigma^2)\mu^2}{2} - \frac{(\zeta + y_\bullet/\sigma^2)^2}{2(\gamma + n/\sigma^2)} - \frac{\log[2\pi/(\gamma + n/\sigma^2)]}{2}\right\}$$

a normal distribution, as expected. □

In the absence of prior knowledge, a flat prior, often called improper because it does not have a finite integral, is frequently used. For this noninformative situation, such a prior can usually be obtained from the conjugate by choosing (limiting) special values of the parameters, $(\boldsymbol{\zeta}, \gamma)$. For the exponential family, Jeffreys' prior is a special case of the corresponding conjugate distribution, usually with $\gamma \to 0$.

Example

For the Poisson distribution, the conjugate gamma distribution for prior ignorance can be taken with $\zeta = \frac{1}{2}$ and $\gamma = 0$, yielding Jeffreys' prior. □

Because $\gamma = 0$, the sample size, n, is not being increased when such an noninformative prior is used.

Summary

Considerable debate exists among statisticians as to the ways in which inferences or decision-making should be carried out, with an especially sharp opposition between Bayesians and frequentists. However, the likelihood function is a common factor underlying all of these approaches. Thus, in this book, we only present results in terms of likelihoods, or, more exactly, deviances, with the corresponding value of the AIC to allow for differing numbers of parameters. From this information, any reader can make the necessary adjustments for his or her preferred method of inference.

Introductory books on likelihood inference and modelling include Edwards (1972), Kalbfleisch (1985a, 1985b), King (1989), and Lindsey (1995a). A more advanced text is Lindsey (1996b). For likelihood-based frequentist inference, see Cox and Hinkley (1974), and Bayesian inference, Box and Tiao (1973), Berger and Wolpert (1988), and Gelman *et al.* (1995).

This page intentionally left blank

Appendix B
Diagnostics

B.1 Model Checking

Comparisons of models based on likelihoods usually only provide global indications of lack of fit of a model of interest. If the latter is found to be wanting, a saturated model would not usually be an acceptable alternative. Instead, unexpected ways in which the data depart from the model must be studied in order to discover means of improving the old model of interest. Thus, for example, if the old model assumes independence among observations, the new model might require some specific form of dependence.

If one is aware of the type of departure, embedding a model in more complex ones will be the most useful procedure to follow. The methods outlined in this appendix are more exploratory. Basically, we shall decompose global measures of goodness of fit, such as the deviance, into the individual terms arising from each observation. This may allow us to discover which aspects of the models of interest are unsatisfactory, leading to the development of more complex or functionally different ones. The study of residuals is generally only useful if the sample size is relatively small, say at most 100 observations or so, and the model under consideration is far from being saturated.

Departures from Models

Two types of departures from a model may be distinguished:

1. the observations on one, or a small number of units, may not be well represented by the model, or

2. the whole data set may show systematic departure from the model.

The study of such departures is known as *diagnostics*. In realistic cases, departures from a model will often be difficult to interpret. As well, if several explanatory variables are present in a model, it will not even be easy to represent the complete model graphically.

B.2 Residuals

The study of departures from a model has classically primarily been concerned with the expected values of the responses under some model and with their differences from the observed responses, called *residuals*. This reflects the fact that this approach was originally developed for normal linear regression models. However, the concept of residual has subsequently acquired several more general definitions that are of much wider application. Thus, the role of residuals is central in diagnostics. As we shall see, for the normal distribution, most of the more recently derived definitions of residuals collapse to the same quantities, aside from standardizing factors.

Plots of residuals can sometimes be useful in detecting departures from the model for many types of response variables. However, when the response can only take a few values, as in Bernoulli trials, they may be of limited usefulness.

B.2.1 Hat Matrix

The estimated expected value of the response in a regression model is

$$\begin{aligned} \widehat{\mathrm{E}[Y_i]} &= \hat{\mu}_i \\ &= \hat{y}_i \end{aligned}$$

This is usually, although somewhat misleadingly, called the *fitted value*. It could mistakenly be taken to imply that all values "should" be at their expected value, with no random variation. That would never be the case in a nondeterministic, probability-based, *statistical* model. Indeed, it may not even be near the most common observed values if the distribution is very skewed.

For models with a linear structure, we can show that

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$$

so that $\mathbf{H}$ is called the *hat matrix*, because it puts the hat on y. This matrix is idempotent and symmetric; its trace is equal to p, the dimension of the parameter vector, and its elements cannot exceed one.

The reciprocals of the diagonal elements, h_{ii}, called the *effective replication*, can be roughly interpreted as the number of observations providing

information about $\hat{y}_i$. If this value is close to unity, the observation is isolated or self-estimating. Because of the continuity (the smoothing) of an unsaturated model, neighbouring points provide additional information in predicting a given response.

Examples

1. For normal linear regression,

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \end{aligned} \tag{B.1}$$

so that

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \tag{B.2}$$

The variance of $\hat{y}_i$ is $h_{ii}\sigma^2$.

2. For a generalized linear model, $\mathbf{X}$ in the above hat matrix is replaced by $\mathbf{V}^{\frac{1}{2}}\mathbf{X}$, where the weights, from Equation (A.3), are

$$\mathbf{V} = \text{diag}[\text{J}_{\eta_i\eta_i}(\eta)] \tag{B.3}$$

The hat matrix,

$$\mathbf{H} = \mathbf{V}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}^T\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{\frac{1}{2}} \tag{B.4}$$

now generally depends on the parameter values through $\mathbf{V}$. □

B.2.2 Kinds of Residuals

Deviance Residuals

For independent observations, the deviance is a sum of terms. Then, the standardized *deviance residuals* (Pregibon, 1981) decompose the goodness of fit of a model, measured by the likelihoods comparing the model of interest embedded in a saturated model, as in Section A.1.5. These terms may indicate which individual observations contribute most to the lack of fit. They are defined as the square root of the (corrected) contribution of the ith observation to the deviance:

$$\varepsilon_i^D = \frac{\text{sign}(\tilde{\eta}_i - \hat{\eta}_i)\sqrt{2\text{l}(\tilde{\eta}_i; y_i) - 2\text{l}(\hat{\eta}_i; y_i)}}{\sqrt{1 - h_{ii}}}$$

where h_{ii} is the ith diagonal element of the hat matrix and $\tilde{\eta}_i$ is the value of the linear structure, η, that maximizes the unconstrained likelihood for the data, the saturated model. For exponential dispersion models with unknown dispersion parameter, and the generalized linear models based on them, the deviance can be calculated for fixed dispersion parameter and then an estimate of that parameter supplied.

Examples

1. For the normal distribution, the deviance residuals are

$$\varepsilon_i^D = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\sigma}^2(1-h_{ii})}} \tag{B.5}$$

for fixed σ^2.

2. For the logistic regression, the deviance residuals are

$$\varepsilon_i^D = \frac{\pm\sqrt{2y_i \log\left(\frac{y_i}{n_i\hat{\pi}_i}\right) + 2(n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - n_i\hat{\pi}_i}\right)}}{\sqrt{1-h_{ii}}}$$

where the sign is chosen to be the same as that of $y_i - n_i\hat{\pi}_i$. □

Fitted Value Residuals

The *raw fitted value residual* for each response is simply its difference from its fitted value:

$$\varepsilon_i^R = y_i - \hat{y}_i \tag{B.6}$$

These are the classical, well-known residuals.

In normal linear regression models, for which such an approach is most useful, the variance of the response variable is assumed constant for all values of the explanatory variable. However, the variance of the raw residuals is not constant, as an explanatory variable changes, but is larger for responses near the mean of that variable. Thus, it is useful to standardize the raw fitted value residuals by dividing them by their standard error to obtain a standardized *studentized* (fitted value) *residual*:

$$\varepsilon_i^F = \frac{y_i - \hat{\mu}_i}{\sqrt{(1-h_{ii})\widehat{\text{var}}[Y_i]}}$$

This is also sometimes called the *standardized Pearson residual* because $(y_i - \hat{y}_i)^2/\widehat{\text{var}}[Y_i]$ is the contribution of the ith observation to the Pearson (score) statistic.

Examples

1. For the normal regression model, from Equations (B.1), (B.2), and (B.6), the variance of the raw residuals is given by

$$\text{var}[\boldsymbol{\varepsilon}^R] = \sigma^2(\mathbf{I}_n - \mathbf{H})$$

where $\mathbf{I}_n$ is the identity matrix. Thus, the studentized residuals are

$$\varepsilon_i^F = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\sigma}^2(1-h_{ii})}}$$

They are the (standardized) proportional contributions of the raw residuals to the standard deviation and are identical to the deviance residuals for this model, given in Equation (B.5).

2. For a logistic regression model, the studentized residuals are

$$\varepsilon_i^F = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)(1-h_{ii})}}$$

□

Except for the normal distribution, these residuals will generally have a skewed distribution.

Score Residuals

Maximum likelihood estimates for a model are obtained by solving the score equations. For independent responses, these involve a sum of terms, one for each observation, set equal to zero. Thus, we can inspect the individual terms of the sum, called *score residuals*, to see which are furthest from zero. As above, these will be standardized by correcting for their standard deviation:

$$\varepsilon_i^S = \frac{\mathrm{U}_i}{\sqrt{(1-h_{ii})v_i}}$$

where U_i is the term for the ith individual in the score and v_i is the ith diagonal element of the weight matrix in Equation (B.3).

In a generalized linear model, if the score for the linear structure, η, is used, the score residuals are identical to the studentized fitted value residuals given above.

Likelihood Residuals

Another possibility is to compare the deviance for a fitted model for the complete set of observations with that when each observation, in turn, is omitted. This requires substantial calculation, but may be approximated by the *likelihood residuals*,

$$\varepsilon_i^L = \mathrm{sign}(\tilde{\eta}_i - \hat{\eta}_i)\sqrt{h_{ii}(\varepsilon_i^S)^2 + (1-h_{ii})(\varepsilon_i^D)^2} \qquad \text{(B.7)}$$

a weighted average of deviance and score residuals (Williams, 1987).

In certain circumstances, those discussed in Section A.1.5, the sum of squares of these various residuals provides a global measure of lack of fit.

B.2.3 Residual Plots

Any of these residuals can be plotted against a variety of statistics and other indices, each providing different information about departures from the model.

In an *index plot*, the residuals are shown against the corresponding observation number. Ordering in this way may make identification of departures from the model easier. If the order has intrinsic meaning, for example, as the order of collection of the data in time, the plot may indicate systematic variability in this sense. In this case, it may also be helpful to plot the residuals against their *lagged* values at one or more previous time points.

Residuals can be plotted against the estimated means or estimated linear structure. They may also be plotted against each of the available explanatory variables, including those not in the model under consideration.

Finally, a *normal probability* or Q–Q (quantile) *plot* shows the standardized residuals, arranged in ascending order, against an approximation to their expected values, that is given by a standard normal distribution, $\Phi^{-1}[(i-3/8)/(n+1/4)]$. If the model fits well, this should yield a straight line at 45°. If the distribution of residuals is too skewed, the line will not pass through the origin, and if it is too long-tailed, the line will be curved. Remember, however, that residuals for nonnormal models are generally skewed.

All of these graphics can be inspected, both to look for patterns in the residuals and to detect certain extreme values.

B.3 Isolated Departures

When only a very few observations do not fit the model, several possibilities may be considered.

- there may be some error in choosing certain members of the population sampled or it may not be homogeneous for the factors considered;
- there may be some error in recording the results, either on the part of people doing the recording or transcribing it or on the part of the individuals concerned, for example, when respondents do not understand a question;
- some rare, but possible, occurrence may have been observed;
- the model may not be sufficiently well specified to account for completely acceptable observations, thus, pointing to unforeseen aspects of the phenomenon under study.

If there is no error, one will eventually have to decide if the departure is important enough to modify the model to take it into account.

B.3.1 Outliers

Any individual observation that departs in some way from the main body of the data is called an *outlier*. This implies that extreme observations can only be determined in relation to some model.

Example

The data set $\{240, 194, 215, 194, 450, 240, 215, 215\}$ appears to contain the outlier, 450. However, these are weekly salaries of a random sample of employees, where the large figure is for a manager and the others for junior staff, a proportion reflecting the pay structure of the company where the sample was taken (McPherson, 1989). □

Outliers may be due to extreme values of the response variable or of one or more of the explanatory variables. The possibility that the ith observation is an outlier can be studied by fitting the model without that observation. This yields a reduction in deviance for the possibility that it is an outlier. If one wishes to check all observations in this way, the approximation using the likelihood residuals of Equation (B.7) can considerably reduce the number of calculations required.

In complex situations, it is rarely wise simply to eliminate an outlier, unless it is known to be an error. In such case, it should, if possible, be corrected! Eliminating one outlier and refitting the model will quite often result in a second outlier appearing, and so on. It is usually preferable either to find out why the model cannot easily accommodate the observation or to accept it as a rare value.

Note, however, that the definition of an observation may not always be clear in this context. If Bernoulli trials are aggregated as frequencies in a contingency table, should one trial or one frequency be omitted?

B.3.2 Influence and Leverage

An *influential* observation is one that, if changed by a small amount or omitted, will modify substantially the parameter estimates of the model. It is an observation that may have undue impact on conclusions from the model. However, it may not be an outlier, in the sense that it may not be too far from the main body of data. It may have a small residual.

Leverage is one indication of how much influence an observation has. The general definition of the leverage of the jth observation on the ith fitted value is the magnitude of the derivative of the ith fitted value with respect to the jth response value. In a generalized linear model, this is given by the hat matrix, $\mathbf{H}$. Thus, a measure of the leverage effect of the observation, i, in the determination of $\hat{\mu}_i$ is the diagonal element of the hat matrix, h_{ii}, with $0 \leq h_{ii} \leq 1$. Note, however, from Equations (B.3) and (B.4), that this depends both on the explanatory variables and on the parameter estimates.

It is a measure of the distance of that observation from the remaining ones, in the space of these variables. The trace of $\mathbf{H}$ is equal to p, the dimension of the vector of regression parameters, $\boldsymbol{\beta}$, so that the average leverage is p/n. Values much larger than, say twice, this should be examined.

Cook's distance (Cook, 1977) is used to examine how each observation affects the complete set of parameter estimates. The estimates, with and without each observation, can be compared using

$$\mathrm{C}_i = \frac{1}{p}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})^T \mathbf{X}^T \mathbf{V} \mathbf{X} (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})$$

where $\hat{\boldsymbol{\beta}}_{(i)}$ is the parameter estimate without the ith observation. This statistic measures the squared distance between $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{(i)}$. Again, to avoid refitting the model n times, one with each observation removed, this distance can be approximated by

$$\mathrm{C}_i \doteq \frac{h_{ii}(\varepsilon_i^F)^2}{p(1-h_{ii})}$$

It is a combination of leverage and residuals and is most usefully presented as a plot against index values.

B.4 Systematic Departures

Systematic departures from a model can sometimes be detected from the residual plots already described. Certain patterns may appear when the residuals are plotted against some other statistic or against explanatory variables.

In a very general regression model, we would have

$$g(\mu_i) = \eta(\mathbf{x}_i, \boldsymbol{\beta})$$

where μ_i is the mean of the random variable following some probability distribution. Misspecification of such a model may come about in a number of ways:

- an incorrect choice of probability distribution,
- an incorrect specification of the way in which the mean changes with the explanatory variables,
  - the systematic component, $\eta(\cdot)$, may be misspecified,
  - the link function, $g(\cdot)$, may not be appropriate,
- missing explanatory variables,

- incorrect functions of the explanatory variables in the model, including missing interactions among them, or not enough such different functions,
- dependence among the observations, for example over time.

These can be verified by fitting the appropriate models and comparing the likelihoods, as in Section A.1.5.

Summary

A number of good books on diagnostics exist, although they concentrate primarily on normal linear models. See, for example, Cook and Weisberg (1982) and Barnett and Lewis (1984). More generally, for generalized linear models, see Pregibon (1981), Pierce and Schafer (1986), Williams (1987), and Davison and Tsai (1992).

This page intentionally left blank

References

[1] Aalen, O.O. (1978) Nonparametric inference for a family of counting processes. *Annals of Statistics* **6**, 701–726.

[2] Aalen, O.O. (1989) A linear regression model for the analysis of life times. *Statistics in Medicine* **8**, 907–925.

[3] Agresti, A. (1990) *Categorical Data Analysis*. New York: John Wiley.

[4] Aitkin, M. and Clayton, D. (1980) The fitting of exponential, Weibull, and extreme value distributions to complex censored survival data using GLIM. *Journal of the Royal Statistical Society* **C29**, 156–163.

[5] Aitkin, M., Francis, B., Hinde, J., and Anderson, D. (1989) *Statistical Modelling in GLIM.* Oxford: Oxford University Press.

[6] Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In Petrov, B.N. and Csàki, F., eds., *Second International Symposium on Inference Theory*, Budapest: Akadémiai Kiadó, pp. 267–281.

[7] Altham, P.M.E. (1978) Two generalizations of the binomial distribution. *Journal of the Royal Statistical Society* **C27**, 162–167.

[8] Altham, P.M.E. (1984) Improving the precision of estimation by fitting a model. *Journal of the Royal Statistical Society* **B46**, 118–119.

[9] Altman, D.G. (1991) *Practical Statistics for Medical Research.* London: Chapman & Hall.

[10] Andersen, A.H., Jensen, E.B., and Schou, G. (1981) Two-way analysis of variance with correlated errors. *International Statistical Review* **49**, 153–167.

[11] Andersen, E.B. (1991) *Statistical Analysis of Categorical Data.* Berlin: Springer-Verlag.

[12] Andersen, P.K., Borgan, O., Gill, R.D., and Keiding, N. (1993) *Statistical Models Based on Counting Processes.* Berlin: Springer-Verlag.

[13] Andrews, D.F. and Herzberg, A.M. (1985) *Data. A Collection of Problems from Many Fields for the Student and Research Worker.* Berlin: Springer-Verlag.

[14] Anscombe, F.J. (1981) *Computing in Statistical Science through APL.* Berlin: Springer-Verlag.

[15] Barndorff-Nielsen, O. (1978) *Information and Exponential Families in Statistical Theory.* New York: John Wiley.

[16] Barnett, V. and Lewis, T. (1984) *Outliers in Statistical Data.* New York: John Wiley.

[17] Barnhart, H.X. and Sampson, A.R. (1995) Multiple population models for multivariate random length data – with applications in clinical trials. *Biometrics* **51**, 195–204.

[18] Baskerville, J.C., Toogood, J.H., Mazza, J., and Jennings, B. (1984) Clinical trials designed to evaluate therapeutic preferences. *Statistics in Medicine* **3**, 45–55.

[19] Bates, D.M. and Watts, D.G. (1988) *Nonlinear Regression Analysis and its Applications.* New York: John Wiley.

[20] Bennett, G.W. (1988) Determination of anaerobic threshold. *Canadian Journal of Statistics* **16**, 307–316.

[21] Berger, J.O. and Wolpert, R.L. (1988) *The Likelihood Principle: A Review, Generalizations, and Statistical Implications.* Hayward, California: Institute of Mathematical Statistics.

[22] Berkson, J. (1944) Application of the logistic function to bio-assay. *Journal of the American Statistical Association* **39**, 357–365.

[23] Berkson, J. (1951) Why I prefer logits to probits. *Biometrics* **7**, 327–339.

[24] Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society* **B36**, 192–236.

[25] Beveridge, W. (1936) Wages in the Winchester manors. *Economic History Review* **7**, 22–43.

[26] Birch, M.W. (1963) Maximum likelihood in three-way contingency tables. *Journal of the Royal Statistical Society* **B25**, 220–233.

[27] Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975) *Discrete Multivariate Analysis: Theory and Practice.* Cambridge: MIT Press.

[28] Bliss, C.I. (1935) The calculation of the dosage-mortality curve. *Annals of Applied Biology* **22**, 134–167.

[29] Blossfeld, H.P., Hamerle, A., and Mayer, K.U. (1989) *Event History Analysis.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

[30] Box, G.E.P. and Cox, D.R. (1964) An analysis of transformations. *Journal of the Royal Statistical Society* **B26**, 211–252.

[31] Box, G.E.P. and Tiao, G.C. (1973) *Bayesian Inference in Statistical Analysis.* New York: John Wiley.

[32] Box, G.E.P. and Tidwell, P.W. (1962) Transformations of the independent variables. *Technometrics* **4**, 531–550.

[33] Brown, B.M. and Maritz, J.S. (1982) Distribution-free methods in regression. *Australian Journal of Statistics* **24**, 318–331.

[34] Brown, L.D. (1986) *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory.* Hayward, California: Institute of Mathematical Statistics.

[35] Burridge, J. (1981) Empirical Bayes analysis of survival time data. *Journal of the Royal Statistical Society* **B43**, 65–75.

[36] Buxton, J.R. (1991) Some comments on the use of response variable transformations in empirical modelling. *Journal of the Royal Statistical Society* **C40**, 391–400.

[37] Coleman, J.S. (1964) *Introduction to Mathematical Sociology.* New York: The Free Press of Glencoe.

[38] Collett, D. (1991) *Modelling Binary Data.* London: Chapman & Hall.

[39] Cook, R.D. (1977) Detection of influential observations in linear regression. *Technometrics* **19**, 15–18.

[40] Cook, R.D. and Weisberg, S. (1982) *Residuals and Influence in Regression.* London: Chapman & Hall.

[41] Cox, D.R. (1972) Regression models and life-tables. *Journal of the Royal Statistical Society* **B34**, 187–220.

[42] Cox, D.R. and Oakes, D. (1984) *Analysis of Survival Data.* London: Chapman & Hall.

[43] Cox, D.R. and Hinkley, D.V. (1974) *Theoretical Statistics.* London: Chapman & Hall.

[44] Cressie, N. (1986) Kriging nonstationary data. *Journal of the American Statistical Association* **81**, 625–634.

[45] Cressie, N. (1989) Geostatistics. *American Statistician* **43**, 197–202.

[46] Cressie, N. (1993) *Statistics for Spatial Data.* New York: John Wiley.

[47] Crouchley, R., Davies, R.B., and Pickles, A.R. (1982) Identification of some recurrent choice processes. *Journal of Mathematical Sociology* **9**, 63–73.

[48] Davison, A.C. and Tsai, C.L. (1992) Regression model diagnostics. *International Statistical Review* **60**, 337–353.

[49] de Angelis, D. and Gilks, W.R. (1994) Estimating acquired immune deficiency syndrome incidence accounting for reporting delay. *Journal of the Royal Statistical Society* **A157**, 31–40.

[50] de Jong, P. (1988) The likelihood of a state space model. *Biometrika* **75**, 165–169.

[51] de Jong, P. and Greig, M. (1985) Models and methods for pairing data. *Canadian Journal of Statistics* **13**, 233–241.

[52] Decarli, A., Francis, B., Gilchrist, R. and Seeber, G.U.H., eds. (1989) *Statistical Modelling.* Berlin: Springer-Verlag

[53] Derman, C., Gleser, L.J., and Olkin, I. (1973) *A Guide to Probability Theory and Application.* New York: Holt, Rinehart, and Winston.

[54] Diggle, P.J. (1990) *Time Series. A Biostatistical Introduction.* Oxford: Oxford University Press.

[55] Diggle, P.J., Liang, K.Y., and Zeger, S.L. (1994) *The Analysis of Longitudinal Data.* Oxford: Oxford University Press.

[56] Dinse, G.E. (1982) Nonparametric estimation for partially-complete time and type of failure data. *Biometrics* **38**, 417–431.

[57] Dobson, A.J. (1990) *An Introduction to Generalized Linear Models.* London: Chapman & Hall.

[58] Dyke, G.V. and Patterson, H.D. (1952) Analysis of factorial arrangements when the data are proportions. *Biometrics* **8**, 1–12.

[59] Edwards, A.W.F. (1972) *Likelihood. An Account of the Statistical Concept of* Likelihood *and its Application to Scientific Inference.* Cambridge: Cambridge University Press.

[60] Efron, B. (1986) Double exponential families and their use in generalized linear regression. *Journal of the American Statistical Association* **81**, 709–721.

[61] Fahrmeir, L. (1989) Extended Kalman filtering for nonnormal longitudinal data. In Decarli *et al.*, pp. 151–156.

[62] Fahrmeir, L., Francis, B., Gilchrist, R., and Tutz, G., eds. (1992) *Advances in GLIM and Statistical Modelling.* Berlin: Springer-Verlag.

[63] Fahrmeir, L. and Tutz, G. (1994) *Multivariate Statistical Modelling Based on Generalized Linear Models.* New York: Springer-Verlag.

[64] Feigl, P. and Zelen, M. (1965) Estimation of exponential survival probabilities with concomitant information. *Biometrics* **21**, 826–838.

[65] Feller, W. (1957) *An Introduction to Probability Theory and its Applications.* New York: John Wiley.

[66] Fingleton, B. (1984) *Models of Category Counts.* Cambridge: Cambridge University Press.

[67] Fisher, R.A. (1922) On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society* **222**, 309–368.

[68] Fisher, R.A. (1934) Two new properties of mathematical likelihood. *Proceedings of the Royal Society* **A144**, 285–307.

[69] Fisher, R.A. (1958) *Statistical Methods for Research Workers.* Edinburgh: Oliver and Boyd. (First edition, 1925)

[70] Fisher, R.A. (1959) *Statistical Methods and Scientific Inference.* Edinburgh: Oliver and Boyd. (First edition, 1956)

[71] Fisher, R.A. (1960) *Design of Experiments.* Edinburgh: Oliver and Boyd. (First edition, 1935)

[72] Fitzmaurice, G.M. and Laird, N.M (1993) A likelihood-based method for analysing longitudinal binary responses. *Biometrika* **80**, 141–151.

[73] Fleming, T.R. and Harrington, D.P. (1991) *Counting Processes and Survival Analysis.* New York: John Wiley.

[74] Francis, B., Green, M., and Payne, C. (1993) *The GLIM System. Release 4 Manual.* Oxford: Oxford University Press.

[75] Fry, F.E.J. and Hart, J.S. (1948) Cruising speed of goldfish in relation to water temperature. *Journal of the Fisheries Research Board of Canada* **7**, 169–175.

[76] Gamerman, D¿ (1991) Dynamic Bayesian models for survival data. *Journal of the Royal Statistical Society* **C40**, 63–79.

[77] Gehan, E.A. (1965) A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika* **52**, 203–223.

[78] Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1995) *Bayesian Data Analysis.* London: Chapman & Hall.

[79] Gilchrist, R., ed. (1982) *GLIM 82.* Berlin: Springer-Verlag.

[80] Gilchrist, R., Francis, B. and Whittaker, J., eds (1985) *Generalized Linear Models.* Berlin: Springer-Verlag.

[81] Glasser, M. (1967) Exponential survival with covariance. *Journal of the American Statistical Association* **62**, 561–568.

[82] Haberman, S.J. (1978) *Analysis of Qualitative Data. Volume 1. Introductory Topics.* San Diego: Academic Press.

[83] Haberman, S.J. (1979) *Analysis of Qualitative Data. Volume 2. New Developments.* San Diego: Academic Press.

[84] Hand, D. and Crowder, M. (1996) *Practical Longitudinal Data Analysis.* London: Chapman & Hall.

[85] Hand, D.J., Daly, F., Lunn, A.D., McConway, K.J., and Ostrowski, E. (1994) *A Handbook of Small Data Sets.* London: Chapman & Hall.

[86] Harkness, R.D. and Isham, V. (1983) A bivariate spatial point pattern of ants' nests. *Journal of the Royal Statistical Society* **C32**, 293–303.

[87] Harvey, A.C. (1989) *Forecasting, Structural Time Series Models and the Kalman Filter.* Cambridge: Cambridge University Press.

[88] Harvey, A.C. and Fernandes, C. (1989) Time series models for count or qualitative observations. *Journal of Business and Economic Statistics* **7**, 407–423.

[89] Hay, J.W. and Wolak, F.A. (1994) A procedure for estimating the unconditional cumulative incidence curve and its variability for the human immunodeficiency virus. *Journal of the Royal Statistical Society* **C43**, 599–624.

[90] Healy, M.J.R. (1988) *GLIM. An Introduction.* Oxford: Oxford University Press.

[91] Healy, M.J.R. and Tillett, H.E. (1988) Short-term extrapolation of the AIDS epidemic. *Journal of the Royal Statistical Society* **A151**, 50–61.

[92] Heckman, J.J. and Willis, R.J. (1977) A beta-logistic model for the analysis of sequential labor force participation by married women. *Journal of Political Economy* **85**, 27–58.

[93] Heitjan, D.F. (1991a) Generalized Norton–Simon models of tumour growth. *Statistics in Medicine* **10**, 1075–1088.

[94] Heitjan, D.F. (1991b) Nonlinear modeling of serial immunologic data: a case study. *Journal of the American Statistical Association* **86**, 891–898.

[95] Huet, S., Bouvier, A., Gruet, M.A., and Jolivel, E. (1996) *Statistical Tools for Nonlinear Regression. A Practical Guide with S-Plus Examples.* Berlin: Springer-Verlag.

[96] Hurley, M.A. (1992) Modelling bedload transport events using an inhomogeneous gamma process. *Journal of Hydrology* **138**, 529–541.

[97] Jarrett, R.G. (1979) A note on the intervals between coal-mining disasters. *Biometrika* **66**, 191–193.

[98] Jennrich, R.I. and Schluchter, J.R. (1986) Unbalanced repeated measures with structured covariance matrices. *Biometrics* **42**, 805–820.

[99] Jones, B. and Kenward, M.G. (1989) *Design and Analysis of Cross-over Trials.* Chapman & Hall.

[100] Jones, R.H. (1993) *Longitudinal Data Analysis with Serial Correlation: A State-space Approach.* London: Chapman & Hall.

[101] Jones, R.H. and Ackerson, L.M. (1990) Serial correlation in unequally spaced longitudinal data. *Biometrika* **77**, 721–731.

[102] Jones, R.H. and Boadi-Boateng, F. (1991) Unequally spaced longitudinal data with AR(1) serial correlation. *Biometrics* **47**, 161–175.

[103] Jørgensen, B. (1987) Exponential dispersion models. *Journal of the Royal Statistical Society* **B49**, 127–162.

[104] Jørgensen, B. (1993) *The Theory of Linear Models.* London: Chapman & Hall.

[105] Kalbfleisch, J.D. and Prentice, R.L. (1980) *The Statistical Analysis of Failure Time Data.* New York: John Wiley.

[106] Kalbfleisch, J.G. (1985a) *Probability and Statistical Inference.* Volume 1: *Probability.* New York: Springer-Verlag.

[107] Kalbfleisch, J.G. (1985b) *Probability and Statistical Inference.* Volume 2: *Statistical Inference.* New York: Springer-Verlag.

[108] Kaplan, E.L. and Meier, P. (1958) Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53**, 457–481.

[109] Keene, O.N. (1995) The log transformation is special. *Statistics in Medicine* **14**, 811–819.

[110] Kempton, R.A. and Howes, C.W. (1981) The use of neighbouring plot values in the analysis of variety trials. *Journal of the Royal Statistical Society* **C30**, 59–70.

[111] King, G. (1989) *Unifying Political Methodology. The Likelihood Theory of Statistical Inference.* Cambridge: Cambridge University Press.

[112] Kitagawa, G. (1987) Non-Gaussian state–space modeling of nonstationary time series. *Journal of the American Statistical Association* **82**, 1032–1063.

[113] Klotz, J. (1973) Statistical inference in Bernoulli trials with dependence. *Annals of Statistics* **1**, 373–379.

[114] Lambert, P. (1996) Modelling of non-linear growth curves on series of correlated count data measured at unequally spaced times: a full likelihood based approach. *Biometrics* **52**, 50-55.

[115] Lawless, J.F. (1982) *Statistical Models and Methods for Lifetime Data.* New York: John Wiley.

[116] Lawless, J.F. and Nadeau, C. (1995) Some simple robust methods for the analysis of recurrent events. *Technometrics* 37, 158–168.

[117] Lindsey, J.K. (1971) *Analysis and Comparison of Some Statistical Models.* University of London, Ph.D. thesis.

[118] Lindsey, J.K. (1972) Fitting response surfaces with power transformations. *Journal of the Royal Statistical Society* **C21**, 234–247.

[119] Lindsey, J.K. (1974a) Comparison of probability distributions. *Journal of the Royal Statistical Society* **B36**, 38–47.

[120] Lindsey, J.K. (1974b) Construction and comparison of statistical models. *Journal of the Royal Statistical Society* **B36**, 418–425.

[121] Lindsey, J.K. (1989) *The Analysis of Categorical Data Using GLIM.* Berlin: Springer-Verlag.

[122] Lindsey, J.K. (1992) *The Analysis of Stochastic Processes Using GLIM.* Berlin: Springer-Verlag.

[123] Lindsey, J.K. (1993) *Models for Repeated Measurements.* Oxford: Oxford University Press.

[124] Lindsey, J.K. (1995a) *Introductory Statistics: The Modelling Approach.* Oxford: Oxford University Press.

[125] Lindsey, J.K. (1995b) *Modelling Frequency and Count Data.* Oxford: Oxford University Press.

[126] Lindsey, J.K. (1995c) Fitting parametric counting processes by using log-linear models. *Journal of the Royal Statistical Society* **C44**, 201–212.

[127] Lindsey, J.K. (1996a) Fitting bivariate intensity functions, with an application to modelling delays in reporting acquired immune deficiency syndrome. *Journal of the Royal Statistical Society* **A159**, 125–131.

[128] Lindsey, J.K. (1996b) *Parametric Statistical Inference.* Oxford: Oxford University Press.

[129] Lindsey, J.K., Alderdice, D.F., and Pienaar, L.V. (1970) Analysis of nonlinear models – the nonlinear response surface. *Journal of the Fisheries Research Board of Canada* **27**, 765–791.

[130] Lindsey, J.K. and Jones, B. (1996) A model for cross-over trials evaluating therapeutic preferences. *Statistics in Medicine* **15** 443–447.

[131] Lindsey, J.K. and Lambert, P. (1995) Dynamic generalized linear models and repeated measurements. *Journal of Statistical Planning and Inference* **47**, 129–139.

[132] Lindsey, J.K. and Laurent, C. (1996) Estimating the proportion of epithelial cells affected by exposure to ethylene oxide through micronuclei counts. *Journal of the Royal Statistical Society* **D45**, 223–229.

[133] Lindsey, J.K. and Mersch, G. (1992) Fitting and comparing probability distributions with log linear models. *Computational Statistics and Data Analysis* **13**, 373–384.

[134] McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models.* 2nd ed. London: Chapman & Hall.

[135] McPherson, G. (1989) The scientists' view of statistics — a neglected area. *Journal of the Royal Statistical Society* **A152**, 221–240 (with discussion).

[136] Morgan, B.J.T. (1992) *Analysis of Quantal Response Data.* London: Chapman & Hall.

[137] Nelder, J.A. (1961) The fitting of a generalization of the logistic curve. *Biometrika* **17**, 89–100.

[138] Nelder, J.A. (1962) An alternative form of a generalized logistic equation. *Biometrics* **18**, 614–616.

[139] Nelder, J.A. (1966) Inverse polynomials, a useful group of multi-factor response functions. *Biometrics* **22**, 128–141.

[140] Nelder, J.A. and Wedderburn, R.W.M. (1972) Generalized linear models. *Journal of the Royal Statistical Society* **A135**, 370-384.

[141] Oliver, F.R. (1970) Estimating the exponential growth function by direct least squares.*Journal of the Royal Statistical Society* **C19**, 92–100.

[142] Parzen, E. (1979) Nonparametric statistical data modelling. *Journal of the American Statistical Association* **74**, 105–131.

[143] Pierce, D.A. and Schafer, D.W. (1986) Residuals in generalized linear models. *Journal of the American Statistical Association* **81**, 977–986.

[144] Plackett, R.L. (1965) A class of bivariate distributions. *Journal of the American Statistical Association* **60**, 516–522.

[145] Pollock, K.H., Winterstein, S.R., and Conroy, M.J. (1989) Estimation and analysis of survival distributions for radio-tagged animals. *Biometrics* **45**, 99–109.

[146] Pregibon, D. (1981) Logistic regression diagnostics. *Annals of Statistics* **9**, 705–724.

[147] Priestley, M.B. (1981) *Spectral Analysis and Time Series.* San Diego: Academic Press.

[148] Rasch, G. (1960) *Probabilistic Models for some Intelligence and Attainment Tests.* Copenhagen: Danmarks Paedogogiske Institut.

[149] Ripley, B.D. (1981) *Spatial Statistics.* New York: John Wiley.

[150] Ripley, B.D. (1988) *Statistical Inference for Spatial Processes.* Cambridge: Cambridge University Press.

[151] Ross, G.J.S. (1990) *Nonlinear Estimation.* Berlin: Springer-Verlag.

[152] Royston, P. and Thompson, S.G. (1995) Comparing non-nested regression models. *Biometrics* **51**, 114–127.

[153] Sandland, R.L. and McGilchrist, C.A. (1979) Stochastic growth curve analysis. *Biometrics* **35**, 255–271.

[154] Scallon, C.V. (1985) Fitting autoregressive processes in GLIM. *GLIM Newsletter* **9**, 17–22.

[155] Searle, S.R. (1971) *Linear Models.* New York: John Wiley.

[156] Seber, G.A.F. and Wild, C.J. (1989) *Nonlinear Regression.* New York: John Wiley.

[157] Seeber, G.U.H., Francis, B.J., Hatzinger, R., and Steckel-Berger, G. (1995) *Statistical Modelling.* Berlin: Springer-Verlag.

[158] Selvin, S. (1991) *Statistical Analysis of Epidemiologic Data.* Oxford: Oxford University Press.

[159] Senn, S. (1993) *Cross-over Trials in Clinical Research.* New York: John Wiley.

[160] Snedecor, G.W. and Cochran, W.G. (1967) *Statistical Methods.* Ames: Iowa State University Press.

[161] Sokal, R.R. and Rohlf, F.J. (1969) *Biometry. The Principles and Practice of Statistics in Biological Research.* San Francisco: W.H. Freeman.

[162] Stoyan, D., Kendall, W.S., and Mecke, J. (1987) *Stochastic Geometry and its Applications.* New York: John Wiley.

[163] Strauss, D. (1992) The many faces of logistic regression. *American Statistician* **46**, 321–327.

[164] Stuart, A. (1953) The estimation and comparison of strengths of association in contingency tables. *Biometrika* **40**, 105–110.

[165] Thierens, H., Vral, A., and de Ridder, L. (1991) Biological dosimetry using the micronucleus assay for lymphocytes: interindividual differences in dose response. *Health Physics* **61**, 623–630.

[166] Tjur, T. (1982) A connection between Rasch's item analysis model and a multiplicative Poisson model. *Scandinavian Journal of Statistics* **9**, 23–30.

[167] Tong, H. (1990) *Non-linear Time Series. A Dynamical System Approach.* Oxford: Oxford University Press.

[168] Tweedie, M.C.K. (1947) Functions of a statistical variate with given means, with special reference to Laplacian distributions. *Proceedings of the Cambridge Philosophical Society* **43**, 41–49.

[169] Upton, G.J.G. (1978) *The Analysis of Cross-Tabulated Data.* New York: John Wiley.

[170] Upton, G.J.G. and Fingleton, B. (1985, 1989) *Spatial Data Analysis by Example.* Vol. I. *Point Pattern and Quantitative Data.* Vol. II. *Categorical and Directional Data.* New York: John Wiley.

[171] van der Heijden, P.G.M., Jansen, W., Francis, B., and Seeber, G.U.H. (1992) *Statistical Modelling.* Amsterdam: North-Holland.

[172] Wallach, D. and Goffinet, B. (1987) Mean squared error of prediction in models for studying ecological and agronomic systems. *Biometrics* **43**, 561–573.

[173] Wei, L.J. and Lachin, J.M. (1984) Two-sample asymptotically distribution-free tests for incomplete multivariate observations. *Journal of the American Statistical Association* **79**, 653–661.

[174] West, M. and Harrison, P.J. (1989) *Bayesian Forecasting and Dynamic Models.* Berlin: Springer-Verlag.

[175] West, M., Harrison, P.J., and Migon, H.S. (1985) Dynamic generalized linear models and Bayesian forecasting. *Journal of the American Statistical Association* **80**, 73–97.

[176] Wilkinson, G.N. and Rogers, C.E. (1973) Symbolic description of factorial models for analysis of variance. *Journal of the Royal Statistical Society* **C22**, 392–399.

[177] Williams, D.A. (1987) Generalized linear model diagnostics using the deviance and single case deletions. *Journal of the Royal Statistical Society* **C36**, 181–191.

[178] Williams, E.J. (1959) *Regression Analysis.* New York: John Wiley.

[179] Zippin, C. and Armitage, P. (1966) Use of concomitant variables and incomplete survival information in the estimation of an exponential survival parameter. *Biometrics* **22**, 665–672.

Index

Springer Texts in Statistics *(continued from page ii)*

Nguyen and Rogers: Fundamentals of Mathematical Statistics: Volume I: Probability for Statistics
Nguyen and Rogers: Fundamentals of Mathematical Statistics: Volume II: Statistical Inference
Noether: Introduction to Statistics: The Nonparametric Way
Peters: Counting for Something: Statistical Principles and Personalities
Pfeiffer: Probability for Applications
Pitman: Probability
Rawlings, Pantula and Dickey: Applied Regression Analysis
Robert: The Bayesian Choice: A Decision-Theoretic Motivation
Robert and Casella: Monte Carlo Statistical Methods
Santner and Duffy: The Statistical Analysis of Discrete Data
Saville and Wood: Statistical Methods: The Geometric Approach
Sen and Srivastava: Regression Analysis: Theory, Methods, and Applications
Shao: Mathematical Statistics
Shumway and Stoffer: Time Series Analysis and Its Applications
Terrell: Mathematical Statistics: A Unified Introduction
Whittle: Probability via Expectation, Third Edition
Zacks: Introduction to Reliability Analysis: Probability Models and Statistical Methods