THE ANALYSIS OF CATEGORICAL LONGITUDINAL DATA

Thesis submitted in accordance with
the requirements of the University of
Liverpool for the degree of
Doctor in Philosophy

by

SIMON CHARLES FEAR

April 1998

To Thomas

Acknowledgements

Thanks to Maddy for suggesting I study mathematics, and Melanie and my wife, Gill, for tolerating the consequences. For steering me from a number of options into a career in statistics I extend my gratitude to my undergraduate tutors at the University of Exeter, Drs A.G. Munford and T.C. Bailey.

I would like to thank especially my thesis supervisors at the University of Liverpool, Professor P.J. Brown and Dr D.Y. Downham, for their guidance and encouragement, and I thank Professor R.J. Bhansali for suggesting many major improvements to earlier versions of the manuscript.

This work was supported by a grant from the EPSRC.

Contents

Abstract

A full likelihood-based approach to modelling categorical longitudinal data is studied, in which all marginal expectations and interactions are parametrized. The dependence-ratio model of Ekholm *et al.* (*Biometrika*, **82,** 847–854, 1995) is extended to polytomous data and a more general parametrization. I also introduce and criticise a modification of the model of Fitzmaurice and Laird (*Biometrika*, **80,** 141–151, 1993). A method is given for fitting the original and modified Fitzmaurice and Laird models to unbalanced data, without resort to computationally expensive imputation.

When fitting certain fully marginal models, we need to obtain cell probabilities corresponding to specified odds ratios. It is shown that we must use numerical rather than analytical techniques. An algorithm is presented that is shown by simulation to be up to two orders of magnitude faster than the Newton–Raphson iteration proposed by Glonek and McCullagh (*JRSS*, **B57,** 533–546, 1995). This considerably extends the practical limit on the number of observations one can model. In certain circumstances a solution is obtained that cannot be found using Newton–Raphson iteration. The opposite also occurs; neither algorithm always converges.

Markov chain models are also studied, and a fitting method is given that lifts certain restrictions of previous literature presentations. Such models are developed in great generality and are considered for multivariate processes also; here we model probabilities conditional not only on history but also on other simultaneously observed outcome variables, which may be of different data types.

Such 'timepoint-wise factorized' models offer an intuitively appealing approach to the problem of modelling data given that some subjects have dropped out of a trial, which is almost inevitable in longitudinal studies. It is shown that one cannot use maximum likelihood techniques to assess the degree of bias due to informative dropout. However, if this need not be considered, a very flexible model, able to allow for different types of dropout simultaneously, is proposed and examined.

Chapter 1

Introduction

1.1 Longitudinal data analysis

Consider an outcome variable, Y_t, measured on each of several occasions ('timepoints') t, for several units or individuals. The vector of outcomes for the uth unit is denoted

$$\mathbf{Y}_u = (Y_{u1}, Y_{u2}, \ldots Y_{uT})',$$

where there are T timepoints. The methods are developed for constant T, though in Chapters 5 and 6 non-constant T values are considered.

It will be assumed throughout that there is a set of explanatory variables, $\mathbf{X}_{ut}$, for each timepoint, $t = 1, 2, \ldots, T$, and that such variables may vary with time.

In the most general setting, such as occurs with routine observational patient records in hospitals, the observation times are not the same for all subjects. Such cases are not discussed here in any detail. Thus, attention is restricted to discrete and fixed time models that can be described theoretically by a multivariate distribution rather than as realizations of some underlying stochastic process.

Longitudinal data are a subset of the more general case of dependent data, where the classical assumption of independence does not hold. Even when the data are not repeated measures, there may be a natural ordering to the subscripts: in human sibling studies, excluding multiple births, children would be naturally ordered by age.

In animal litter studies such distinction may be impossible or meaningless. The general case lies outside the scope of this thesis, although wherever possible I will indicate when the longitudinal models described here may be used more widely.

The principle method of analysis throughout this thesis will be likelihood-based multivariate regression: the analyses use the model

$$\mathbf{g}(\mathrm{E}[\mathbf{Y}]) = \boldsymbol{\eta}(\mathbf{Y}, \mathbf{X}, \boldsymbol{\gamma}), \tag{1.1}$$

where $\mathbf{g}$ and $\boldsymbol{\eta}$ are specified vector-valued functions, the explanatory variables $\mathbf{X}$ may be manifest or latent, and $\boldsymbol{\gamma}$ is a vector of parameters, some of which may be nuisance parameters. In general $\mathbf{g}$ is a function of the expected value (but sometimes also higher moments) of $\mathbf{Y}$, as in the classical generalized linear model (GLM), and $\boldsymbol{\eta}$ may incorporate a description of the assumed error. The seemingly unusual appearance of $\mathbf{Y}$ on both sides of the relationship is clarified below.

In common with the GLM, the function(s) $\mathbf{g}()$ is called the *link function*; almost always here $\mathbf{g}()$ denotes a vector of such functions. Similarly $\boldsymbol{\eta}()$ is a vector of linear functions known as the *linear predictor(s)*.

For longitudinal and/or dependent data analysis, models are often split into three or occasionally four main types (see, for example, Diggle *et al.*, 1994, and Neuhaus, 1992). These are the marginal, mixed (or random-effects), and transitional approaches, with the optional but rare inclusion of response-conditional models. This terminology is described before suggesting an alternative classification.

In the literature, $\mathbf{g}$ is usually of the form

$$\mathbf{g}(\mathrm{E}[\mathbf{Y}]) = (g_1(\mathrm{E}[Y_1]), g_2(\mathrm{E}[Y_2]), \ldots g_T(\mathrm{E}[Y_T]), g_\times(\mathrm{E}[Y_\times]))'. \tag{1.2}$$

The overall model is a collection of individual models for expected values at each timepoint, plus (possibly) a dependency model, here denoted $g_\times(\mathrm{E}[Y_\times])$. A *marginal model* is one for which at each timepoint

$$g_t(\mathrm{E}[Y_{ut}]) = \eta_t(\mathbf{X}_{ut}, \boldsymbol{\gamma}), \tag{1.3}$$

together with a dependency model to correct for biased estimates of standard errors, where g_t is now explicitly a function of the marginal (univariate) expectation of Y_{ut}. Such models are described by McCullagh and Nelder (1989) and were first implemented by Liang and Zeger (1986) and Zeger and Liang (1986). The description *marginal* is traditionally reserved for the model for $\mathrm{E}[Y_t]$; the dependency model may take any of several forms, as discussed briefly below and in detail in Chapter 2.

A marginal model is applicable when timepoints are fixed, discrete, and the same for all units. However, in most data sets there are missing observations for some timepoints for some units. An observation y_{u2} denotes an observation on the uth unit at some fixed time 2, and might be the first or second actual observation on that unit. Certain predetermined timepoints might be omitted from the final analysis altogether, if there are many missing data. Missing data are considered in detail in Chapters 5 and 6.

The marginal model is normally contrasted most sharply with the *random-effects* model, for which again assuming (1.2),

$$g_t(\mathrm{E}[Y_{ut}]) = \eta_t(\mathbf{X}_{ut}, \boldsymbol{\gamma}_u). \tag{1.4}$$

Note the presence of subscript u on $\boldsymbol{\gamma}$ here and the implicit assumption of latent variables among the explanatory $\mathbf{X}_{ut}$, the so-called random effects. Alternatively we may consider those elements of $\boldsymbol{\gamma}_u$ that vary with each unit as random variables and then $\boldsymbol{X}_{ut}$ can be represented as an observed/fixed design matrix; the effect is just the same. Such models are inherently overparametrized, so that there is need for a plausible set of restrictions. The strategy often used in the literature is that of setting

$$\boldsymbol{\gamma}_u = \boldsymbol{\beta} + \mathbf{b}_u$$

and then imposing some distributional assumptions on the "random effects" $\mathbf{b}_u$ to gain identifiability. (The above description of the approach is due to Diggle *et al.*, 1994, but embodies a vast literature including the seminal work of Laird and Ware,

1982.) The assumed distribution is often Normal.

The random effects in this model may be estimated explicitly, or more often the random effects are integrated from the likelihood or determined by sufficient statistics (for example, Conaway, 1992), to give a model more naturally written as explicitly conditional on the random effects:

$$g_t(\mathrm{E}[Y_{ut} \mid \mathbf{b}_u]) = \eta_t(\mathbf{X}_{ut}, \boldsymbol{\gamma}, \mathbf{b}_u). \tag{1.5}$$

Such models are sometimes called *conditional.* In emphasising the distinction between conditional and marginal models in the interpretation of the parameters $\boldsymbol{\gamma}$, Zeger *et al.* (1988) used the terms "population averaged" and "subject specific" for marginal and conditional estimates, respectively. Subject-specific effects might not be random, although this is the common usage.

In a *transitional* model, rather than model functions of the univariate expectations, the transitional probabilities from outcome y_{ut} at time t to outcome $y_{u(t+1)}$ at time $t+1$ are parametrized. This thesis is almost exclusively concerned with polytomous data, for which such a model of transitional probabilities is directly equivalent to a model for the conditional expectations: in Chapter 4, the model

$$g_t(\mathrm{E}[Y_{ut} \mid Y_{u1}, \ldots Y_{u(t-1)}]) = \eta_t(Y_{u1}, \ldots Y_{u(t-1)}, \mathbf{X}_{u1} \ldots \mathbf{X}_{ut}, \boldsymbol{\gamma}), \tag{1.6}$$

is considered. It is common to assume that the links and linear predictors are sufficiently well chosen as to render the need for an interaction model $g_\times(\mathrm{E}[Y_\times])$ redundant. Finally, the response-conditional model of Rosner (1992b) is similar to the above transitional formulation, except that dependency on all the other observations on each subject, rather than only on previous observations, is modelled:

$$g_t(\mathrm{E}[Y_{ut} \mid \{Y_{us} : s \neq t\}]) = \eta_t(\{Y_{us} : s \neq t\}; \{X_{ur} \; \forall r\}; \boldsymbol{\gamma}). \tag{1.7}$$

While this model may be of great importance for general dependent data structures, it clearly makes no intuitive sense for longitudinal data: dependency on future values

is not relevant for most longitudinal models. This model is accordingly not discussed further.

The description *conditional* has been used above in two distinct senses: conditional on unobserved random effects, and conditional on observed values. There is no particular reason why we should not condition on *both*, to give a transitional model with random effects. Thus, the hierarchy of model choice advocated here is that one first chooses between response-conditional and univariate-expectation models (with allowance for dependence), and then decides whether to include random effects. From this point of view, the standard random-effects model is more marginal than conditional, in that univariate (and in this sense marginal) expectations are modelled; the essential difference between this and the standard marginal model is that effectively certain of the explanatory variables, $\mathbf{b}_u$, are unobserved in the 'conditional' model.

In summary, for longitudinal data there are two main types of model — marginal and response-conditional — in either of which there may be latent variables (subject-specific effects). This classification was proposed by Ware *et al.* (1988).

1.1.1 Special considerations for categorical data

When the outcome $\mathbf{Y}$ is a discrete random variable, certain of the inherent simplifications that arise with Normal data no longer hold.

Importantly, with any non-Gaussian data the marginal estimates of mean and covariance parameters are closely linked: the fundamental simplification of classical ANOVA — the independence of the estimates of means and standard deviations — fails to hold. Linear models of continuous data with different covariance structures will give similar results with regard to estimates of means (indeed, asymptotically identical), but for discrete data, because of the lack of independence of estimators, different models can lead to quite different (indeed asymptotically non-equivalent) estimators. With discrete data the fundamental difficulties are in the interpretation and in the estimated parameter values, which are closely tied to the assumptions about the nature of dependency.

Similarly, for linear random-effects models with Gaussian errors, estimates of the

fixed-effect coefficients are robust to assumptions about the distribution of the random effects; for discrete data this no longer holds. This observation is reminiscent of the frequently quoted interpretation of random effects as accounting for otherwise unexplained covariance; since for discrete data the assumed covariance structure affects the estimates of marginal, fixed effects, so random-effects assumptions correspond to covariance assumptions and thus must affect fixed effects.

A related issue adds enormous computational, and appreciable interpretational, difficulties. For T-variate Gaussian data the entire dependency structure is encapsulated in a $T \times T$ dispersion (covariance) matrix, but for T-variate binary data the equivalent matrix is of size $(2^T - 1) \times (2^T - 1)$, which is enormous for even moderately large T. For polytomous data a complete description of the covariance structure is cumbersome even for reasonably small T. The computational burden of a general treatment often outweighs the possible advantages it might bestow, so simplifications are of great importance and frequent consideration.

Another difference between Gaussian and polytomous models is that for Gaussian data the link function to the marginal expectation is often the identity function, but for binary data this is often, at least for the univariate/marginal links, the logit (or probit, or complementary log-log) link to the probability π_t of success at time t. For polytomous data we use the "baseline" extension,

$$\log\left(\frac{\pi_t}{\pi_0}\right)$$

or the adjacent-category link to

$$\log\left(\frac{\pi_t}{\pi_{t-1}}\right).$$

This latter form, or the more familiar cumulative logit link or even a link to the mean score is fully discussed in Agresti (1990) and is expanded where applicable below. Transitional versions of these models are created simply by the presence of observed historical values within the linear predictors for these same links. Importantly no further new development is needed for the univariate marginal models that are the backbone of models for longitudinal data; everything known about the properties of

such links for univariate data holds here too, with the possible introduction of bias from the dependency assumptions.

1.1.2 Modelling dependency

A key distinction between longitudinal and cross-sectional data is the almost certain lack of independence between the measurements on each individual or unit. The three- (or four-) way classification of models in the literature is perhaps most clearly justified if we consider the methodology of dealing with intra-subject dependencies rather than the univariate marginal, and random-effect-conditional marginal model that is often the focus. Three distinct strategies emerge. For each of these strategies, subjects are assumed to be mutually independent; we are only concerned with within-subject dependence.

The easiest conceptual approach is that now linked to random-effects models. This is simply to assume that observations on a subject are independent given the random effects: more formally,

$$\mathrm{cov}(Y_{ui}, Y_{uj} \mid \mathbf{b}_u) = 0 \quad \forall i,\ j.$$

Because polytomous data are of primary concern, the terms 'independent' and 'uncorrelated' are frequently used synonymously; formally the dispersion matrix is assumed to be diagonal. This assumption, often known as *local independence*, is adopted by most advocates of random-effects modelling. A notable attractive exception (Conaway, 1989; Conaway, 1990) does not accept that local independence is a reasonable assumption in the longitudinal data setting. Conaway's solution is to incorporate previous responses as covariates and so to advocate a model that is random-effects with conditioning on history added: I would describe this model as transitional with additional random effects. It is of course not *necessary* to assume local independence once random effects have been included; one could parametrize covariance in the manner described below, and in Chapter 2, *and* have random effects.

A second way to deal with within-subject dependence is to take a response-conditional or transitional approach, since this factorizes the joint likelihood into a set of univariate

likelihoods that are functionally independent, at least when the parameters are not shared across timepoints. Thus, there is no need for a specific dependency model, one of the many attractions of such conditioning, provided of course that such an analysis addresses the underlying question. Discussion of these models, including what happens when parameters *are* shared between timepoints, is given in Chapter 4.

The third main approach is to look towards a more completely specified model for the joint likelihood and to specify the interaction terms linked to $g_\times(Y_\times)$, equation (1.2), estimating these parameters alongside those for the marginal means. This approach, arguably the natural extension of the GLM to dependent data, is explored in Chapter 2. There is here no study of the addition of random effects within the marginal and/or interaction models. In this case one would obtain perhaps the most natural extension of the generalized linear mixed model (GLMM), but this is a topic for future research.

In pre-empting my discussion of such models in Section 1.2 and Chapter 2 below, the approaches of Stram *et al.* (1988) and the GEE models of Liang and Zeger (see later for full references) are considered here. In both of these approaches dependency is regarded as strictly nuisance, and point estimates of marginal effects are obtained from models assuming independence across timepoints; only when estimating the errors in such estimates is the dependency utilized. Even crude (marginal) estimates which ignore correlation are nearly optimal, but dependency is important when assessing the precision of estimates (Zeger, 1988). Similarly Fitzmaurice and Laird (1993) stress that univariate marginal estimates are robust to dependency misspecification. If only the estimates of effects on such univariate distributions are needed, the independence assumption may offer a practical compromise between computational simplicity and accuracy, but much more may be needed.

Two of the above three methods for dealing with the dependence structure do not model dependency directly. This might seem to be something of an anomaly for a repeated-measures design. Louis (1988) writes

> "If covariance parameters are not of interest, then the burden of proof will be to show why a longitudinal study is necessary."

But in the same article, Louis (1988) writes that longitudinal studies, when used to address cross-sectional questions, provide

> "increased precision, more accurate measurement of covariates, and protection from certain types of selection and dropout bias"

Fully marginal models are introduced in Chapter 2 to address the specific modelling of covariance structure, at least for data where T is sufficiently small for the analysis to be computationally practicable. The transitional, or Markov model (Chapter 4) *does* have a type of parametrized dependence, although this is expressed as a univariate effect that is interpretable for descriptions and predictions, rather than, as for the marginal approach, for descriptions of the distribution. The classical random-effects approach with its assumption of local independence sidesteps the issue entirely and is open to criticism if intra-subject dependence is of interest and cannot be brushed aside as a set of nuisance parameters. Indeed, if longitudinal data have been collected for the purpose of the analysis of change, the classical random-effects model is quite inappropriate. Such considerations are important when choosing a modelling strategy, and are considered in the next section.

1.1.3 Choice of longitudinal model

Each type of model described above has its advocates. Thus a newcomer to longitudinal data analysis might reasonably ask "which type of model is best for my particular problem?". Obviously the reasons for collecting the data and the questions being addressed influence the model choice. Following the spirit of generalization of many particular models into the unifying framework of the generalized linear model, one might think that there would be some "grand unified" longitudinal model that could be fitted to provide answers to any relevant question posed. This is not the case however:

> "The megalomanical strategy of fitting a grand unified model, supposedly capable of answering any conceivable question that might be posed, is, in our view, dangerous, unnecessary and counterproductive. It violates that

basic principle of applied statistics, the avoidance of unnecessary modelling."

(Melinda Drum and Peter McCullagh, in a comment on the review of Fitzmaurice *et al.*, 1993).

Consider a study (Neuhaus, 1992) of the sexual behaviour of a San Francisco based cohort where the main outcome variable was whether or not a person had engaged in unsafe sex during the past month, the question being repeated once per year for five years. Neuhaus gives three problems, each calling for an entirely different modelling strategy, as follows:

- Does the prevalence of unsafe behaviour depend on age? This type of question might be studied by a cross-sectional study, which indicates the method of choice for a cohort study: we require a marginal model. Indeed a first crude approximation might be to base one's estimates for prevalence on the data obtained in the first wave of the study only, that is to say, to ignore four fifths of the data and perform only ordinary cross-sectional analysis. Such an estimate lacks power because it fails to use all the available information, and is subject to bias when compared to the estimate for time-1 prevalence obtained by extracting the time-1 marginal from a model of the joint distribution. A non-technical explanation for the bias in the simplistic approach is that a single cross section cannot allow for an age-cohort effect, equivalent to prevalence being related to cohort rather than to age per se; a particular generation may continue to practise unsafe sex while a younger generation, say, might not. Only in the longitudinal design and analysis is it possible to disentangle the effects of age and cohort.

- Does the probability of engaging in unsafe sex change after an individual receives HIV antibody test results? Neuhaus identifies this as a subject-specific problem, for which he advocates a random-effects approach. I would favour a transitional approach here as the question more directly relates to change with respect to previous behaviour. However, if the question is to be interpreted more precisely as "is the behaviour of post-test individuals different on average from that of

pre-test" then a marginal model is plausible. Whether or not random effects would be justified in this marginal model depends on whether one wishes to describe the relationship in the population at large or per individual.

- Does the probability of engaging in unsafe sex depend on previous sexual behaviour? This question is instantly framed in transitional terms; the concern is the effect of previous outcome on present behaviour. Although less intuitive, this question can be addressed by a fully marginal model, with attention focussed on suitable marginal interaction parameters.

It would seem clear that marginal models — or at least, *fully* marginal models — can answer almost all of the likely questions. Moreover, marginal models are not longitudinal-data specific: they can be used to analyse the more general case of multivariate dependent data. However, marginal models in the usual meaning do not address the problem of subject-specific interpretations, but marginal models with some covariates unobserved can be considered to be standard random-effects models. An advocate of random-effects models might argue that these, rather than the marginal 'simplification', were the most general model.

If a random-effects model is essentially a type of marginal model, then the same strengths must apply. But random-effects should not be preferred to classical (population averaged) marginal models merely because they are generally easier to fit, under the assumption of local independence, and perhaps easier to interpret in the absence of a plethora of dependence parameters. The decision to make before analysis is whether population-averaged coefficient estimates are required, or whether individual (subject-specific) effects are more important. In this latter case, in the longitudinal setting, the transitional modelling approach, which is subject-specific in its incorporation of previous values but otherwise population-averaged, should be a strong candidate for analysis and interpretation.

While advocating the use of marginal, including subject-specific marginal, models, I do not advocate the artificiality of answering questions such as the third in Neuhaus (1992) by the use of marginal models. Here interpretation in a transitional framework

is the most natural, and that decision should be taken before computational and other practical details are considered.

A further question, not posed by Neuhaus (1992) but common in the longitudinal setting, is that of the prediction of a future value given past history. A marginal model cannot be used directly for this purpose; such a model might be used to construct the modelled joint density, then to re-interpret this in a transitional framework. This approach is not sensible if prediction is the only or primary object of the analysis.

Despite the above considerations when choosing a model, three main strands in the choice of model in practice can be identified. These may be broadly categorized as *good* — the study is designed to answer a clearly formulated problem with the model predetermined; *compromise* — given what is generally a substantial amount of collected data, such as complete patient records, decide what the question is (or should have been) and analyse accordingly; and *bad* — fit all the models for which software is available, and interpret what they tell you. Many consultancy episodes tend towards the second and even the third of these categories.

Optimality, in the sense that the ideal model is to be fitted, and practicality, in the sense that many models, especially for polytomous data, are impossible to fit with current computing power, must be balanced. One might, therefore, be forced to assume simplified forms of dependency structure, such as homogeneity or even independence, while strongly believing them to be unlikely. Matrix size, and the inevitable sparseness of data in high-dimensional contingency tables, may force such compromises.

The choice between subject-specific and population-averaged models is perhaps the thorniest issue because either might be justified and the choice between the two is often down to personal choice. In this thesis population-averaged marginal models are considered in detail but the non-random-effects transitional model that is also considered here is partly subject-specific in any case. Arguably, the local independence assumption is less tenable for binary, and perhaps all polytomous data, than for Gaussian data; however, it is not *necessary* to assume such independence given random effects. Time limitations have prevented me from developing the multivariate GLMM

sans conditional independence here.

Some argue that only subject-specific effects have any real meaning, but Ware *et al.* (1988, p. 104) argue that very often marginal and subject-conditional estimates are the same anyway, within estimation errors. For example, suppose that we wish to answer the question of whether or not a given drug has an effect, as assessed by analysing a crossover trial (e.g. Section 2.3.3). Inference is based on whether or not the parameter relating to drug administration, β, say, is significant. In general,

$$|\beta_{\mathrm{PA}}| < |\beta_{\mathrm{SS}}| \tag{1.8}$$

where β_{PA} is the estimate from a population-averaged model, and β_{SS} is that from a random-effects model (Neuhaus *et al.*, 1991). Thus if we fit a population-averaged model and find that β_{PA} is significant, it is almost sure that β_{SS} will be likewise; unless we are actually interested in the *size* of the effect, we are almost sure to reach the same conclusion from either model. Similarly, even if the size of such a parameter is of interest, a population-averaged model can be fitted because it gives a lower bound for the subject-specific equivalent.

1.2 Some simpler models in the literature

1.2.1 The Koch model

The key reference to this model is Koch *et al.* (1977), which includes an extensive references list representing work on the ideas crystallized by Grizzle *et al.* (1969). Importantly, in the former paper a development of a general methodology for the analysis of multivariate categorical data is attempted although the title claimed only to tackle repeated measurements.

Data are presented as counts of the number of occurrences of each of the possible response profiles. Since the data are categorical, there are a finite number of profiles. Profile i has estimated probability $\pi_i = n_i/n$, where n_i is the number of subjects with response profile i, and $n = \sum n_i$. The modelling focusses on suitably chosen contrasts

of the π_i: that is, on some vector $\mathbf{F} = A\boldsymbol{\pi}$. Suitable choices of A readily yield models for cell probabilities, marginal probabilities or marginal expectations. A more general form than that given in the Koch *et al.* (1977) paper is readily derived: namely

$$\mathbf{F}(\boldsymbol{\pi}) = A_2 f(A_1 \boldsymbol{\pi})$$

where the A_i are linear operators and f is a possibly nonlinear function, such as log or exponential. The appendix to the original paper uses different notation to express this, and more complex, models. The above nomenclature allows simple expressions for the standard logistic transformation, in terms of differences specified in A_2 of logs of suitable summations in $A_1\boldsymbol{\pi}$ of cell probabilities.

In this framework, hypotheses of interest are framed as $C\mathbf{F} = 0$ for a suitable contrast matrix C, and the test statistic is

$$(C\mathbf{F})'(CV_FC)^{-1}(C\mathbf{F})$$

which is distributed as a χ^2 variate on degrees of freedom equal to the number of rows of C under the null hypothesis. The matrix V_F, the dispersion matrix for $\mathbf{F}$, is

$$V_F = \begin{pmatrix} A & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & A \end{pmatrix} (\mathrm{diag}(\hat{\pi}_i) - \hat{\boldsymbol{\pi}}\hat{\boldsymbol{\pi}}') \begin{pmatrix} A & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & A \end{pmatrix}'.$$

The Koch model is inherently marginal in the sense developed above, and cannot be adapted to subject specificity because of the nature of the model and because at no stage is provision made for continuous covariates, as there is with all the later marginal models. Indeed the Koch model stands in the same relationship with the general marginal model as ANOVA does to regression (Zeger, 1988). This latter shortcoming has led to the recent developments in marginal modelling described throughout this thesis.

Another disadvantage of the Koch model is that modelling each profile could be equiv-

alent to fitting a model to the individual cells of a very large multidimensional contingency table; usually such tables are sparse, which raises questions about the validity of maximum likelihood asymptotic theory. In a general review paper Landis *et al.* (1988) proposed a Mantel–Haenszel strategy to overcome the problem, and discussed strategies for dealing with missing data.

An alternative proposal to increase the efficiency of estimation in this setting (Zeger, 1988) is to use the fitted probabilities from a first-time Koch-model fit to compute a *fitted* (rather than moment-estimated) multinomial covariance matrix for each cell, and then to recalculate the model coefficients weighting by the fitted, rather than the observed, covariance matrix. I know of no implementation of this technique.

1.2.2 The beta-binomial model

Early developments in this field make the application of beta-binomial mixed distributions to longitudinal data contemporaneous with that of Koch and co-workers. Several papers (Griffiths, 1973; Williams, 1975) predate the more commonly cited work of Crowder (1978), and differ in application; the first use of the beta-binomial distribution is attributed to Skellam (1948).

The basic idea, when applied to clustered and/or longitudinal binary data, is to assume that each repeated measurement on subject u has the same subject-constant probability π_u of success, but that π_u varies across subjects according to a $\beta(a, b)$ distribution, and is modelled by, for example,

$$\text{logit } \pi_u = \alpha + \boldsymbol{\beta}' \mathbf{x}_u, \tag{1.9}$$

Classically the model is expressed in terms of π_u following a beta distribution with common mean μ and variance $\delta\mu(1-\mu)$: the original use of the beta-binomial distribution was to allow for under/overdispersion in binomial observations. The covariates $\mathbf{x}_u$ in (1.9) are constant for each subject. For polytomous data, the model extends to multinomial–Dirichlet (Crowder, 1978).

This model has the advantage over the Koch model of allowing for continuous covari-

ates, but it still cannot allow for time-varying covariates $\mathbf{x}_{ut}$. Moreover, the model imposed by the beta-binomial assumption, at least in its simplest original form, restricts pairs of responses on a subject to having equal correlation. Steps were taken in this direction (Prentice, 1986), but attention has shifted towards GEE, and more general random-effects, models. A series of papers by Rosner (1984,1989,1992a,1992c), propose response-conditional models with beta-binomial assumptions to account for random effects, rather than beta-binomial models.

The beta-binomial can be thought of as ANOVA for proportions, representing the subject-constants π_u (Crowder, 1978). This is easily seen if subjects (ANOVA 'groups') in the same 'supergroup' (treatment group) have the same intercept, rather than a separate parameter α_u for each individual, although this restriction is not strictly necessary. The conceptual simplicity of this interpretation leaves it as a model of choice if the application is sufficiently straightforward. Consideration of the ANOVA aspect of the model may have led Crowder to the proposition of binary proportions as Normal variates (Crowder, 1985), an idea that continues to resurface (e.g. Rochon, 1996).

Whereas the Koch model is inherently marginal in nature, the beta-binomial model is inherently subject-specific, though always estimable provided the number of observations is two or more; ordinary random effects need not be considered. Another way to consider the idea of (1.9) is as ordinary, univariate regression on a summary measure of subject response. Which leads us, entirely out of historical sequence, to two-stage models.

1.2.3 Two-stage models — derived variables

The idea of derived variables — that is, a scalar summary of the vector outcome — was well known to Wishart in 1938 (Diggle *et al.*, 1994). Attempts to develop a rigorous theory of such methods are not repeated. The inadequacy of ANOVA on a derived variable to deal with problems involving longitudinal data has been discussed by Diggle *et al.* (1994) and is not pursued here. Instead I follow Crowder and Hand (1990), whereby a natural link between the process of two-stage modelling and that of introducing formal random effects is exploited.

To fit the two-stage model frequently cited as Korn and Whittemore (1979), we fit in the first stage a model such as

$$g(\mathrm{E}[Y_{ut}]) = \alpha_u + \beta_u t \tag{1.10}$$

for each subject u. Hopefully, the outcome vector will be large enough for this to be tenable, and we can find an interpretable model that is linear in the predictor. If β_u is not significant for 'most' of the subjects, then we might be justified in taking α_u as a derived variable for univariate analysis: the mean response for each subject being taken as a derived variable. More generally, and especially when there are time varying covariates, we would want not a univariate second-stage analysis but a further by-subject analysis

$$g(\mathrm{E}[Y_{ut}]) = \alpha_u + \boldsymbol{\gamma}_u' \mathbf{x}_{ut}. \tag{1.11}$$

Thus, we could in the first place have fitted

$$g(y_{ut}) = \alpha_u + \beta_u t + \boldsymbol{\gamma}_u' \mathbf{x}_{ut} \tag{1.12}$$

and then tested if the model could be reduced to (1.11) by standard variable selection techniques. Similarly, if in the standard first stage the α_u were not 'sufficiently' different from each other, one would again look to covariate effects additional to the slopes, β_u, and return to model (1.12) with fixed α.

Equation (1.12) is a subject-specific model that is directly estimable without the addition of random effects provided there are at least three time points and at least as many observations as parameters. Suppose that the model does not fit well enough and it is proposed to allow for time-varying intercepts and slopes:

$$g(\mathrm{E}[Y_{ut}]) = \alpha_{ut} + \boldsymbol{\gamma}_{ut}' \mathbf{x}_{ut} \tag{1.13}$$

where the time-varying intercepts α_{ut} generalize the trend-based 'intercepts' $\alpha_u + \beta_u t$ of (1.10) to allow for nonlinear time effects. This model now has more parameters

than observations and can only be fitted by imposing artificial constraints on the parameters, which is commonly done by assuming that the parameters come from some pre-assigned distribution. This is the standard random-effects formulation. In this light, equation (1.12) is essentially nothing other than an estimable random effects model without artificial constraints (or fixed effects). This approach offers further insight into the philosophical debate on the precise interpretation of random versus fixed effects. A fixed effect can be considered a model simplification of an assumed underlying subject-specific effect. An effect that is more or less the same for all subjects must be assumed to be more or less the same for the population but is still, nevertheless, a subject-specific effect. This approach contrasts with one often advocated: that random effects represent a set of unmeasured or unmeasurable population-averaged covariates.

The first stage of the Korn and Whittemore procedure is to fit equation (1.12). In the second stage the covariate effect parameters γ_u are averaged over subjects. If there are different numbers of observations between subjects, or simple forms of missing data, the outcome vectors might be weighted. The averaging process assumes that the individual effects, γ_u, and the variances of these, say Σ_u, are distributed multivariate Normally with mean γ and variance Σ. If the variances, which are essentially nuisance parameters, are assumed known, then γ is simply a weighted sum of the individual fitted γ_u. This is certainly plausible if the $\hat{\gamma}_u$ are obtained by maximum likelihood, and at least asymptotically are Normally distributed.

A different second stage could be advocated if from a preliminary analysis of (1.10) β_u can be taken as zero or constant across units. Then α_{ut}, the 'random' intercept, need not be estimated by random-effects methods, but can be simply taken as the value obtained in stage one. Then when fitting (1.12) or its fixed (or mixed) effect counterpart, α_u can be handled as an offset (to use GLIM terminology), with a potentially large reduction in the number of computations.

Although I have shown that contemplating derived-variable models can lead conceptually to random-effects formulations, simpler analyses in terms of the derived variables can be carried out. It is possible to take either the intercepts, α_u, or the slopes, β_u, as

the derived variable for the univariate analysis. Another common choice, at least for continuous measures, is "area under the curve" but this is less attractive for discrete, especially binary, data. Amongst the disadvantages of considering slopes as derived variables is that no distinction is preserved between subjects for whom the change is equal but the baseline is different: for example the distinction between low and high-risk patients is lost, in the lung function example of Buist and Vollmer (1988). Slope may be a poor summary with values far from the within-subject trend being of intrinsic interest. Potentially enormous amounts of information are lost in the summary process. Also, in the simple two-stage procedure, data from the goodness-of-fit of the summary measure are not carried forward to the second-stage analysis. Diem and Liukkonen (1988) proposed an alternative method in which rather than model

$$\beta_u = X\gamma + \epsilon_u$$

directly, where ϵ_u is a Normal$(0,\sigma_u^2)$ error, let δ_u represent the error in the estimation of the observed slopes, b_u,

$$b_u = \beta_u + \delta_u$$

and then estimate γ in

$$b_u = X\gamma + \epsilon_u + \delta_u,$$

and so the heterogeneity of variance of slopes is accommodated. For their data, this model fitted as well as a random-effects model. Although this is not always the case, the errors δ_u more or less represent random effects in the standard formulation and the method of estimation differs only slightly from the usual method.

1.2.4 The two-stage procedure of Stram *et al.*

Stram *et al.* (1988) take perhaps the simplest approach possible to multivariate dependent data. Separate marginal models are fitted for each time point, by temporarily assuming full independence, and then dependence is accommodated when considering overall group effects over time.

An obvious advantage of this approach is the simplicity with which the univariate models may be modelled using standard methods for fitting and calculating the variance matrix of exponential-family GLMs or standard quasi-likelihood extensions to this theory. Moreover, one can fit to unbalanced data, and to data when there are missing observations, provided that the data are not missing informatively (defined in Chapter 6). A strong disadvantage, however, is that the efficiency of parameter estimates decreases with increasing departure from independence. For this reason, such models are perhaps best restricted to where neither predictions nor marginal estimates are required per se, but rather one is concerned with whether or not there is a difference in mean score between groups or more generally some covariate effect.

Stram *et al.* (1988) consider ordered categorical data and claim that by 1988 (or 1985, when the paper was first submitted) there were already sufficient flexible and general methods for the analysis of continuous repeated measurements. However in the same year Wei and Stram (1988) showed the generality of the method for all types of data.

Changed to conform with the nomenclature here, the parameters γ of the linear predictors for the univariate marginal expectations are modelled through link functions

$$g(\mu_t) = X_t\gamma_t, \quad t = 1, 2, \ldots T$$

by solving the full- or quasi-likelihood independent score equations

$$\mathbf{U}_t(\gamma_t) = \frac{\partial \ell}{\partial \gamma_t} = \mathbf{0},$$

where ℓ is the log likelihood. Each of these yields a maximum-likelihood estimate γ_t that is asymptotically Normal with asymptotic variance equal to the information matrix

$$\mathcal{I}_t(\gamma_t) = -\frac{\partial \mathbf{U}_t}{\partial \gamma_t'}.$$

Appealing to the multivariate Central Limit theorem, the joint asymptotic dispersion

of two sets of estimates γ_s and γ_t, is

$$\left(\sum_{u=1}^{n} \mathbf{D}_{su}\mathbf{D}'_{su}\right)^{-1}\left(\sum_{u=1}^{n} \mathbf{D}_{su}\mathbf{D}'_{tu}\right)\left(\sum_{u=1}^{n} \mathbf{D}_{tu}\mathbf{D}'_{tu}\right)^{-1}$$

where $\mathbf{D}_s$ and $\mathbf{D}_t$ are weighted functions of $\mathbf{U}_s$ and $\mathbf{U}_t$ evaluated at the maximum likelihood estimates of γ (Stram *et al.*, 1988). Implicit in this derivation is the assumption that the models for the various timepoints all have the same form (the same number of parameters). Ware *et al.* (1988) state succinctly that the asymptotic covariance matrix of $\hat{\gamma}$ can be estimated empirically. They use the so-called sandwich variance estimator (Section 1.4.5).

To test for the significance of an effect, one considers the corresponding estimates across all timepoints. Stram *et al.* (1988) consider group effects, but parameters related to continuous covariates may be tested similarly. This becomes a problem in multiple statistical inference. In similar vein a linear trend in components of γ over time might be tested.

In these analyses we are establishing whether a covariate does or does not have an effect, or that a covariate has an effect increasing by a certain proportion with each time interval. However, the fitted marginal models should be interpreted with caution because their estimates are inefficient when independence does not hold. The same can be said of the very closely related GEE method, discussed more fully and compared with full-likelihood models in Chapter 2.

The Stram procedure has the same aim as the Korn and Whittemore method; both methods are designed solely to assess the population effect of covariates. In the Stram method, a set of models, one for each timepoint, is defined; in the Korn and Whittemore method, the halfway stage is a set of models, one for each subject.

1.2.5 Independent increments models

Louis (1988) describes these models but provides no references. There seem to be no published reports of using this attractively simple technique. Perhaps it is so easily implemented with standard software that nobody has specifically mentioned the

approach. First differences are assumed to provide independent observations, which are analysed by univariate models. The local independence assumption of the standard random-effects model is implicitly assumed and is intuitively plausible for continuous outcomes though less plausible for discrete data. The method of differencing has one unfortunate consequence: covariates that are constant across timepoints drop out of the model by cancellation. Thus, while such strategies as the Koch and beta-binomial models suffer from the inability to model time-varying covariates, the independent increments model suffers from the opposite in that *only* time-varying covariates can be modelled.

1.3 The random effects model

Random effects are introduced in Sections 1.1 and 1.2, but more discussion is needed because of the importance of such models. The inclusion of random effects into models helps when dealing with certain model inadequacies; in particular, random effects can be used to model between-subject heterogeneity that is not otherwise accounted for. A random effect is a latent, or unobserved variable. If the model assumes the correct form for the distribution of such a variable, it may be strengthened. However, if a missing variable is non-Normal, binary say, then ignoring it, or treating it as Normal, can give an incorrect error structure that might have serious repercussions in interpreting the fit and predictions of the model (Louis, 1988).

When examining the local independence assumption, dependency between outcomes is explainable plausibly, or partially, by mutual dependency on some set of explanatory variables. The local independence assumption allows one to argue that *all* the dependence is explained in this way, and that in general a single variable suffices for this purpose. Conversely, modelled covariance structures, with more dependency than that accounted for by univariate dependency on explanatory variables, can be considered as surrogates for unmeasured covariates (Louis, 1988). By this latter argument we could assert that ordinary marginal models with full dependence structure are in fact subject-specific. They are equivalent to a random-effects formulation with

the random effects integrated/summed out, as in the REML fit of the Rasch model (Section 1.3.1) with dependency models playing the part of sufficient statistics for the REML fit.

These issues have been discussed for assumed Gaussian outcomes and random effects (Louis, 1988). If the validity of a measured Gaussian covariate is tenuous, then it should not be included in the model but instead modelled as a random effect. No inferences can be drawn about the effect of that covariate, but improved inferences in the other covariates are major benefits. Louis (1988) further asserts that if there is a choice between ordinary and conditional maximum likelihood (that is, ML versus REML) then REML produces less biased estimates of covariance parameters. Whether this holds for categorical outcomes is a subject for future study.

1.3.1 Random-effects models in general

A random-effects model is one in which the linear predictor includes parameters at the individual level that are not estimable by the data but only after some arbitrary constraint has been imposed, such as assuming that such effects are multivariate Normal. Given the random effects we model

$$g_t(\mathrm{E}[Y_{ut} \mid \mathbf{b}_{ut}]) = X_{ut}\boldsymbol{\gamma}_t + Z_{ut}\mathbf{b}_{ut} \tag{1.14}$$

for $t = 1, 2, \ldots T_u$ and $u = 1, 2, \ldots N$, where X_{ut} and Z_{ut} are design matrices for the fixed effects parameters, $\boldsymbol{\gamma}_t$, and the random effects, $\mathbf{b}_u t$, respectively.

There is no model for interaction terms, because it is implicit in all such models discussed in the literature that the distributions of the Y_{ut} given $\mathbf{b}_{ut}$ at each timepoint are independent (the so-called local independence assumption).

If there is no intrinsic interest in the estimates of the random effects we may use REML techniques, basing inference on the conditional likelihood of the data given sufficient statistics for the random effects. However, this begs the question of whether it is worth introducing such effects in the first place. One answer to this criticism is that conditional-likelihood (REML) models use *only* the longitudinal information —

based on comparisons within subjects — rather than the cross-sectional information, to estimate the fixed effects γ. A marginal model uses both cross-sectional and longitudinal information for these estimates, which are therefore subject to two different forms of potential bias.

If we do not use conditional likelihood methods, we have a true random-effects formulation and need to make some distributional assumptions about the effects before we can proceed. If the outcomes $\mathbf{Y}$ and the random effects (with expectation zero) are assumed to be multivariate Normal, then for an identity link function (g above) we have the model of Laird and Ware (1982).

More generally we can assume that (1.14) gives the standard link function to a generalized linear model for an outcome for which $\mathbf{Y}$, given $\mathbf{b}$, is a member of the exponential family. Almost always the random effects are assumed to be multivariate Normal, which may be as good an idea as any, given that this assumption is never testable. However, recently Lee and Nelder (1996) have developed methods for extending the possible random effect distributions to include the 'conjugate' of the exponential-family outcome: the term is used in quotation marks to emphasise that this is not synonymous with a Bayes conjugate. By the further assumption of local independence, such models might be fitted to longitudinal data in the usual way.

1.3.2 Random effects for binary data

Soon after the pioneering work on random effects models for Normal outcomes, the methodology was extended to serial observations with binary response by Stiratelli *et al.* (1984). Interestingly, the algorithm used to overcome the problem of the intractability of the integral, namely, the use of conditional modes rather than conditional means and the use of approximations for the score equations, predates its application to the more general setting of the mixed GLM (and hence to continuous variables) by nearly a decade: see Breslow and Clayton (1993).

Stiratelli *et al.* (1984) use a logistic model for mean response, with the assumption that the parameters of this model are Normally distributed in the population. Specifically,

we link to the vector of log odds ratios for individual u using

$$\boldsymbol{\lambda}_u = X_u\boldsymbol{\gamma} + Z_u\mathbf{b}_u; \tag{1.15}$$

the random effects $\mathbf{b}_u$ are arbitrarily taken to be multivariate Normal with mean 0 and dispersion matrix D, say. Moreover, in the empirical Bayes approach that is intrinsic to the strategy of Stiratelli *et al.* (1984), we assume a diffuse prior for $\boldsymbol{\gamma}$ (specifically, multivariate Normal with mean 0 and dispersion G, letting G^{-1} tend to zero).

Writing $\pi_{ut} = \Pr(Y_{ut} = 1)$ the likelihood contribution of the uth subject is

$$L_u = \prod_{t=1}^{T_u} \pi_{ut}^{y_{ut}}(1-\pi_{ut})^{1-y_{ut}}, \tag{1.16}$$

(recall that this is based on the assumption of local independence throughout) and so the observed overall likelihood is

$$L(\boldsymbol{\gamma}, D) = \prod_{u=1}^{N} \int L_u \exp\left\{-\tfrac{1}{2}\mathbf{b}_u' D^{-1}\mathbf{b}_u\right\} |D|^{-1} \, \mathrm{d}\mathbf{b}_u. \tag{1.17}$$

If a closed-form solution existed, one could compute $\hat{\boldsymbol{\gamma}}$ and $\hat{D}$ by maximizing this, and then obtain standard Bayes estimates

$$\hat{\mathbf{b}}_u = \mathrm{E}[\mathbf{b}_u \,|\, \mathbf{y}_u, \hat{\boldsymbol{\gamma}}, \hat{D}]; \tag{1.18}$$

$$\hat{\boldsymbol{\gamma}} = \mathrm{E}[\boldsymbol{\gamma} \,|\, \mathbf{y}, \hat{D}, G^{-1} = 0]. \tag{1.19}$$

As there is no closed form for any of these integrals, including those implicit in the expectations, Stiratelli *et al.* (1984) substitute conditional modes for conditional expectations and an approximation of (1.17) to obtain $\hat{D}$.

Stiratelli *et al.* (1984) frequently refer to the two-stage procedure of Korn and Whittemore (1979), justifying the need for a more flexible model at the expense of considerably increased computational overhead. I have already discussed the first of these issues above when illustrating how the two-stage model can be viewed as "poor man's random effects". Thus I concentrate on the computational problem, which is consid-

erable: the EM fitting algorithm (Stiratelli *et al.*, 1984) converged only after "several hundred" steps. With current technology, this is less of a problem than it was in 1984; but the method might be more widely applied as part of the day-to-day armoury of the applied statistician. These days the models can be fitted in packages such as SAS, using `proc mixed`, but this is not an identical formulation to that of Stiratelli *et al.* (1984) and might be rather less efficient, as it is less finely tuned to the particular problem of longitudinal binary data. Recent versions of the Multilevel software (for example, ML*n*) might be applicable if nesting holds.

Another common technique in numerical computation, the Gibbs sampler, now enables potentially highly-multiple integrals to be evaluated routinely, and so much of the work of Stiratelli *et al.* (1984) in overcoming the obstacle of integrating expression (1.17) might now be redundant. However, it is unclear that a 'black box' approach offers solutions that are preferable to those resulting from the careful insight of Stiratelli *et al.* (1984); in fact, it is not even yet known what precise criteria the MCMC algorithm requires in order to converge at all.

1.3.3 Random effects for multinomial data

The key papers of Stiratelli *et al.* (1984), Gilmour *et al.* (1985) and Anderson and Aitkin (1985) discussed only binary data. Even by the time of Agresti's comprehensive review of models for repeated ordered categorical response data (Agresti, 1989), extensions to polytomous data were still only conjectural. An important idea is to model

$$g_t(\mathrm{E}[Y_{ut,k} \mid b_u]) = \alpha_k + b_u + \alpha_t \tag{1.20}$$

(with the addition of fixed effects $X\boldsymbol{\beta}$ as applicable) for a suitable link function $g()$, such as cumulative logit, adjacent-category logit, or mean (see Agresti, 1989; Agresti, 1990). Here again u indexes subject/unit, t indexes time and I introduce k to denote the cutoff point of the presumed underlying continuous variable, that is, the multinomial category in question. Again we assume local independence; there is no interaction model. Inference is assumed to be concerned with tests for marginal ho-

mogeneity: in this context, we are assessing whether there is a time effect in addition to, or interacting with, covariate effects.

This approach has now been implemented (e.g. Agresti and Lang, 1993; Agresti, 1993). A conditional-likelihood implementation was given by Conaway (1989).

1.4 Nomenclature and definitions

1.4.1 The polynomial exponential family

The ordinary multivariate exponential family is the family of distributions for which the probability (density) function may be written, in canonical form,

$$f(\mathbf{y};\boldsymbol{\zeta}) = \exp\{\boldsymbol{\zeta}'\mathbf{s}(\mathbf{y}) + c(\mathbf{y}) - C(\boldsymbol{\zeta})\}, \tag{1.21}$$

where $\mathbf{s}(\mathbf{y})$ is a vector of sufficient statistics for the canonical parameters $\boldsymbol{\zeta}$, $c(\cdot)$ is a shape function (not depending on any unknown parameters), and C is a normalizing constant (constant, that is, given constant parameters) namely

$$C = \log\left\{\sum \exp\{\boldsymbol{\zeta}'\mathbf{s}(\mathbf{y}) + c(\mathbf{y})\}\right\} \tag{1.22}$$

where summation is over all possible values of Y: integration replaces summation for continuous variables.

A subset of these distributions is the *linear* exponential family, for which

$$\mathbf{s}(\mathbf{y}) = \mathbf{y}.$$

This subfamily includes all the common members of the ordinary univariate exponential family. For a multivariate distribution with $\mathbf{s}(\mathbf{y}) = \mathbf{y}$, independence of the components of Y is immediate. There might appear to be no need to model the data in this more complicated way. However, it is sometimes convenient to do so (e.g. in Section 5.3).

The simplest useful multivariate family is the so-called *quadratic* exponential family

(Gourieroux *et al.*, 1984), which has been discussed in the context of longitudinal data by many authors (e.g. Zhao and Prentice, 1990; Prentice and Zhao, 1991; Fitzmaurice and Laird, 1993; Fitzmaurice *et al.*, 1993; McCullagh, 1994). In this distribution we retain the above linear terms but also introduce a vector of pairwise products $\mathbf{w}$, i.e.

$$\mathbf{w} = (y_1^2, y_2^2, \ldots, y_T^2, y_1y_2, \ldots, y_1y_T, \ldots, y_{T-1}y_T)'.$$

In particular, following the notation of Fitzmaurice and Laird (1993), we write

$$f(\mathbf{y}; \boldsymbol{\Psi}, \boldsymbol{\Omega}) = \exp\{\boldsymbol{\Psi}'\mathbf{y} + \boldsymbol{\Omega}'\mathbf{w} + c(\mathbf{y}) - C(\boldsymbol{\Psi}, \boldsymbol{\Omega})\}. \tag{1.23}$$

The most obvious member of this family is the ordinary multivariate Normal distribution, which in any dimension T has sufficient statistics $\mathbf{y}$ and $\mathbf{w}$.

For multivariate binary data, using the form (1.23), $c(\cdot)$ is expressed as a linear combination of products of three or more of the elements of $\mathbf{y}$ (Zhao and Prentice, 1990). Taking some such sufficient statistics into the shape function, however, violates the original definition of the quadratic exponential family (Gourieroux *et al.*, 1984), where $\mathbf{s}(\mathbf{y}) = (\mathbf{y}', \mathbf{w}')'$ only. In other words, for a true member of the quadratic exponential family, the three-way and higher-order canonical parameters should be zero or non-existent (as for a bivariate distribution, or multivariate Gaussian).

It is, however, convenient to take some (or all) of the higher-order $\boldsymbol{\zeta}$ terms into the shape function. Whether or not to do so depends on the intended inference; any parameters or interactions regarded as nuisance only are arguably more naturally grouped into a 'non-parametric' shape function — especially if, as in Liang *et al.* (1992) and Prentice and Zhao (1991), inference will intentionally be only semi-parametric.

Effectively the choice is between full- or quasi-likelihood. A full likelihood approach must allow for all the possible $\boldsymbol{\zeta}$ parameters. For example, if it happens that there are no two- or higher-way interaction terms in the full distribution (such as for univariate distributions), the distribution belongs to the linear exponential family. If, however,

it is decided to simply ignore the interaction terms, we assume membership of a quasi-linear family.

The distinction is less clear than this in practice. We might find from the likelihood ratio of a quadratic family fit to that of a linear family fit that the interaction parameters could be taken as zero. The fit of this latter model is then full-likelihood based, but is identical to the fit of a quasi-linear model (one for which we never even considered interaction terms). On the other hand, we shall see in Chapter 2 that in semi-parametric models we are forced to introduce some value for at least some of the interaction terms (for example, zero for an independence model) because these enter the score equations. Consequently, we cannot in practice fully achieve the quasi-likelihood aim; we always fit a constrained version of the full likelihood (at least, for polytomous data).

There is an obvious hierarchy of models characterized by the highest-order crossproduct in a naturally extended definition of $\mathbf{w}$. To this end we redefine the general form (1.23) as the *polynomial exponential family* of distributions (PEF; McCullagh, 1994). For this to hold we now formally take $\mathbf{w}$ as a vector of two- and higher-way products of terms of $\mathbf{y} = (y_1, y_2, \ldots y_T)'$, up to and including the full T-way product (though possibly some $\boldsymbol{\Omega}$ terms are zero). In this context I refer to the *order* of a model as being the highest-order interaction parametrized (i.e. not taken to be zero, or taken entirely into the shape function).

In the interests of simplicity and symmetry, it is convenient to concatenate the sufficient-statistic and parameter vectors into

$$\mathbf{z} = (\mathbf{y}', \mathbf{w}')' \quad \text{and} \quad \boldsymbol{\xi} = (\boldsymbol{\Psi}', \boldsymbol{\Omega}')'$$

and write (1.23) more concisely as

$$f(\mathbf{y}; \boldsymbol{\xi}) = \exp\{\boldsymbol{\xi}'\mathbf{z} + c(\mathbf{z}) - C(\boldsymbol{\xi})\}, \tag{1.24}$$

which is the parametrization I shall generally assume implicitly throughout this thesis.

It is useful to note that for polytomous distributions

$$C(\boldsymbol{\xi}) = -\log \mathrm{P}(Y_1 = Y_2 = \cdots Y_T = 0). \tag{1.25}$$

Throughout Chapter 2 the parameters $\boldsymbol{\xi}$ are variationally independent: none of them is constrained a priori to be a function of another — though choice of linear predictors may impose such constraints — and there are no structural zeros in the joint probability table. Structural zeros are considered in Chapter 5.

1.4.2 A note on the use of the term *order*

Frequently pairwise interactions are called 'first-order', which is sensible when applied to interactions, there being only one interaction between each pair. However this is not the easiest terminology to apply to the hierarchy of odds ratios for discrete data. Instead, therefore I refer throughout to an nth-order interaction as an interaction between n variables, and an nth-order odds ratio as the ratio relating n variables. This has the unfortunate consequence that what I call a first-order interaction is not an interaction in any normal sense.

Also by analogy with the terminology GEE1 and GEE2 (Liang *et al.*, 1992), I define a GEEn model as being any model of order n in the above sense, to be fitted using the generalized estimating equations approach discussed in Section 2.6.

1.4.3 Subdistributions and reproducibility

When observation vectors are not all the same size, or when some are only partially observed, it is convenient to contrast full observations with these 'sub-observations'. The corresponding distribution is called here a 'subdistribution'.

I use superscript set notation to pick out the elements of the full set of potential values that appear in the sub-observation and write $\mathbf{Y}^{\mathcal{A}} = \{Y_i \,|\, i \in \mathcal{A}\}$, where $\mathcal{A} \subseteq \mathcal{T} = \{1, 2, \ldots, T\}$ assuming a full observation is of size T.

For polytomous data a full observation has probability function

$$\mathrm{P}(\mathbf{Y}) = \exp\{\boldsymbol{\xi}'\mathbf{z} - C\}.$$

For convenience, if the set superscript is missing, then a full observation or distribution is implicit. The subdistribution of an incomplete vector $\mathbf{Y}^{\mathcal{A}}$ for $\mathcal{A} \subset \mathcal{T}$ is then

$$\mathrm{P}(\mathbf{Y}^{\mathcal{A}}) = \sum_{\text{missing } \mathbf{Y}^{\mathcal{T}\setminus\mathcal{A}}} \exp\{\boldsymbol{\xi}'\mathbf{z} - C\},$$

where $\setminus$ denotes set difference. A direct argument shows that this subdistribution, being the marginal of a polynomial exponential family distribution, must again be polynomial exponential, and so can also be written in canonical form as

$$\mathrm{P}(\mathbf{Y}^{\mathcal{A}}) = \exp\{\boldsymbol{\xi}^{\mathcal{A}\prime}\mathbf{z}^{\mathcal{A}} - C^{\mathcal{A}}\}$$

for $\boldsymbol{\xi}^{\mathcal{A}}$ the canonical parameters of the marginal distribution. In general, $\boldsymbol{\xi}^{\mathcal{A}} \neq \boldsymbol{\xi}^{\mathcal{T}}$ except for very degenerate cases.

In modelling terms, if the full vector $\mathbf{Y}$ satisfies a marginal model, then all subsets $\mathbf{Y}^{\mathcal{A}}$, $\mathcal{A} \subset \mathcal{T}$, satisfy the corresponding subset of the model. This useful property is known as *reproducibility* (Liang *et al.*, 1992) or *upward compatibility* (McCullagh, 1989).

1.4.4 Constraints on the canonical parameters $\boldsymbol{\xi}$, and model selection

The linear predictors may be deliberately chosen to ensure that no a priori constraints are imposed on the canonical parameters, $\boldsymbol{\xi}$ in (1.24). Most simply, for an identity-link model (discussed in Section 2.4), writing

$$\boldsymbol{\xi} = \boldsymbol{\alpha} + X\boldsymbol{\beta}, \tag{1.26}$$

where X is the design matrix for explanatory variables, provided there is a separate parameter in $\boldsymbol{\alpha}$ for every corresponding parameter in $\boldsymbol{\xi}$, for all orders of interaction,

then the full PEF distribution is modelled. I refer to such a model as '$\boldsymbol{\xi}$-unconstrained' or equivalently, here, 'intercept-unconstrained'.

This is a useful baseline, for likelihood-ratio testing for example, but certainly most models of interest will violate such lack of constraint. Any marginal model that imposes intercept constraints in a marginal framework, e.g. sharing an intercept for the means at different timepoints, will be $\boldsymbol{\xi}$-constrained, though the nature of such constraints is not easily understood or usefully written in terms of $\boldsymbol{\xi}$ constraints. Similarly, in Markov models (Chapters 4 and 6), even an apparently unconstrained model will not be $\boldsymbol{\xi}$-unconstrained if the link function, modelling the dependence on previous values, is inappropriate. As for marginally-linked models, it can be very difficult to determine what constraints on $\boldsymbol{\xi}$ are imposed by the model chosen, and in practice I do not attempt to do so.

Imposing as few restrictions on $\boldsymbol{\xi}$ as possible is of some benefit. The high-order odds ratios, in particular, will commonly be regarded as nuisance parameters, but one might argue that they should still ideally be modelled as precisely as possible, because the robustness of the ordinary parameter-estimate variance matrix depends on the proximity of the modelled distribution to the true distribution. This is an argument against merely setting high-order $\boldsymbol{\xi}$ to zero for computational and interpretive simplicity, or other strategies such as setting all ratios of like order to be the same. Likelihood-ratio tests are also adversely affected by the number of unmodelled nuisance parameters — an issue when choosing between various models for first-order margins, for example.

It is useful to be able to refer to and distinguish the intercept parameters, $\boldsymbol{\alpha}$, which define an 'intercept model', as distinct from the parameters of presumably more interest, i.e. $\boldsymbol{\beta}$, which define an 'explanatory model'. I shall maintain this notational distinction throughout. When the distinction is not needed, I denote the full set of parameters as $\boldsymbol{\gamma}$.

When testing for nonsignificant departure from model saturation, we can consider dropping or equating some of the $\boldsymbol{\alpha}$, for distribution simplification, or some of the $\boldsymbol{\beta}$, for covariate simplification, within a single, unified analysis-of-deviance framework. The two sets of parameters are not mathematically distinct. The introduction of

the concept of 'intercept model' clarifies the way in which parameters in the linear predictors serve to constrain the parametric model for the underlying distribution.

In forward selection we need to consider what we mean by the *null* model. A completely null model, with $\xi_a \equiv \xi_b \quad \forall a, b$, and all $\beta = 0$, is not likely to be of any interest, even for comparisons. The most natural analogue of the univariate-GLM null model is a $\boldsymbol{\xi}$-unconstrained model with no explanatory variables. However in the literature it is more common to start from the strongly $\boldsymbol{\xi}$-constrained independence model.

1.4.5 Variance estimators

The ordinary estimate for the dispersion matrix of maximum likelihood parameter estimates is the evaluated Fisher information matrix

$$\mathcal{I} = \sum \mathrm{E}[\mathbf{U}_u \mathbf{U}_u'] \tag{1.27}$$

where $\mathbf{U}_u = \partial \ell_u / \partial \boldsymbol{\gamma}$ is the score contribution of subject u. However, this only gives accurate assessment of variance if the dependency structure for the distribution of the data is specified correctly. Often, for longitudinal data with several timepoints, this cannot be achieved and the full distribution may be poorly specified. Liang and Zeger (1986) proposed to use the so-called *robust* (to model misspecification) or *sandwich* estimator:

$$\mathcal{S} = \mathcal{I}^{-1} \left(\sum \mathbf{U}_u \mathbf{U}_u' \right) \mathcal{I}^{-1}. \tag{1.28}$$

This estimator is also that used by Stram *et al.* (1988).

A cruder approximation was proposed (for use within Fisher scoring iterations) by Azzalini (1994):

$$\mathcal{A} = \sum \mathbf{U}_u \mathbf{U}_u'. \tag{1.29}$$

A justification for this approximation (not stated explicitly by Azzalini) is that over a large sample the expectation and average of any statistic will converge, and hence the sum of the evaluations approaches the sum of the expectations.

Chapter 2

Marginal and canonically-linked models

A *marginal model*, as defined on page 2, equation (1.3), is one in which there are linear predictors for the marginal means of the observations Y_t at timepoints $t = 1, 2, \ldots, T$. The *fully marginal* models discussed below, in Section 2.3, have linear predictors linked, perhaps indirectly, to *all* the marginal expectations, including the higher-order expectations. I describe and generalize the approach of Ekholm *et al.* (1995) in Section 2.3.1, and that of Molenberghs and Lesaffre (1994), Glonek and McCullagh (1995), and independently, myself, in Section 2.3.2.

The mixed parametrization of Fitzmaurice and Laird (1993), discussed in Section 2.5, and the GEE and related methods of Section 2.6, offer marginal, but not fully marginal, models. The new model introduced in Section 2.4, the canonically-linked model, is not marginal. It is discussed in this chapter because it is a natural simplification of the model of Fitzmaurice and Laird.

Before turning to the particular models, we discuss the idea of full linkage in general terms in Section 2.1, and define and discuss the nomenclature for various types of odds ratio in Section 2.2.

2.1 Extending the GLM to multivariate data

The single-parameter GLM for univariate data is characterized by the use of a link function from a linear predictor in terms of the parameters of interest to some specified function of the canonical parameter of the distribution. Other parameters are essentially nuisance parameters and do not enter into the heart of the model, the linear predictor.

This philosophy underpins McCullagh and Nelder's marginal modelling approach to multivariate data (see Liang *et al.*, 1992). Interest is in the univariate margins of the joint distribution for $\mathbf{Y}$; the dependence structure, and any scale parameters, are regarded as nuisance and are estimated separately. In Liang and Zeger's GEE model (now more properly called GEE1), there are no marginal scale parameters (at least for discrete data) and the 'nuisance' interactions, parametrized by some $\boldsymbol{\alpha}$ in their notation, are essentially extrinsic to the model; of course the estimated values of the interactions affect those of the parameters of interest, $\boldsymbol{\beta}$. Such models extend the GLM inasmuch as the classical GLM has a single link function, and the multivariate marginal version has a separate link to a function of the mean at each timepoint with some exogenous correction for dependency to prevent biased estimates. The strategy of Stram *et al.* (1988), discussed in Section 1.2.4, is explicitly two-stage: the marginal models are fitted separately, as if the data were independent, and are then corrected for dependency using the sandwich variance estimator (Section 1.4.5).

An alternative approach is to account for *all* the canonical parameters that need to be modelled, an idea not often exploited in univariate, multiparameter GLMs but now common in the multivariate setting (Fitzmaurice and Laird, 1993; Molenberghs and Lesaffre, 1994; Glonek and McCullagh, 1995). This enables marginal modelling to proceed as before, while simultaneously defining a specific model (not necessarily marginal) for the interaction terms. This is invaluable not only when the interaction is regarded as being of interest rather than nuisance, but also serves to emphasise just how sensitive the marginal part of the model is to interaction mis-specification.

A further advantage of this fully-linked approach is that we are led to use maximum

likelihood for all the parameter estimates, allowing use of the existing machinery for hierarchical model selection. The generalized estimating equation approach, with its exogenous interaction estimates, is essentially quasi-likelihood, so that this battery of established techniques cannot be used directly. (Technically, the term quasi-likelihood is misused here, because in the original definition quasi-likelihood refers to the integral of the estimating equations, which may not exist in the multivariate setting — see Liang *et al.*, 1992. Nevertheless this label is frequently used in the literature.) Recent and ongoing research uses approximate, empirical and/or projected likelihood methods to overcome this shortcoming when assessing models that have been fitted using quasi-likelihood techniques (Owen, 1988; Owen, 1991; McLeish and Small, 1992; Li, 1993; Li and McCullagh, 1994; Tsou and Royall, 1995; Hanfelt and Liang, 1995). However, I avoid the need to consider and evaluate such approximations since, apart from the discussion of the generalized estimating equation method at the end of this chapter, the models I shall present and discuss below are all fully-linked/full-likelihood.

2.2 The multivariate polytomous distribution and nomenclature

Consider distributions belonging to the polynomial exponential family (Section 1.4.1). If the variables $\mathbf{y}$ in (1.24) are continuous, and the shape function is identically zero, we obtain the standard multivariate Normal distribution, which belongs to the quadratic exponential family and only requires pairwise covariance specification. When the data are discrete and the shape function is identically zero, and all crossproducts up to the full interaction $y_1 y_2 \cdots y_T$ are included in $\mathbf{z}$ (with corresponding parameters $\boldsymbol{\xi}$), we obtain the multivariate polytomous distribution, which belongs to the polynomial exponential family of order T. This is the only distribution considered in detail in this thesis.

The crossproduct vector $\mathbf{z}$ for polytomous data coded such that each y_t takes values $0, 1, 2, \ldots$ contains only true *cross*products; terms such as y_1^2, $y_2^2 y_3$, etc., do not appear. For binary data coded 0/1 this is obvious because $y^2 \equiv y$, but for polytomous data

multiplication does not give similar group closure. The reason for the lack of squared terms is because the joint probability, conveniently thought of as a table and denoted π, is completely specified by $k^T - 1$ odds ratios, where k is the number of categories. The set of canonical parameters $\boldsymbol{\xi}$ is precisely such a minimal set of odds ratios (i.e. just sufficient to fully specify the probability table), as assumed in Sections 2.2.1 and 2.2.2 and discussed in Appendix A2.2.

2.2.1 Binary data

In this simplest of cases, the canonical parameters $\boldsymbol{\xi}$ are *zero-conditional* log odds ratios (log CORs): that is, the logs of the odds ratios (denoted $\boldsymbol{\chi} = \mathrm{e}^{\boldsymbol{\xi}}$) *conditional on all other observations being zero.* Thus, the first-order COR is

$$\chi_i = \frac{\mathrm{P}(Y_i = 1 \mid Y_t = 0,\ \forall t \neq i)}{\mathrm{P}(Y_i = 0 \mid Y_t = 0,\ \forall t \neq i)}$$

and for pairwise interactions

$$\chi_{ij} = \frac{\mathrm{P}(Y_i = Y_j = 1 \mid Y_t = 0,\ \forall t \neq i,j)\mathrm{P}(Y_i = Y_j = 0 \mid Y_t = 0,\ \forall t \neq i,j)}{\mathrm{P}(Y_i = 1, Y_j = 0 \mid Y_t = 0,\ \forall t \neq i,j)\mathrm{P}(Y_i = 0, Y_j = 1 \mid Y_t = 0,\ \forall t \neq i,j)}.$$

Higher-order CORs are ratios of lower-order odds ratios: for example for a 3-way COR

$$\chi_{ijk} = \frac{1_k\mathrm{COR}_{ij}}{\chi_{ij}}$$

where the term $1_k\mathrm{COR}_{ij}$ refers to a pairwise odds ratio that is identical to the ordinary COR, χ_{ij}, except that rather than condition on *all* the other observations being zero, we condition on all other observations except y_k being zero, with $y_k = 1$. There are several instances below of introducing a new index, here k, to define sets of ratios inductively. The particular indexing is quite arbitrary; it is not implied that y_k is a later observation than y_i or y_j, and the same applies in higher dimensions too. Throughout we assume $i \neq j \neq k \cdots$, and this gives us the minimal set of ratios discussed above and by, for example, Bishop *et al.* (1975, p. 42 ff.). From the above definition ξ_{ij} is the same as ξ_{ji} for all i and j, which allows the subscripts to be

written in ascending order without loss of generality; the same convention is adopted for higher order interactions.

When there is a single subscript, ξ_i is a zero-conditional logit, referred to here as a 'first-order log COR' to avoid having to continually refer to 'logits and log odds ratios'. This convention also stresses the strong symmetry that lies at the heart of the algorithmic techniques of Sections 2.4.5, 2.5.1 and Chapter 3. It is often expedient to consider within the set of odds ratios a *zeroth* order ratio, defined as unity, which represents the restriction that the cells of the probability table sum to one.

For two or more subscripts the obvious notation is followed: ξ_{13} denotes the log COR of Y_1 and Y_3, and z_{13} denotes the product $y_1 y_3$.

When referring to an element of the vector $\boldsymbol{\xi}$, or $\boldsymbol{\chi}$, further notation is needed to allow for an unspecified number of subscripts. Script capital letters, beginning at $\mathcal{A}$, denote a *set* of subscripts, with $\mathcal{A} \subseteq \mathcal{T} = \{1, 2, \ldots, T\}$ for outcome vectors of size T (*cf* Section 1.4.3 where a superscript denotes a set of variables). Single-element subsets are included in this notation, as is the full interaction term, $\xi_{\mathcal{T}}$. Thus, $\xi_{\mathcal{B}}$ means any one element of $\boldsymbol{\xi}$ (of any order, unless $\mathcal{B}$ is specified), whereas ξ_i refers to a first-order log COR (logit) and ξ_{ij} to a second-order, etc. Although $\xi_{\mathcal{A}}$ refers to elements of the $\boldsymbol{\xi}$ vector, $\boldsymbol{\xi}_u$ refers to the entirety of the ξ vector for the uth subject. This is unambiguous in context as subscript u is reserved exclusively for subject (i.e. unit) and bold face is used for vectors.

Although the $\boldsymbol{\xi}$ are the canonical parameters, we are more likely to wish to interpret and parametrize the *marginal* odds ratios, which I designate by $\boldsymbol{\Lambda}$, or even more frequently their logs, denoted $\boldsymbol{\lambda}$. Again it is expedient to blur the terminological distinction between marginal logits,

$$\lambda_i = \log \frac{\mathrm{P}(Y_i = 1)}{\mathrm{P}(Y_i = 0)}$$

and higher-order log odds ratios,

$$\lambda_{ij} = \log \frac{\mathrm{P}(Y_i = Y_j = 1)\mathrm{P}(Y_i = Y_j = 0)}{\mathrm{P}(Y_i = 1, Y_j = 0)\mathrm{P}(Y_i = 0, Y_j = 1)},$$

and to refer to all the $\boldsymbol{\Lambda}$ (or $\boldsymbol{\lambda}$) terms as marginal (log) odds ratios, denoted MORs. The subscript conventions described above apply here also.

There are the same number of marginal $\boldsymbol{\lambda}$ as there are zero-conditional $\boldsymbol{\xi}$, with the same subscripts, but the joint probability distribution cannot be written in closed form in terms of $\boldsymbol{\lambda}$ when there are more than two variables since this involves inverting the multivariate logistic transform (Chapter 3).

The full interaction ratios, $\Lambda_{\mathcal{T}}$ and $\chi_{\mathcal{T}}$, are identical and synonymous. This term may be regarded as being marginal or conditional, there being no further variables available on which to condition, in the conditional definition. This parameter is always orthogonal to the lower-order $\Lambda_{\mathcal{A}}$ (Molenberghs and Lesaffre, 1994).

In Chapter 5 we will consider *specified-conditional* odds ratios (SCORs)

$$\text{SCOR}_i = \frac{\text{P}(Y_i = 1 \mid \text{specified values of } Y_t,\ t \neq i)}{\text{P}(Y_i = 0 \mid \text{specified values of } Y_t,\ t \neq i)};$$

an ordinary COR is a SCOR for specified values all zero. To avoid possible confusion, 'ordinary' CORs, χ, are generally here referred to as *zero-conditional* ratios.

2.2.2 Polytomous data

Let variable Y_t take k_t values, which are coded here as $0, 1, 2, \ldots, (k_t - 1)$: some authors use codes from 1 to k_t. The probability table, $\boldsymbol{\pi}$, of the joint distribution of $\mathbf{Y}$, has $\prod_{t=1}^{T} k_t$ cell probabilities and may be characterized by any of several minimal sets of odds ratios: Agresti (1990) discusses the bivariate case and Bishop *et al.* (1975) demonstrate certain aspects of the multivariate extension. The minimal set giving the canonical parameters in equation (1.24) is the set of log CORs containing the all-zero-index cell, $\pi_{00\ldots0}$, i.e. $\text{P}(Y_1 = Y_2 = \cdots = Y_T = 0)$. These ratios are called here *zero-based* CORs, which extends the nomenclature *zero-conditional* CORS.

A simple example and an appeal to induction, given in Appendix A2.2, demonstrates why this set of ratios must be the same as the set of canonical parameters in the polynomial exponential family representation in all dimensions.

Although the canonical parameters are of theoretical importance, we will probably be

more interested in *marginal* ratios, such as

$$\lambda_{ij\cdots}^{rs\cdots} = \frac{\lambda_{i\cdots}^{r\cdots|y_j=s}}{\lambda_{i\cdots}^{r\cdots|y_j=0}},$$

conditional on the levels of Y_j, where the ratios are based on a marginal table collapsed over the unindexed variables.

There is no particular reason to prefer one minimal set of such ratios over another in the marginal case, other than for ease of interpretation. The CORs can never be expressed in terms of the marginal ratios in closed form whatever set is used (Bishop *et al.*, 1975; Fitzmaurice and Laird, 1993).

If there are more than a few possible values k_t or if T becomes large, there are an enormous number of odds ratios, which will be very difficult to interpret whatever form is chosen. This problem has been addressed for ordinal data, for which the natural ordering suggests simplifications (Agresti, 1989; Molenberghs and Lesaffre, 1994). A different approach (Qu *et al.*, 1992, 1995) views the categorical outcome as the discretization of an underlying multivariate Normal variable and models the correlation structure of the underlying distribution, which drastically reduces the number of interaction parameters, and effectively removes the constraints that complicate their marginal discrete counterparts.

2.3 The fully marginal model

The fully marginally linked model could be considered to be the essence of marginal modelling. Perhaps the most obvious linkage is a straightforward extension of the GLM to vector outcomes: for the uth unit, let

$$\mathbf{g}(\mathrm{E}[\mathbf{Z}_u]) = X_u\boldsymbol{\gamma}, \tag{2.1}$$

where $\mathbf{Z} = (\mathbf{Y}', \mathbf{W}')'$ is the vector of variables and crossproducts introduced in the definition of the polynomial exponential family (1.24), X is a design matrix, and $\boldsymbol{\gamma}$ represents the parameters to be estimated. However, there is a problem in determining

a form for the components of the link function $\mathbf{g}()$ for polytomous data: the higher-order expectations are severely constrained by the first-order marginal expectations (Prentice and Zhao, 1991; Liang *et al.*, 1992; and Section 3.2.6), as are the third-order by the second, and so on, making the problem essentially intractable.

In an alternative approach (Prentice and Zhao, 1991) that does not overcome the problem, the first-order marginal expectations are linked by

$$\mathbf{g}^{(1)}(\mathrm{E}[\mathbf{Y}_u]) = X_u^{(1)}\boldsymbol{\gamma}^{(1)}$$

and the triangular elements of the dispersion matrix are linked by

$$\mathbf{g}^{(2)}(\mathrm{vec}\,\{\mathrm{E}[\mathbf{Z}_u\mathbf{Z}_u'] - \boldsymbol{\nu}_u\boldsymbol{\nu}_u'\}) = X_u^{(2)}\boldsymbol{\gamma}^{(2)},$$

where $\boldsymbol{\nu}_u = \mathrm{E}[\mathbf{Z}_u]$; i.e. in the bivariate case, we link to μ_1, μ_2 and $\mathrm{cov}(y_1, y_2)$ directly. Although such links are perhaps more natural to interpret than the raw link to expectations about the origin, we would need nonlinear predictors for the higher moments involving functions of the lower moments to ensure that predicted values lay in the correct region.

None of the fully marginal approaches described here or in the literature has been able to overcome this problem in general. Perhaps the best compromise, due to Lipsitz *et al.* (1991) and implemented by, for example, Liang *et al.* (1992), links to the marginal odds ratios, $\boldsymbol{\Lambda}$. The attraction is that interpretation is easy (in low dimension), and in the context of bivariate polytomous data the constraint problem is overcome — fixing all the Λ_1^r and Λ_2^s imposes no constraint on possible values of Λ_{12}^{rs}. With bivariate data, the marginal Λ_{12}^{rs} are equivalent to the canonical χ_{12}^{rs}, since there is no third variable on which to condition for these zero-conditional ratios. Such canonical parameters may lie anywhere on the real line, as shown formally in Section 2.4.5; I am unaware of any reference in the literature to this. Similarly, with T-variate data, the Tth-order odds ratio is unconstrained. However, all the intermediate marginal ratios *are* constrained (see Section 3.2.6). The constraints on $\boldsymbol{\Lambda}$ are not as strong as are the

constraints on marginal expectations and/or covariances, and in practice modelling without inbuilt constraints seems to always yield valid results, at least provided the true parameter values are not too close to the limits of the parameter space, just as in the univariate case direct linear predictors can be considered for probabilities near 0.5 when in general a logit link should be used.

2.3.1 Dependence ratios

A (marginal) *dependence ratio* is defined to be

$$\tau_{12\cdots t} = \frac{\nu_{ij\cdots t}}{\nu_i \nu_j \cdots \nu_t}, \tag{2.2}$$

including for convenience the univariate (degenerate) ratios

$$\tau_t = \nu_t, \tag{2.3}$$

where $\nu_{ij\cdots t} = \mathrm{E}[Y_i Y_j \cdots Y_t]$ are moments about the origin; such a ratio is unity if there is (marginal) independence, and for the bivariate case is greater than unity if there is positive correlation and less than unity if there is negative correlation. A multidimensional analogue of correlation is required for $T > 2$. However, such ratios are more highly constrained than simply to be positive. Consider bivariate binary data; $\nu_{12} = \pi_{12}$, $\nu_1 = \pi_{1+}$ and $\nu_2 = \pi_{+1}$ (plus for summation), and so

$$\nu_{12} \leq \min(\nu_1, \nu_2) = \nu_{\min};$$

the inequality is strict for non-degenerate distributions. Immediately

$$\frac{\nu_{12}}{\nu_{\max}} \leq \frac{\nu_{\min}}{\nu_{\max}}$$

so that

$$\tau_{12} \leq \frac{1}{\nu_{\max}}. \tag{2.4}$$

Tighter constraints apply to higher-order ratios with T-variate data (mentioned but not given by Ekholm *et al.*, 1995).

Ekholm *et al.*'s approach is to link linear predictors to the logarithms of the dependence ratios,

$$\boldsymbol{\rho} = \log \boldsymbol{\tau}.$$

In the development of the model Ekholm *et al.* (1995) intended to obtain estimates using GLIM4 (Anders Ekholm, personal communication; algorithmic details in Ekholm and Green, 1994). Rather than fit general models, the $\boldsymbol{\rho}$ are constrained to follow certain simple symmetrical relationships, such as setting all ratios of like order equal to each other ("horizontal homogeneity"). Although no attempt is made to explicitly include constraints to ensure that relationships such as (2.4) hold, inconsistent estimates do not appear to arise in practice, as for odds ratios.

One claimed advantage of dependence ratios is the ease of interpretation even with increasing order, which compares favourably with any form of odds ratios. A second advantage is computational. Description of the method to obtain cell probabilities $\boldsymbol{\pi}$ from odds ratios $\boldsymbol{\Lambda}$ takes all of Chapter 3, which should be contrasted with the following.

Denote logs of expectations as $\boldsymbol{\epsilon} = \log \boldsymbol{\nu}$. The relationship between these and the log ratios is simply the linear form $\boldsymbol{\epsilon} = M\boldsymbol{\rho}$, where M is easily derived from (2.2), specifically

$$\epsilon_{\mathcal{A}} = \rho_{\mathcal{A}} - \sum_{i \in \mathcal{A}} \rho_i,$$

where the set notation for subscripts is as introduced on page 38, and summation runs over only univariate ratios with an index in $\mathcal{A}$.

For trivariate binary data, with $\boldsymbol{\rho}$ ordered $\boldsymbol{\rho} = (\rho_1, \rho_2, \rho_3, \rho_{12}, \rho_{13}, \rho_{23}, \rho_{123})'$, and similarly for $\boldsymbol{\epsilon}$, we have

$$M = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ -1 & -1 & 0 & 1 & 0 & 0 & 0 \\ -1 & 0 & -1 & 0 & 1 & 0 & 0 \\ 0 & -1 & -1 & 0 & 0 & 1 & 0 \\ -1 & -1 & -1 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{2.5}$$

The inverse, M^{-1}, is very like M: wherever there is -1 in the off-diagonals of M, M^{-1} has $+1$. The general form of M in the T-variate case is

$$M_T = \begin{pmatrix} I_T & 0 \\ N & I_S \end{pmatrix}, \tag{2.6}$$

where I_T and I_S are identity matrices ($S = 2^{(T-1)}$ for binary data), and N is a suitable $S \times T$ matrix of ones and zeros: an element is -1 if the corresponding column corresponds to an index of the $\nu_{\mathcal{A}}$ for the row and otherwise it is zero.

I now derive the score equations for the general link formulation

$$\mathbf{g}(\boldsymbol{\tau}_u) = X_u \boldsymbol{\gamma} \tag{2.7}$$

for subject u. Ekholm *at al.*'s models are special cases of this formulation for particular choices of design matrix X and parameters $\boldsymbol{\gamma}$, i.e. for restricted values of $\boldsymbol{\tau}$.

I restrict consideration to the following standard link functions: the univariate ratios, i.e. univariate expectations, have a logit link, while the higher-order ratios have a log link (that is, identity to the log ratios). Consequently the corresponding $\boldsymbol{\tau}$ estimates are positive, but may fail to satisfy constraints such as (2.4).

The score equations and information matrix

For polytomous data, the probability function for the uth observation vector is the order-T polynomial exponential family form given in equation (1.24), with $c(\mathbf{z}) \equiv 0$.

Assuming independence between subjects, the log likelihood is

$$\ell = \sum_u \ell_u = \sum_u \{\boldsymbol{\xi}_u' \mathbf{z}_u - C(\boldsymbol{\xi}_u)\}. \tag{2.8}$$

Thus the score contribution for subject u is, by the chain rule,

$$\mathbf{U}_u(\boldsymbol{\gamma}) = \frac{\partial \boldsymbol{\eta}_u'}{\partial \boldsymbol{\gamma}} \frac{\partial \boldsymbol{\rho}_u'}{\partial \boldsymbol{\eta}_u} \frac{\partial \boldsymbol{\epsilon}_u'}{\partial \boldsymbol{\rho}_u} \frac{\partial \boldsymbol{\nu}_u'}{\partial \boldsymbol{\epsilon}_u} \frac{\partial \boldsymbol{\xi}_u'}{\partial \boldsymbol{\nu}_u} \frac{\partial \ell_u}{\partial \boldsymbol{\xi}_u} = X_u' A M' \Delta_u V_u^{-1} (\mathbf{z}_u - \boldsymbol{\nu}_u), \tag{2.9}$$

where $\boldsymbol{\eta}$ are the linear predictors. Working through (2.9) from right to left, and dropping the subscript u, the term

$$\frac{\partial \ell}{\partial \boldsymbol{\xi}} = \mathbf{z} - \boldsymbol{\nu}$$

follows from (1.22):

$$\frac{\partial C}{\partial \boldsymbol{\xi}} = \frac{1}{\mathrm{e}^C} \sum \mathbf{z} \exp\{\boldsymbol{\xi}'\mathbf{z}\} = \sum \mathbf{z} \exp\{\boldsymbol{\xi}'\mathbf{z} - C\} = \boldsymbol{\nu}. \tag{2.10}$$

Hence

$$\frac{\partial \boldsymbol{\nu}'}{\partial \boldsymbol{\xi}} = \sum \mathbf{z}(\mathbf{z} - \boldsymbol{\nu})' \exp\{\boldsymbol{\xi}'\mathbf{z} - C\} = V,$$

the dispersion matrix. Assuming a non-degenerate distribution, this must be non-singular, so that $\partial \boldsymbol{\xi}'/\partial \boldsymbol{\nu}$ is simply the inverse. See, for example, Zhao and Prentice (1990) and Liang *et al.* (1992).

Since $\boldsymbol{\nu} = \mathrm{e}^{\boldsymbol{\epsilon}}$, we obtain

$$\frac{\partial \boldsymbol{\nu}'}{\partial \boldsymbol{\epsilon}} = \Delta = \mathrm{diag}\{\nu_{\mathcal{A}}\}$$

where $\mathcal{A}$ runs over all possible sets of indices, and since $\boldsymbol{\epsilon} = M\boldsymbol{\rho}$ for M as above,

$$\frac{\partial \boldsymbol{\epsilon}'}{\partial \boldsymbol{\rho}} = M'.$$

The term $A = \partial \boldsymbol{\rho}'/\partial \boldsymbol{\eta}$, where $\boldsymbol{\eta}$ are the linear predictors, allows for the mixed links under consideration, i.e. identity for all second- and higher-order log ratios, but logit

for the first-order terms. The following general result is also used later: if

$$\text{logit } a = \log(a/(1-a)) = b,$$

then

$$a = \frac{e^b}{1+e^b} \quad \text{and} \quad \frac{\partial a}{\partial b} = a(1-a). \tag{2.11}$$

Consequently, the derivative matrix A takes the block diagonal form

$$A = \begin{pmatrix} B & 0 \\ 0 & I \end{pmatrix}$$

where I is a suitably sized identity matrix and B is as follows: since for first-order terms

$$\rho_i = \log \tau_i = \log \frac{e^{\eta_i}}{1+e^{\eta_i}},$$

we have, using (2.11) and the chain rule,

$$B = \text{diag}\left\{\frac{\tau_i(1-\tau_i)}{\tau_i}\right\} = \text{diag}\left\{1-\tau_i\right\}.$$

Note that

$$D = AM'\Delta = \frac{\partial \boldsymbol{\nu}'}{\partial \boldsymbol{\eta}}$$

is more efficiently computed in closed form than by actual matrix multiplication. Finally, (2.9) is obtained by premultiplication by $\partial\boldsymbol{\eta}'/\partial\boldsymbol{\gamma} = X'$.

Assuming the canonical $\boldsymbol{\xi}$ are variationally independent, exactly the same functional form obtains for the unconstrained likelihood as for the constrained forms of Ekholm. Only the form of the design matrix X, and the corresponding size and form of the parameter vector $\boldsymbol{\gamma}$, are different. Although highly constrained forms can be more efficiently calculated than by the above scheme, the penalty to pay for the improved efficiency is the necessity to write computer code specific to each model. In any data-fitting exercise, we want to compare the fits for several combinations of $\boldsymbol{\rho}$ and

explanatory variable restrictions: a single program implementing the above full scheme will serve to fit them all, for a variety of designs $\boldsymbol{\eta} = X\boldsymbol{\gamma}$. Furthermore, it is extremely difficult to express even the simplest of $\boldsymbol{\rho}$-specified constraints (for example, horizontal homogeneity above) in terms of constraints on $\boldsymbol{\xi}$, rendering a direct likelihood-constrained approach intractable in general.

The expected information matrix is easily derived from the score function $\mathbf{U}(\boldsymbol{\gamma})$. Each subject contributes

$$\mathcal{I}(\boldsymbol{\gamma}) = \mathrm{E}[\mathbf{U}\mathbf{U}'] = X'AM'\Delta V^{-1}\Delta MAX = X'DV^{-1}D'X, \tag{2.12}$$

the overall information being the sum of these terms. Given $\mathbf{U}$ and $\mathcal{I}$, the model may be fitted by Fisher scoring. During the iterations, Azzalini's approximation might be used in place of $\mathcal{I}$ (see Section 1.4.5).

Polytomous data

In principle the extension of the above to polytomous data is straightforward, but there is risk of an explosion of parameters. Keeping the same conventions as before (Section 2.2.2), k_i values for Y_i are indexed $0, 1, 2, \ldots, (k_i - 1)$, and using subscripts for the variables in question, and superscripts for their values, denote the marginal expectations, more easily viewed as marginal probabilities, as

$$\nu_{ij\cdots k}^{rs\cdots t} = \mathrm{P}(Y_i = r, Y_j = s, \ldots, Y_k = t). \tag{2.13}$$

By natural extension of the ideas of Ekholm *et al.* define a dependence ratio for polytomous data as

$$\tau_{ij\cdots k}^{rs\cdots t} = \frac{\nu_{ij\cdots k}^{rs\cdots t}}{\nu_i^r \nu_j^s \cdots \nu_k^t}, \tag{2.14}$$

with the convention for univariate terms

$$\tau_i^r = \nu_i^r = \mathrm{P}(Y_i = r). \tag{2.15}$$

In the binary case all superscripts are unity and may be dropped.

The score equations are essentially identical to the binary form (2.9) developed above, except that wherever ν_1 or some such term appears we now substitute the vector

$$\boldsymbol{\nu}_1 = (\nu_1^1, \nu_1^2, \ldots, \nu_1^{k_1-1})'.$$

This adaptation is straightforward but tedious to write out in full.

The dependence ratios defined above can be adapted to deal with ordinal, as distinct from nominal, data. If we replace the essentially unordered $\boldsymbol{\nu}$ by an ordered set of cumulative probabilities

$$\upsilon_{ij\cdots k}^{rs\cdots t} = \mathrm{P}(Y_i \leq r, Y_j \leq s, \ldots, Y_k \leq t), \tag{2.16}$$

we can define a set of "cumulative dependence ratios"

$$\kappa_{ij\cdots k}^{rs\cdots t} = \frac{\upsilon_{ij\cdots k}^{rs\cdots t}}{\upsilon_i^r \upsilon_j^s \cdots \upsilon_k^t}. \tag{2.17}$$

Any of the standard cumulative link models can be used to ensure the ordering over the univariate logits; unlike nominal data one can foresee problems in specifying the necessary constraints if unconstrained ratios for the higher orders are used.

2.3.2 All marginal odds ratios

Although the dependence ratios of the above discussion are arguably easier to interpret, marginal logit and pairwise marginal odds ratios are the most commonly quoted measures of location and association throughout the literature. Thus, we consider fully marginally linked models of the form

$$\boldsymbol{\lambda}_u = X_u \boldsymbol{\gamma}, \tag{2.18}$$

where $\boldsymbol{\lambda}_u$ is the full set of marginal log odds ratios described in Section 2.2.1, and X_u is the design matrix, for the uth subject. This model allows fitted values of $\boldsymbol{\lambda}$ to lie

anywhere on the real line, and so fails to incorporate the constraints on $\boldsymbol{\lambda}$ that ensure a valid probability distribution (see Section 3.2.6). This procedure is followed in the hope that extreme cases such as the trivariate binary

$$\left.\begin{aligned} \lambda_1 = \lambda_2 = \lambda_3 = 0.5 \\ \lambda_{12} = 0.5,\ \lambda_{13} = 1,\ \lambda_{23} = 20 \end{aligned}\right\}, \tag{2.19}$$

for which no probability table exits, will not arise in practice nor will occur as intermediate values in an iterative fitting process.

Although model (2.18) is conceptually tidy, the high-order ratios involved are very difficult to interpret in any intuitive sense; a third-order odds ratio is often baffling to the non-specialist, and one needs to be able to clearly visualize a hypercube in order to readily understand the effect of varying values of still higher order ratios. The argument in favour of this type of model is not in the ease of interpretation, but in reproducibility (defined in Section 1.4.3), assuming any missing data are missing at random (defined in Section 6.2). Dependence-ratio models share this feature, but the ubiquity of the odds ratio as the measure of first choice makes these less attractive. One might posit a mixed model, linking to first- and second-order odds ratios and then dependence ratios for higher orders, but such a model is likely to be algorithmically unattractive (*cf* Section 2.5.2), and one might prefer instead the mixed model of Fitzmaurice and Laird (1993), discussed in Section 2.5.

The conceptual problem of interpreting high-order dependencies, however specified, remains unresolved. There appears to be no easy way to present such results, although marginal odds ratios may be the easiest to explain. Another unresolved problem is that high-order dependencies must be modelled from high-order sub-tables, which increasingly contain small, even zero, observed values with increasing T, unless the sample size is very large. Furthermore, if our model of choice proposes common odds ratios, say $\lambda_{ij} = \lambda_{kl} \quad \forall i, j, k, l$, it is known that maximum likelihood estimation of such common ratios is fragile; statistics such as the conditional ratios of Mantel–Haenszel are designed to overcome this problem, but are not maximum likelihood estimates. In the spirit of the existing literature on marginal modelling of binary data, I ignore

these limitations and discuss maximum likelihood estimation.

The score equations are now derived in two ways. The first method is the most common in the literature, but does not readily generalize to arbitrary dimension, as now shown.

The score equations (standard formulation)

For polytomous data with links (2.18) and likelihood as in the previous section (equation 2.8), the score contribution for the uth subject is, by the chain rule,

$$\mathbf{U}(\gamma) = \frac{\partial \ell}{\partial \gamma} = \frac{\partial \lambda'}{\partial \gamma}\frac{\partial \nu'}{\partial \lambda}\frac{\partial \xi'}{\partial \nu}\frac{\partial \ell}{\partial \xi} = X'DV^{-1}(\mathbf{z} - \nu), \tag{2.20}$$

where the derivation of all terms except D has been given in the derivation of the dependence-ratio model (Section 2.3.1).

The form of D. D is (block) diagonal for the first-order terms, because by definition each λ_i is a function of the corresponding ν_i only (and vice versa). For logit $\nu_i = \lambda_i$, we have established in (2.11) that

$$\frac{\partial \nu_i}{\partial \lambda_i} = \nu_i(1 - \nu_i) = \mathrm{var}(Z_i).$$

The completion of D for higher orders is considerably more complicated and several approaches are explored. Anticipating the discussion of unbalanced data in Section 5.2, let us use the notion of subdistributions introduced in Section 1.4.3: for a sub-observation $\mathbf{y}^{\mathcal{A}}$, reproducibility of the marginal polynomial exponential family distributions ensures that the subdistribution is again in the polynomial exponential family, of order the number of subscripts in $\mathcal{A}$. We may write this in the canonical form

$$\mathrm{P}(\mathbf{Y}^{\mathcal{A}} = \mathbf{y}^{\mathcal{A}}) = \exp\left\{(\xi^{\mathcal{A}})'\mathbf{z}^{\mathcal{A}} - C^{\mathcal{A}}(\xi^{\mathcal{A}})\right\};$$

$\xi^{\mathcal{A}} \neq \xi$ of the full distribution and the $\xi^{\mathcal{A}}$ are zero-conditional log odds ratios of the subdistribution; $\mathbf{x}^{\mathcal{A}}$ is the vector of observed $\mathbf{y}^{\mathcal{A}}$ and their pairwise crossproducts.

The marginal expectations of this subdistribution are the ordinary marginal expectations of the full distribution (by reproducibility); in particular,

$$\nu_{\mathcal{A}} = \nu_{\mathcal{A}}^{\mathcal{A}} = \sum_{\text{all } Y^{\mathcal{A}}} z_{\mathcal{A}}^{\mathcal{A}} \exp\left\{(\boldsymbol{\xi}^{\mathcal{A}})'\mathbf{z}^{\mathcal{A}} - C^{\mathcal{A}}(\boldsymbol{\xi}^{\mathcal{A}})\right\}$$

so that

$$\frac{\partial \nu_{\mathcal{A}}}{\partial \xi_{\mathcal{A}}^{\mathcal{A}}} = \text{var}(Z_{\mathcal{A}}^{\mathcal{A}}) = \text{var}(Z_{\mathcal{A}})$$

again by reproducibility of marginal moments, and similarly for $\mathcal{B}$ a proper subset of $\mathcal{A}$

$$\frac{\partial \nu_{\mathcal{A}}}{\partial \boldsymbol{\xi}_{\mathcal{B}}} = \nu_{\mathcal{A}}(1 - \nu_{\mathcal{B}}) = \text{cov}(Z_{\mathcal{A}}, Z_{\mathcal{B}}).$$

Although $\xi_{\mathcal{A}}^{\mathcal{A}} \equiv \lambda_{\mathcal{A}}$, $\partial \nu_{\mathcal{A}}/\partial \xi_{\mathcal{A}}^{\mathcal{A}}$ and $\partial \nu_{\mathcal{A}}/\partial \lambda_{\mathcal{A}}$ are not equal in general: the first statement is that the full-order log *zero-conditional* odds ratio (of the subdistribution) is equivalent to the log *marginal* odds ratio, because there are no variables left to condition on for the conditional ratio. (In an earlier version I presumed that the partial derivatives equated.) Holding the $\xi_{\mathcal{R}}^{\mathcal{A}}$ fixed for the differentiation is not equivalent to holding the $\lambda_{\mathcal{R}}$ fixed, where $\mathcal{R}$ indexes the remaining elements of the vector. Thus, this approach is discontinued and an alternative sought.

Liang *et al.* (1992) claim that D is available 'routinely', while Molenberghs and Lesaffre (1994) are more explicit in indicating that the terms may be found by implicit differentiation. This does not illustrate the scale of the problem for data with larger than trivariate observations. The implicit differentiation is suggested by the comparative ease with which the marginal odds ratios may be written in terms of the marginal expectations, that is

$$\Lambda_{\mathcal{A}} = \Lambda_{\mathcal{A}}(\nu_{\mathcal{B}} \,|\, \mathcal{B} \subseteq \mathcal{A}). \tag{2.21}$$

Such expressions may, in theory, be inverted to yield $\nu_{\mathcal{A}}$ in terms of $\lambda_{\mathcal{B}}$ and $\lambda_{\mathcal{A}}$, or equivalently in terms of $\nu_{\mathcal{B}}$ and $\lambda_{\mathcal{A}}$, for $\mathcal{B}$ the proper subsets of $\mathcal{A}$. These are polynomial expressions of order 2^a, where a is the number of elements of $\mathcal{A}$. Simplifying assumptions may be made when $T \leq 3$ (Liang *et al.*, 1992) that make such expressions tractable, but the general, full model cannot plausibly be expressed in this way; see

Section 3.2.

We might attempt to work from

$$\frac{\partial \boldsymbol{\Lambda}'}{\partial \boldsymbol{\lambda}} = \frac{\partial \boldsymbol{\nu}'}{\partial \boldsymbol{\lambda}} \frac{\partial \boldsymbol{\Lambda}'}{\partial \boldsymbol{\nu}}, \tag{2.22}$$

where the left-hand side is simply $\mathrm{diag}(\boldsymbol{\Lambda})$ and $E = \partial \boldsymbol{\Lambda}'/\partial \boldsymbol{\nu}$ can be expressed algebraically possibly using a computer program such as Maple. Providing the matrix E is invertible we then have

$$D = \frac{\partial \boldsymbol{\nu}'}{\partial \boldsymbol{\lambda}} = \mathrm{diag}(\Lambda) E^{-1}. \tag{2.23}$$

This method requires hard coding of the fitting routine for each possible problem dimension, which is not attractive, apart from being likely to be very inefficient in higher dimensions.

In search of a better method, we may consider the following. For $\mathcal{T}$ indexing the full-order interaction,

$$\frac{\partial \Lambda_{\mathcal{T}}}{\partial \boldsymbol{\nu}} \equiv \frac{\partial \xi_{\mathcal{T}}}{\partial \boldsymbol{\nu}};$$

this is *not* true for lower orders. For two sets of indices $\mathcal{A}$ and $\mathcal{B}$,

$$\frac{\partial \nu_{\mathcal{A}}}{\partial \lambda_{\mathcal{B}}} = 0 \quad \text{if } \mathcal{B} \not\subseteq \mathcal{A} \tag{2.24}$$

because each $\nu_{\mathcal{A}}$ is a function of (only) those $\lambda_{\mathcal{B}}$ for $\mathcal{B} \subseteq \mathcal{A}$. Thus, in theory, we can save a considerable amount of time by observing that D is upper triangular, with further explicit zero entries in the upper half. Consider

$$\frac{\partial \Lambda_{\mathcal{T}}}{\partial \boldsymbol{\lambda}} = \frac{\partial \boldsymbol{\nu}'}{\partial \boldsymbol{\lambda}} \frac{\partial \Lambda_{\mathcal{T}}}{\partial \boldsymbol{\nu}} \tag{2.25}$$

that is,

$$\mathbf{e}_{2^T-1} = D\mathbf{a}, \tag{2.26}$$

where the vector $\mathbf{e}$ is zero except for the last $[(2^T - 1)\text{st}]$ element, which is unity. The

vector **a** is obtained from

$$\frac{\partial \Lambda_{\mathcal{T}}}{\partial \boldsymbol{\nu}} = \frac{\partial \xi_{\mathcal{T}}}{\partial \boldsymbol{\nu}} = \frac{\partial \boldsymbol{\xi}'}{\partial \boldsymbol{\nu}} \frac{\partial \xi_{\mathcal{T}}}{\partial \boldsymbol{\xi}} = V^{-1}\mathbf{b}$$

where **b** is zero except for the last element, which is $\chi_{\mathcal{T}}$. This term cancels with the equivalent term $\Lambda_{\mathcal{T}}$ on the left-hand side of (2.25) to leave **e** as above. In other words, **a** is the last column of the inverse of the variance matrix, V.

This system is insufficient to determine D as some entries must be calculated before we can complete with back substitution. Consider the bivariate case: the system is explicitly

$$\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{\partial \nu_1}{\partial \lambda_1} & 0 & \frac{\partial \nu_{12}}{\partial \lambda_1} \\ 0 & \frac{\partial \nu_2}{\partial \lambda_2} & \frac{\partial \nu_{12}}{\partial \lambda_2} \\ 0 & 0 & \frac{\partial \nu_{12}}{\partial \lambda_{12}} \end{pmatrix} \begin{pmatrix} V^{13} \\ V^{23} \\ V^{33} \end{pmatrix}, \tag{2.27}$$

where V^{ij} is the (i,j)th element of V^{-1}. We substitute the terms otherwise obtainable into the first two entries of the diagonal, before obtaining

$$\begin{aligned} \frac{\partial \nu_{12}}{\partial \lambda_1} &= \nu_1(1-\nu_1)\left(-\frac{V^{13}}{V^{33}}\right) \\ \frac{\partial \nu_{12}}{\partial \lambda_2} &= \nu_2(1-\nu_2)\left(-\frac{V^{23}}{V^{33}}\right) \\ \frac{\partial \nu_{12}}{\partial \lambda_{12}} &= \frac{1}{V^{33}} \end{aligned}$$

The trivariate case raises a new problem, the necessity for an exogenous explicit form for the diagonal terms $\partial\nu_{12}/\partial\lambda_{12}$, etc. (These may perhaps be supplied by the answer to solving each bivariate system.)

As this approach is leading nowhere, I consider an alternative expression for the likelihood used in this context by Glonek and McCullagh (1995).

The score equations (alternative formulation)

Rather than consider the observed data for each subject as a vector of outcomes $\mathbf{y} = (y_1, y_2, \ldots, y_T)'$, consider the outcome as a single multinomial observation based on the cells of the joint probability table, $\boldsymbol{\pi}$. Although still interested in the marginal means and odds ratios, we obtain score equations from a different starting point. Denote the multinomial observation by the vector $\mathbf{m}$, of length 2^T (for binary outcomes), but write an element of this as $m_{\mathcal{A}}$ using the set notation as above: in the bivariate case, $m_{12} = 1$ represents an outcome of $y_1 = y_2 = 1$; the all-zero outcome has $m_0 = 1$. Note that one, and only one, element of $\mathbf{m}$ is unity and the rest are zero for each subject.

Denoting the logarithms of the probabilities $\boldsymbol{\pi}$ by $\mathbf{p}$, and using the same subscript notation where applicable, the likelihood contribution of the uth subject is

$$\ell_u = \mathbf{m}'\mathbf{p} = m_{\mathcal{A}} p_{\mathcal{A}} \tag{2.28}$$

for a particular outcome $m_{\mathcal{A}} = 1$. Then the score contribution, assuming fully marginal links, is

$$\frac{\partial \ell_u}{\partial \boldsymbol{\gamma}} = \frac{\partial \boldsymbol{\lambda}'}{\partial \boldsymbol{\gamma}} \frac{\partial \mathbf{p}'}{\partial \boldsymbol{\lambda}} \frac{\partial \ell_u}{\partial \mathbf{p}} \tag{2.29}$$

$$= X' \left(\frac{\partial s'}{\partial \mathbf{p}} \right)^{-1} \mathbf{m} \tag{2.30}$$

where $s(\mathbf{p}) = \boldsymbol{\lambda}$ is the transformation from log cell probabilities to log odds ratios defined in Section 3.1, with derivatives given in Section 3.3.1. We can, then, *relatively* easily obtain

$$\frac{\partial \boldsymbol{\lambda}'}{\partial \mathbf{p}} = \frac{\partial s'}{\partial \mathbf{p}},$$

which can be inverted provided the transformation is nonsingular and all the partial derivatives exist to yield $\partial \mathbf{p}'/\partial \boldsymbol{\lambda}$. This always holds in non-degenerate cases (Glonek and McCullagh, 1995). Unfortunately we cannot write $\mathbf{p} = s^{-1}(\boldsymbol{\lambda})$, which could be differentiated directly, in closed form in general (see Chapter 3). Nevertheless, given that we can obtain $\partial s'/\partial \mathbf{p}$ (Section 3.3.1), the formulation (2.30) does not suffer

the algorithmic difficulties of the standard form (2.20), and enables the fitting of fully marginal models, with log-odds-ratio links, in arbitrary dimension — at least in theory.

Calculation of ν or p from given λ

This transformation is potentially very demanding of computer time, at least when $T > 2$: the whole of Chapter 3 is devoted to this topic. The transformation is needed when evaluating $(\mathbf{z} - \boldsymbol{\nu})$, D and V^{-1} in (2.20) in the standard approach, and for evaluating $\mathbf{p}$, whence $\partial s'/\partial \mathbf{p}$, in the alternative form (2.30).

Information matrix and model fitting

The (expected) information matrix can be derived easily from the score function $\mathbf{U}(\boldsymbol{\gamma})$ in both approaches. In the standard approach, each subject contributes

$$\mathcal{I}(\boldsymbol{\gamma}) = \mathrm{E}[\mathbf{U}\mathbf{U}'] = X'DV^{-1}DX, \tag{2.31}$$

the overall information being the sum of these terms. In the alternative approach we formulate this as

$$\mathcal{I}(\boldsymbol{\gamma}) = \mathrm{E}[\mathbf{U}\mathbf{U}'] = X' \left[\frac{\partial s'}{\partial \mathbf{p}}\right]^{-1} \mathrm{E}[\mathbf{M}\mathbf{M}'] \left[\left(\frac{\partial s'}{\partial \mathbf{p}}\right)'\right]^{-1} X, \tag{2.32}$$

where $\mathrm{E}[\mathbf{M}\mathbf{M}'] = \mathrm{diag}(\boldsymbol{\pi})$ since

$$\mathrm{E}[\mathbf{M}\mathbf{M}'] = \sum_{m} \begin{pmatrix} m_0^2 & m_0 m_1 & \dots \\ m_0 m_1 & m_2^2 & \dots \\ \vdots & & \ddots \end{pmatrix} \prod_{\mathcal{A}} \pi_{\mathcal{A}}^{m_{\mathcal{A}}};$$

only one $m_{\mathcal{A}}$ is nonzero in each term of the expansion.

Given $\mathbf{U}$ and $\mathcal{I}$, the model may be fitted by Fisher scoring. In simulation studies, I have always been able to use the much simpler form $\mathcal{A} = \sum \mathbf{U}_u \mathbf{U}_u'$ in place of its

Table 2.1: Cerebrovascular deficiency data. Here outcome 0 represents abnormal reaction to treatment, and outcome 1 normal reaction.

	Outcomes for the (first, second) period			
Group	(0, 0)	(0, 1)	(1, 0)	(1, 1)
AB	6	0	6	22
BA	9	4	2	18

A for placebo, B for active drug

expectation, $\mathcal{I}$; the full form of $\mathcal{I}$ is used only after convergence, when reporting the parameter variance matrix.

Polytomous data

Except that there are more cell probabilities, the score and information matrix contributions are unchanged from (2.30) and (2.32), respectively, for unordered categories. Notation and algorithmic considerations are given in Section 3.8. Ordered categories are considered by Glonek and McCullagh (1995).

2.3.3 Examples

Example 1 — cerebrovascular deficiency

The following data are from a 2×2 crossover trial on cerebrovascular deficiency, quoted in Jones and Kenward (1989). This data set (Table 2.1) has been frequently used as an example although it has appeared in slightly corrupted form, discussed below.

Zhao and Prentice (1990) and Fitzmaurice and Laird (1993) report the same table but reverse the meaning of the outcome and so give parameter estimates that are opposite in sign to those quoted below. Zeger and Liang (1992) and Diggle *et al.* (1994) quote the same table but state that the codes are A for active and B for placebo. Although such discrepancies might be construed as rendering any interpretation essentially meaningless, the various formulations give the same answer to the

key question; do the treatment and/or period have any significant effect on outcome? As an exercise in variable selection, the true meaning of the codes is irrelevant, but that is *not* the clinician's view!

Example 1A The standard form of the marginal model for these data is as follows: with covariates x_1 for treatment assignment ($x_1 = 0$ if treatment A or $x_1 = 1$ if treatment B) and x_2 for period ($x_2 = 0$ if period 1 or $x_2 = 1$ if period 2)

$$\text{logit P}(Y_i = 1) = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2; \tag{2.33}$$

$$\log \text{MOR}_{12} = \alpha_{12}. \tag{2.34}$$

Instead we can code the covariates together as x_{1i}, denoting the treatment (formerly x_1) at timepoint i; then the first-order models are

$$\lambda_1 = \alpha_0 + \beta_1 x_{11}, \tag{2.35}$$

$$\lambda_2 = (\alpha_0 + \beta_2) + (\beta_1 + \beta_3) x_{12},. \tag{2.36}$$

Noting that by the design of the experiment $x_{12} = 1 - x_{11}$, model (2.36) becomes

$$\lambda_2 = (\alpha_0 + \sum \beta_i) - (\beta_1 + \beta_3) x_{11} = \alpha_2' + \beta_2' x_{11}, \tag{2.37}$$

which is now a "separate regression lines" formulation. The interpretation of this form of the model is so different that it is deferred to Example 1B.

Proceeding by standard analysis of deviance, the standard model reduces to

		SE.robust	SE.info
alpha_0	1.0788097	0.2807553	0.2807553
beta_1	-0.5600159	0.2356943	0.2356943
alpha_12	3.4011974	0.8316650	0.8316650

where `SE.robust` is obtained from the 'sandwich' variance estimator (Section 1.4.5). Here these and the information-based standard error estimates are identical, which occurs because this model is saturated (for data collapsed to ignore period).

Interestingly, the period parameter, β_2, has been dropped in the selection procedure,

though it is odd to consider a model ignoring period when the data are longitudinal and when there is dependence over time. This might explain why the period-ignored model is not advocated in previous reports. Here, the time dependence is modelled as $\lambda_{12} = \alpha_{12} \neq 0$, rather than in the first-order model.

Rather more worrying is the very different overall interpretation that results from fitting the alternative parametrization introduced above and now described.

Example 1B To avoid any confusion with the parameters α and β of the previous description, rewrite the links as

$$\lambda_1 = a_1 + b_1 x_{11}, \tag{2.38}$$
$$\lambda_2 = a_2 + b_2 x_{11}, \tag{2.39}$$
$$\lambda_{12} = a_{12}, \tag{2.40}$$

where again x_{11} is treatment at period 1 (coded zero for treatment A).

Again proceeding by analysis of deviance, we reduce the model to

		SE.robust	SE.info
a_1	1.5694579	0.4121230	0.4197657
b_1	-1.1680850	0.4564135	0.4605571
a_2	0.6486954	0.2573044	0.2573044
a_12	3.5378033	0.8099391	0.8626233

that is, b_2 (only) is dropped. Now from the (a, b) parameters, we infer that outcome in the second period is independent of treatment — or more precisely: cannot be shown to be otherwise — thus period 1 and 2 effects are modelled differently. This accords with a glance at the following table for outcome equal to 1 (normal reaction to treatment):

	Period	
Group	1	2
AB	28	22
BA	20	22

This inference contrasts with the inferences of all previous analyses, in which there is

a single model for observation mean, always depending on treatment. That inference is preferred on intuitive grounds, and the model is simpler. Unfortunately, the two models are not nested, though the similarity of the evaluated likelihoods suggests that a formal test for preference is unlikely to be conclusive.

The (a, b) model aliases treatment effect, which is of prime clinical interest, as period effect, which is of secondary (if any) medical interest. The clinically important inference if the (a, b) model is preferred is that participation in the study probably induced a propensity to common outcome independent of treatment, although treatment appears to have an effect on first-timers.

How could such different conclusions be obtained by seemingly the same method? In fact the (a, b) model finally selected corresponds to dropping the original β_3, but then instead of dropping either β_1 or β_2 altogether we drop β_1 from the time-2 model *only*. This is not the sort of choice we make with the standard selection procedures for univariate generalized linear models. This approach is effectively a relabelling of the parameters with the corresponding deviance change assessed on one degree of freedom, while the model still has the same number of parameters as in the original formulation. It may be noted that dropping β_1 from the time-2 model only induces a much smaller change in deviance than does dropping β_2, or dropping β_1 altogether. This could be considered a precautionary tale against "black box" approaches to variable selection, most of which would advocate dropping the parameter giving the least change in the deviance.

Example 2 — the 6 cities data

This second example is again a familiar one in the literature, and being a example of a 4-wave discrete-outcome longitudinal study, provides an example of the feasibility of the proposed algorithms for problems of reasonably high dimension. Another fully marginal analysis was recently presented by Glonek and McCullagh (1995).

The Six Cities data set represents a repeated binary response: yearly wheezing status (yes/no) on a cohort of children aged 7 years at the start of the study. As in the previous example I am informed (G. Fitzmaurice, personal communication) that any

Table 2.2: Six cities wheeze data. N represents no reported wheeze.

No maternal smoking					Maternal smoking				
			Age 10					Age 10	
Age 7	Age 8	Age 9	N	Y	Age 7	Age 8	Age 9	N	Y
N	N	N	237	10	N	N	N	118	6
		Y	15	4			Y	8	2
	Y	N	16	2		Y	N	11	1
		Y	7	3			Y	6	4
Y	N	N	24	3	Y	N	N	7	3
		Y	3	2			Y	3	1
	Y	N	6	2		Y	N	4	2
		Y	5	11			Y	4	7

clinical implications drawn from an analysis of the published data set are to be taken with extreme caution, since the available data represent an idealized situation in which a truly ordinal or even continuous outcome (wheeze) is recorded only as a dichotomy, only the 537 complete-record cases are presented and a key explanatory variable (mother's smoking) is treated as fixed although it is known to vary over time. Thus, the primary purpose of the following discussion is to compare methodology and *potential* differences in inferences.

The data set on the presence of wheeze at each age, and the explanatory variable of interest, maternal smoking, are summarized in Table 2.2 (Zeger *et al.*, 1988; Fitzmaurice and Laird, 1993).

Fitzmaurice and Laird (1993) present the analysis of a so-called 'saturated' model, in which there is saturation for the mean, with no constraints on the higher-order interactions. This model is not intercept-unconstrained in my definition (Section 1.4.4), because it has an intercept parameter common to the models for all four timepoints, which allows for a linear trend with age/timepoint, but is not an unrestricted form such as if age were a factor. However for comparison and as a starting point for backwards selection this model is fitted: in the current notation

$$\lambda_t = \alpha_0 + \beta_1(\text{age}) + \beta_2(\text{smokes}) + \beta_3(\text{age.smokes}), \quad t = 1, 2, 3, 4$$

$$\lambda_{\mathcal{A}} = \alpha_{\mathcal{A}}, \quad \mathcal{A} = \{1,2\}, \ldots, \{1,2,3,4\}.$$

Here age is taken as age minus 9 years (as in previous analyses in the literature). This centring is important unless starting values are very carefully chosen, since linear predictor values in excess of ± 3 give cell probabilities of zero or one to four decimal places, so that poor starting values or intermediate values will lead to failure of convergence for an apparently degenerate distribution.

The obtained fits for this 'saturated' model are as follows:

	estimated	SE.robust	SE.info
alpha_0	-1.90684196	0.11902508	0.11836386
beta_1	-0.16350163	0.05587540	0.05686830
beta_2	0.30776342	0.18798019	0.18890268
beta_3	0.08491753	0.08780011	0.08856262
alpha_12	2.00304438	0.26251198	0.26096563
alpha_13	1.74963235	0.26965664	0.26729976
alpha_14	2.07459385	0.26979057	0.27922043
alpha_23	2.47007135	0.28886314	0.27943656
alpha_24	2.05560914	0.28003753	0.28210986
alpha_34	0.09296585	0.61997793	0.62697781
alpha_123	-0.27899195	0.61472962	0.60687325
alpha_124	2.08649992	0.28760370	0.28862929
alpha_134	-0.23515451	0.61756146	0.62223226
alpha_234	0.10296020	0.66072021	0.66221609
alpha_1234	0.12054453	1.41728650	1.42261997

The higher-order parameters cannot be directly compared to values given in previous literature since these sources did not use marginal odds ratios, except for Glonek and McCullagh (1995), who did not present the values for this model. Whereas Fitzmaurice and Laird (1993) found all 3rd and 4th order estimate z-tests (that is, estimate/standard error) suggestive that the parameters could be taken as zero, in the fully marginal formulation we find that α_{124} is apparently far from zero. A clinical interpretation of this result is not easy: the log odds ratio for the time-1 and time-2 table, collapsed over time-3 value, is twice as great when the time-4 outcome is 1 as when the time-4 outcome is 0 — but the clinical implications are not obvious.

Of the 2nd order parameters, the 'odd man out' is α_{34} in the marginal model above, whereas in the conditional form of Fitzmaurice and Laird (1993) all the pairwise ratios are roughly equal, except for the zero-conditional equivalent of α_{24}.

Of course inferences from univariate z-tests are always dubious since they are based on a possibly wildly inaccurate assumption that the joint confidence region is spherical, so potential simplifications of the above model were studied more formally by analysis of deviance. The full model has fitted log likelihood -793.1496, and for the model with $\alpha_{1234} = 0$ (see Appendix A2.3.3) the log likelihood is -793.1531. The model in which all 3rd-order ratios are set equal to zero has evaluated log likelihood -818.447, giving deviance change of approximately 50 on 4 degrees of freedom. From the z-tests, it is α_{124} that is an unlikely candidate to be dropped, but I suggest it is unrealistic to propose a model incorporating only one of the interactions of a certain order.

Goodness of fit can also be assessed as in Fitzmaurice and Laird (1993) by calculating

$$G^2 = 2\sum \left\{ \text{observed} \times \log \frac{\text{observed}}{\text{expected}} \right\}, \tag{2.41}$$

where summation is over all cells in the multinomial table. This gives $G^2 = 7.9$ on 15 d.f. for the 'saturated' model, $G^2 = 7.91$ on 16 d.f. when dropping α_{1234}, but $G^2 = 58.5$ on 20 d.f. for the model dropping 3-way interactions. As these values indicate, there is a sudden jump from very good to very poor predictions in the attempted simplification.

Fitzmaurice and Laird (1993) found that a very simple model fitted well in their parametrization: they were able to assume a common pairwise zero-conditional ratio and drop all higher-order interactions. They could argue that this highlights the advantage of their modelling approach, but it is possible to have data for which marginal odds ratios may be simply modelled while zero-conditional ratios may not (see Section 2.4.4). For comparison I studied the fit of models that are the marginal analogues of their zero-conditional formulations, such as a lag-one Markov structure (all interactions except $\alpha_{i(i+1)}$ set to zero), a common marginal pairwise ratio, and a full independence model (Appendix A2.3.3). None of these fitted sufficiently well by analysis of deviance or G^2 criteria. Although full details of these analyses are not given here, I report that most of the poorly fitting models were very slow to converge, and indeed the common-pairwise model failed to converge at all from any of several

attempted sets of starting values; problems arose specifically in finding the probability table for the maternal smoking group, and this was not even helped by saturating the pairwise ratio model to

$$\lambda_{\mathcal{A}} = \alpha_{\mathcal{A}} + \beta_{\mathcal{A}}(\text{smokes}).$$

The independence model converged in only 4 iterations from an all-zero start vector, despite being the worst fit. Convergence criteria are apparently linked with the conditioning of the odds-ratio-to-probability conversion problem dealt with at length in Chapter 3, but formal quantification of such criteria is not attempted here.

2.4 Use of the canonical link

Since the zero-conditional log odds ratios may take values over the whole of the real line, it is possible to model simply

$$\boldsymbol{\xi} = X\boldsymbol{\gamma}.$$

Although this model is not marginal, it is a full likelihood model and is introduced for comparison with the primarily marginal mixed models in Section 2.5. By a suitable choice of zeros in the design matrices $X_i = (\mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_{12\ldots T}')'$ the parameters $\boldsymbol{\gamma}$ may be effectively split into sets relating only to ratios of each particular order. The introduction of further notation to indicate such grouping of parameters would mar the simplicity of the theoretical presentation below, but in practice reparametrization might be helpful for ease of interpretation and for computational convenience.

The score contribution for subject u is readily obtained from (2.8):

$$\mathbf{U}_u(\boldsymbol{\gamma}) = \frac{\partial \ell_u}{\partial \boldsymbol{\gamma}'} = \frac{\partial \boldsymbol{\xi}_u'}{\partial \boldsymbol{\gamma}} \frac{\partial \ell_u}{\partial \boldsymbol{\xi}_u} = X_u'(\mathbf{z}_u - \boldsymbol{\nu}_u), \tag{2.42}$$

where $\boldsymbol{\nu}_u = \mathrm{E}[\mathbf{Z}_u]$ are the *marginal* expectations about the origin, as in the score function for marginal models in the previous section. This provides very much simpler score equations than previously considered either here or in the literature.

The negative derivative

$$\mathcal{I}(\gamma) = \sum_{u=1}^{n} X_u' \mathrm{cov}(\mathbf{Z}_u) X_u \tag{2.43}$$

is obtained by straightforward application of the chain rule, noting that $\partial \nu' / \partial \xi = \mathrm{cov}(\mathbf{Z})$, as derived in Section 2.3, or by considering $\sum \mathrm{E}[\mathbf{U}_u \mathbf{U}_u']$. The observed and expected information matrices are equal in this case. This scheme is good for *any* polynomial exponential family model with identity link, including models for continuous data and in fact even for mixed discrete and continuous outcomes.

2.4.1 Partitions of the design matrix

Care needs to be taken deriving the information matrix for partitioned design matrices; less calculations are saved than might be hoped at first. Let us write the design matrix as

$$X = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}$$

and similarly partition the linear model parameters as

$$\gamma' = (\gamma_{\mathrm{a}}', \gamma_{\mathrm{b}}')'$$

where, say, the γ_{a} are parameters in the linear predictors for the first-order logits, and the γ_{b} are linked to the higher-order log ratios. The score contribution of the uth subject, dropping the subscript u for convenience,

$$\mathbf{U}(\gamma) = \sum X'(\mathbf{z} - \nu)$$

partitions as

$$\begin{aligned} \mathbf{U}_{\mathrm{a}}(\gamma_{\mathrm{a}}) &= \sum A'(\mathbf{y} - \mu) \\ \mathbf{U}_{\mathrm{b}}(\gamma_{\mathrm{b}}) &= \sum B'(\mathbf{w} - \eta), \end{aligned}$$

where $\boldsymbol{\mu} = \mathrm{E}[\mathbf{Y}]$, $\boldsymbol{\eta} = \mathrm{E}[\mathbf{W}]$, and $\mathbf{z}$ is partitioned as $(\mathbf{y}', \mathbf{w}')'$. The derivative of each of these is easily derived but for brevity only that for $\boldsymbol{\gamma}_\mathrm{a}$ is given:

$$\frac{\partial \mathbf{U}'}{\partial \boldsymbol{\gamma}_\mathrm{a}} = \frac{\partial \boldsymbol{\xi}_\mathrm{a}'}{\partial \boldsymbol{\gamma}_\mathrm{a}} \frac{\partial \boldsymbol{\mu}'}{\partial \boldsymbol{\xi}_\mathrm{a}} \frac{\partial \mathbf{U}'}{\partial \boldsymbol{\mu}} = -\sum A'(V_{\mathrm{aa}} \quad V_{\mathrm{ab}}) \begin{pmatrix} A \\ 0 \end{pmatrix} = -\sum A' V_{\mathrm{aa}} A \tag{2.44}$$

where the V_{ab} are the obvious conforming partitions of the full variance matrix V. If one followed Fitzmaurice and Laird (1993) rather too blindly, it might seem that the Newton–Raphson parameter updates could be obtained from

$$\boldsymbol{\gamma}_\mathrm{a}^{(s+1)} = \boldsymbol{\gamma}_\mathrm{a}^{(s)} + \left(\sum A' V_{\mathrm{aa}} A\right)^{-1} \left(\sum A'(\mathbf{y} - \boldsymbol{\mu})\right) \tag{2.45}$$

with an analogous formula for the $\boldsymbol{\gamma}_\mathrm{b}$ updates. But this is quite wrong. The parameter sets $\boldsymbol{\gamma}_\mathrm{a}$ and $\boldsymbol{\gamma}_\mathrm{b}$ are not orthogonal, unlike their analogues in the Fitzmaurice and Lard model described in Section 2.5. While the marginally-linked and conditionally-linked parameters are orthogonal, higher-order parameters within either set are not orthogonal to lower-order parameters. In fact the information matrix here is

$$\mathcal{I}(\boldsymbol{\gamma}) = \sum \begin{pmatrix} A' V_{\mathrm{aa}} A & A' V_{\mathrm{ab}} B \\ B' V_{\mathrm{ab}} A & B' V_{\mathrm{bb}} B \end{pmatrix}$$

which is certainly not block diagonal.

Only having to invert this $\mathcal{I}(\boldsymbol{\gamma})$, once per Fisher scoring step, is better than having to also invert the full variance matrix V, for every covariate pattern within each scoring step, for the fully marginal model. This becomes of increasing importance for larger outcome vectors: $\mathcal{I}(\boldsymbol{\gamma})$ is number-of-parameters square, whereas V for a full interaction model is $k^T - 1$ square, where k is the number of categories, and T is the number of timepoints.

The incorrect form (2.45) corresponds very closely to using GEE1 (in the form of Liang *et al.*, 1992, equation 12) instead of GEET; this is equivalent to assuming a

linear exponential model instead of a polynomial model of order T. Whereas if GEE1 is used instead of GEET the penalty is only in lack of efficiency for the γ_a estimates, if one attempts to fit a conditional odds-ratio model using (2.45) it will generally fail to converge.

At a fundamental level, zero-conditional odds ratio models are worse affected by variance mis-specification than are fully marginal models, which in turn are worse affected than are mixed-marginal models. One should consider carefully before imposing constraints on $\boldsymbol{\xi}$ in a fully zero-conditional model.

2.4.2 Advantages of all-canonical links

This model is extremely easy to fit (using, say, Newton–Raphson) in comparison with other approaches discussed here, especially when any natural partitioning of the design matrix can be exploited. Importantly, no matrix inversion is required when calculating the score function, $\mathbf{U}$, whereas whenever any part of the model is marginally linked, inversion of at least part of the dispersion matrix V is needed. This is not true for the information matrix $\mathcal{I}$, but this is a minor problem by comparison, as already shown; $\mathcal{I}$ needs to be inverted only once for each Newton–Raphson iteration, unlike the inversion of V that is needed to calculate *each* subject's contribution to the score in marginal and mixed models.

It is still necessary to calculate the marginal expectations from the conditional odds ratios, but this is easier (and quicker) than for fully marginal or Fitzmaurice and Laird (1993) models; an algorithm is described in Section 2.4.5.

2.4.3 Disadvantages of all-canonical links

None of the estimated parameters $\boldsymbol{\gamma}$ have the easy and familiar immediate interpretation offered by those describing first- and second-order marginal odds ratios (although some do have interpretation in terms of conditional independencies). This is a smaller problem for higher-order ratios where interpretation of either variety of ratio is beyond ordinary intuition. However, the canonical-link model suffers from having *none* of its linear predictors immediately interpretable. Even though there is a useful interpreta-

tion of zero-conditional odds ratios for certain types of multivariate problems, with longitudinal data it is counterintuitive to seek to model a log odds ratio given that all *future* observations are zero. Such models are particularly poor when the aim of the modelling exercise is the prediction of outcomes; but this particular criticism could be made of any model that does not feature history conditioning.

Another potentially quite serious drawback of the canonically-linked model is that conditional odds-ratio models, including the Fitzmaurice and Laird hybrid model discussed in Section 2.5, are not reproducible (Section 1.4.3). Thus the formulation as given above can only be used directly if all subjects have the same number of observations. Modifications to handle unbalanced designs and missing values are considered in Section 5.2.

These models are extremely sensitive to mis-specification of the dispersion structure and, when a poor model is chosen, fitting may be numerically impossible. Given the simplicity and speed of fitting, one might always fit to unconstrained $\boldsymbol{\xi}$, with the interaction terms regarded as mere nuisance terms.

2.4.4 Marginal inference from zero-conditional fits

Consider the analogous fully marginal model $\boldsymbol{\lambda} = X\boldsymbol{\gamma}_{\mathrm{M}}$ and zero-conditional model $\boldsymbol{\xi} = X\boldsymbol{\gamma}_{\mathrm{C}}$, where analogous models are those with identical design matrices X. Suppose that the sole or primary object of the analysis is to assess whether some group or continuous covariate, x_1 say, has a significant effect on outcome.

It is interesting to conjecture whether such a covariate effect, judged as significant according to an analysis of deviance on stepwise selection of the simpler conditional logit model, will necessarily remain significant as assessed by the same analysis of the analogous marginal model. The conjecture is not that analogous models fit equally well, but rather that the finally selected model includes the covariate (or not) according to whether it is significant (or not) irrespective of the type of model, marginal or conditional. In fact the simplistic idea that analogous models are likely to be of much practical use is wrong, and the following counterexamples indicate a better approach than analogous models to the question of covariate dependency.

Suppose we observe the following cell counts in a simple comparison of binary response pattern between two groups coded by the dummy variable x_1:

	$x_1 = 0$		$x_1 = 1$		Pooled	
	$y_2 = 0$	$y_2 = 1$	$y_2 = 0$	$y_2 = 1$	$y_2 = 0$	$y_2 = 1$
$y_1 = 0$:	5	5	5	5	10	10
$y_1 = 1$:	5	5	5	20	10	25

Here the zero-conditional first-order ratios (logits) are equal (specifically, $\xi_1 = \xi_2 = 0$) and do not depend on group: the parameter estimate of β_1 in the model

$$\xi_i = \alpha_i + \beta_1 x_1$$

can be taken as zero. Thus group effect will be dropped from the first-order conditional model. The maximum likelihood fits match exactly those of the observed, suitably collapsed tables.

However, in the marginal formulation we find $\lambda_1 = \lambda_2 = 0$ within the group $x_1 = 0$ while $\lambda_1 = \lambda_2 = \log\ 25/10$ in the other group. The analogous model to that above,

$$\lambda_i = \alpha_i + \beta_1 x_1,$$

has point estimate $\hat{\beta}_1 = \log\ 2.5$ and this parameter would not be dropped by stepwise or backwards selection; assume this for the purposes of illustration, although sample size would have to be larger to be entirely sure of it.

Thus while group effect disappears from the first-order model in the conditional formulation, it would not be dropped from the model for the pairwise conditional logs odds ratios, which are 0 and $\log 4$, respectively, for the first and second groups, but $\log 2.5$ for the collapsed table. That is, x_1 contributes significantly to a model that adequately models the observed dispersion structure, though it will not appear in the model for the first-order zero-conditional ratios. Conversely, once x_1 is included in the model for the marginal logits, it need not necessarily appear in the marginal interaction model (although in this particular example it will appear here too, since the marginal and zero-conditional ratios are equivalent for bivariate data).

It may not seem immediately apparent why one should want to fit a zero-conditional model at all if one is interested only in marginal inference. The attraction is the speed of the fit, and that for reasonably large numbers of observations the marginal algorithm may not be numerically feasible. Consider the transform from fitted zero-conditional $\hat{\boldsymbol{\xi}} = X\hat{\boldsymbol{\gamma}}_{\mathrm{C}}$ using $\hat{\boldsymbol{\pi}} = T(\hat{\boldsymbol{\xi}})$, the odds ratio to cell probability transform discussed in Section 2.4.5, to $\hat{\boldsymbol{\lambda}} = s(\hat{\mathbf{p}})$, the further transform from log cell probabilities to log marginal odds ratios (Section 3.1.5). We can then fit

$$\hat{\boldsymbol{\lambda}} = X\boldsymbol{\gamma}_{\mathrm{M}}$$

by, say, ordinary least squares, to obtain estimates of the marginal parameters $\boldsymbol{\gamma}_{\mathrm{M}}$. Here X is assumed to be the same for both models, although this might be relaxed provided the transformation remains one to one. In view of the appearance of covariate effects at different levels of the zero-conditional and marginal fits illustrated above, X should be at least intercept-unconstrained.

Because of the complexity and nonlinearity of the transform from $\hat{\boldsymbol{\xi}}$ to $\hat{\boldsymbol{\lambda}}$ the form of confidence regions for the marginal parameter estimates $\boldsymbol{\gamma}_{\mathrm{M}}$ is hard to assess. This problem is not addressed here in detail, but considerations of whether or not to include a covariate will have been dealt with at the zero-conditional stage; sometimes only pointwise marginal estimates are needed. Alternatively, we might obtain a crude interval by studying the range of values in the marginal interpretation as we vary the zero-conditional $\boldsymbol{\gamma}_{\mathrm{C}}$ between the extremes of the zero-conditional confidence intervals.

2.4.5 Calculating probabilities from conditional odds ratios

In evaluating the score equations (2.42) we require marginal expectations, $\boldsymbol{\nu}$, but only have direct estimates of the zero-conditional parameters $\boldsymbol{\xi} = X\boldsymbol{\gamma}$. Following the approach of Fitzmaurice and Laird (1993), we first find the cell probabilities, $\boldsymbol{\pi}$, and from these read off the marginal expectations.

Shortcuts may be taken when, for example, all the high-order $\boldsymbol{\xi}$ are constrained to be zero, but a fully general algorithm is presented here.

The principle of the algorithm is straightforward, and simpler and faster than that mentioned in Fitzmaurice and Laird (1993) — see Section 2.5.1. Odds ratios with appropriate multipliers are entered into appropriate cells of the probability table, with a one in position $(0, 0, \ldots)$. The whole is then divided by the sum of all cells to give the desired probabilities.

I now present this technique as a formal algorithm and show, in Appendix A2.4.5, that it gives valid results in all dimensions. The fact that *marginal* odds ratios are subtly constrained (Section 3.2.6) suggests such proof is worth giving.

Denote a probability as, for example, $\pi_{10111} = \Pr(Y_1 = 1, Y_2 = 0, Y_3 = 1, Y_4 = 1, Y_5 = 1)$, and let $c_{\cdot}$ denote a cell in a table indexed as for the probability table. Further define a subscript mapping from a subset $\mathcal{B} \subseteq \mathcal{T} = \{1, 2, \ldots, T\}$ to cell indices $s(\mathcal{B})$: if $\mathcal{B} = \{i_1, i_2, \ldots, i_t\}$, then $s(\mathcal{B})$ is the binary code having zeros everywhere except for ones in positions $i_1, i_2, \ldots, i_t$. Recall that $\boldsymbol{\chi} = \exp \boldsymbol{\xi}$. Then we can find probabilities using the following algorithm:

A1 For all $\mathcal{B} \subseteq \mathcal{T}$ (including the empty set, yielding the empty product, unity, and for which we explicitly define $\chi_0 = 1$)

$$c_{s(\mathcal{B})} = \prod_{\mathcal{C} \subseteq \mathcal{B}} \chi_{\mathcal{C}}.$$

A2 Then again for all $\mathcal{B} \subseteq \mathcal{T}$

$$\pi_{s(\mathcal{B})} = c_{s(\mathcal{B})} / \sum_{\mathcal{C} \subseteq \mathcal{T}} c_{s(\mathcal{C})}.$$

The proof that this gives a correct probability table, given in Appendix A2.4.5, has two corollaries: firstly, the canonical parameters are indeed variationally independent, unlike their marginal counterparts, and secondly, each set of parameters $\boldsymbol{\xi}$ corresponds to a unique probability table.

2.5 Mixed parametrizations

The approach of Fitzmaurice and Laird (1993) includes marginal modelling and exploits the canonical parametrization. Linear models are fitted to the marginal logits, as in fully marginal or GEE models, but they use the identity link to the zero-conditional log odds ratios for higher-order interactions, under the general assumption that higher-order interactions are a nuisance rather than of intrinsic interest.

In theory one could set up a model that was marginal up to some higher order and conditional thereafter, but this may be almost as difficult to fit as a fully marginal model (see Section 2.5.2). Therefore here consideration is given only to the mixed parametrization of Fitzmaurice and Laird (1993).

To emphasise the two sets of parameters, write the probability function as

$$\mathrm{P}(\mathbf{y};\ \boldsymbol{\Psi}, \boldsymbol{\Omega}) = \exp\{\boldsymbol{\Psi}'\mathbf{y} + \boldsymbol{\Omega}'\mathbf{w} - C(\boldsymbol{\Psi}, \boldsymbol{\Omega})\}, \tag{2.46}$$

where $\mathbf{w}$ is a vector of crossproducts of elements of $\mathbf{y}$, the $\boldsymbol{\Psi}$ are the first-order zero-conditional ratios (that is, logits), and $\boldsymbol{\Omega}$ the remaining zero-conditional, canonical parameters. The first-order model is

$$g(\mu_i) = X_{\mathrm{M}}\boldsymbol{\gamma}_{\mathrm{M}}, \quad i = 1, 2, \ldots, T$$

for some link function $g()$, which is here assumed to be the logit, of the marginal univariate means. The model is completed by setting

$$\boldsymbol{\Omega} = X_{\mathrm{C}}\boldsymbol{\gamma}_{\mathrm{C}}.$$

Simplifications such as common odds ratios, or high-order interactions set to zero, can be considered.

Intrinsic to the Fitzmaurice and Laird method is the orthogonality of the parameter sets $\boldsymbol{\gamma}_{\mathrm{M}}$ and $\boldsymbol{\gamma}_{\mathrm{C}}$; the Fisher information matrix is block diagonal. Thus the score equations (and Fisher-scoring steps) decompose, giving score contribution from each

subject

$$\mathbf{U}(\gamma_{\mathrm{M}}) = X_{\mathrm{M}}'\Delta V_{\mathrm{MM}}^{-1}(\mathbf{y}-\boldsymbol{\mu}) \tag{2.47}$$

$$\mathbf{U}(\gamma_{\mathrm{C}}) = X_{\mathrm{C}}'[\mathbf{w}-\boldsymbol{\nu}-V_{\mathrm{CM}}V_{\mathrm{MM}}^{-1}(\mathbf{y}-\boldsymbol{\mu})] \tag{2.48}$$

where $\boldsymbol{\nu}$ denotes the marginal expectations about the origin of the crossproducts $\mathbf{W}$, Δ is a diagonal matrix of univariate variances, and the variance matrix V decomposes as follows:

$$\begin{pmatrix} V_{\mathrm{MM}} & V_{\mathrm{MC}} \\ V_{\mathrm{CM}} & V_{\mathrm{CC}} \end{pmatrix} = \begin{pmatrix} \mathrm{cov}(\mathbf{Y}) & \mathrm{cov}(\mathbf{Y},\mathbf{W}) \\ \mathrm{cov}(\mathbf{W},\mathbf{Y}) & \mathrm{cov}(\mathbf{W}) \end{pmatrix}.$$

In particular, only V_{MM}, a $T \times T$ matrix, needs to be inverted in evaluating each score contribution, as compared to a fully marginal model where all of V needs inverting each time (or indeed to a fully zero-conditional model, where nothing at all needs inverting during the evaluation of $\mathbf{U}$). Most of the computational burden is in calculating the marginal expectations $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ from the sets of odds ratios (Section 2.5.1).

A disadvantage noted by Fitzmaurice and Laird (1993) is common to all models using links to any of the canonical parameters: even though it is first-order marginal the model is not reproducible and so cannot deal with unbalanced designs and/or missing values without considerable modification (Fitzmaurice *et al.*, 1994). When the design *is* balanced and the model may be fitted, it is likely that inference will be most concerned with marginal means, with interactions merely nuisance. These are far more quickly found by the Fitzmaurice and Laird approach than by fully marginal approaches. Moreover, the orthogonality of the two sets of parameters means that estimates of the marginal parameters are scarcely less efficient even under extreme mis-specification of the conditional dispersion structure, which is certainly not true of either of the 'pure' parametrizations.

2.5.1 Calculating probabilities from mixed odds ratios

For the mixed-parameter method of Fitzmaurice and Laird (1993), one needs to generate the probability table from the univariate marginal expectations, μ_i, or their logits, λ_i, and the zero-conditional higher-order logits, $\xi_{\mathcal{A}}$, where $\mathcal{A}$ has at least two elements.

An extra stage of calculation is required compared with the scheme in Section 2.4.5. Firstly a table is filled such that it has the required higher-order ratios $\chi_{\mathcal{A}}$: in the notation of Fitzmaurice and Laird, this is called $S(\Omega)$. Secondly, iterative proportional fitting (IPF) is used to scale the table to conform to the first-order margins μ_i. In the process, on convergence the cell entries are assured to sum to unity, giving us the required probability table.

As is well known (e.g. Bishop *et al.*, 1975), IPF on first-order margins has the property that higher-order marginal ratios are not altered; higher-order *conditional* ratios can be shown to be similarly unaffected.

Filling $S(\Omega)$ to have the required conditional odds ratios is a simple adaptation of algorithm step A1 of Section 2.4.5. We simply temporarily set all the first-order χ_i equal to unity. Actually, we could choose any values, since these first-order ratios have no effect on the higher-order ratios being set in the table, as demonstrated in Section 2.4.5, but unity has the greatest computational advantage in that actual multiplication need not be performed.

The second stage, IPF of the first-order margins, replaces step A2 of Section 2.4.5. That step could be considered as IPF of the *zeroth-order* margin: namely, the probabilities sum to one.

The proof that this scheme gives the required unique probabilities is almost established by the proof in Appendix A2.4.5 that step A1 is correct for all dimensions T. The second step is not controversial except for the question as to whether there are cases for which the marginal μ_i are incompatible with the conditional $\chi_{\mathcal{A}}$. But this question is answered in Bishop *et al.* (1975) where it is shown that we can scale any contingency table, here $S(\Omega)$, to fit arbitrary first-order margins (because what Bishop *et al.* call

the configurations C_i — that is, the univariate marginal totals — which follow directly from the given μ_i, do not overlap, and the margins all sum to the same total, namely unity). Thus we can make the table conform to the given μ_i provided only these lie strictly between zero and one, as they must do if we use a suitable link function, such as the logit.

2.5.2 Calculating probabilities when some higher-order ratios are also marginal

In principal the mixed-ratios algorithm can be adapted to where we specify second-order (or even higher) ratios marginally. In the first stage, all the marginally specified ratios' counterparts, that is χ_i, χ_{ij}, etc., up to the required order, are set to unity, just as we set χ_0 and χ_i to unity above. In the IPF stage, we fit to the configurations specified by the marginal ratios.

It is this latter stage that raises problems. Firstly, we cannot guarantee convergence, because certain combinations of marginal odds ratios are incompatible. Because of the way in which they are specified, the overlapping configurations are consistent in the sense of Bishop *et al.* (1975), which is necessary but not sufficient for convergence. Secondly, in the models we are fitting, the configurations per se are not specified, but rather the set of odds ratios from which they are to be derived. One approach is to calculate the complete multivariate configurations (as marginal probability tables) from the given marginal ratios, although this is time-consuming (Chapter 3). However, once we had the complete configurations, standard IPF could proceed directly.

Another approach is to fit, iteratively, and in descending order of given interaction, to pseudo-configurations, which are marginal tables having only one of the required ratios each. Although they violate consistency, such tables appear to work well in practice (G. Molenberghs, private communication).

Given that neither method is concise or rapid, we might instead consider a different approach, and solve the nonlinear system $f(\pi) = (\Lambda', \chi')'$ using a suitably modified version of the iterative scheme proposed for the pure marginal problem itself (Chapter 3). Unfortunately, the symmetry of the fully marginal version is less apparent for

mixed parametrizations. We might here simply follow the lead of Glonek and McCullagh (1995) and use straightforward Newton–Raphson for suitably formulated $f(\pi)$, although as we will see in Chapter 3 this can be prohibitively slow.

2.6 GEE and related methods

It has become widely accepted that the generalized estimating equations (GEE) approach described below may be applied without regard to the true sample likelihood, whenever it is desired to fit a marginal model to longitudinal, or otherwise correlated, data. This is equivalent to modelling only the marginal mean and pairwise covariances (or odds ratios) for some multivariate outcome. To help understand the rationale of the GEE algorithm we first consider when the GEE2 equation gives the maximum likelihood solution to the marginal model problem (Zhao and Prentice, 1990; Zhao and Prentice, 1991).

2.6.1 The quadratic exponential assumption

In assuming all the three- and higher-way interactions in (1.24) or equivalently (2.46) are identically zero, we assume our observations are from a quadratic exponential family distribution, rather than from an order-T polynomial family. The score function is then

$$\mathbf{U}(\gamma) = \sum_u \frac{\partial \nu_u'}{\partial \gamma} \frac{\partial \xi_u'}{\partial \nu_u} \frac{\partial \ell_u}{\partial \xi_u} \tag{2.49}$$

for the model

$$\lambda_{\mathcal{A}} = \mathbf{x}_{\mathcal{A}}' \gamma_{\mathcal{A}}$$

as $\mathcal{A}$ runs through single and double indices only. By our assumptions the potential 3- and 4-way interaction terms in the dispersion matrix V, where

$$V^{-1} = \frac{\partial \xi'}{\partial \nu},$$

are taken to be zero, and so V^{-1} is assumed block diagonal, imposing orthogonality between the first- and second-order γ parameter estimates, provided of course they

are distinct.

Equation (2.49), restricted to first- and second-order ratios only, is precisely the GEE estimating equation proposed by Liang and Zeger (1986). Does the use of this equation for polytomous data really imply an assumption of simplified likelihood (to quadratic exponential) or do we accept the spirit of the original proposal and merely assert that whatever the true likelihood, the GEE score is Godambe-optimal? A further subtle claim is that the 3- and higher-way interactions are taken into a shape function which is then estimated non-parametrically.

If the full likelihood can be maximized, as for polytomous data for short series of observations, then that approach should be preferred, as advocated above. In the following review of the principal approaches to GEE modelling I concentrate on (first-order) marginal models, ignoring other approaches such as the latent polychoric covariance approach (Qu *et al.*, 1992; Qu *et al.*, 1995) and the use of additive (non-linear) links in analogues of generalized additive models (Yee and Wild, 1996; Wild and Yee, 1996).

2.6.2 Choice of parametrization

A fundamental design consideration is whether to link to (i) the mean and pairwise (marginal) odds ratios (Lipsitz *et al.*, 1991; Liang *et al.*, 1992) or (ii) the mean and pairwise covariances (Zhao and Prentice, 1990; Zhao and Prentice, 1991; Prentice and Zhao, 1991). Denoting the parameters of interest in the linear predictor for the mean as $\boldsymbol{\gamma}_{(1)}$ and those in the predictor for association as $\boldsymbol{\gamma}_{(\times)}$, we have functions

$$\begin{aligned} \boldsymbol{\mu} &= \boldsymbol{\mu}(\boldsymbol{\gamma}_{(1)}), \\ \boldsymbol{\eta} &= \boldsymbol{\eta}(\boldsymbol{\gamma}_{(1)}, \boldsymbol{\gamma}_{(\times)}) \end{aligned}$$

where $\boldsymbol{\mu}$ are univariate marginal means and $\boldsymbol{\eta}$ are the expectations about the origin of pairwise crossproducts, $\eta_{ij} = \mathrm{E}[Y_i Y_j]$. In the Prentice & Zhao case,

$$\eta_{ij} = \sigma_{ij} + \mu_i \mu_j,$$

where $\boldsymbol{\sigma} = \boldsymbol{\sigma}(\boldsymbol{\gamma}_{(\times)})$.

The score contribution of each subject, equation (2.49), is

$$\begin{pmatrix} \partial\ell/\partial\boldsymbol{\gamma}_{(1)} \\ \partial\ell/\partial\boldsymbol{\gamma}_{(\times)} \end{pmatrix} = D'V^{-1}\mathbf{f}, \tag{2.50}$$

where V is the dispersion matrix,

$$D = \begin{pmatrix} \partial\boldsymbol{\mu}'/\partial\boldsymbol{\gamma}_{(1)} & 0 \\ \partial\boldsymbol{\eta}'/\partial\boldsymbol{\gamma}_{(1)} & \partial\boldsymbol{\eta}'/\partial\boldsymbol{\gamma}_{(\times)} \end{pmatrix} \quad \text{and} \quad \mathbf{f} = \begin{pmatrix} \mathbf{y} - \boldsymbol{\mu} \\ \mathbf{w} - \boldsymbol{\eta} \end{pmatrix}, \tag{2.51}$$

for $\mathbf{w}$ the vector of pairwise crossproducts of $\mathbf{y}$.

Prentice and Zhao make a further transformation to express this in terms of

$$\mathbf{s} = (s_{11}, s_{12}, s_{13}, \ldots, s_{nn})'$$

where

$$s_{ij} = (y_i - \mu_i)(y_j - \mu_j)$$

is the pairwise empirical covariance, and the predicted covariances, σ_{ij}, to give

$$\tilde{D} = \begin{pmatrix} \partial\boldsymbol{\mu}'/\partial\boldsymbol{\gamma}_{(1)} & 0 \\ \partial\boldsymbol{\sigma}'/\partial\boldsymbol{\gamma}_{(1)} & \partial\boldsymbol{\sigma}'/\partial\boldsymbol{\gamma}_{(\times)} \end{pmatrix} \quad \text{and} \quad \tilde{\mathbf{f}} = \begin{pmatrix} \mathbf{y} - \boldsymbol{\mu} \\ \mathbf{s} - \boldsymbol{\sigma} \end{pmatrix}, \tag{2.52}$$

with

$$\tilde{V} = \begin{pmatrix} \text{var}(\mathbf{Y}) & \text{cov}(\mathbf{Y}, \mathbf{S}) \\ \text{cov}(\mathbf{S}, \mathbf{Y}) & \text{var}(\mathbf{S}) \end{pmatrix}. \tag{2.53}$$

This can be written as a linear transformation $\mathbf{f} = A\tilde{\mathbf{f}}$, $V = A\tilde{V}A'$, $D = A\tilde{D}$, for an obvious choice of the matrix A (Prentice and Zhao, 1991, p. 839).

2.6.3 GEE1 and GEE2

In the original formulation now known as GEE1 in the terminology following Liang *et al.* (1992), we substitute for the above terms the simpler block-diagonal matrices

$$\bar{D} = \begin{pmatrix} \partial\boldsymbol{\mu}'/\partial\boldsymbol{\gamma}_{(1)} & 0 \\ 0 & \partial\boldsymbol{\eta}'/\partial\boldsymbol{\gamma}_{(\times)} \end{pmatrix} \tag{2.54}$$

and

$$\bar{V} = \begin{pmatrix} \text{var}(\mathbf{Y}) & 0 \\ 0 & \text{var}(\mathbf{W}) \end{pmatrix} \tag{2.55}$$

with obvious analogous substitutions for the Prentice and Zhao formulation. Because of this block diagonal form the system can now be decomposed into two sets of equations, with the 'top half', estimates for $\boldsymbol{\gamma}_{(1)}$, still depending on $\boldsymbol{\gamma}_{(\times)}$ through $\text{var}^{-1}(\mathbf{Y})$.

Thus, as originally proposed, GEE1 is a two-stage algorithm; first solve the 'top half' for $\boldsymbol{\gamma}_{(1)}$, using the current estimates of $\boldsymbol{\gamma}_{(\times)}$, then re-estimate $\boldsymbol{\gamma}_{(\times)}$ by any of several methods (none of which is the same as solving the single system described above) generally based on Pearson-type residuals using the current values of $\boldsymbol{\gamma}_{(1)}$ in the Liang and Zeger approach. Prentice proposed other methods and advocated the form given earlier.

These older methods are not discussed further here as the the single-system formulation above appears to be becoming generally accepted (at least by Liang, Zeger, Prentice and Zhao), though even the current Statlib S-PLUS GEE library is based on old GEE1.

In preferring GEE1 over GEE2, we "ignore some functions involving higher-order interactions since they contain 'little relevant information'" (V.P. Godambe in the Discussion of Liang *et al.*, 1992). GEE1 gives consistent estimates of $\boldsymbol{\gamma}_{(1)}$, but these are less efficient than those obtained by GEE2 if the dependence of pairwise interactions in $\boldsymbol{\gamma}_{(\times)}$ is misspecified. On the other hand, the increased efficiency of the GEE2 estimates is at the cost of introducing bias if due to misspecification $\text{E}[\mathbf{S}] \neq \boldsymbol{\sigma}$ or $\text{E}[\mathbf{W}] \neq \boldsymbol{\eta}_u$.

This consideration prompts Prentice and Zhao (1991) to advocate taking $\tilde{D}_u$ as block diagonal regardless of the postulated covariance model.

Liang *et al.* (1992) suggest that if interest is in $\gamma_{(1)}$ with $\gamma_{(\times)}$ considered primarily a set of nuisance parameters, GEE1 may be preferable apart from its simplicity and speed, while GEE2 should be used when dependence is the object of the study. I would argue that in the first case the model of Fitzmaurice and Laird is greatly to be preferred, because of the true parameter orthogonality; and in the second case, I would suggest a fully marginal model (at least in the absence of software to fit a mixed marginal/zero-conditional form).

Chapter 3

Algorithms for marginal models

The fully marginal models discussed in Section 2.3.2 link parameters of interest, γ, to marginal log odds ratios

$$\lambda_u = X_u \gamma$$

for each subject u with covariate matrix X_u. However, the likelihood, and hence the score equations (2.30), are expressed in terms of the vector of log probabilities, $\mathbf{p}$, rather than λ. The whole of this chapter is concerned with the difficult algebraic and numerical problem of obtaining the required $\mathbf{p}$ from given λ, or equivalently, for ease of discussion and presentation, finding the probability table, $\pi = \exp\{\mathbf{p}\}$, from given marginal odds ratios (MORs), $\Lambda = \exp\{\lambda\}$.

In Section 3.1, nomenclature and a new, recursive definition of the system of equations to be solved is given. The need for numerical techniques to solve the problem is highlighted by detailed consideration of the analytic solution in Section 3.2.

The numerical technique adopted by Glonek and McCullagh (1995) is a standard application of Newton–Raphson iteration; this is reviewed in Section 3.3. An alternative algorithm based on residual correction (or equivalently here quasi-Newton–Raphson) techniques is presented in Section 3.4. This algorithm was developed independently of the publication of Glonek and McCullagh (1995); it is often not an improvement, being slower than Newton–Raphson when the number of timepoints, T, is small, though quicker when $T \geq 6$.

Both of these algorithms are prohibitively slow when T is large, as is evident from the results of simulations presented in Section 3.7. Moreover, it is seen that both can fail to find a solution even when one exists. In an effort to overcome these obstacles, I have developed and studied two further algorithms, one (called SR) based on residual correction, the other (denoted SQb) on a modified Gauss–Seidel technique. These are presented in Sections 3.5 and 3.6, respectively. Both are considerably quicker than Newton–Raphson — though they also do not always converge — with SQb preferred in almost all cases both on grounds of speed and on probability of convergence. This is demonstrated and discussed in Section 3.7.

The numerical and analytical techniques are extended to polytomous, unordered outcomes in Section 3.8. In Sections 3.1–3.7, it is assumed that the outcome variables are binary, which greatly simplifies the discussion.

3.1 Precise formulation of the MOR problem

In the following subsections, we define terms and set out a recursive definition of the system of equations we will need to solve to find $\boldsymbol{\pi}$ from $\boldsymbol{\Lambda}$ (or $\mathbf{p}$ from $\boldsymbol{\lambda}$); that is, we give a new description of what Glonek and McCullagh (1995) call the multivariate logistic transform. With this groundwork established, the thrust of this chapter — finding the inverse of the transform — begins in Section 3.2.

3.1.1 Subscript notation

Denote by π_a the cell probability

$$\pi_a = \mathrm{P}(\quad (Y_1, Y_2, \ldots, Y_T)' = (a_1, a_2, \ldots, a_T)' \quad)$$

for the T-variate variable Y, where $a_i \in \{0, 1\}$, $i = 1, 2, \ldots, T$. For convenience, any particular vector subscript a is written as a string of binary digits rather than as a vector over $\mathbf{Z}_2$. Less conventionally, for reasons given in the next paragraph, this string may be read *from right to left*, considered as a binary number, and then the subscript is written as the decimal representation of this subsequent number, ignoring

any leading zeros. For example, the trivariate cell π_{011} and the 4-variate π_{0110} are both denoted π_6 in the decimal subscript system.

There are three reasons for adopting this unusual convention. Firstly, it usefully allows us to write, for example, π_6 as the probability that Y_2 and Y_3 are one and all other variables are zero, without having to specify how many other variables there are. Secondly, of some use when computer programming, the decimal subscript is the index of a one-dimensional array representation of the probability table, using the standard conventions that the first element is subscripted zero and that first subscript changes fastest. Finally, the decimal subscripts give the sequence of the elements of the vector $\boldsymbol{\pi}$ in the recursive definition of the logistic transform in Section 3.1.5.

The same conventions apply to the subscripts of the *logarithms* of the cell probabilities used frequently below, denoted $\mathbf{p}$.

The subscripts of the MORs themselves, $\boldsymbol{\Lambda}$, or equivalently their logarithms, $\boldsymbol{\lambda}$, follow a different logic, through necessity and to obey the standard conventions. Here, the number of subscripts indicates the order of the interaction, and the (decimal) values of the subscripts indicate which Y-variables are under consideration. Thus, λ_i is the log odds ratio (logit) for Y_i, λ_{ij} is the marginal log odds ratio between Y_i and Y_j, and so on. This convention is as followed in Chapter 2. Again assume without loss of generality that the subscripts are ordered $i < j < \cdots < T$, without repeats.

As a further aid to symmetry, define $\Lambda_0 = 1$ as the 'zeroth order' interaction; this will serve as the condition that cell probabilities sum to unity. On the log scale, likewise define $\lambda_0 = 0$. The subscript 0 is not used in any other combination.

3.1.2 Tilde notation

The system of equations to be solved, derived in the following three subsections, and denoted $S(\boldsymbol{\pi}) = \boldsymbol{\Lambda}$, involves the operations of addition, multiplication and division. Except for the convention that $\mathbf{a}/\mathbf{b}$ denotes componentwise division of vectors, standard notation will suffice to define the system $S(\boldsymbol{\pi})$ recursively.

However, in the approach of Glonek and McCullagh (1995), see Section 3.3, and in calculation of the score equations (2.30), one works directly with log probabilities and

odds ratios, so that it is also useful to write the problem on this scale, i.e. as $s(\mathbf{p}) = \boldsymbol{\lambda}$ (see following three subsections for the definition of this system also). To avoid the complexity of notation involved in writing down logarithms of sums of exponentials, and thus greatly tidy the recursive definition of $s(\mathbf{p})$, define the binary operator $\not\sim$ ('tildeplus') as follows:

$$a \not\sim b = \log(\mathrm{e}^a + \mathrm{e}^b).$$

It is easy to show that tildeplus is associative and commutative, and that ordinary addition is distributive over it; these properties are used in Section 3.1.4 and 3.1.5. Also, for $a > b$,

$$a \not\sim b = a + \log(1 + \mathrm{e}^{b-a}), \tag{3.1}$$

which is computationally more efficient, since only one exponentiation is needed, and additionally since $\mathrm{e}^{b-a} < 1$, the logarithm is in standard form for series expansion. Equation (3.1) also shows that for any a, b,

$$a \not\sim b \leq \max(a, b) + \log\ 2, \tag{3.2}$$

with equality iff $a = b$, a result we will use in Section 3.4.3. Some further aspects are considered in Appendix A3.1.2.

3.1.3 The univariate case

To motivate the form of the general definition of the multivariate logistic transform in Section 3.1.5, we first consider the univariate system, which although itself trivial, illustrates the symmetry of the larger, nontrivial systems. Given the log odds ratio (logit) λ_1 (or its exponential, Λ_1) what probability lies in each of the cells? Of course, this problem can be solved directly without recourse to representation as a system of equations, since the inverse of the logit transform may be written explicitly, to give π_1; then $\pi_0 = 1 - \pi_1$ follows immediately.

But let us here write this out as a system of two equations in two unknowns:

$$\pi_0 + \pi_1 = \Lambda_0 \tag{3.3}$$

$$\pi_1/\pi_0 = \Lambda_1, \tag{3.4}$$

where $\Lambda_0 \equiv 1$ as defined at the end of Section 3.1.1. This system of equations is more concisely denoted

$$S_1(\boldsymbol{\pi}) = \boldsymbol{\Lambda}. \tag{3.5}$$

The subscript to S indicates the number of outcome variables.

Equivalently, for the log cell probabilities, on substituting for $\mathbf{p} = \log(\boldsymbol{\pi})$ and taking logarithms of both sides of equations (3.3) and (3.4), also writing $\log(e^{p_0} + e^{p_1})$ as $p_0 \curlywedge p_1$, we obtain

$$\begin{aligned} p_0 \curlywedge p_1 &= \lambda_0 \\ -p_0 + p_1 &= \lambda_1. \end{aligned}$$

This system is denoted by $s_1(\mathbf{p}) = \boldsymbol{\lambda}$ (lower case for logs).

3.1.4 The bivariate case

To further motivate the general T-variate formulation of Section 3.1.5, the bivariate or 4-cell system is now given explicitly. On the natural scale, and using the decimal subscript convention of Section 3.1.1,

$$\pi_0 + \pi_1 + \pi_2 + \pi_3 = \Lambda_0 \tag{3.6}$$

$$(\pi_1 + \pi_3)/(\pi_0 + \pi_2) = \Lambda_1 \tag{3.7}$$

$$(\pi_2 + \pi_3)/(\pi_0 + \pi_1) = \Lambda_2 \tag{3.8}$$

$$(\pi_0\pi_3)/(\pi_1\pi_2) = \Lambda_{12}, \tag{3.9}$$

denoted $S_2(\pi) = \Lambda$. On the log scale, using tilde notation, the system $s_2(\mathbf{p}) = \boldsymbol{\lambda}$ is

$$
\begin{aligned}
p_0 \tilde{+} p_1 \tilde{+} p_2 \tilde{+} p_3 &= \lambda_0 && (3.10)\\
-(p_0 \tilde{+} p_2) + (p_1 \tilde{+} p_3) &= \lambda_1 && (3.11)\\
-(p_0 \tilde{+} p_1) + (p_2 \tilde{+} p_3) &= \lambda_2 && (3.12)\\
p_0 - p_1 - p_2 + p_3 &= \lambda_{12}. && (3.13)
\end{aligned}
$$

Some further subscript notation facilitates the partitioning of the 4-vector into

$$
\mathbf{p} = \begin{pmatrix} \mathbf{p_0} \\ \mathbf{p_1} \end{pmatrix}
$$

where $\mathbf{p_0} = (p_0, p_1)'$ is effectively identical to the $\mathbf{p}$ of the univariate case, and $\mathbf{p_1} = (p_2, p_3)'$ are the 'new' elements. Bold face is used for both the symbol and subscript of $\mathbf{p_1}$ to distinguish it from the element p_1 (which is in $\mathbf{p_0}$). This notation facilitates later extension to polytomous variables (Section 3.8).

Recalling that in the preceding subsection we have already defined the function

$$
s_1(\mathbf{p_0}) = s_1 \begin{pmatrix} p_0 \\ p_1 \end{pmatrix} = \begin{pmatrix} p_0 \tilde{+} p_1 \\ -p_0 + p_1 \end{pmatrix} \qquad (3.14)
$$

and further introducing the concept of componentwise tildeplus for vectors,

$$
\mathbf{p_0} \tilde{+} \mathbf{p_1} = \begin{pmatrix} p_0 \\ p_1 \end{pmatrix} \tilde{+} \begin{pmatrix} p_2 \\ p_3 \end{pmatrix} = \begin{pmatrix} p_0 \tilde{+} p_2 \\ p_1 \tilde{+} p_3 \end{pmatrix}, \qquad (3.15)
$$

the bivariate system can be written

$$
s_2(\mathbf{p}) = \begin{pmatrix} s_1(\mathbf{p_0} \tilde{+} \mathbf{p_1}) \\ -s_1(\mathbf{p_0}) + s_1(\mathbf{p_1}) \end{pmatrix} = \boldsymbol{\lambda}. \qquad (3.16)
$$

As the notation is unfamiliar, let me clarify that in (3.16) the term $s_1(\mathbf{p_0} \tilde{+} \mathbf{p_1})$

represents the operation of s_1 as in (3.14), but acting on the components of $\mathbf{p_0} \not\sim \mathbf{p_1}$ given in (3.15), i.e. in (3.14) substitute $p_0 \not\sim p_2$ for p_0, and $p_1 \not\sim p_3$ for p_1. Equation (3.10) is recovered exactly on noting the commutativity of tildeplus. It is vital for the emerging symmetry that the components of $\mathbf{p}$ and $\boldsymbol{\lambda}$ are introduced in the order given here and stated in Section 3.1.5.

The unlogged equivalent version is

$$S_2(\boldsymbol{\pi}) = \begin{pmatrix} S_1(\boldsymbol{\pi_0} + \boldsymbol{\pi_1}) \\ S_1(\boldsymbol{\pi_1}) \,/\, S_1(\boldsymbol{\pi_0}) \end{pmatrix} = \boldsymbol{\Lambda}. \tag{3.17}$$

where operation S_1 is defined by (3.5) to act on two components according to (3.3) and (3.4). Expression (3.17) is readily verified by expansion. The division sign is used to denote *componentwise* division, here and throughout.

3.1.5 The general case

This is obtained by induction. Suppose that the problem for T variables is

$$S_T(\boldsymbol{\pi}) = \begin{pmatrix} S_{(T-1)}(\boldsymbol{\pi_0} + \boldsymbol{\pi_1}) \\ S_{(T-1)}(\boldsymbol{\pi_1}) \,/\, S_{(T-1)}(\boldsymbol{\pi_0}) \end{pmatrix} = \boldsymbol{\Lambda}, \tag{3.18}$$

as is shown to hold for $T = 2$ in the preceding subsection. Here

$$\boldsymbol{\pi_0} = (\pi_0, \pi_1, \ldots, \pi_{[2^{(T-1)}-1]})'$$

is the vector of probabilities for the $(T-1)$-variate case — except for an additional trailing zero in the binary form of the subscripts — and

$$\boldsymbol{\pi_1} = (\pi_{2^{(T-1)}}, \ldots, \pi_{2^T-1})'$$

are the 'new' probabilities (those where $Y_T = 1$ rather than zero). To illustrate: in the trivariate case, using binary subscript notation,

$$\boldsymbol{\pi} = \begin{pmatrix} \boldsymbol{\pi_0} \\ \boldsymbol{\pi_1} \end{pmatrix} = \begin{pmatrix} (\pi_{000}, \pi_{100}, \pi_{010}, \pi_{110})' \\ (\pi_{001}, \pi_{101}, \pi_{011}, \pi_{111})' \end{pmatrix}.$$

Note that in the decimal subscript notation (Section 3.1.1) the sequence is simply $\pi_0, \pi_1, \ldots \pi_7$.

To preserve symmetry, the order of introduction of the odds ratios $\boldsymbol{\Lambda}$ is equally important, and obeys the following scheme. For $T = 1$, $\boldsymbol{\Lambda} = (\Lambda_0, \Lambda_1)'$, while for $T = 2$, $\boldsymbol{\Lambda} = (\Lambda_0, \Lambda_1, \Lambda_2, \Lambda_{12})'$. This generalizes to

$$\boldsymbol{\Lambda} = (\Lambda_0, \Lambda_1, \Lambda_2, \Lambda_{12}, \Lambda_3, \Lambda_{13}, \Lambda_{23}, \Lambda_{123}, \ldots). \tag{3.19}$$

The 'new' interactions with Y_T are ordered by appending T to the subscripts for the $(T-1)$-variate case. For convenience (and following accepted usage) the subscript 0 is dropped except for Λ_0 itself. This convention is as followed in Glonek and McCullagh (1994).

We can now build the system for $T + 1$ variables. First consider the the equations for the 'top half' of the problem — that is, for those components of $\boldsymbol{\Lambda}$ that appeared in the T-variate system, i.e. $(\Lambda_0, \ldots, \Lambda_{123\ldots T})'$. Since these are marginal odds, they are found by collapsing the probability table over the $(T + 1)$st variable, i.e. from the probabilities $\boldsymbol{\pi_0} + \boldsymbol{\pi_1}$ of the $(T + 1)$-variable table. Hence

$$S_T(\boldsymbol{\pi_0} + \boldsymbol{\pi_1}) = (\Lambda_0, \ldots, \Lambda_{123\ldots T})' \tag{3.20}$$

by the induction hypothesis that S_T is the appropriate transformation for a T-variable probability table, here $\boldsymbol{\pi_0} + \boldsymbol{\pi_1}$, to its marginal log odds ratios.

The remaining equations of the new system concern ratios Λ_T, $\Lambda_{1(T+1)}, \ldots, \Lambda_{123\ldots T(T+1)}$,

the elements of which are by definition

$$\frac{(\text{odds ratios between } \mathbf{Y}^{\mathcal{A}} \in \{Y_1, \ldots, Y_T\}, \text{ given that } Y_{T+1} = 1)}{(\text{odds ratios between } \mathbf{Y}^{\mathcal{A}} \in \{Y_1, \ldots, Y_T\}, \text{ given that } Y_{T+1} = 0)} \tag{3.21}$$

as $\mathcal{A}$ runs though all subsets (beginning with the empty set, giving the marginal logit) in the order determined above for $\mathbf{\Lambda}$. Again by the induction hypothesis, the numerators are found by $S_T(\boldsymbol{\pi}_1)$ and the denominators by $S_T(\boldsymbol{\pi}_0)$. Hence

$$S_T(\boldsymbol{\pi}_1)/S_T(\boldsymbol{\pi}_0) = (\Lambda_T, \Lambda_{1(T+1)}, \ldots \Lambda_{123\cdots(T+1)})', \tag{3.22}$$

as always here assuming componentwise division.

Thus the general T-variate case is as stated in (3.18), by the combination of equations (3.20) and (3.22), and the preceding demonstration of the validity for $T = 2$.

On taking logs, the sequence of logged systems s_T is readily seen to grow according to

$$s_{T+1}(\mathbf{p}) = \begin{pmatrix} s_T(\mathbf{p_0} \not\sim \mathbf{p_1}) \\ -s_T(\mathbf{p_0}) + s_T(\mathbf{p_1}) \end{pmatrix} \tag{3.23}$$

where the log probabilities $\mathbf{p}$ are partitioned into halves $\mathbf{p_0}$ and $\mathbf{p_1}$ by direct analogy with the case for unlogged $\boldsymbol{\pi}$.

For completeness define s_0 and S_0 as the identity function, enabling us to make the decompositions (3.18) and (3.23) on even the univariate system.

The technical report by Glonek and McCullagh (1994) appears to be the only previous reference that gives an explicit, recursive, general form for the multivariate logistic transform. They write the system in the form

$$C \log(L\boldsymbol{\pi}) = \boldsymbol{\lambda}$$

(in my notation for $\boldsymbol{\pi}$ and $\boldsymbol{\lambda}$), where the matrices C and L are defined in terms of the direct products of the related matrices for the problem in one less dimension. The form $s(\mathbf{p}) = \boldsymbol{\lambda}$ is simpler and quicker to evaluate (flop counts for both methods are given in Section 3.7.2). However, Glonek and McCullagh's description is a little more

general than mine, in that they also allow for ordinal data; I consider only binary data (above) and unordered polytomous data (in Section 3.8).

3.2 Analytic solutions and considerations

Given a probability table, the mapping $S(\boldsymbol{\pi})$ specifies only one set of odds ratios, but we are concerned here with the inverse problem: obtaining probabilities from odds ratios, i.e. finding $\boldsymbol{\pi} = S^{-1}(\boldsymbol{\Lambda})$. This mapping is *not* unique, as seen at several points below, unless we specify the constraint that all probabilities are positive (the first equation, for $\Lambda_0 = 1$, then ensures the probabilities cannot exceed one, so this is not a necessary further constraint). Such lack of uniqueness is generally ignored in previous publications (e.g. Liang *et al.*, 1992, Glonek and McCullagh, 1995). Better documented is that for certain $\boldsymbol{\Lambda}$ there is no solution to the constrained problem. There are further constraints to be imposed, on the odds ratios, that are difficult to write explicitly and to interpret.

The term *valid solution* is used here to denote a solution that meets all the implicit and explicit constraints: a solution that represents a well-defined probability table. By conjecture (Darroch, 1962; discussed in Appendix A3.2), only one, if any, of the many solutions to the unconstrained problem is valid, as we shall discuss further below.

The structure of this section is as follows. In Sections 3.2.1 and 3.2.2 we find $S^{-1}(\boldsymbol{\Lambda})$ analytically for $T = 1$ and 2, respectively; then in Section 3.2.3 we illustrate the extreme difficulty of solving when $T = 3$, showing also that for $T \geq 4$ one is forced to turn to numerical techniques. Successive solutions for $T = 1$, 2 and 3 suggest a hierarchical approach discussed in Section 3.2.4. Numerical techniques are likely to suffer from the existence of multiple solutions to the unconstrained problem (Section 3.2.5); moreover, some systems have no valid solution at all (Section 3.2.6).

3.2.1 The univariate solution

The univariate system is trivial to solve. The quickest solution is already given, at the beginning of Section 3.1.3. Proceeding more directly by isolating π_1 in equation

(3.4) then substituting in (3.3), we have

$$\pi_0 = \frac{\Lambda_0}{1+\Lambda_1}, \tag{3.24}$$

$$\pi_1 = \frac{\Lambda_0\Lambda_1}{1+\Lambda_1}. \tag{3.25}$$

Note that there is is a unique solution to the unconstrained problem even when $\Lambda_0 \neq 1$. This will seldom concern us, since we must have $\Lambda_0 = 1$ for a valid probability table, but we will consider circumstances in which $\Lambda_0 \neq 1$ in Section 3.6 (page 119). Guaranteed uniqueness of the solution fails to hold for $T \geq 2$, as we now see.

3.2.2 The bivariate solution

To tidy the form of the bivariate inversion $S^{-1}(\mathbf{\Lambda})$, to find $\boldsymbol{\pi}$, the system $S(\boldsymbol{\pi})$ is made as linear as possible by inverting the logit in equations (3.7) and (3.8), giving expressions in terms of the marginal means, μ_1 and μ_2. For given values of Λ_1 and Λ_2, the values of μ_1 and μ_2 are easily calculated. This gives

$$\pi_{00} + \pi_{10} + \pi_{01} + \pi_{11} = 1, \tag{3.26}$$

$$\pi_{10} + \pi_{11} = \mu_1, \tag{3.27}$$

$$\pi_{01} + \pi_{11} = \mu_2, \tag{3.28}$$

$$\frac{\pi_{11}\pi_{00}}{\pi_{10}\pi_{01}} = \Lambda_{12} \tag{3.29}$$

By successive substitution,

$$\pi_{10} = \mu_1 - \pi_{11}, \tag{3.30}$$

$$\pi_{01} = \mu_2 - \pi_{11}, \tag{3.31}$$

$$\pi_{00} = 1 - \mu_1 - \mu_2 + \pi_{11} \tag{3.32}$$

and so we obtain the following quadratic equation in probability π_{11}:

$$q(\pi_{11}) = (1-\Lambda_{12})\pi_{11}^2 + [1-(1-\Lambda_{12})(\mu_1+\mu_2)]\pi_{11} - \Lambda_{12}\mu_1\mu_2 = 0 \tag{3.33}$$

Other choices for substitution give rise to similar, quadratic expressions in terms of π_{00}, π_{01} or π_{10} as desired, but such expressions are no simpler than (3.33). The solution of a quadratic is only avoided when $\Lambda_{12} = 1$, which is the case for independent variables, when the solution is trivial. When $\Lambda_{12} \neq 1$, equation (3.33) has two distinct real solutions, as illustrated shortly below; complex roots cannot occur. Thus the general, unconstrained problem does not have a unique solution even for two variables, a fact which is frequently overlooked.

The constraint that the solution must represent a valid probability table, $\pi_{ij} \geq 0$ (strict inequality assuming a non-degenerate distribution), might suggest that it is enough to find a solution to (3.33) in the range (0, 1) — indeed this was claimed by Liang *et al.* (1992, p. 13). But there can be two solutions in this interval (though not both admissible), as seen by example: if $\mu_1 = \mu_2 = 1/4$ and $\Lambda_{12} = 4$, then $\pi_{11} = (5 \pm \sqrt{13})/12$, approximately $\{0.116,\ 0.717\}$. In this case, only the smaller root gives rise to a valid table, a result we now generalize.

To ascertain which of the solutions to the quadratic gives a valid solution in all components, we note that for $\pi_{11} > 0$, (3.30)–(3.32) introduce the requirement

$$a = \max\{0,\ \mu_1 + \mu_2 - 1\} < \pi_{11} < \min\{\mu_1,\ \mu_2\} = b. \tag{3.34}$$

Darroch's (1962) result (Appendix A3.2) shows that precisely one of the solutions lies in this range. More practically, I now derive a simple method of predetermining which one.

The quadratic $q(\pi_{11})$ of equation (3.33) is negative at the lower bound and positive at the upper (Darroch, 1962). Then by reference to Figure 3.1, considering the sign of the leading coefficient, namely $1 - \Lambda_{12}$, it is clear that

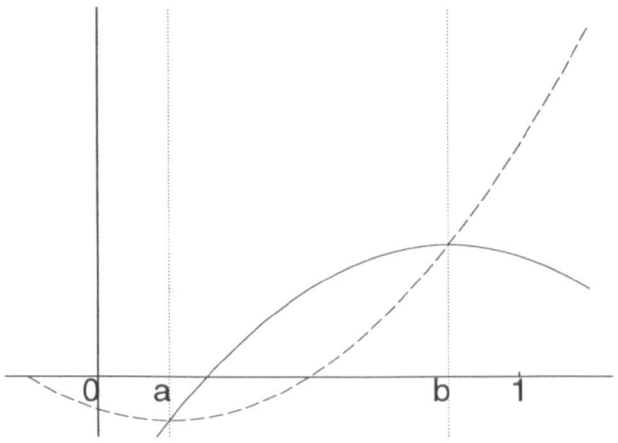
if $\Lambda_{12} > 1$, valid solution is the smaller root

if $\Lambda_{12} = 1$, degenerate case, $\pi_{11} = \mu_1\mu_2$

if $\Lambda_{12} < 1$, valid solution is the larger root

Liang *et al.* (1992) only give the solution with smaller root. Of course, when we expect positive association, as they do in all their examples, and in general for repeated

Figure 3.1: Possible quadratic curves with a single root between points $a < b$, both lying between 0 and 1.



measures data, this is likely to be the right choice.

3.2.3 The trivariate problem and beyond

We have previously considered the 'explosion' in the number of parameters as T grows; the algebraic complexity of the solution to the system grows similarly quickly. The problems are compounded because for $T \geq 3$ there may be no solution to the constrained problem at all, and the conditions under which this obtains cannot be easily specified in terms of the given odds ratios themselves (Section 3.2.6).

Something of the complexity of the trivariate problem is revealed if one uses the computer algebra system Maple to solve the system directly. Even on a Unix mainframe, this attempt crashes because the subexpressions exceed available buffer space.

We can, however, do rather better by hand. Proceeding as for the two-variable case in Section 3.2.1, firstly simplify by inverting the first-order logits to give expressions in μ_i, $i = 1 \ldots 3$: this creates the system

$$\pi_{000} + \pi_{100} + \pi_{010} + \pi_{110} + \pi_{001} + \pi_{101} + \pi_{011} + \pi_{111} = 1 \quad (3.35)$$

$$\pi_{100} + \pi_{110} + \pi_{101} + \pi_{111} = \mu_1 \quad (3.36)$$

$$\pi_{010} + \pi_{110} + \pi_{110} + \pi_{111} = \mu_2 \quad (3.37)$$

$$(\pi_{000} + \pi_{001})(\pi_{110} + \pi_{111})/(\pi_{010} + \pi_{011})(\pi_{100} + \pi_{101}) = \Lambda_{12} \quad (3.38)$$

$$\pi_{001} + \pi_{101} + \pi_{011} + \pi_{111} = \mu_3 \quad (3.39)$$

$$(\pi_{000} + \pi_{010})(\pi_{101} + \pi_{111})/(\pi_{001} + \pi_{011})(\pi_{100} + \pi_{110}) = \Lambda_{13} \quad (3.40)$$

$$(\pi_{000} + \pi_{100})(\pi_{011} + \pi_{111})/(\pi_{001} + \pi_{101})(\pi_{011} + \pi_{111}) = \Lambda_{23} \quad (3.41)$$

$$(\pi_{100}\pi_{010}\pi_{001}\pi_{111})/(\pi_{000}\pi_{110}\pi_{101}\pi_{011}) = \Lambda_{123}. \quad (3.42)$$

Substitute first for π_{100}, π_{010} and π_{001} by (3.36), (3.37) and (3.39), then for π_{000} by (3.35), i.e.

$$\pi_{100} = \mu_1 - \pi_{110} - \pi_{101} - \pi_{111} \quad (3.43)$$

$$\pi_{010} = \mu_2 - \pi_{110} - \pi_{011} - \pi_{111} \quad (3.44)$$

$$\pi_{001} = \mu_3 - \pi_{101} - \pi_{011} - \pi_{111} \quad (3.45)$$

$$\pi_{000} = 1 - \mu_1 - \mu_2 - \mu_3 + 2\pi_{111} + \pi_{110} + \pi_{011} + p_{101} \quad (3.46)$$

After this particular substitution (only) we obtain expressions for the two-way odds ratios — (3.38), (3.40) and (3.41) — each in terms of π_{111} and only *one* other probability (namely that for which subscript positions i and j are unity, and the remaining position is zero, for each Λ_{ij}). These can then be expressed as quadratics with coefficients in terms of π_{111} (and known Λ) only. For example, from (3.38)

$$(1 - \Lambda_{12})\pi_{110}^2 + [2(1 - \Lambda_{12})\pi_{111} + (1 - (1 - \Lambda_{12})(\mu_1 + \mu_2))]\pi_{110} + q(\pi_{111}) = 0, \quad (3.47)$$

where $q()$ is precisely the quadratic function in equation (3.33).

The expression of the roots of these equations using the standard formula is unprepossessing, but on scrutiny reduces to

$$\pi_{110} = \alpha_{12} - \pi_{111}, \quad \pi_{101} = \alpha_{13} - \pi_{111}, \quad \pi_{011} = \alpha_{23} - \pi_{111}, \quad (3.48)$$

where

$$\alpha_{ij} = \tfrac{1}{2}\left(\mu_i + \mu_j \pm \frac{\delta_{ij}^{1/2} - 1}{1 - \Lambda_{ij}}\right), \quad \Lambda_{ij} \neq 1, \tag{3.49}$$

where δ_{ij} is the determinant of the relevant quadratic. Expressions (3.49) do not simplify greatly in the general case, though importantly they do not include π_{111} or any other unknowns. As an exceptional case, if $\Lambda_{ij} = 1$, for any pair (i, j), then (3.49) is replaced by

$$\alpha_{ij} = \mu_i \mu_j.$$

Observe that $\alpha_{ij} = \nu_{ij}$, the marginal expectation; this is the probability in the (1,1)-cell of the marginal bivariate subsystem for Y_i and Y_j.

Consequently, only one of the two roots provides a valid solution to the constrained problem; the other root yields an invalid subtable with 'probabilities' outside the range $(0, 1)$. The root that should be chosen is as prescribed in the previous section; (3.49) is the root of (3.33) after summation over the relevant third subscript.

Finally, substituting for π_{110}, π_{011} and π_{101} into (3.42) gives a quartic in π_{111} with coefficients in terms of the known odds ratios and known α_{ij} only; given the solution to this, the remaining probabilities are determined by back substitution. As noted by previous authors (e.g. Molenberghs and Lesaffre, 1994), the equation is unattractive, and an algebraic solution, while possible, is inelegant.

The quartic equation is given here for completeness and to show how unwieldy this expression is. The coefficients, with c_n denoting the coefficient of π_{111}^n, are

$$\begin{aligned}
c_4 &= 1 - \Lambda_{123}, \\
c_3 &= -(1 - \mu_1 - \mu_2 - \mu_3 + 2(\alpha_{12} + \alpha_{13} + \alpha_{23}))(1 - \Lambda_{123}) + 1, \\
c_2 &= [(1 - \mu_1 - \mu_2 - \mu_3 + \alpha_{12} + \alpha_{13} + \alpha_{23})(\alpha_{12} + \alpha_{13} + \alpha_{23}) \\
&\quad + \alpha_{12}\alpha_{13} + \alpha_{12}\alpha_{23} + \alpha_{13}\alpha_{23}](1 - \Lambda_{123}) \\
&\quad + \mu_1\mu_2 + \mu_1\mu_3 + \mu_2\mu_3 - (1 + \mu_1)\alpha_{23} - (1 + \mu_2)\alpha_{13} - (1 + \mu_3)\alpha_{12}, \\
c_1 &= -[(1 - \mu_1 - \mu_2 - \mu_3 + \alpha_{12} + \alpha_{13} + \alpha_{23})(\alpha_{12}\alpha_{13} + \alpha_{12}\alpha_{23} + \alpha_{13}\alpha_{23}) \\
&\quad + \alpha_{12}\alpha_{13}\alpha_{23}](1 - \Lambda_{123}) + \alpha_{12}\alpha_{13} + \alpha_{12}\alpha_{23} + \alpha_{13}\alpha_{23} + 2\alpha_{12}\alpha_{13}\alpha_{23} + \mu_1\mu_2\mu_3 \\
&\quad - (\mu_2 + \mu_3 - \alpha_{23})\mu_1\alpha_{23} - (\mu_1 + \mu_3 - \alpha_{13})\mu_2\alpha_{13} - (\mu_1 + \mu_2 - \alpha_{12})\mu_3\alpha_{12},
\end{aligned}$$

$$c_0 = -(1-\mu_1-\mu_2-\mu_3+\alpha_{12}+\alpha_{13}+\alpha_{23})\alpha_{12}\alpha_{13}\alpha_{23}\Lambda_{123}$$

I have collected terms as multiples of $1-\Lambda_{123}$ where applicable, which shows how the equation simplifies, at least a little, when the high-order ratio is unity. Liang *et al.* (1992) always make this simplifying assumption.

Although these coefficients can indeed be substituted into a generic solution to the quartic, to determine π_{111}, there is very little simplification of the expanded terms (just as the determinant of the earlier quadratics, δ_{ij}, is no simpler than $c_1^2-4c_2c_0$). We are essentially forced to adopt numeric techniques, unless a neater analytic solution to the quartic is found. Moreover, we now face the extra problem of determining which of the four possible solutions is valid.

For $T=4$, analogous successive substitution would lead us into having to solve six quadratics, four quartics (each time choosing the correct solution), and finally an octic, for which there is no explicit formula. Indeed, in this case (and beyond) no general formula can be written, though this does not preclude the existence of a formula for our particular problem. However, instead of seeking such a formula, I more practically emphasise the need to turn from analytical to purely numerical solutions.

3.2.4 The hierarchical approach

The successive substitutions in the trivariate case above were chosen to simplify the algebra, rather than for statistical meaning, our concern being to find a solution by the shortest computational route. If the order of substitutions is changed, one does not arrive at quadratics in only two cell probabilities and a quartic in one, but instead one obtains simultaneous quadratics and quartics in four probabilities. In fact Maple fails to find a solution precisely because it appears to follow one of these wrong paths irrespective of the order of the input equations.

However, if we review the successive substitutions for the trivariate problem, we see we have firstly solved the three marginal bivariate problems (in finding the $\alpha_{ij}=\nu_{ij}$), and secondly via the quartic equation distributed these marginal probabilities to individual

cells (in finding $\nu_{123} = \pi_{111}$ and then using back substitution). This is now seen to be identical to the scheme proposed by Liang *et al.* (1992); they were seeking expectations directly, but the transform between $\boldsymbol{\pi}$ and $\boldsymbol{\nu}$ is trivial (i.e. linear). Liang *et al.* (1992) also note that the high-order equations need to be solved numerically, although one need not make their assumptions of high-order independence.

As a refinement of the numerical technique proposed by Liang *et al.* (1992), Molenberghs and Lesaffre (1994) point out that when solving the high-order equation one can efficiently find the unique solution between the limits a and b of equations (A9) and (A10) in Appendix A3.2 (*cf* 3.34), by using Newton–Raphson starting from $(a+b)/2$, which should ensure speedy convergence to the valid solution. But this is not useful for problems in higher dimensions, because the limits a and b, and any required derivatives, are prohibitively difficult to calculate.

It would seem, therefore, that algebraic complexity forces the hierarchy to end at $T = 3$.

3.2.5 Solutions to the unconstrained system

As seen in Section 3.2.3, if the constraints leading to a valid probability table are not made explicitly, then there are a potentially enormous number of solutions. For the bivariate case there are potentially two solutions (Section 3.2.1); for the trivariate case, there are not four, but potentially 32 solutions: up to two choices for each α_{ij} give rise to up to eight different quartic equations, each having up to four solutions. Often, many solutions are complex. If any $\Lambda_{ij} = 1$ there are multiple roots (more particularly, then there is only one choice of α_{ij}). For independent variables, with all $\Lambda_{ij} = \Lambda_{123} = 1$, there is only one solution.

A general formula for the number of solutions can be derived, but there is little need for one. Importantly, there can be very many solutions, which is unfortunate for iterative schemes hoping to find one of them rather than another; in Section 3.7.3 it is shown that Glonek and McCullagh's (1995) Newton–Raphson technique can set off towards an invalid solution if the start value is poor (a generic problem of the method). My algorithm SQb, Section 3.6, can suffer likewise.

In iterative schemes one wants to be assured of convergence, and convergence to the right solution. This problem is poorly addressed in the literature, and I can offer no analytical results for $T \geq 3$. However the results of extensive simulations reported in the following sections of this chapter shed some light on this matter.

3.2.6 Constraints on Λ

As the problem is meaningless if Λ terms are not strictly positive, positivity is assumed throughout. In fitting marginal models with an identity link to the logits, negative Λ components cannot occur.

The constraints that must be met to give a valid table are naturally expressed in terms of marginal expectations (Darroch, 1962; Glonek and McCullagh, 1994) written here μ_i and ν_{ij} for first- and second-order moments. For example, for the trivariate problem, one such constraint is

$$\mu_2 - \nu_{12} - \nu_{23} + \nu_{13} > 0,$$

which on using (3.49), recalling that $\nu_{ij} = \alpha_{ij}$, becomes

$$\pm\frac{\delta_{13}^{1/2} - 1}{1 - \Lambda_{13}} \mp \frac{\delta_{12}^{1/2} - 1}{1 - \Lambda_{12}} \mp \frac{\delta_{23}^{1/2} - 1}{1 - \Lambda_{23}} > 0.$$

As noted by Glonek and McCullagh (1994), this expression does not simplify to anything that may be easily interpreted.

In the wider application of fitting marginal models, poor starting or intermediate values for parameters γ in the linear predictors may generate a set of $\lambda = X\gamma$ for which there is no valid probability table. Worse still, necessary conditions for this event in terms of λ are not readily available.

Solving the bivariate marginal problems, giving the required marginal expectations, can test whether a valid solution exists, and this can be done if the chosen algorithm fails to find a solution. However, this is not readily extended to higher dimensions because of the algebraic complexity of the marginal expectation constraints themselves.

An old method for determining the number of roots of a polynomial between two points a and b is to use Sturm sequences (for a modern description, see Dobbs, 1980). I attempted this method here to ascertain the existence of a real solution in (0, 1) for the $(1, 1, \ldots, 1)$-cell for a given set of odds ratios, in the hope that the constraints might become clearer in this alternative formulation. Ideally, one would want to test for a solution not merely in $(0, 1)$ but between the limits (a, b) of Appendix A3.2, but as stated at the end of Section 3.2.4, calculating these limits is itself impractical. Once the coefficients of the high-order polynomial are calculated, testing for sign changes over $(0, 1)$, or if possible over (a, b), involves comparatively little computation. I do not report details for two reasons: firstly, algebraic expressions for the coefficients are not currently available for $T \geq 4$, and secondly, one seeks an algebraic, not numerical, formulation of the tabulated sign changes to classify sets of ratios with no solution. But even for $T = 3$, the Sturm sequence process gives expressions far more complicated than those rejected as useful above.

3.3 The Newton–Raphson method

Glonek and McCullagh (1995) use Newton–Raphson iteration to solve the nonlinear system of equations $s(\mathbf{p}) = \boldsymbol{\lambda}$ for $\mathbf{p}$. Here this is to iterate according to

$$\mathbf{p}^{(n+1)} = \mathbf{p}^{(n)} - \left[\frac{\partial s'}{\partial \mathbf{p}} \Big|_{\mathbf{p}=\mathbf{p}^{(n)}} \right]^{-1} \left\{ s(\mathbf{p}^{(n)}) - \boldsymbol{\lambda} \right\}. \tag{3.50}$$

For problems of small dimension this approach is feasible (although about 6 times slower than Algorithm SQb of Section 3.6 even for $T = 3$; see Section 3.7). The cost of evaluation, and inversion, of the derivative matrix, which is not great for $T \leq 3$, is offset by quadratic convergence, provided there is convergence.

For problems with many variables, however, the cost of evaluating and inverting a huge derivative matrix ($2^T \times 2^T$) is prohibitive, and this has led to my seeking a more efficient alternative. In Section 3.4 we study a quasi-Newton–Raphson method — i.e. a fixed matrix replaces the true derivative matrix and is chosen so as to be

trivially invertible. The subsequent loss of quadratic convergence, however, makes this particular method much less efficient for small T; it is only faster for $T \geq 6$. In Sections 3.5 and 3.6 we will consider two further algorithms that are always faster (as shown by simulations reported in Section 3.7), and increasingly so with increasing T.

With modern computers one sometimes questions the need to accelerate numerical methods. But consider a large-sample data set with 9 or 10 timepoints, taking, quite plausibly on current hardware, a whole day to fit using algorithm SQb within the Fisher scoring loops (bear in mind that if there are continuous covariates, a separate set of odds ratios has to be inverted for every individual, every Fisher scoring iteration). We will see in Section 3.7 that the SQb algorithm is at least 175 times faster than Newton–Raphson for problems of this size. Thus is would currently take approximately six months to fit the same model using the Newton–Raphson technique.

Even more important than speed however, is the observation that there are sets of odds ratios for which Newton–Raphson is unable to find a solution even when one exists. This can be due to poor starting values, which is a well-known shortcoming of the method, though to be fair this problem occurred only rarely in the simulations. A way to overcome this is to use one of my alternative algorithms to limited precision to obtain good starting values.

But failure to converge may also be due to numeric singularity of the derivative matrix at or near the solution, despite Glonek and McCullagh's proof of its analytic non-singularity. The particular cases in which this occurs all have specified odds ratios that are extremely diverse in magnitude, and the corresponding solution $\mathbf{p}$ values are likewise extremely diverse in magnitude. Sometimes, at such extremes, all the algorithms considered here will fail. Depending upon the particular set of ratios, indeed sometimes Newton–Raphson succeeds when all my alternative algorithms fail. However, there are sets of odds ratios *not* invertible by Newton–Raphson that *are* invertible by SQb and/or other algorithms. Thus, algorithms other than Newton–Raphson are clearly needed, even if a dramatic increase in speed is not regarded as essential.

3.3.1 Calculating the derivatives of $s(\mathbf{p})$

We now consider the calculation of

$$\frac{\partial s(\mathbf{p})'}{\partial \mathbf{p}}$$

which is required for calculating the score function (2.30) of Section 2.3.2 for fully marginal models, and more particularly in the present context when using the Newton–Raphson algorithm.

First, a technicality: rather than $\partial s'/\partial \mathbf{p}$, which involves $s(\mathbf{p})'$ as a row vector, here I instead give $\partial s/\partial \mathbf{p}'$, which leaves $s(\mathbf{p})$ as a column vector, making it easier to refer back to the columnar definitions of the system in Section 3.1. The required $\partial s'/\partial \mathbf{p}$ is the transpose of the matrices shown here.

I now develop a general form for the derivative inductively. First note the general result that

$$\frac{\partial}{\partial a}(a \nsim b) = \frac{\partial}{\partial a}\log(\mathrm{e}^a + \mathrm{e}^b) = \frac{\mathrm{e}^a}{\mathrm{e}^a + \mathrm{e}^b}. \tag{3.51}$$

Then for the univariate case,

$$\frac{\partial s_1(\mathbf{p})}{\partial \mathbf{p}'} = \frac{\partial}{\partial \mathbf{p}'}\begin{pmatrix} p_0 \nsim p_1 \\ -p_0 + p_1 \end{pmatrix} = \begin{pmatrix} \frac{\mathrm{e}^{p_0}}{\mathrm{e}^{p_0}+\mathrm{e}^{p_1}} & \frac{\mathrm{e}^{p_1}}{\mathrm{e}^{p_0}+\mathrm{e}^{p_1}} \\ -1 & 1 \end{pmatrix} \tag{3.52}$$

which can be expressed in terms of the unlogged cell probabilities as

$$\frac{\partial s_1(\mathbf{p})}{\partial \mathbf{p}'} = \begin{pmatrix} \frac{\pi_0}{\pi_0+\pi_1} & \frac{\pi_1}{\pi_0+\pi_1} \\ -\frac{\pi_0}{\pi_0} & \frac{\pi_1}{\pi_1} \end{pmatrix}. \tag{3.53}$$

The unusual form for the second row is chosen to illustrate the emerging symmetry,

which becomes clearer in the bivariate case:

$$\frac{\partial s_2(\mathbf{p})}{\partial \mathbf{p}'} = \begin{pmatrix} \frac{\pi_0}{\pi_0+\pi_1+\pi_2+\pi_3} & \frac{\pi_1}{\pi_0+\pi_1+\pi_2+\pi_3} & \frac{\pi_2}{\pi_0+\pi_1+\pi_2+\pi_3} & \frac{\pi_3}{\pi_0+\pi_1+\pi_2+\pi_3} \\ -\frac{\pi_0}{\pi_0+\pi_2} & \frac{\pi_1}{\pi_1+\pi_3} & \frac{\pi_2}{\pi_0+\pi_2} & \frac{\pi_3}{\pi_1+\pi_3} \\ -\frac{\pi_0}{\pi_0+\pi_1} & -\frac{\pi_1}{\pi_0+\pi_1} & \frac{\pi_2}{\pi_2+\pi_3} & \frac{\pi_3}{\pi_2+\pi_3} \\ \frac{\pi_0}{\pi_0} & -\frac{\pi_1}{\pi_1} & -\frac{\pi_2}{\pi_2} & \frac{\pi_3}{\pi_3} \end{pmatrix}. \tag{3.54}$$

.

In order to generalize, introduce the operation on two scalars a and b

$$\Delta_b^a = \frac{b}{a} \tag{3.55}$$

and extend this to 2-vectors $\mathbf{c} = (c_0, c_1)'$ and $\mathbf{d} = (d_0, d_1)'$, say, as

$$\Delta_{\mathbf{d}}^{\mathbf{c}} = \begin{pmatrix} \frac{d_0}{c_0+c_1} & \frac{d_1}{c_0+c_1} \\ -\frac{d_0}{c_0} & \frac{d_1}{c_1} \end{pmatrix} = \begin{pmatrix} \Delta_{d_0}^{c_0+c_1} & \Delta_{d_1}^{c_0+c_1} \\ -\Delta_{d_0}^{c_0} & \Delta_{d_1}^{c_1} \end{pmatrix}. \tag{3.56}$$

Finally extend this to general vectors, of length a power of two, according to the right-hand expression above, letting subscripts zero and one refer to a vector consisting of the first and second halves, respectively, of the elements of the vector in question.

We have shown

$$\frac{\partial s_T(\mathbf{p})}{\partial \mathbf{p}} = \Delta_{\boldsymbol{\pi}}^{\boldsymbol{\pi}} \tag{3.57}$$

for $T = 1$ and 2 by simple substitution. To see why the formula must hold in general, consider (3.23), repeated here for convenience:

$$s_{T+1}(\mathbf{p}) = \begin{pmatrix} s_T(\mathbf{p_0} \not\sim \mathbf{p_1}) \\ -s_T(\mathbf{p_0}) + s_T(\mathbf{p_1}) \end{pmatrix} \tag{3.58}$$

Then in general

$$\frac{\partial s_{T+1}(\mathbf{p})}{\partial \mathbf{p}'} = \begin{pmatrix} \frac{\partial s_T(\mathbf{p_0} \not\sim \mathbf{p_1})}{\partial \mathbf{p_0}} & \frac{\partial s_T(\mathbf{p_0} \not\sim \mathbf{p_1})}{\partial \mathbf{p_1}} \\ -\frac{\partial s_T(\mathbf{p_0})}{\partial \mathbf{p_0}} & \frac{\partial s_T(\mathbf{p_1})}{\partial \mathbf{p_1}} \end{pmatrix}.$$

In our systems the terms such as $\mathbf{p_0} \not\sim \mathbf{p_1}$ are the tildesums of log probabilities, which

are simply the logs of sums of probabilities, and these sums of probabilities are explicit in the Δ notation. Thus, equation (3.57) holds in full generality.

Expression (3.57) may be calculated readily in languages allowing recursion such as S-PLUS and C. The C function `sdiff.c` given in Appendix A3.3.1 is one such implementation.

3.4 The SM algorithm

The following iterative scheme may be considered either as an application of the method of residual correction or as a quasi Newton–Raphson scheme. I have been unable to demonstrate clear conditions for convergence but in simulations this algorithm only very rarely fails to converge to the solution when one exists. However, convergence is very slow; this section concludes with methods for addressing this.

3.4.1 A residual correction approach

As an alternative to the Newton–Raphson approach of Section 3.3, we now consider using the method of residual correction. I first describe this method in quite general terms, since it may be unfamiliar and will be referred to again in Section 3.5. The particular choices that specify the SM algorithm from the general scheme are given below.

We wish to find a solution $\mathbf{p}^*$ to a nonlinear problem (e.g. that defined in equation 3.23):

$$s(\mathbf{p}) = \boldsymbol{\lambda}, \tag{3.59}$$

where s is not directly invertible, so that the equation cannot be straightforwardly cast into a fixed-point problem. To overcome this obstacle, we introduce an invertible function, M, hopefully 'close' to s in some sense. The difference between evaluations of M and s is denoted

$$\mathbf{c}(\mathbf{p}) = M(\mathbf{p}) - s(\mathbf{p}) \tag{3.60}$$

which is a 'residual' term 'correcting' for the approximation of $s(\mathbf{p})$ by $M(\mathbf{p})$. On

rearranging (3.60) to isolate $s(\mathbf{p})$ and substituting in the equation to be solved, (3.59) can be written $M(\mathbf{p}) = \boldsymbol{\lambda} + \mathbf{c}(\mathbf{p})$, whence, since M is chose to be invertible,

$$\mathbf{p} = M^{-1}[\lambda + \mathbf{c}(\mathbf{p})]. \tag{3.61}$$

This manoeuvre has, as desired, re-expressed (3.59) as a fixed-point problem: the required solution satisfies $\mathbf{p}^* = M^{-1}[\lambda + \mathbf{c}(\mathbf{p}^*)]$.

The standard iterative approach to such a problem (Burden and Faires, 1985) is to iterate according to $\mathbf{p}^{(n+1)} = M^{-1}[\lambda + \mathbf{c}(\mathbf{p}^{(n)})]$. In general, however, there is no simpler expression for $\mathbf{c}(\mathbf{p})$ than its definition (3.60), and it may now be eliminated:

$$\mathbf{p}^{(n+1)} = M^{-1}[\lambda + M(\mathbf{p}^{(n)}) - s(\mathbf{p}^{(n)})]. \tag{3.62}$$

In the SM algorithm, we choose M to be not only invertible but also linear. Equation (3.62) then becomes

$$\boxed{\mathbf{p}^{(n+1)} = [I - M^{-1}s]\,\mathbf{p}^{(n)} + M^{-1}\boldsymbol{\lambda}} \tag{3.63}$$

where operator notation is used for s.

The choice of a trivially invertible matrix M, characterizing the SM algorithm, is given in Section 3.4.3.

In (3.62) and (3.63) we have eliminated the correction terms that motivated our approach. Instead it can be useful to concentrate on these, and eliminate $\mathbf{p}$. Substituting for $\mathbf{p}$ in (3.60) according to (3.61) gives the fixed-point formulation

$$\boxed{\mathbf{c}^{(n+1)} = \mathbf{c}^{(n)} + \boldsymbol{\lambda} - s(M^{-1}(\mathbf{c}^{(n)} + \boldsymbol{\lambda}))} \tag{3.64}$$

At convergence, to $\mathbf{c}^{\infty}$ say, $\mathbf{p}^* = M^{-1}(\boldsymbol{\lambda} + \mathbf{c}^{\infty})$ is the required solution to (3.59).

3.4.2 A quasi Newton–Raphson approach

As mentioned in Section 3.3, the derivative matrix in (3.50) is difficult to invert because of its size, at least for large T. A solution is to substitute some simpler, fixed matrix, say M^{-1}, for the inverse derivative matrix (Burden and Faires, 1985). Substituting M^{-1} for $\partial s/\partial \mathbf{p}$ in (3.50) yields exactly the scheme (3.63) developed above.

Despite the equivalence of these two approaches, we will in this Section consider the method as being residual correction, particularly because the accelerator steps in Section 3.4.4 are based on consideration of successive correction terms.

3.4.3 Choice of M

The choice of M is free, but to be applicable M should be 'close' to s (in the residual correction approach) or $\partial s/\partial \mathbf{p}$ (for quasi Newton–Raphson). The transformation M is described here as a *pseudo-loglinearization* (PLL) of s. The PLL transform replaces $\not\sim$ wherever it occurs in the system s by $+$, while existing addition and subtraction are left unchanged. Although this definition is likely to be more widely applicable than to the current problem, the range of applications is not considered here. The terminology PLL comes from a further equivalent definition avoiding the definition of $\not\sim$: after taking logs of an original system e^s we substitute $\log a + \log b$ for any $\log(a+b)$, etc.

The resulting matrix here is indeed not extremely 'close' to s, but nevertheless convergence to the potentially large correction terms is generally obtained in practice. The absolute value of the difference between the true forms $a \not\sim b$ and the substituted $a + b$ is bounded by

$$\left|(a \not\sim b) - (a+b)\right| = \left| -a \not\sim -b\right| \leq \max(-a,-b) + \log\ 2;$$

the right-hand side of this inequality, from equation (3.2), is positive when a and b are log probabilities (hence negative).

The approximation is worst for the first line of equations in $s(\mathbf{p})$, when the difference

is bounded only by

$$\left|\log \sum e^{p_i} - \sum \log\, p_i\right| \leq \max(-p_i) + \log\, 2^T.$$

For the rest of the equations, in practice the inaccuracy in the positive terms is roughly cancelled by similar inaccuracy in the negative terms — see for example Figure 3.3.

Despite the lack of precision of the approximation of s by a PLL form M, there are two important reasons for studying it further: firstly it is easily defined recursively, making computer implementation simple for general T; secondly, it has an extremely simple inverse as seen shortly below.

As in the recursive definition of the MOR system s_T in terms of s_{T-1}, unsurprisingly the PLL forms M_T exhibit a similar recursive pattern. For example, the PLL matrix for the univariate system s_1 is

$$M_1 = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix},$$

while that for $T = 2$ is readily seen to be, in block form,

$$M_2 = \begin{pmatrix} M_1 & M_1 \\ -M_1 & M_1 \end{pmatrix}.$$

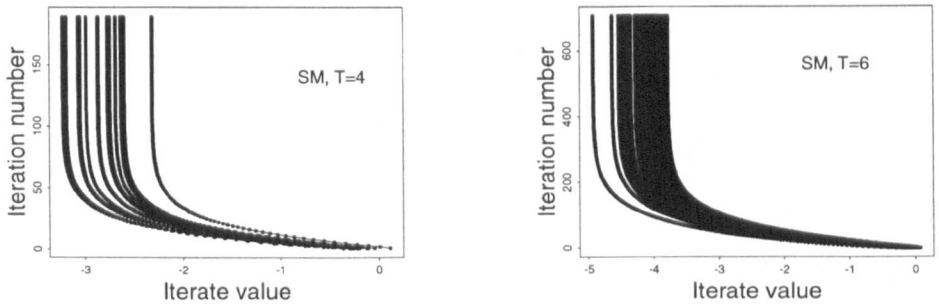
It is straightforward to derive the general case:

$$M_{T+1} = \begin{pmatrix} M_T & M_T \\ -M_T & M_T \end{pmatrix}. \tag{3.65}$$

Importantly, it is easily shown using mathematical induction that, provided M_1 has the given form, M_T satisfies

$$M_T^{-1} = \frac{1}{2^T} M_T' \tag{3.66}$$

Figure 3.2: Trace plots for p convergence to 6 d.p.: values of successive iterates plotted against iteration number, with successive values of each component joined by straight lines (barely visible on these particular examples).



for all T; prime denotes transpose.

This particular choice of M is of enormous numerical benefit, central to the usefulness of the algorithm, because it avoids the need for matrix inversion. The recursion formula (3.65) may be rewritten as

$$M_{T+1}^{-1} = \frac{1}{2}\begin{pmatrix} M_T^{-1} & -M_T^{-1} \\ M_T^{-1} & M_T^{-1} \end{pmatrix} = \frac{1}{2^T}\begin{pmatrix} M_T' & -M_T' \\ M_T' & M_T' \end{pmatrix}. \tag{3.67}$$

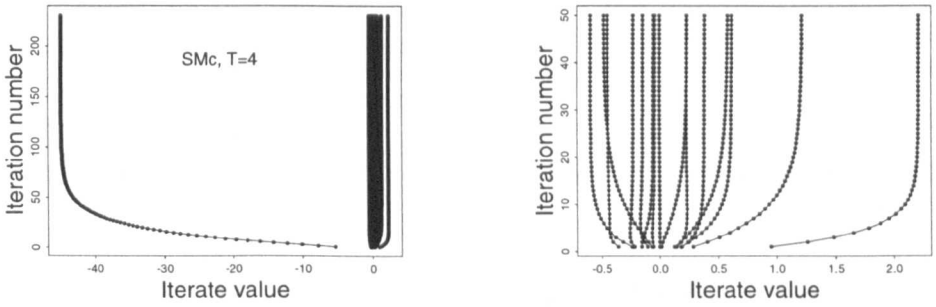
3.4.4 Convergence and accelerator steps

The trace plots in Fig. 3.2 illustrate typical convergence of $\mathbf{p}$ using the SM algorithm. Iterates change fast initially, reaching convergence to within one or two decimal places of the answer after roughly 2^{T+1} iterations. Thereafter, convergence is always slow.

To study convergence behaviour analytically, it is helpful to work in terms of the residual corrections, i.e. to consider the scheme (3.64). Note that (3.64) involves terms $s(M^{-1}(\mathbf{x}))$, which is considered more easily than the expression $M^{-1}(s(\mathbf{x}))$ within the formulation (3.63).

If there is convergence, to $\mathbf{c}^\infty$ say, then $\mathbf{p} = M^{-1}(\mathbf{c}^\infty + \boldsymbol{\lambda})$ is the required solution. In practice it can take up to 30% more iterations to obtain $\mathbf{c}$ convergence to the same

Figure 3.3: Trace plots for **c** convergence, for **c** a vector of sixteen components, i.e. $T = 4$. The right-hand figure is a detail of the first 50 iterations not including the dominant c_0 term.



precision as for **p** convergence for the same $\boldsymbol{\lambda}$.

The correction terms, at convergence, are more diverse in value than are **p**. Figure 3.3 is typical of results in all dimensions, with one component dominating the others in magnitude; this is always the c_0 term. Initial convergence is not monotonic for all components; the pattern can be more extreme when $\boldsymbol{\lambda}$ values differ greatly in magnitude.

However, after possibly a 'burn-in' period of comparatively few iterations, we see that for **c** iterates (as for **p** iterates), the behaviour is very like that of obtaining the sum of a geometric progression (GP) by adding its successive terms. This observation motivates the accelerator steps introduced later in this section, but let us first consider some analytic results for $T = 1$ and $T = 2$.

The univariate problem always has an explicit solution so is a trivial application of the algorithm; nevertheless it is informative to note that after expansion and factorization (3.64) becomes

$$c_0^{(n+1)} = \tfrac{1}{2}c_0^{(n)} + \phi_0(\lambda_1), \tag{3.68}$$

$$c_1 \equiv 0, \tag{3.69}$$

where

$$\phi_0(\lambda_1) = -\left(-\tfrac{1}{2}\lambda_1 \nsim \tfrac{1}{2}\lambda_1\right) = -\log\left(2\cosh\tfrac{1}{2}\lambda_1\right),$$

and $\lambda_0 \equiv 0$ is omitted. Expanding the sequence (3.68) and noting $\mathbf{c}^{(0)} = \mathbf{0}$ gives

$$c_0^{(n+1)} = \frac{1}{2}c_0^{(n)} + \phi_0 = \frac{1}{2}\left(\frac{1}{2}c_0^{(n-1)} + \phi\right) + \phi = \cdots = \left(1 + \frac{1}{2} + \frac{1}{4} + \cdots\right)\phi$$

so that c_0^{∞} is found as the sum of a GP, with first term ϕ_0 and common ratio $\frac{1}{2}$, i.e. $c_0^{\infty} = 2\phi$.

In the bivariate case, after some algebra exploiting the distributive property of addition over tildeplus, iterates are

$$\begin{aligned}
c_0^{(n+1)} &= \tfrac{1}{4}c_0^{(n)} - \phi_0(c_1^{(n)}, c_2^{(n)}, \lambda_1, \lambda_2, \lambda_{12}) && (3.70)\\
c_1^{(n+1)} &= \tfrac{1}{2}c_1 + \tfrac{1}{2}\lambda_1 - \phi_1(c_2^{(n)}, \lambda_2, \lambda_{12}) && (3.71)\\
c_2^{(n+1)} &= \tfrac{1}{2}c_2 + \tfrac{1}{2}\lambda_2 - \phi_2(c_1^{(n)}, \lambda_1, \lambda_{12}) && (3.72)\\
c_{12} &\equiv 0 && (3.73)
\end{aligned}$$

where

$$\begin{aligned}
\phi_0 &= \tfrac{1}{4}(-c_1 - c_2 - \lambda_1 - \lambda_2 + \lambda_{12}) \nsim \tfrac{1}{4}(c_1 - c_2 + \lambda_1 - \lambda_2 - \lambda_{12})\\
&\quad \nsim \tfrac{1}{4}(-c_1 + c_2 - \lambda_1 + \lambda_2 - \lambda_{12}) \nsim \tfrac{1}{4}(c_1 + c_2 + \lambda_1 + \lambda_2 + \lambda_{12}),\\
\phi_1 &= \Psi\left(\tfrac{1}{4}(c_2 + \lambda_2 + \lambda_{12})\right) - \Psi\left(\tfrac{1}{4}(c_2 + \lambda_2 - \lambda_{12})\right),\\
\phi_2 &= \Psi\left(\tfrac{1}{4}(c_1 + \lambda_1 + \lambda_{12})\right) - \Psi\left(\tfrac{1}{4}(c_1 + \lambda_1 - \lambda_{12})\right),
\end{aligned}$$

where

$$\Psi(a) = a \nsim -a. \tag{3.74}$$

When $\lambda_{12} = 0$, for independent variables, ϕ_1 and ϕ_2 are zero, so that (3.71) and (3.72) find the sum of a GP.

Otherwise we may substitute (3.72) in (3.71), assuming convergence, to give us the logarithm of a quadratic equation which can be solved easily, and then by back sub-

stitution find all correction terms analytically. This is not better than using the ordinary solution of Section 3.2.1, nor does it prove that the iterative scheme will converge. Moreover, even if a proof can be given, it could not be used directly to show convergence for $T \geq 3$, which is the important application.

Both iterative schemes (3.63) and (3.64) have the same derivative except for a transposition which does not affect the following argument. To prove convergence it would suffice to show

$$\left\| I - M^{-1} \frac{\partial s}{\partial \mathbf{p}} \right\| < 1, \tag{3.75}$$

for some matrix norm, evaluated at all $\mathbf{p}$ generated within iterations.

It might be possible to make some headway with the very cumbersome expressions in (3.75) using the infinity-norm, but it is easy to find numerical examples with $T \geq 3$ where the infinity-norm exceeds unity. Thus, we have to look to the smallest possible norm, the spectral radius, denoted here ρ, if we are to show convergence.

Unfortunately, it is not feasible to write a closed-form expression for the spectral radius for $T \geq 3$, since this requires obtaining expressions for all the eigenvalues (which are found to be complex, since the matrix of real values is not symmetric). We may nevertheless find ρ numerically for a large number of different $\mathbf{p}$ values and consider the results of this in lieu of formal proof.

The results in Table 3.1 suggest that (3.75) may hold analytically, but certainly not always numerically. It emerges that the lower bound is apparently

$$\rho = \frac{2^T - 1}{2^T}.$$

However for large values of T this rarely obtains, unless components of $\mathbf{p}$ are similar in magnitude (i.e. the variables are near independence). It is rare to find ρ in excess of unity except for large T and/or extremely diverse cell probabilities. The simulations reported in Section 3.7 below suggest this may be numerical artifact rather than proof of non-convergence. This is however of little consolation given that one must proceed numerically.

Table 3.1: Minimum, median and maximum evaluations of the spectral radius of $I - M^{-1}(\partial s/\partial \mathbf{p})$, evaluated at $\mathbf{p}$ values generated by initially letting $\mathbf{p}$ be generated as a random sample from a $\beta(a, b)$ distribution then dividing each component by the sum of the generated set. For the strongly U-shaped generator, the spectral radius can exceed unity. Summaries are for 1000 simulated $\mathbf{p}$ values for each combination.

		2^T					
Generator		4	8	16	32	64	128
$\beta(3,3)$	min	0.7500	0.8750	0.9375	0.9687	0.9844	0.9922
	med	0.7500	0.8750	0.9375	0.9688	0.9844	0.9922
	max	0.8234	0.9417	0.9709	0.9900	0.9956	0.9985
$\beta(1,1)$	min	0.7500	0.8750	0.9375	0.9687	0.9844	0.9922
	med	0.7500	0.8750	0.9581	0.9866	0.9959	0.9988
	max	0.9966	0.9956	0.9998	0.9998	0.9999	1.0000
$\beta(0.1,0.1)$	min	0.7500	0.8750	0.9375	0.9951	1.0000	1.0000
	med	0.7500	0.9905	1.0000	1.0000	1.0000	1.0000
	max	1.0000	1.0000	1.0021	1.0320	1.0713	1.3864

For the examples considered — see Table 3.1, the conjectured lower bound for ρ, and the simulations in Section 3.7 — the SM algorithm takes more iterations to converge with increasing T and as the spectral radius becomes close to unity.

Algorithm SMϕ and its modifications

The factorizations of $s(M^{-1}(\mathbf{c} + \boldsymbol{\lambda}))$ within (3.68) and (3.70)–(3.72) suggest the following:

Conjecture 1 *For any number T of variables, for each component of the vector of correction terms, except the last, which is identically zero, successive iterates are obtained as*

$$c_i^{(n+1)} = r_i c_i^{(n)} + \phi_i^{(n)}, \tag{3.76}$$

where

1. $r_i = 1/m_i$ *given the following recursive definition of the vector*

$$\mathbf{m}_{T+1} = (2\mathbf{m}'_T, \mathbf{m}'_T)' \qquad \textit{with} \quad \mathbf{m}_1 = (2, 1)';$$

2. $\phi_i^{(n)}$ *is a function of possibly all the* $\boldsymbol{\lambda}$ *and possibly all the other correction terms, but* not $c_i^{(n)}$, *for all* n.

Justification. Equation (3.76) holds exactly for $T = 1$, when ϕ is constant (see above). I have verified by expansion that the conjecture holds not only for $T = 2$ (as above) but also for $T = 3$. More precisely, I have verified that for a given set of values c_i then all ϕ_i are independent of the corresponding c_i, but that within the iterations of the SM algorithm, the $c_j^{(n)}$ within $\phi_i^{(n)}$, for at least one $j \neq i$, are functions of $c_i^{(n-1)}$. Thus (3.76) can be considered a "first order" approximation only. For $T \geq 4$, the algebra is too daunting. In addition, I have found that the following algorithm, SMϕ, based on this conjecture, indeed converges whenever the raw SM algorithm converges, for all $T \leq 8$ (though I have not done many simulations at this extreme).

Algorithm SMϕ is based both on Conjecture 1 and on the observation that in all simulations the sequence $\phi_i^{(n)}$ converges faster than does the sequence $c_i^{(n)}$, for all i. By Conjecture 1, for each component, at iteration $n + 1$ we find

$$c_i^{(n+1)} = \phi_i^{(n)} + r_i\phi_i^{(n-1)} + r_i^2\phi_i^{(n-2)} + \cdots + r_i^j\phi_i^{(n-j)} + \cdots \tag{3.77}$$

and in the limit, if there is convergence,

$$c_i^{\infty} = \phi_i^{\infty} + r_i\phi_i^{\infty} + r_i^2\phi_i^{\infty} + \cdots = \frac{\phi_i^{\infty}}{1 - r_i}. \tag{3.78}$$

Assume that at step $n + 1$ the ϕ_i terms appear to have converged; then

$$c_i^{\infty} \approx \frac{\phi_i^{(n)}}{1 - r_i}. \tag{3.79}$$

We find by simulation that successive approximations to c_i^{∞} calculated from the correction terms of the raw SM algorithm generally reach convergence after roughly half

the number of iterations required for convergence of correction terms; see Appendix A3.4.

Algorithm SMϕ denotes making the "safe" acceleration to c_i^∞ once successive SM approximations to it have converged. It is computationally faster, but not equivalent, to check for convergence of ϕ_i; c_i^∞ is estimated to less precision than is ϕ_i since $1/(1-r_i) > 1$ for all components (under Conjecture 1). In practice, I have found it satisfactory to use $\boldsymbol{\phi}$ convergence to one more decimal place than is required for $\mathbf{c}$ convergence. It is clear that this scheme converges whenever SM does, and converges to the same limit. However, even halving the number of SM steps leaves too many to compete with other algorithms.

Algorithm "SMϕ!" denotes taking the unsafe step of using the projected c_i^∞ as the correction term for the next iteration, i.e. given $c_i^{(n)}$, find $c_i^{(n+1)}$ in the usual way, but then take

$$c_i^{(n+2)} = (c_i^{(n+1)} - r_i c_i^{(n)})/(1-r_i).$$

The exclamation mark denotes danger: although when $T \leq 4$ we frequently find a spectacular improvement over SMϕ, for extreme sets of odds ratios, and generally for large T, there is no convergence. Often two or more candidate values appear to 'flip' between two or more values to ever increasing precision; unfortunately none are the correct answer. The obvious approach to this particular problem of considering the mean of each successive pair of SMϕ! projections as a new next iterate fails to converge on sets of $\boldsymbol{\lambda}$ for which unmodified SMϕ! converges very well.

A compromise is possible and has been implemented as Algorithm SMϕ_ϵ. The first modification is to introduce a quantity ϕ_ϵ, or a vector of such values, and at each iteration check for ϕ convergence to precision ϕ_ϵ, componentwise. If this is attained, take an SMϕ! acceleration step on each such component, otherwise continue with an ordinary SM step. So far, this is essentially the same as SMϕ! with the addition of a "burn-in" to precision ϕ_ϵ. This is in itself a generally useful addition because the first SM iterates can be far from the final values and may even move away from them (despite the observation of eventual convergence in almost all cases); see Fig. 3.3.

The second modification is to decrease individual components of ϕ_ϵ by some amount d_ϵ after every SMϕ! acceleration of the corresponding correction term. This is done here by multiplying affected components by some fixed $d_\epsilon < 1$, with the effect that acceleration is disabled after $\lfloor \ln(c_\epsilon/\phi_\epsilon)/\ln d_\epsilon \rfloor$ accelerations have been made, where c_ϵ is the precision required for $\mathbf{c}$ terms. This overcomes the observed oscillation problem because eventually no further acceleration steps are taken and final convergence is under raw SM (if it is not already attained under SMϕ steps). Thus, provided SM converges, SMϕ_ϵ will also converge.

Although SMϕ_ϵ is safe in this sense, the total number of iterations might exceed that of SM, if a poor set of projections leads away from the solution before a final pure SM phase begins. But simulations show it to be a great improvement in all cases; see Appendix A3.4.

The optimum choice of ϕ_ϵ and d_ϵ differs according to the particular set of $\boldsymbol{\lambda}$ values at hand. However, by trial and error initial values of $\phi_\epsilon = 0.01$, for all components, and $d_\epsilon = 0.75$, work well for most $\boldsymbol{\lambda}$ when $T \leq 9$, with $c_\epsilon = 1 \times 10^{-6}$.

The four variants of the SM algorithm are compared in Appendix A3.4, where it is seen that SMϕ_ϵ is preferred over the other three. In the remainder of the main text, only SMϕ_ϵ will be discussed.

Aitken acceleration

Even ignoring Conjecture 1, from inspection of successive iterates in all simulations, all the series appear to converge as a geometric progression (GP), at least after a burn-in that depends on the dimensionality of the problem.

Assuming that n is sufficiently large that we are in the GP tail, then writing Σ_b for the c_i value at the end of burn-in,

$$c_i^{(n+2)} = \Sigma_b + a_i + a_i q_i + \ldots + a_i q_i^{m-2} + a_i q_i^{m-1} + a_i q_i^m, \tag{3.80}$$

$$c_i^{(n+1)} = \Sigma_b + a_i + a_i q_i + \ldots + a_i q_i^{m-2} + a_i q_i^{m-1}, \tag{3.81}$$

$$\vdots$$

for some, unspecified m and constant a_i.

The ratios q_i are found by writing

$$c_i^{(n+2)} - c_i^{(n+1)} = a_i q_i^m, \tag{3.82}$$

$$c_i^{(n+1)} - c_i^{(n)} = a_i q_i^{m-1}. \tag{3.83}$$

Division gives

$$q_i = \frac{c_i^{(n+2)} - c_i^{(n+1)}}{c_i^{(n+1)} - c_i^{(n)}}, \tag{3.84}$$

which is independent of a_i and m.

The GP assumption is that iterate values converge to

$$c_i^{\infty} = c_i^{(n+1)} + \sum_{j=0}^{\infty} b_i q_i^j, \tag{3.85}$$

where by (3.82) and (3.83),

$$b_i = c_i^{(n+2)} - c_i^{(n+1)} = a_i q_i^m.$$

Though the series cannot be assumed to be a true GP we obtain an estimate

$$c_i^{\infty} \approx c_i^{(n+1)} + \frac{b_i}{1 - q_i}. \tag{3.86}$$

Using this projected value as the next iterate can be recognised as Aitken (sometimes called Steffenson) acceleration, although this is usually only presented in textbooks in univariate applications (e.g. Burden and Faires, 1985).

For $T \leq 4$ this acceleration can work well (though not as well as the accelerators discussed above). However when T exceeds 5, convergence is observed to be no better than for raw SM iterations.

Such acceleration is also observed to do very badly when applied to the series of $\mathbf{p}$ iterates using (3.63); convergence is far worse or not obtained when $T \geq 3$.

Aitken acceleration is accordingly not considered further.

3.5 The SR algorithm

In the SM algorithm, the linear approximation M^{-1} to the true nonlinear inverse s^{-1} is quickly evaluated but is not accurate, especially in the first component. We now consider using a better, nonlinear approximation. Since the system is defined recursively, it is natural to find and define such an inverse recursively.

In this section we will work with the unlogged system S. Letting $\pi = (\pi_0', \pi_1')'$ and $\Lambda = (\Lambda_0', \Lambda_1)'$ be partitioned as in Section 3.1.5, the system $S_{T+1}(\pi) = \Lambda$ decomposes as

$$S_T(\pi_0 + \pi_1) = \Lambda_0 \tag{3.87}$$

$$S_T(\pi_1)/S_T(\pi_0) = \Lambda_1 \tag{3.88}$$

where division and multiplication are componentwise. The SR algorithm uses the approximation

$$S(\pi_0 + \pi_1) \approx S(\pi_0) + S(\pi_1), \tag{3.89}$$

which is similar in spirit to the PLL transform of Section 3.4.3, but makes less false replacements of logs of sums by sums of logs. Having made this change to (3.87) the system has block solution

$$S_T(\pi_0) = \frac{\Lambda_0}{1 + \Lambda_1}, \qquad S_T(\pi_1) = \frac{\Lambda_0 \Lambda_1}{1 + \Lambda_1}; \tag{3.90}$$

compare the true univariate inverse in equations (3.24) and (3.25) of Section 3.2.1. Each system in (3.90) is then approximately inverted using the same method, and so on recursively, until we reach univariate systems S_1, for which the analytic inverse is determined.

Denote the approximate inverse found by recursive application of (3.89) in (3.90) as R^{-1}. Although R^{-1} approximates S^{-1} quite well — actually, remarkably well in lower dimensions — it is not precise enough to be a simple substitute. In the SR algorithm we again use the method of residual correction (Section 3.4.1) to find $\mathbf{C}$

such that $\boldsymbol{\pi} = R^{-1}(\boldsymbol{\Lambda}\mathbf{C})$ is the solution of $S(\boldsymbol{\pi}) = \boldsymbol{\Lambda}$, i.e. $\boldsymbol{\pi} = S^{-1}(\boldsymbol{\Lambda})$. For the unlogged system, multiplicative correction terms (based on ratio residuals) are found to be analytically and numerically preferable to additive terms, i.e. in the theoretical development of Section 3.4.1, (3.60) becomes

$$\mathbf{C}(\boldsymbol{\pi}) = R(\boldsymbol{\pi}) \,/\, S(\boldsymbol{\pi}),$$

where R is the inverse of R^{-1} (this exists in simple closed form, but is not given here). The multiplicative version of (3.64) used here is

$$\mathbf{C}^{(n+1)} = \boldsymbol{\Lambda}\mathbf{C}^{(n)} / SR^{-1}(\boldsymbol{\Lambda}\mathbf{C}^{(n)}). \tag{3.91}$$

This algorithm is not as succinctly expressed in terms of $\boldsymbol{\pi}$ or its logarithm, $\mathbf{p}$, directly (unlike the SM algorithm). Because R^{-1} is not linear, both R^{-1} and its inverse would have to be evaluated in the multiplicative equivalent of (3.63):

$$\boldsymbol{\pi}^{(n+1)} = R^{-1}\left(\frac{\boldsymbol{\Lambda} R(\boldsymbol{\pi})}{S(\boldsymbol{\pi})}\right).$$

As for the quality of the approximation, for $T = 2$, only one component of $S(R^{-1}(\mathbf{x}))$ is not mapped to itself; there is only one correction term to calculate. It is shown on page 282 of Appendix A3.5 that (3.91) always converges, to the valid solution, in this case.

For $T \geq 3$, the first, and last two, correction terms are equivalent to unity, which is to say, the approximation is less good with increasing T. Convergence properties are correspondingly worse; see Appendix A3.5 and Section 3.7.3.

A modification of the algorithm, denoted SRquad, is to introduce a term that gives the analytic solution of the quadratic equation for the correction term for the bivariate case. In other words, take $R_2^{-1} = S_2^{-1}$ exactly and do not recurse as far as $R_1^{-1} = S_1^{-1}$. This 'quadratic correction' is applied to every 4-vector encountered in the recursion. When $T = 2$ this modified R^{-1} process, denoted R_q^{-1}, is of course the exact inverse, and there are no residual corrections to be made. For $T \geq 3$ the first, and last four,

components need no residual correction; for brevity, the details are omitted. As before, R_q^{-1} approximates S^{-1} increasingly badly with increasing T. Nevertheless, simulations reported in Appendix A3.5 show the SRquad algorithm is a great improvement over the raw SR form, for $T \leq 6$, even though the 'quadratic corrections' are applied to subsystems for which this is not the true analytic correction at all.

Other types of corrections have been considered extensively. Although various block-wise and partial corrections have been tried, none have worked as well as the above SRquad process.

As seen in Section 3.7.3 and Appendix A3.5, neither SR nor SRquad are sure to converge for $T \geq 3$ even when a valid solution exists. However, when they do converge, the usual situation for nonextreme tables, they do so much more quickly than the SM algorithm of the previous section, and also considerably more quickly than Newton-Raphson.

3.6 The SQ algorithm

Although SR and SRquad are generally much quicker than SMϕ_ϵ, and Newton-Raphson, they do not always converge. We now study a different algorithm that is both quicker and observed to converge more often. Part of this speed gain is no doubt due to the reduction of the problem to the size of that of one less variable, as we shall show below.

Here, in solving $S_T(\boldsymbol{\pi}) = \boldsymbol{\Lambda}$, instead of using approximate inverses, we assume that a numeric (or even analytic) solution to the inverse problem for $T-1$ variables can be found. Then we can write, exploiting the recursive definition of the problem, from (3.88),

$$\boldsymbol{\pi}_1 = S_{T-1}^{-1}(\boldsymbol{\Lambda}_1 S_T(\boldsymbol{\pi}_0)), \tag{3.92}$$

which in (3.87) gives

$$\boldsymbol{\pi}_0 + S_{T-1}^{-1}(\boldsymbol{\Lambda}_1 S_{T-1}(\boldsymbol{\pi}_0)) = S_{T-1}^{-1}(\boldsymbol{\Lambda}_0). \tag{3.93}$$

However, even assuming S_{T-1}^{-1} can be calculated exactly, $\boldsymbol{\pi}_0$ is not found easily, because it is not explicit in (3.93). System (3.93) can be solved analytically, at least for $T = 2$, but gives a system of two simultaneous quadratic equations; substituting, evaluating the coefficients and solving is much less efficient than simply using the bivariate analytic solution of Section 3.2.1.

We will thus proceed iteratively, after finding a suitable fixed-point formula of the general form

$$\boldsymbol{\pi}^{(n+1)} = \mathbf{g}(\boldsymbol{\pi}^{(n)}). \tag{3.94}$$

Naively we might attempt to use (3.93) directly and set

$$\boldsymbol{\pi}_0{}^{(n+1)} = S^{-1}(\boldsymbol{\Lambda}_0) - S^{-1}(\boldsymbol{\Lambda}_1 S(\boldsymbol{\pi}_0{}^{(n)})). \tag{3.95}$$

Unfortunately this is found to generate negative components unless initial estimates are very good, and such components stay negative.

Instead, the SQ algorithm uses

$$\boldsymbol{\pi}_0{}^{(n+1)} = \boldsymbol{\pi}_0{}^{(n)} \left(\frac{S^{-1}(\boldsymbol{\Lambda}_0)}{\boldsymbol{\pi}_0{}^{(n)} + S^{-1}(\boldsymbol{\Lambda}_1 S(\boldsymbol{\pi}_0{}^{(n)}))} \right), \tag{3.96}$$

which is obtained from (3.93) by division by its left-hand side, then multiplying through by $\boldsymbol{\pi}_0$. By using division instead of subtraction, the generation of negative values, as in (3.95), is avoided. Note that the term in large parentheses in (3.96), unity on convergence by construction, decreases the value of the next iterate when the denominator is too large, and increases it when it is too small (for each component) — which while by no means being a proof is encouraging for convergence.

In particular, note that the number of elements in the vector scheme (3.96) is only half that of other algorithms. On convergence, to $\boldsymbol{\pi}_0{}^{\infty}$ say, $\boldsymbol{\pi}_1 = S^{-1}(\boldsymbol{\Lambda}_1 S(\boldsymbol{\pi}_0{}^{\infty}))$ follows from (3.88), as indeed does the less computationally expensive $\boldsymbol{\pi}_1 = S^{-1}(\boldsymbol{\Lambda}_0) - \boldsymbol{\pi}_0{}^{\infty}$ from (3.87).

For discussive purposes, and in mimicry of the computational steps required, scheme (3.96) can be re-expressed as follows: given current estimates $\boldsymbol{\pi}_0{}^{(n)}$ and $\boldsymbol{\pi}_1{}^{(n)}$, calcu-

late

$$\mathbf{x}^{(n+1)} = \boldsymbol{\pi_1}^{(n)}/\boldsymbol{\pi_0}^{(n)}, \tag{3.97}$$

whence

$$\boldsymbol{\pi_0}^{(n+1)} = \frac{S^{-1}(\boldsymbol{\Lambda_0})}{1 + \mathbf{x}^{(n+1)}} \tag{3.98}$$

and then find

$$\boldsymbol{\pi_1}^{(n+1)} = S^{-1}(\boldsymbol{\Lambda_1} S(\boldsymbol{\pi_0}^{(n+1)})). \tag{3.99}$$

This scheme can now be seen as an adaptation of the Gauss–Seidel acceleration method for linear systems (Burden and Faires, 1985), but applied blockwise — i.e. first evaluating the top half of $\mathbf{g}(\boldsymbol{\pi})$ in (3.94), using the latest available updates of $\boldsymbol{\pi_1}$ to modify $\boldsymbol{\pi_0}$, then updating $\boldsymbol{\pi_1}$ using $\boldsymbol{\pi_0}$ just obtained.

A sufficient condition for convergence of a fixed-point scheme is

$$\left| \frac{\partial g_i(\boldsymbol{\pi})}{\partial \pi_j} \right| \leq \frac{K}{2^T} \tag{3.100}$$

for each $j = 1, 2, \ldots, 2^T$ and each component function g_i, for some $K < 1$ (Burden and Faires, 1985). Unfortunately, this cannot in general be evaluated here, since the functions g_i involve elements of the inverse transform S^{-1}, which cannot practically be written in closed form, at least for $T \geq 3$ — the only cases of real interest. I observe that since convergence is not always attained in simulations (Section 3.7), it is clear that the sufficient condition does not hold globally.

A feature of the SQ algorithm is its calculation, in (3.99), of $S^{-1}(\mathbf{a})$ for $\mathbf{a}$ *not* a set of odds ratios; all components of $\mathbf{a}$ are positive, but the first is not necessarily unity, so the answer is not a probability table. From the discussion in Section 3.2, there are multiple solutions $S^{-1}(\mathbf{a})$, some with negative and/or complex values. However, we do not in practice observe convergence to a solution that causes the overall algorithm to fail, except in two related circumstances: poor starting values, and extremely diverse odds ratios, when iterate values may reach machine zero or infinity. Such numerical errors aside, it would appear that Darroch's conjecture (Appendix A3.2) holds in greater generality than originally proposed.

3.6.1 Calculating S^{-1}; algorithm SQb

There remains the question of calculating S^{-1} in (3.93) and (3.92). In the SQ algorithm I use the same method recursively, also noting that $S^{-1}(\mathbf{\Lambda_0})$ need only be found once. This works well, but is wasteful in recalculating the inverse in (3.99) each time. The obvious modification, denoted SQa, is to start with the values from the previous iteration, i.e. $S^{-1}(\mathbf{\Lambda_1} S(\boldsymbol{\pi_0}^{(n)}))$; the closer we are to the true values of $\boldsymbol{\pi_0}$, the greater the saving.

A less obvious further modification, denoted SQb, is to run the recursive evaluations to less precision than required overall, because it is wasteful to calculate a quasi-exact inverse for the wrong term. But as convergence is approached, increasingly greater precision is needed, and ultimately the recursive evaluations must be calculated to one more significant figure than required overall if we are to avoid numerical problems. My solution is to converge subsystems to precision δ/d_2, where δ is the current infinity norm of $\boldsymbol{\pi_0}^{(n+1)} - \boldsymbol{\pi_0}^{(n)}$, and to converge $S^{-1}(\mathbf{\Lambda_0})$ to precision ϵ/d_1, where ϵ is the desired overall precision. The simulations assume $d_1 = 2$ and $d_2 = 5$ except where stated otherwise.

Algorithm SQb offers great improvements over both SQ and SQa as seen in Appendix A3.6. In the comparisons in Section 3.7, only SQb is considered.

An implementation of algorithm SQb in C code is given in Appendix A3.6.

3.6.2 Starting values and an alternative formulation

As mentioned in the discussion around equation (3.100), the SQ algorithm(s) are sensitive to starting values. Using those most easily calculated, the independence values, i.e. $\pi_i = 1/2^T \; \forall i$, works far worse in practice than using the following scheme (a detailed comparison is not given here).

First, instead of arbitrary values ignoring $\mathbf{\Lambda}$ let us use values satisfying one of the two sets of original equations. Since $S^{-1}(\mathbf{\Lambda_0})$ is already calculated, the obvious choice is (3.87), i.e. let us satisfy $\boldsymbol{\pi_0} + \boldsymbol{\pi_1} = S^{-1}(\mathbf{\Lambda_0})$. To determine how much of the marginal probability collapsed over Y_T — i.e. $S^{-1}(\mathbf{\Lambda_0})$ — should be apportioned to each of $\boldsymbol{\pi_0}$

and $\boldsymbol{\pi_1}$, we note that Λ_T (which is the first component of $\boldsymbol{\Lambda_1}$) readily gives μ_T, the marginal mean of Y_T. Specifically,

$$\mu_t = \frac{\Lambda_T}{1+\lambda_T}, \quad 1-\mu_T = \frac{1}{1+\Lambda_T}.$$

At the solution we also require $\sum \boldsymbol{\pi_1} = \mu_T$ and $\sum \boldsymbol{\pi_0} = 1-\mu_T$, where summation is over the elements of each vector. Values that satisfy both (3.87) and the Y_T marginal totals simultaneously are

$$\boldsymbol{\pi_0} = \frac{S^{-1}(\boldsymbol{\Lambda_0})}{1+\Lambda_T}, \quad \boldsymbol{\pi_1} = \frac{\Lambda_T S^{-1}(\boldsymbol{\Lambda_0})}{1+\Lambda_T} = \Lambda_T \boldsymbol{\pi_0}.$$

These values, incidentally, render a first iteration of (3.98) redundant: on eliminating $\mathbf{x}$, by (3.97), (3.98) becomes

$$\boldsymbol{\pi_0}^{(n+1)} = \boldsymbol{\pi_0}^{(n)} \left(\frac{S^{-1}(\boldsymbol{\Lambda_0})}{\boldsymbol{\pi_0}^{(n)} + \boldsymbol{\pi_1}^{(n)}} \right).$$

Thus the first step taken is (3.99).

These are the starting values used for all the simulations reported in Appendix A3.6 and Section 3.7. Despite the above rationale for their choice, the splitting of the marginal probability $S^{-1}(\boldsymbol{\Lambda_0})$ by a constant factor does not always give good results. It works extremely well when μ_T (hence also Λ_T) is very small: then at the solution the subtable $\boldsymbol{\pi_0}$, which contains almost all the probability, is very close to $S^{-1}(\boldsymbol{\Lambda_0})$, so that the starting value is very close to the solution.

When $\mu_T \approx 0.5$, however, there is no such guarantee, and indeed convergence is observed to be worse for such tables (detail omitted).

At the other extreme, for μ_T close to one, $\boldsymbol{\pi_1} \approx S^{-1}(\boldsymbol{\Lambda_0})$, so that starting values for the tiny probabilities $\boldsymbol{\pi_0}$ can be very poor (while those for $\boldsymbol{\pi_1}$ are very good). For such cases it is possible to re-express the algorithm by first isolating $\boldsymbol{\pi_0}$ rather than $\boldsymbol{\pi_1}$ in (3.88):

$$\boldsymbol{\pi_1}^{(n+1)} = \boldsymbol{\pi_1}^{(n)} \left(\frac{\boldsymbol{\pi_0}^{(n)} + \boldsymbol{\pi_1}^{(n)}}{S^{-1}(\boldsymbol{\Lambda_0})} \right), \tag{3.101}$$

$$\boldsymbol{\pi}_0^{(n+1)} = S^{-1}\left(\frac{S(\boldsymbol{\pi}_1^{(n+1)})}{\boldsymbol{\Lambda}_1}\right), \tag{3.102}$$

and in such cases this is observed to work extremely well (details omitted). I have implemented code (not presented here) that chooses scheme (3.98)–(3.99) when $\mu_T < 0.5$ and (3.101)–(3.102) otherwise, but its results are disappointing. Except in very extreme cases, for example $\mu_T > 0.99$, there is surprisingly little gain. In most cases, and especially in real applications, μ_T will not be this extreme, and (3.97)–(3.99) works quite adequately.

Real gains are apparently only to be made by better apportioning the marginal probability $S^{-1}(\boldsymbol{\Lambda}_0)$, according to odds ratios in addition to Λ_T. However, to do so is to develop another algorithm in its own right, which is left for future research. In the remainder of this chapter, alternative formulations and starting values are not considered. The simulations in Appendix A3.6 and Section 3.7 all use the above starting values and the standard form (3.96).

3.7 Comparison of algorithms

Algorithms SMϕ_ϵ, SR, SQb and Newton–Raphson (NR) are compared with each other on both flop count and robustness (in the sense of whether convergence is reached, and if so, whether it is to the desired precision) in Section 3.7.3; variations within versions of the same algorithm are reported in Appendices A3.4–A3.6. The mechanism of simulation is considered in Section 3.7.1 before describing the details of flop counts in Section 3.7.2. A measure of extremity of sets of odds ratios is defined within Section 3.7.1; this measure is shown to be a fair if not good predictor of algorithm performance.

The Summary to Section 3.7.3, beginning on page 144, shows that after much detailed consideration there is a very simple strategy for choosing which algorithm to use — in particular, always use SQb first.

3.7.1 Simulating Λ values

Since one wishes to study $S^{-1}(\mathbf{\Lambda})$, naively we might generate sets of values $\mathbf{\Lambda}$ directly. Bias towards values greater than unity can be avoided by simulating the log ratios, $\boldsymbol{\lambda}$, say from a uniform distribution on $(-a, a)$, with λ_0 set to zero. Unfortunately, however, this method generates sets of odds ratios for which there are no valid solutions, a problem which becomes more acute in higher dimension.

Instead, a probability table, $\boldsymbol{\pi}$, can be generated and by setting $\mathbf{\Lambda} = S(\boldsymbol{\pi})$ the existence of a valid analytic solution is guaranteed. Even then, if the generated $\boldsymbol{\pi}$ values are diverse in magnitude, there can be numerical problems when running the algorithms. For example, the following apparently unexceptional $\boldsymbol{\pi}$ values were generated from a set of random uniform variates on (0, 1), with each term subsequently divided by the sum:

$$\pi_0 = 0.0024,\ \pi_1 = 0.2156,\ \pi_2 = 0.1975,\ \pi_3 = 0.0005,$$
$$\pi_4 = 0.2592,\ \pi_5 = 0.0134,\ \pi_6 = 0.1240,\ \pi_7 = 0.1874.$$

The corresponding odds ratios are

$$\Lambda_0 = 1,\ \Lambda_1 = 0.7149,\ \Lambda_2 = 1.038,\ \Lambda_{12} = 0.6677,$$
$$\Lambda_3 = 1.4037,\ \Lambda_{13} = 0.4846,\ \Lambda_{23} = 1.2577$$

but

$$\Lambda_{123} = \frac{\pi_1\pi_2\pi_4\pi_7}{\pi_0\pi_3\pi_5\pi_6} = 9.39 \times 10^5.$$

Unfortunately such diversity in the odds ratios occurs quite often and many examples are more extreme: the diversity occurs when in the expression for the odds ratio there is one or more very small value in the denominator and no small values in the numerator (or vice versa). It is unclear whether the diversity is purely a numerical artifact or whether we should expect such enormously divergent values for high-order odds ratios in general. Simply setting the high-order ratios to unity, as in the GEE estimation procedure of Liang *et al.* (1992), is not obviously always valid.

Tables with less extreme ratios are generated by narrowing the range before division by the sum: that is, reducing the diversity of the probabilities $\boldsymbol{\pi}$. This is achieved

here by using the beta distributions such as $\beta(3,3)$ and $\beta(5,5)$. Such tables could be more representative of the data encountered in practice. Conversely, by using strongly U-shaped distributions such as $\beta(0.1,0.1)$ in the simulations, diverse odds ratios are generated and can be studied when considering relative robustness.

In the first simulations, considered in Appendices A3.4–A3.6, a set of 420 tables was generated for each size $T = 3,\ldots,7$. Samples of size 60 were drawn from each of the $\beta(i,i)$, distributions for i in $\{0.1, 0.5, 0.75, 1, 3, 5, 10\}$. These 420 tables were grouped into four sets, of increasing 'extremity', regardless of the initial generator. 'Extremity' here (as discussed further shortly) was taken to be the ratio of the largest to the smallest odds ratio in $S(\boldsymbol{\pi})$ and is denoted

$$\eta = \frac{\Lambda_{\max}}{\Lambda_{\min}}. \tag{3.103}$$

The whole set was then divided equally using the quartiles as cutpoints.

With increasing T, table 'extremity' increases for the same $\beta(i,i)$ generator. This can lead to tables near the median of the sets simulated here having at least one odds ratio in excess of a million. Again, whether these should be reclassified as extreme, or accepted as 'normal', is open to doubt and more work is needed.

The arbitrary cutoff points, the quartiles of the first simulation, were replaced in subsequent simulations by the values 100, 500 and 50 000. Although hardly less arbitrary, these cutoff points are based on close scrutiny of the tables where convergence was slow or not obtained in the first simulations.

As seen in Section 3.7.3, the measure of extremity η in (3.103), though *ad hoc*, is a good predictor of Newton–Raphson convergence: when $\eta > 50\,000$, the derivative matrix is mostly numerically singular, causing fatal error, while for $\eta < 50\,000$ this did not occur in any simulation. For other algorithms, η is a somewhat crude measure of extremity, because convergence can be poor for tables not judged to be extreme, and good for 'extreme' tables. In study not detailed here, the ratio $\Lambda_{\max}/\Lambda_{\min}$ is a much better predictor than is $\Lambda_{\max}$ alone, confirming that a more precise predictor of convergence would take into account more than two of the odds ratios.

This is not pursued here, since I cannot offer a clear theoretical framework on which to base such a measure (*cf* the difficulty of specifying conditions for there being a valid solution, Section 3.2.6). Moreover, ultimately η need not be taken account of when choosing which algorithm to use in practice; see Summary subsection on page 144.

3.7.2 Flop counts

Methods of counting

For the SM, SR and Newton–Raphson algorithms it is convenient to calculate the number of floating-point operations (flops) per iteration, and to count the number of iterations to convergence. For the SQ algorithm this is not possible since it calls itself recursively an unpredetermined number of times within each 'outer' loop; in this case a modified version of the program is run, summing additions and multiplications during iterations.

The following pairs of operations are treated as equivalent: addition and subtraction, division and multiplication. These two types are totalled separately here, although on many modern processors manufacturers claim there is little difference between them. It is less clear how to deal with exponentiation and taking logarithms. On RISC machines, such as most university mainframes, such operations are done in software, and the number of additions/subtractions and divisions/multiplications depends on the compiler; source code is generally not made available to the user. Let us assume that exponentiation is based on the evaluation of a series expansion, and assume that the necessary number of multiplications is k_1. Similarly, let the necessary number of multiplications in evaluating a logarithm be k_2. Reasonable estimates for software evaluations are $k_1 = 20$ and $k_2 = 50$ (V. Alexandrov, Department of Computer Science, University of Liverpool; personal communication). As exponentiation is an on-chip operation on some mainframes, however, meaningful comparisons might be made with $k_1 = k_2 = 1$.

In practice the issue is not critical here. Low values would considerably increase the relative performance of SMϕ_ϵ over that reported here, but not to such an extent

that it would be better than other algorithms; the effect on the Newton–Raphson flop count is very minor, since few such operations are required in comparison to the number of ordinary multiplications and additions. The other algorithms do not involve logarithms or exponentials. The evaluation of tildeplus, notationally useful but computationally expensive — see equation (3.1) — is avoided below.

Flops in evaluating $S(\pi)$ and $s(\mathbf{p})$

Referring to equation (3.18), for π a vector of length 2^T, the top half requires $2^{(T-1)}$ additions before recursing, and the bottom half $2^{(T-1)}$ divisions after recursing. Each of the three systems $S_{(T-1)}$ has the same requirements with $2^{(T-2)}$ replacing $2^{(T-1)}$ in the counting, and so on until we reach three systems S_0. Hence the total number of divisions is T terms of the geometric progression

$$2^{(T-1)} + 3^1 \times 2^{(T-2)} + 3^2 \times 2^{(T-3)} + \cdots + 3^{(T-1)} = 3^T - 2^T,$$

which is identical to the total number of additions.

If S is evaluated non-recursively many of the divisions can be replaced by multiplication, but this makes little if any practical difference.

Unless $\not\propto$ is ever implemented as a floating-point CPU instruction (an unlikely event), it is more efficient to evaluate the log system, s, by taking logs of the evaluated unlogged system, at a cost of only 2^T exponentials of the argument passed, and 2^T logarithms of the result.

Glonek and McCullagh (1994) evaluate the form $C \log(L \exp \mathbf{p})$. This takes 2^T logarithms and 2^T exponentials, and two matrix multiplications of vectors, requiring up to $(2^T)^2$ multiplications and additions each. The term "up to" is used because the block diagonality of C might be exploited to reduce one of the calculations. But no matter how efficiently this is done, it takes more operations of all kinds to evaluate the Glonek and McCullagh form than it does to evaluate the form (3.18).

Flops in evaluating M^{-1} for the SM algorithm

Referring to equation (3.67), we see $M^{-1}\mathbf{x}$ can be evaluated without performing any multiplication or division until the vector has been calculated, after which comes division of all elements by 2^T.

The process thus requires only $(2^T)^2$ additions or subtractions and 2^T multiplications or divisions.

Flops in evaluating R^{-1} and R_q^{-1} for the SR algorithms

Working on the unlogged recursive system (3.90), which avoids evaluating tildeplus, at the first stage we calculate $\mathbf{\Lambda_0}/(\mathbf{1}+\mathbf{\Lambda_1})$ and $\mathbf{\Lambda_0\Lambda_1}/(\mathbf{1}+\mathbf{\Lambda_1})$, where division is componentwise and the $\mathbf{\Lambda_i}$ are vectors of length $2^{(T-1)}$. The common denominator requires $2^{(T-1)}$ additions of one, which are considered here as ordinary floating additions, although they could be done more efficiently by CPU increment instructions. There are $2^{(T-1)}$ componentwise divisions in $\mathbf{\Lambda_0}/(\mathbf{1}+\mathbf{\Lambda_1})$, and a further $2^{(T-1)}$ multiplications for the bottom half.

On recursion, these counts are repeated for each half of the system, with $2^{(T-2)}$ replacing $2^{(T-1)}$, until each half is a scalar, for which one division is performed with a further multiplication on the bottom half. Since there are progressively two halves with $2\times 2^{(T-2)}$ divisions/multiplications, then four quarters each with $2\times 2^{(T-3)}$, etc., there are $2\times 2^{(T-1)}$ flops at each stage, giving a total count of $T2^T$ divisions/multiplications. Similarly there are $T2^{(T-1)}$ additions of one in the total.

The 'quadratic correction' is equivalent to finding the analytic solution to the bivariate problem, and as such is evaluated most quickly as in Section 3.2.1. If the square root operation is equivalent to 20 multiplications, the correction is obtained in 37 multiplications and 7 additions. The remaining counts are as above except that one recurses only $T-1$ levels before reaching 4-vectors to which the quadratic correction is applied. The tally is $(T+\frac{33}{4})2^T$ multiplications and $(T+\frac{5}{2})2^{(T-1)}$ additions.

No exponentials or logarithms are required.

Flops in evaluating $\partial s/\partial \mathbf{p}$

With reference to Section 3.3.1, at the first stage of recursion we need to calculate $2^{(T-1)}$ additions for the superscripts of the subsystems, and one sign change, which is disregarded. Each of the four subsystems generates $2^{(T-2)}$ additions, if one does not avoid unnecessary repetitions of the same calculation. This is continued until each Δ is operating on scalars, when there is one division.

There are thus $(2^T)^2$ divisions in total, and the number of additions is T terms of the geometric progression

$$2^{(T-1)} + 4^1 \times 2^{(T-2)} + 4^2 \times 2^{(T-3)} + \cdots + 4^{(T-1)} = 2^{(T-1)}(2^T - 1).$$

If the log probabilities $\mathbf{p}$ have not yet been converted to $\boldsymbol{\pi}$, there are a further 2^T exponentiations to perform, but in the summary counts below, calculating $\boldsymbol{\pi}$ is already accounted for.

In the form proposed by Glonek and McCullagh (1994) one needs here to evaluate $CD^{-1}L$ where $D = \mathrm{diag}(L\boldsymbol{\pi})$. Excluding the 2^T exponentiations needed, i.e. assuming that $\boldsymbol{\pi} = \exp(\mathbf{p})$ is already calculated elsewhere, and avoiding unnecessary multiplication by zero in calculating $D^{-1}L$, the process needs $(2^T)^2$ divisions. Then one must multiply the result by C, with a potential cost of $(2^T)^3$ multiplications and additions (though this could be reduced by exploiting the block diagonality of C). No matter how one estimates this last stage, the process is not as fast as evaluating the matrix using my recursive method.

Flops per SM iteration

Referring to expression (3.63), $M^{-1}\boldsymbol{\lambda}$ is evaluated only once, contributing 2^T divisions and $(2^T)^2$ additions as 'baseline cost' for the first iteration only.

We must evaluate $s(\mathbf{p})$ for the logged system here, which costs 2^T logs and exponentials plus $3^T - 2^T$ divisions and the same number of additions in evaluating the unlogged $S(\boldsymbol{\pi})$. Multiplying this result by $I - M^{-1}$ costs only 2^T divisions, with $(2^T)^2 + 2^T$ additions/subtractions.

Allowing for the 2^T additions of elements of precalculated $M^{-1}\boldsymbol{\lambda}$, the tally is as follows:

$$3^T + (k_1 + k_2)2^T \text{ div/multi}, \quad 3^T + (2^T)^2 + 2^T \text{ add/sub} \tag{3.104}$$

per iteration, where k_1 and k_2 are multiplication costs of exponentials and logs, respectively.

The SMϕ algorithms use (3.64) rather than (3.63), but the flop count is the same for both expressions. However, we now also have to evaluate the vector $\boldsymbol{\phi}$ (2^T multiplications, 2^T subtractions), the absolute difference of $\boldsymbol{\phi}^{(n+1)} - \boldsymbol{\phi}^{(n)}$ (2^T subtractions, sign changes ignored), and perform the accelerations. In SMϕ, acceleration is made only once; in SMϕ!, at every step. In SMϕ_ϵ the number of accelerations per component is unknown *a priori*; the C routine used in my simulations — function `smphi` in Appendix A3.4 — keeps a cumulative count of accelerator steps, on which the results presented below are based.

Flops per SR iteration

With reference to expression (3.91), evaluating $\boldsymbol{\Lambda}\mathbf{C}$ costs at most 2^T multiplications; a fully efficient implementation need only perform multiplication on components where corrections are not fixed at unity. There are then at most 2^T divisions by components of $SR^{-1}(\boldsymbol{\Lambda}\mathbf{C})$, evaluation of which costs $3^T + (T-1)2^T$ multiplications and $3^T + (T-2)2^{(T-1)}$ additions. At its least efficient, the total is

$$3^T + (T+1)2^T \text{ div/multi}, \quad 3^T + (T-2)2^{(T-1)} \text{ add/sub}. \tag{3.105}$$

For algorithm SRquad, this becomes

$$3^T + \left(T + \tfrac{29}{4}\right)2^T \text{ div/multi}, \quad 3^T + \left(T + \tfrac{1}{2}\right)2^{(T-1)} \text{ add/sub}. \tag{3.106}$$

Flops per Newton–Raphson iteration

With reference to expression (3.50), here we need to evaluate firstly $s(\mathbf{p}) - \boldsymbol{\lambda}$, $3^T + (k_1 + k_2 - 1)2^T$ multiplications and 3^T additions, and then evaluate $\partial s/\partial \mathbf{p}$, which

costs at worst $(2^T)^2$ divisions with $\left((2^T)^2 - 2^T\right)/2$ additions (grouping terms to build a polynomial in 2^T). Next, solving a linear system, efficiently and ignoring pivoting manipulations, costs $((2^T)^3 + 3(2^T)^2 - 2^T)/3$ multiplications and $(2(2^T)^3 + 3(2^T)^2 - 5(2^T))/6$ additions (Burden and Faires, 1985). Finally, there are 2^T subtractions, giving a tally

$$\left.\begin{array}{r} \frac{1}{3}(2^T)^3 + 2(2^T)^2 + \left(k_1 + k_2 - \frac{4}{3}\right)(2^T) + 3^T \text{ div/multi}, \\ \frac{1}{3}(2^T)^3 + (2^T)^2 - \frac{1}{3}2^T + 3^T \text{ add/sub.} \end{array}\right\} \tag{3.107}$$

Flops per SQ iteration/convergence

At the 'outer' level of recursion, calculating $\mathbf{x}$ in (3.97) costs $2^{(T-1)}$ divisions, while evaluating (3.98) requires a further $2^{(T-1)}$ divisions, $2^{(T-1)}$ additions, and calculation of $S^{-1}(\mathbf{\Lambda_0})$. The cost of this latter is unknown *a priori*, as is the cost of the inverse in (3.99), though we can count here $2^{(T-1)}$ multiplications by components of $\mathbf{\Lambda_1}$ and $3^{(T-1)} - 2^{(T-1)}$ multiplications and additions in evaluating S_{T-1}.

The immediate cost of an 'outer' iteration is thus

$$3^{(T-1)} + 2^T \text{ div/multi}, \quad 3^{(T-1)} \text{ add/sub.} \tag{3.108}$$

Exceptionally, when $T = 1$ there are 2 multiplications but only one addition in finding the analytic solution. Also, for $T = 2$, in SQ and SQa exact inverses are found, using 37 multiplications and 7 additions (again assuming 20 multiplications in obtaining a square root). In SQb, the exact inverse is found only within calculation of the constant $S^{-1}(\mathbf{\Lambda_0})$; there is overall speed and robustness gain in solving $S^{-1}(\mathbf{\Lambda_1} S(\boldsymbol{\pi_0}))$ to reduced precision.

Total flops including 'inner' recursions are counted in a modified version of the C function **sqbalgor** in Appendix A3.6.

3.7.3 Comparison of flop counts and robustness

While it is impossible to separate entirely the issues of speed of convergence and percentage of tables for which a solution is found, throughout this subsection the

most important aspect for each case is identified and discussed.

The effects of applying the modifications and/or accelerator steps on each of the new algorithms of Sections 3.4–3.6 are discussed in Appendices A3.4–A3.6; here only the 'best' versions of the algorithms are compared. Post-simulation quartile cutoff points for extremity index η are used in the intra-algorithm comparisons, while for ease of interpretation, comparisons between algorithms use the cutoff points η = 100, 500, and 50 000. The presentation of two different cutoff schemes provides more insight into the effect of η than would the use of either alone.

No single algorithm always converges when η exceeds a certain value (depending on T). Beyond a certain limit, none of the algorithms considered here — including Newton–Raphson — ever converge.

Methods

For brevity, only the quickest and/or most robust forms of my algorithms, as demonstrated in Appendices A3.4–A3.6, are here compared with Newton–Raphson. Because the $\mathbf{\Lambda}$ values in the simulations were obtained from known $\boldsymbol{\pi}$, it was possible to assess how close the offered solutions were to the true solutions. All algorithms converged to the same values (if they converged) with the very rare exception of SQb (in Simulation 4 only; details are given there).

The tables that follow denote failure to converge to full desired precision within a specified maximum number of iterations as 'soft' failure; at least some approximation to the true value is found, and if left to run longer, convergence would be obtained. Failures denoted 'fatal' are those where no value at all was returned (e.g. numbers out of range, division by zero, etc.).

To attempt fair comparison, all algorithms were run until convergence to 6 d.p., using the infinity-norm on successive iterate values. This is not entirely satisfactory, because the targets for convergence are different. For Newton–Raphson, log probabilities $\mathbf{p}$ are considered; almost without exception a new iteration is started when there is already convergence to 5 d.p., and this next iteration gives convergence to about 10 d.p. A check against the original simulated $\boldsymbol{\pi}$ values shows convergence to around 17 d.p., i.e.

to full double-precision accuracy. However unfair this may seem in comparing to other schemes, inspection of Tables 3.2–3.5 and 3.7 shows that no substantive difference in conclusions would result in omitting one (or even two) of the final iterations.

Algorithm SMϕ_ϵ converges on correction terms; 6 d.p. convergence here is observed to correspond to between 6 and 10 d.p. convergence for π.

Algorithms SR and SRquad converge on multiplicative correction terms and find π accurate to between 8 and 10 d.p.

As algorithm SQb converges to π directly, the least accurate results of those reported are obtained. However, 8 d.p. precision does not increase substantively the flop count (compare Tables A3 and 3.2) and can actually slightly *improve* the flop count. This is not studied in depth here.

All my algorithms converge quickly to one or two decimal places, rather more slowly to three or four, and finally much more slowly to five or six. Newton–Raphson, on the other hand, converges initially comparatively slowly (depending of course on the quality of the starting value), but once it has found the right direction, it converges very rapidly indeed — provided it is not numerically unstable at the solution.

Probabilities accurate to 6 d.p. appear to be satisfactory when these routines are called within Fisher iterations when fitting marginal models.

Convergence to 6 significant figures rather than decimal places has been considered but rejected. Firstly, except for very extreme tables, probabilities are not so different in magnitude that significant rather than decimal places would make an important difference — and when it *might* make a difference the numerical instability of the algorithms makes the question redundant as there is unlikely to be convergence. Secondly, it has no apparent effect on convergence of Fisher scoring for marginal models, although this claim is not supported by a large-scale simulation.

Calculation of $\mathbf{p}^{(n+1)} - \mathbf{p}^{(n)}$ (or equivalent) is included in the flop counts, to avoid bias for algorithms that take a large number of quick iterations. As in Section 3.7.2, sign changes, comparisons and overhead such as function calls and pivoting (within Newton–Raphson iterations) are not included; these are heavily compiler–implementation dependent. With the possible exception of the latter, for large T,

such overhead is not great in compiled C code.

Simulations for comparison between algorithms

It has already been reported, at the end of Section 3.7.1, that the cutoff points reported in Appendices A3.4–A3.6 are not the best if one is interested in the likely behaviour of a particular algorithm. As a compromise the following comparative simulations were based on tables having odds-ratio extremity η between the following cutoff points: non-extreme ($\eta < 100$), moderate ($100 \leq \eta < 500$), extreme ($500 \leq \eta < 50,000$) and very extreme ($\eta > 50,000$). Exceptionally, Simulation 5 reports results based only on the wider range $\eta < 50\,000$.

Simulation 1: non-extreme ratios (Table 3.2)

The results for so-called non-extreme tables are given in Table 3.2. Not reported there is a single SMϕ_ϵ failure when $2^T = 32$; this is a non-fatal or 'soft' failure, with $\delta = \left\|\pi^{(n)} - \pi\right\| = 3.2 \times 10^{-6}$ after 597K multiplications. Increasing the maximum allowed iterations overcomes this problem, but at a 'flop cost' which can hardly be recommended; SMϕ_ϵ is already the slowest algorithm here. Clearly it is better to use a different algorithm in this range.

There was also an otherwise unreported SQb fail, with $\delta = 1.9 \times 10^{-5}$ after 10K multiplications, when $2^T = 16$. This can be overcome by converging to greater precision at all levels of recursion or by increasing the maximum allowed flop counts. If precision is not vital, one might of course simply use the obtained approximation.

There was one fatal error for SRquad when $2^T = 128$ (it is not recommended here anyway). For smaller T, raw SR fails for $2^T = 16$ (once, fatally), 32 (once fatal, once non-fatal), and 64 (twice fatally, twice non-fatally), while SRquad does not fail.

Algorithms SQb and SRquad are always appreciably quicker than Newton–Raphson, and increasingly so with increasing T. Indeed for $T \geq 6$ even SMϕ_ϵ is on average faster than Newton–Raphson — considerably so for $T = 7$.

In these simulations, for $T \geq 5$ SRquad was somewhat quicker than SQb on average, but occasionally much slower. Since the difference is less than an order of magnitude,

Table 3.2: Flop counts to convergence of algorithms SMϕ_ϵ, SRquad (raw SR for $2^T = 128$), SQb and Newton–Raphson, to 6 d.p. These are summaries of counts for 100 sets of non-extreme odds ratios, with $\eta < 100$.

		SMϕ_ϵ		SR(quad)		SQb		Newton–Raphson		
2^T		multis	adds	multis	adds	multis	adds	multis	adds	iters
8	min	4267	925	428	196	238	150	3500	1068	4
	med	**9739**	**2032**	**646**	**294**	**373**	**246**	**4375**	**1335**	**5**
	upper	18K	3754	891	594	512	345	4375	1335	5
	max	75K	16K	3827	2535	1759	1236	5250	1602	6
16	min	11K	3865	1811	932	1087	779	12K	6852	4
	med	**26K**	**8476**	**2855**	**1463**	**1909**	**1397**	**15K**	**8565**	**5**
	upper	48K	16K	4617	2361	2634	1925	15K	8565	5
	max	272K	89K	12770	6517	7648	5615	21K	12K	7
32	min	7818	1540	1082	490	5440	4112	62K	49K	4
	med	**13K**	**2524**	**1518**	**686**	**7722**	**5882**	**77K**	**61K**	**5**
	upper	19K	3692	1954	882	9299	7066	77K	61K	5
	max	292K	60K	42K	19K	37K	28K	123K	98K	8
64	min	75K	75K	20K	13K	23K	18K	403K	369K	4
	med	**119K**	**116K**	**30K**	**19K**	**32K**	**25K**	**503K**	**461K**	**5**
	upper	183K	177K	36K	23K	41K	31K	503K	461K	5
	max	400K	385K	82K	52K	58K	45K	604K	553K	6
128	min	150K	265K	48K	40K	79K	64K	3.0M	2.9M	4
	med	**219K**	**379K**	**80K**	**66K**	**101K**	**81K**	**3.7M**	**3.6M**	**5**
	upper	265K	455K	90K	74K	115K	92K	3.7M	3.6M	5
	max	449K	761K	963K	791K	143K	114K	4.5M	4.3M	6

and both are clearly in all cases considerably faster than either SMϕ_ϵ or Newton–Raphson, in the interests of simplicity I recommend that SQb should be used in preference to all other algorithms when $\eta < 100$.

Simulation 2: moderate ratios (Table 3.3)

By comparison with those for non-extreme ratios, the flop counts in Table 3.3 are appreciably higher, especially at the maxima.

Most of the reported failures are non-fatal, with SRquad a little more robust than SQb, but generally a fraction slower, and occasionally very much slower than even Newton–Raphson. For large T the speed gain in using SQb in preference to Newton–Raphson is considerable, the compromise being that under SQb approx. 2 out of 1000 runs will not converge to the full 6 d.p. This is a small price to pay for a 25-fold speed gain.

For $T \geq 6$, SMϕ_ϵ continues to be considerably faster than Newton–Raphson at least 75% of the time, but can, on the other hand, be spectacularly slow in some cases.

My recommendation for choice of algorithm here is to first try SQb. In the rather unlikely event of a soft failure, use the approximate solution as starting values for Newton–Raphson. In the very unlikely event of fatal error under SQb — none were seen here — simply use Newton–Raphson from other suitable start values (for suggestions see page 144). The cost of any wasted SQb run is trivial compared to the cost of the Newton–Raphson steps, especially for large T. Though SR generally performs very well, occasional extremely large flop counts exclude its general recommendation (at least, in the absence of a better predictor of performance than η).

Simulation 3: extreme ratios (Table 3.4)

Perhaps the most startling feature of Table 3.4 is in the failure of the Newton–Raphson method for small T. These errors are reported as fatal because a non-invertible derivative matrix was obtained after two or three iterations, despite the analytic proof of nonsingularity given in Glonek and McCullagh (1994). However, failure was a result of poor starting values, not of instability at or near the solution. When re-run from

Table 3.3: Flop counts to convergence of algorithms SMϕ_ϵ, SRquad (raw SR for $2^T = 128$), SQb and Newton–Raphson, to 6 d.p., for $100 < \eta < 500$. Summaries are for 100 sets of ratios, except for $T = 6, 7$, where size was increased to 1000 to better estimate the small percentage of failures. The maxima, and failures, are not necessarily for the most extreme ratios. Rows 'soft' record the percentage of non-fatal failures to converge (i.e. an answer is obtained, though not to the desired precision), while 'fatal' is the percentage where missing values are returned.

		SMϕ_ϵ		SR(quad)		SQb		Newton-Raphson		
2^T		multis	adds	multis	adds	multis	adds	multis	adds	iters
8	min	18K	3754	537	245	370	246	4375	1335	5
	med	**45K**	**9166**	**1736**	**784**	**835**	**576**	**5250**	**1602**	**6**
	upper	58K	12K	2281	1353	1176	817	5250	1602	6
	max	284K	59K	93K	42K	2188	1500	11K	3204	12
	soft	(2)		(1)		(4)		—		
	fatal	—		—		—		—		
16	min	22K	7474	2855	1463	2056	1493	15K	9K	5
	med	**171K**	**56K**	**8336**	**4256**	**4574**	**3353**	**18K**	**10K**	**6**
	upper	300K	99K	14730	7514	6908	5063	21K	12K	7
	max	1.1M	346K	130K	66K	12640	9185	46K	26K	15
	soft	(6)		(1)		(10)		—		
	fatal	—		—		—		—		
32	min	41K	23K	9493	5445	7975	6068	77K	61K	5
	med	**175K**	**97K**	**20K**	**12K**	**15K**	**11K**	**92K**	**73K**	**6**
	upper	479K	265K	39K	23K	21K	16K	108K	85K	7
	max	2.2M	1.2M	295K	169K	62K	46K	139K	110K	9
	soft	(14)		(4)		(7)		—		
	fatal	—		(1)		—		—		
64	min	76K	75K	22K	14K	21K	17K	503K	461K	5
	med	**221K**	**212K**	**45K**	**29K**	**45K**	**35K**	**604K**	**553K**	**6**
	upper	369K	355K	63K	40K	56K	44K	604K	553K	6
	max	5.1M	4.9M	1.4M	891K	497K	373K	1.1M	1.0M	11
	soft	(5.3)		(2.2)		(2.3)		—		
	fatal	—		(0.1)		—		—		
128	min	172K	303K	64K	53K	92K	73K	3.7M	3.6M	5
	med	**391K**	**665K**	**141K**	**116K**	**149K**	**118K**	**3.7M**	**3.6M**	**5**
	upper	653K	1.1M	228K	187K	179K	142K	4.5M	4.3M	6
	max	10.5M	17.8M	3.1M	2.5M	460K	368K	7.4M	7.2M	10
	soft	(0.5)		(10.8)		(0.2)		—		
	fatal	—		—		—		—		

Table 3.4: Flop counts to convergence of algorithms SMϕ_ϵ, SRquad (raw SR for $2^T = 128$), SQb and Newton–Raphson, to 6 d.p., for more extreme ratios, $500 < \eta < 50\,000$. Summaries are for 100 sets of ratios.

		SMϕ_ϵ		SR(quad)		SQb		Newton–Raphson		
2^T		multis	adds	multis	adds	multis	adds	multis	adds	iters
8	min	27K	5K	428	196	244	156	4375	1335	5
	med	**79K**	**16K**	**2826**	**1274**	**1040**	**723**	**6125**	**1869**	**7**
	upper	142K	29K	4870	2193	1599	1114	7000	2136	8
	max	518K	107K	62K	28K	2872	1974	9625	2937	11
	soft	(5)		(1)		(22)		—		
	fatal	—		—		—		(7)		
16	min	67K	22K	4160	2128	2824	2063	15K	9K	5
	med	**369K**	**121K**	**15K**	**7714**	**5908**	**4337**	**18K**	**10K**	**6**
	upper	575K	189K	26K	13K	8782	6341	21K	12K	7
	max	1.2M	388K	204K	104K	22K	16K	34K	19K	11
	soft	(16)		(4)		(27)		—		
	fatal	—		(1)		—		(2)		
32	min	59K	33K	11K	7K	9253	7010	77K	61K	5
	med	**1.0M**	**580K**	**65K**	**37K**	**27K**	**21K**	**108K**	**85K**	**7**
	upper	1.8M	989K	116K	66K	39K	29K	123K	98K	8
	max	2.5M	1.4M	463K	265K	92K	67K	170K	134K	11
	soft	(37)		(5)		(40)		—		
	fatal	—		(6)		(1)		—		
64	min	119K	116K	35K	22K	32K	25K	503K	461K	5
	med	**535K**	**515K**	**173K**	**110K**	**77K**	**60K**	**705K**	**646K**	**7**
	upper	3.1M	3.0M	321K	204K	139K	107K	806K	738K	8
	max	4.9M	4.7M	1.2M	740K	370K	279K	1.2M	1.1M	12
	soft	(54)		(22)		(40)		—		
	fatal	—		(8)		—		—		
128	min	207K	360K	83K	69K	92K	73K	3.7M	3.6M	5
	med	**675K**	**1.1M**	**234K**	**192K**	**199K**	**158K**	**4.5M**	**4.3M**	**6**
	upper	878K	1.5M	356K	293K	244K	191K	4.5M	4.3M	6
	max	8.7M	14.8M	2.6M	2.1M	1.5M	1.2M	7.4M	7.2M	10
	soft	(19)		(39)		(2)		—		
	fatal	—		(—)		(9)		—		

values closer to the solution, convergence was indeed obtained.

A suggestion is to run one of my algorithms until convergence to, say, 2 d.p., and then use these as starting values for Newton–Raphson. (This idea is incorporated in the proposed overall strategy at the end of this section, on page 144.) For the ratios simulated here, with $T = 3$, two iterations of SRquad generally suffices, costing only about 200 multiplications while gaining full robustness.

Alternatively, one can alter the parameters of SQb to eliminate the non-fatal failures altogether. Here, for $T = 3$, letting $d_2 = 1.5$ (see Section 3.6.1) and allowing up to 1000 iterations increases the median flop count to only 1252 multiplications, and the maximum to 21K multiplications. This is then as robust and in almost all cases up to five times quicker than start-value modified Newton–Raphson; unfortunately, there are occasional exceptions.

Indeed the average flop counts and robustness for SQb can be improved for all the examples in Table 3.4. For $T = 7$, for example, setting $d_2 = 1.5$ and increasing maximum allowed iterations to 100 decreases the multiplication count to 80K (min), 165K (median), 212K (upper quartile) and 1.1M (maximum), gives no non-fatal errors, and only four fatal errors.

For $2^T = 128$, a special type of SQb failure is found: the algorithm returns a value with error status 1 (maximum iterations reached, which almost always indicates 'soft' failure) for two cases where there will clearly never be convergence; in one such case, $\left\|\boldsymbol{\pi}^{(n)} - \boldsymbol{\pi}\right\| = 6.8 \times 10^7$. These are reported as fatal failures in the table and can be detected in practice (when $\boldsymbol{\pi}$ is unknown) by comparison of $S(\boldsymbol{\pi})$ evaluated at the 'solution' with the given $\boldsymbol{\Lambda}$.

For large T, convergence is always obtained under Newton–Raphson, but the cost is high in terms of speed. My recommendation is again to first try SQb, and if the algorithm fails, use Newton–Raphson with the approximate SQb solution (if any) as starting values. If SQb fails fatally, generate starting values from a truncated SR run as suggested above. The joint scheme is more efficient on average than is the saving of two or three Newton–Raphson iterations by using the second method only, since SQb will very often converge fully on its own. For $T = 7$, the cost of a wasted SQb

run is similar to a single Newton–Raphson step: 717K multiplications.

If the SQb parameters were set optimally the scheme would be even better; however the nature of the relationship between these parameters, T, and the odds ratios, is not yet known precisely enough to recommend use of other than the default settings.

Simulation 4: very extreme ratios (Table 3.5)

The results in Table 3.5 indicate that no algorithm always converges, and Newton–Raphson is less robust (with respect to fatal failures) than SQb, for $T \leq 6$. For $T \geq 7$, Newton–Raphson appears to be slightly more robust than SQb, but with a 59% failure rate can hardly be called ideal.

These failures are due to the numerical singularity of the derivative matrix evaluated at the solution and are not the result of poor starting values. Good starting values merely ensure the error occurs more quickly.

All the SMϕ_ϵ errors are reported as non-fatal, since the return status indicated maximum allowed iterations were exceeded rather than obvious non-convergence. However, this is misrepresentative because convergence is not obtained in all cases even on setting the maximum iteration count to 10 000, and trace plots show increasing divergence rather than convergence. Not every failure case was thus studied because of the time involved, and because it would not obviously be worthwhile, considering the reported flop counts for maximum iterations set at 1000.

The SRquad and SR algorithms are considerably slower than SQb here, and also less robust. They are thus not considered further.

The non-fatal failures of algorithm SQb are further analysed in Table 3.6, where the maximum number of allowed iterations has been increased. The number of non-fatal failures decreases at the cost of sometimes appreciable increase in fatal failure. The algorithm calculates near-zero probabilities and attempts to divide by them. The true probabilities are close to, but never actually, zero in these simulations so that subtractions can leave the larger probability values unchanged or give numerator/denominator ratios of machine infinity. The problem might be circumvented somewhat by having more than double precision, but is endemic to all algorithms considered here.

Table 3.5: Flop counts to convergence of algorithms SMϕ_ϵ, SRquad (raw SR for $2^T = 128$), SQb and Newton–Raphson, to 6 d.p., for very extreme ratios, $\eta > 50\,000$. Summaries are for 100 sets of ratios.

		SMϕ_ϵ		SR(quad)		SQb		Newton–Raphson		
2^T		multis	adds	multis	adds	multis	adds	multis	adds	iters
8	min	36K	7K	319	147	136	78	5250	1602	6
	med	**100K**	**21K**	**4952**	**2230**	**889**	**618**	**7875**	**2403**	**9**
	upper	274K	56K	15K	7K	1446	999	8750	2670	10
	max	586K	121K	108K	49K	2500	1740	18K	5K	20
	soft	(48)		(20)		(59)		—		
	fatal	—		(20)		(1)		(68)		
16	min	388K	127K	8858	4522	4327	3095	18K	10K	6
	med	**607K**	**200K**	**54K**	**28K**	**8383**	**6137**	**30K**	**17K**	**10**
	upper	961K	316K	121K	62K	11K	8K	36K	20K	12
	max	1.2M	399K	237K	121K	17K	12K	46K	26K	15
	soft	(90)		(21)		(74)		—		
	fatal	—		(48)		(10)		(75)		
32	min	388K	215K	33K	19K	21K	16K	92K	73K	6
	med	**1.1M**	**614K**	**203K**	**116K**	**43K**	**32K**	**123K**	**98K**	**8**
	upper	2.1M	1.2M	311K	178K	51K	38K	154K	122K	10
	max	2.5M	1.4M	549K	314K	94K	68K	200K	159K	13
	soft	(94)		(19)		(63)		—		
	fatal	—		(60)		(27)		(64)		
64	min	—	—	206K	131K	72K	55K	705K	646K	7
	med	**5.2M**	**5.0M**	**519K**	**329K**	**103K**	**79K**	**906K**	**830K**	**9**
	upper	—	—	670K	425K	108K	83K	1.0M	922K	10
	max	—	—	1.2M	738K	191K	146K	1.3M	1.2M	13
	soft	(99)		(6)		(45)		—		
	fatal	—		(83)		(49)		(63)		
128	min	—	—	—	—	563K	441K	5.2M	5.0M	7
	med	—	—	—	—	**639K**	**488K**	**7.4M**	**7.2M**	**10**
	upper	—	—	—	—	—	—	8.2M	7.9M	11
	max	—	—	—	—	640K	504K	15.6M	15.1M	21
	soft	(100)		(93)		(8)		—		
	fatal	—		(7)		(89)		(59)		

Table 3.6: Algorithm SQb applied to the same set of ratios as in Table 3.5, with the maximum number of iterations increased to 2000 for $2^T = 8$ and 16, 500 for $2^T > 16$ (the latter to avoid counts in excess of 100M multiplications). For brevity only the number of multiplications is shown.

	2^T				
	8	16	32	64	128
min	136	4444	21K	72K	563K
med	**3970**	**33K**	**140K**	**615K**	**1.2M**
upper	9246	58K	265K	998K	2.6M
max	151K	768K	1.1M	11.8M	4.4M
soft	2	6	8	2	0
fatal	4	13	40	62	88

The joint algorithm of first running SQb to convergence or fatal error, then running Newton–Raphson if needed, is obviously not robust but might produce an answer; in fact SQb performs well for $T = 3$ and 4 without recourse to the Newton–Raphson algorithm. Using SR to produce approximate starting values is even less robust — here this routine will often crash at the first iteration — and even if it does not, often Newton–Raphson fails at the solution.

Currently the best recommendation would be to use the SQb–Newton–Raphson combination, but not to expect perfect results. A meta-algorithm that might help solve the problem is proposed on page 143, but is not yet implemented. It may be that the best achievable is an approximate solution obtained by this method or by adjusting the parameters for SQb suitably.

Simulation 5: very large T (Table 3.7)

An overview of the problem in even higher dimension is given in Table 3.7. Many of these odds ratio sets are 'extreme' (η approaching 50 000), but this would seem to be an inapt description, judging by the convergence patterns in Tables 3.2 to 3.5.

No attempt has been made to look at even larger dimensions. Anticipating 7 Newton–Raphson iterations is to anticipate 2520 million multiplications until convergence when

Table 3.7: Flop counts to convergence of algorithms SMϕ_ϵ, SR, SQb and Newton–Raphson, to 6 d.p., for very large T, for a wide range of odds ratios, $0 < \eta < 50,000$. Summaries of 100 applications, except for Newton–Raphson, $T = 9$, because of the excessive computer time involved; here only 10 sets of odds ratios were used. There were no Newton–Raphson failures in this small simulation. SQb fatal failures can be eradicated by decreasing the number of allowed iterations.

		SMϕ_ϵ		SR		SQb		Newton–Raphson		
2^T		multis	adds	multis	adds	multis	adds	multis	adds	iters
256	min	478K	1.5M	221K	190K	380K	309K	28.7M	28.3M	5
	med	**957K**	**2.8M**	**638K**	**546K**	**556K**	**448K**	**34.5M**	**34.0M**	**6**
	upper	1.7M	5.1M	1.3M	1.1M	721K	580K	34.5M	34.0M	6
	max	22.6M	66.8M	6.0M	5.2M	1.2M	977K	57.5M	56.6M	10
	soft	(3)		(42)		—		—		
	fatal	—		(20)		(11)		—		
512	min	1.2M	6.2M	1.4M	1.3M	1.5M	1.2M	—		
	med	**1.7M**	**8.5M**	**2.6M**	**2.3M**	**1.8M**	**1.5M**	**317M**	**315M**	**7**
	upper	2.1M	10.8M	3.1M	2.7M	1.9M	1.5M	—		
	max	5.5M	27.8M	4.5M	4.0M	3.8M	3.0M	—		
	soft	(4)		(48)		—		—		
	fatal	—		(2)		(30)		(?)		

$T = 10$, and this is currently impractical. Failure rates might improve as η becomes successively less 'extreme'.

One point is very clear: all my algorithms are enormously quicker than Newton–Raphson for problems in high dimension. For example, for $T = 9$, SQb is on average 175 times faster than Newton–Raphson.

As with previous tables, altering the parameters of the SQb algorithm can improve the reported results even further. For example, reducing the maximum allowed iterations to 20 decreases the failure rate for $T = 9$ to 2 non-fatal and only 3 fatal failures. This may seem counter-intuitive until one considers that if more iterations are allowed within the recursive calls one can get too close to a numerically unstable solution for a sub-iteration based on values not near the solution; when such failures are eliminated the outer loop can proceed to full convergence.

Indeed fatal errors can be avoided altogether in this example by reducing maximum

iterations further, to only 10. The cost is the introduction of 36 non-fatal failures — but even the worst of these is within 5 d.p. of the true solution. Multiplication count median rises to 2.7M, the maximum to 3.9M. The maximum allowed number of iterations could be progressively increased, using successive approximate solutions as start values, but this strategy is left for further study.

The joint algorithm is now first to try SQb, and if this fails, start afresh with SMϕ_ϵ. For large tables, SMϕ_ϵ is much quicker than Newton–Raphson (even noting the high addition counts) and it would appear just as unlikely to suffer fatal error. The SR algorithm, in general quicker, is not robust enough to recommend for general use here. Newton–Raphson, although apparently fully robust, simply takes too long to fit to be seriously considered.

A meta-algorithm for extreme tables

We have seen that for very extreme sets of odds ratios, when $\eta > 50,000$, numerical instability often prevents a solution being obtained under any scheme. Even for less extreme problems, increasing η decreases speed and robustness for all algorithms. A method of addressing this problem is proposed now without any deep study.

Suppose we can easily solve the problem for $\sqrt{\mathbf{\Lambda}}$ (where root is taken component-wise). Since this operation preserves to some extent the relative magnitudes of the interactions — at least they retain the same rank — the solution could be similar to that of the intended system, with somewhat less extreme diversity of cell probabilities. Such a solution would be a good starting value for the original problem, and it will be obtained comparatively quickly (by construction, $\eta_{\text{reduced}} = \sqrt{\eta_{\text{original}}}$).

When applied to SQb, this meta-algorithm has been found to offer convergence in a similar (though always greater) number of flops to that found under raw SQb, but with increased robustness.

Greater robustness, and a little speed, is obtained if one applies the meta-algorithm iteratively, reducing η to 100 or less by successive square-root operations, then finding a solution used as the start value for the square of the final $\mathbf{\Lambda}_{\text{modified}}$, and so on until finding the solution to the original system. An admittedly small number of

simulations suggest the applicability of modifying the starting values so that the ratios of starting values to previous solution is the same as the ratio of the two previous solutions, being a crude approximation to the diversifying effect of squaring the odds ratios: if $\boldsymbol{\pi}^{(1)}$ and $\boldsymbol{\pi}^{(2)}$ are the first and second solutions, start the third process with $\boldsymbol{\pi}^{(\text{start})} = (\boldsymbol{\pi}^{(2)})^2/\boldsymbol{\pi}^{(1)}$ rather than just $\boldsymbol{\pi}^{(2)}$.

One might also use less drastic power transformations than the square root (or indeed any other transformation shrinking the ratios towards unity), gaining robustness in approaching the final solution in smoother steps; but the cost in terms of flop count might be prohibitive.

This approach is not pursued in detail here primarily because this method still cannot solve the major problem for $\eta \geq 50,000$, when there is sometimes instability near the solution (under SQb) or even *at* the solution (for Newton–Raphson, SR and SM). But if one reports the last non-fatal failure of the meta-algorithm, hopefully one has at least a sensible approximation to the true solution. Whether this is useful, say within Fisher iterations when fitting marginal models, remains to be seen.

Summary and discussion

A common theme has emerged in the conclusions of the individual Simulations 1–5 above. Except where noted below, the recommended method for all combinations of T and η is a combined approach as follows:

1. Attempt to obtain a solution using SQb. This generally succeeds, leaving the following redundant. However, in the event of non-fatal failure, retain the approximate solution returned. After fatal failure, simply proceed to the next step.

2. Use Newton–Raphson, with starting values

   (a) obtained from the previous step, if available, or

   (b) obtained from a limited-iteration run of SR(quad) ($3T$ iterations is generally enough to be within 2 d.p. of the true solution), or if this should fail

(c) taken arbitrarily, say all zero (even though this does not represent a valid probability table, it generally works), or all equal and summing to unity (the full independence solution with all means 0.5).

Exceptions are that for $\eta < 100$ one would not ever expect to reach step 2, and for $\eta > 50\,000$ one should not expect an answer at all. Also, for $T \geq 8$, replace the second step, if needed, by using SMϕ_ϵ, because the cost of Newton–Raphson iterations becomes prohibitively expensive at this point.

If great accuracy is required, I have discussed (on page 131) the virtues of Newton–Raphson, perhaps first obtaining start values for this under SR(quad) or SQb to non-fatal failure or reduced precision. This will, however, never be available for extreme η because of inherent numerical instability at the solution.

If precision to even 6 d.p. is not required, then for all simulated ratios considered the optimum is to decrease maximum allowed iterations and use SQb — except if really only 2 d.p. are required, in which case SRquad is always quicker. (Full details have not been given of this study in the interests of brevity.) This approach will generally give non-fatal failures with returned approximate solutions near the true solution, and might even converge to full precision. Simply decreasing the specified precision, rather than the iteration count, does not offer this latter benefit. For example, using the simulated odds ratios for $T = 7$ and $\eta < 50,000$, the SR algorithm is observed to converge to within at worst 2 d.p. of the true solution within 20 iterations in all cases (at a multiplication count of 64K).

The SQb phase of the mixed algorithm approach recommended might be considerably improved by better settings for the control parameters (for intermediate precision determination, and for limiting maximum allowed iterations). The optimum, however, appears to depend on T and the odds ratios in a complex way that is not yet understood. As it is, the defaults used still lead to recommending attempting to use SQb in preference to Newton–Raphson.

The problem of very extreme odds ratios remains unresolved and is likely to remain so, although preliminary results of the meta-algorithm are encouraging in some cases.

I have demonstrated that algorithm SQb is very much faster than Newton–Raphson — by up to two orders of magnitude for problems in large dimension. This will enable the fitting of marginal models on several timepoints, perhaps as many as ten, in an acceptable amount of time. The documentation for the software associated with the paper of Glonek and McCullagh (1995) perforce warns against attempting this.

It is true that SQb does not always converge. However it is equally important to note that in certain circumstances, notably small T and $\eta > 50\,000$, my algorithms offer a solution while Newton–Raphson fails to do so due to numeric instability at the solution.

3.8 The MOR problem and algorithms for polytomous data

3.8.1 The system of equations, $S(\pi)$

The recursive definition of the system of equations $S(\pi)$ in Section 3.1, specifically equation (3.18), is now extended to variables that are polytomous on k categories. Since repeated measures are of major concern, assume that k is constant across timepoints.

Consider firstly a single, ternary variable. Again adopt the convention that the subscripts of probabilities π denote the values taken by the variable (writing the value of y_1 before that of y_2, etc.), while for odds ratios Λ we introduce superscripts for values taken and retain subscripts for variables. The MOR problem is, in terms of multinomial odds ratios,

$$\pi_0 + \pi_1 + \pi_2 = \Lambda_0, \tag{3.109}$$

$$\pi_1/\pi_0 = \Lambda_1^1, \tag{3.110}$$

$$\pi_2/\pi_0 = \Lambda_1^2. \tag{3.111}$$

Here $\Lambda_1^i = \pi_i/\pi_0$ rather than $\pi_i/(1-\pi_i)$. The argument is now developed in terms of such multinomial odds ratios.

Taking care with the order and the form of the equations, the system for two ternary variables can be written as follows:

$$\begin{aligned}
\pi_{00}+\pi_{01}+\pi_{02}+\pi_{10}+\pi_{11}+\pi_{12}+\pi_{20}+\pi_{21}+\pi_{22} &= \Lambda_0, \\
(\pi_{10}+\pi_{11}+\pi_{12}) \,/\, (\pi_{00}+\pi_{01}+\pi_{02}) &= \Lambda_1^1, \\
(\pi_{20}+\pi_{21}+\pi_{22}) \,/\, (\pi_{00}+\pi_{01}+\pi_{02}) &= \Lambda_1^2, \\
(\pi_{01}+\pi_{11}+\pi_{21}) \,/\, (\pi_{00}+\pi_{10}+\pi_{20}) &= \Lambda_2^1, \\
(\pi_{11}/\pi_{01}) \,/\, (\pi_{10}/\pi_{00}) &= \Lambda_{12}^{11}, \\
(\pi_{21}/\pi_{01}) \,/\, (\pi_{20}/\pi_{00}) &= \Lambda_{12}^{21}, \\
(\pi_{02}+\pi_{12}+\pi_{22}) \,/\, (\pi_{00}+\pi_{10}+\pi_{20}) &= \Lambda_2^2, \\
(\pi_{12}/\pi_{02}) \,/\, (\pi_{10}/\pi_{00}) &= \Lambda_{12}^{12}, \\
(\pi_{22}/\pi_{02}) \,/\, (\pi_{20}/\pi_{00}) &= \Lambda_{12}^{22}
\end{aligned}$$

By analogy with the development for binary variables, let $S_0()$ denote the identity and let $S_1(\pi_0, \pi_1, \pi_2)$ be the univariate system. Then with

$$\boldsymbol{\pi_0} = \begin{pmatrix} \pi_{00} \\ \pi_{10} \\ \pi_{20} \end{pmatrix}, \qquad \boldsymbol{\pi_1} = \begin{pmatrix} \pi_{01} \\ \pi_{11} \\ \pi_{21} \end{pmatrix}, \qquad \boldsymbol{\pi_2} = \begin{pmatrix} \pi_{02} \\ \pi_{12} \\ \pi_{22} \end{pmatrix} \tag{3.112}$$

the bivariate system above can be written as

$$S_2(\boldsymbol{\pi_0}, \boldsymbol{\pi_1}, \boldsymbol{\pi_2}) = \begin{pmatrix} S_1(\boldsymbol{\pi_0}+\boldsymbol{\pi_1}+\boldsymbol{\pi_2}) \\ S_1(\boldsymbol{\pi_1}) \,/\, S_1(\boldsymbol{\pi_0}) \\ S_1(\boldsymbol{\pi_2}) \,/\, S_1(\boldsymbol{\pi_0}) \end{pmatrix} = \boldsymbol{\Lambda}. \tag{3.113}$$

By direct analogy with the inductive argument for binary variables, this relationship is readily seen to hold for general T. It is tedious to illustrate even the next step, for which there are 27 equations.

Furthermore, analogous reasoning shows that the fully general problem for variables

polytomous on k categories grows as

$$S_{T+1}(\pi_0, \pi_1, \ldots, \pi_{k-1}) = \begin{pmatrix} S_T(\pi_0 + \pi_1 + \cdots + \pi_{k-1}) \\ S_T(\pi_1) \;/\; S_T(\pi_0) \\ \vdots \\ S_T(\pi_{k-1}) \;/\; S_T(\pi_0) \end{pmatrix} = \mathbf{\Lambda}. \qquad (3.114)$$

The succinctness of this description of the system hides the complexity and more importantly the enormous number of parameters involved, when T and/or k are greater than three.

To review notation: for k-ary variables the vector of cell probabilities $\boldsymbol{\pi}$, when partitioned into k equal divisions, is written

$$\boldsymbol{\pi} = (\boldsymbol{\pi_0}', \boldsymbol{\pi_1}', \ldots, \boldsymbol{\pi_{k-1}})'.$$

This naturally extends the notation for binary variables. The bold subscript represents the value taken by Y_T for all the cell probabilities within the vector.

Similarly, $\mathbf{\Lambda}$ is partitioned as

$$\mathbf{\Lambda} = (\mathbf{\Lambda_0}, \mathbf{\Lambda_1}, \ldots, \mathbf{\Lambda_{k-1}})',$$

but here there is no intrinsic meaning to the bold subscripts other than numerical sequence.

3.8.2 Analytic solutions and considerations

With ternary variables, an algebraic solution is unmanageable even for the bivariate problem. Attempts to obtain a solution using Maple were thwarted by lack of available buffer space. Proceeding as in Sections 3.2.1 and 3.2.3, successive substitution reduces the problem to a system of two simultaneous cubic equations in two cell probabilities: more precisely, one derives a cubic equation in π_{01} involving π_{02} and another cubic in π_{02} involving π_{01}, where the second is derived from the first by swapping the names

of the variables and by substituting 1 for 2 in the superscripts of Λ. These cubic equations are not reproduced here as each requires, before reduction, four sides of A4 paper to print using Maple's highly compressed `lprint` format. (By contrast, the quartic equation given in Section 3.2.3 required one side of A4, before further hand-calculated reduction.)

A solution might be obtained but it is of dubious value. Firstly, as will be seen in Section 3.8.6, the problem can be solved iteratively in less than 1000 floating-point operations. Secondly, by direct analogy with the results for binary variables, even the addition of one further variable yields a system with no closed-form solution, involving linked polynomials of degree 9. A general analytic solution cannot be written.

Again as for binary variables, the constraints on $\boldsymbol{\Lambda}$ that would ensure the existence of a solution are not expressible in any interpretable form. One might conjecture that the bivariate system has a unique solution for *any* positive set of odds ratios. In higher dimensions, extending Darroch's (1962) conjecture, if a solution exists, it is unique. This conjecture is supported in the study of ternary systems in Section 3.8.6, in that convergence to the same solution under three different algorithms suggests uniqueness.

3.8.3 Extension of the SM algorithm

A recursively defined pseudo-loglinear form of (3.114) is easily written, to give the matrix M of the SM algorithm (see Section 3.4.3). Unfortunately, unlike for binary variables, the inverse of this matrix fails to have the simplicity of (3.66): even in the univariate case, we find

$$M = \begin{pmatrix} 1 & 1 & 1 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}; \qquad M^{-1} = \frac{1}{3}\begin{pmatrix} 1 & -1 & -1 \\ 1 & 2 & -1 \\ 1 & -1 & 2 \end{pmatrix}. \tag{3.115}$$

Since the motivation behind the SM algorithm is the avoidance of matrix inversion, I do not consider its extension to polytomous data.

3.8.4 The SR and SQ algorithms extended

Unlike the SM algorithm, both the SR and SQ algorithms extend readily to k-ary data. We consider each in turn.

The SR algorithm extended

By direct analogy with the development of equation (3.90) of Section 3.5, the generalized SR approximation, $S(\sum \boldsymbol{\pi}_i) \approx \sum S(\boldsymbol{\pi}_i)$, gives

$$\left.\begin{aligned} S_T(\boldsymbol{\pi}_0) \approx \frac{\boldsymbol{\Lambda_0}}{1+\boldsymbol{\Lambda_1}+\boldsymbol{\Lambda_2}+\cdots+\boldsymbol{\Lambda_{k-1}}}, \quad S_T(\boldsymbol{\pi}_1) \approx \frac{\boldsymbol{\Lambda_0}\boldsymbol{\Lambda_1}}{1+\boldsymbol{\Lambda_1}+\boldsymbol{\Lambda_2}+\cdots+\boldsymbol{\Lambda_{k-1}}}, \\ \ldots, S_T(\boldsymbol{\pi}_{k-1}) \approx \frac{\boldsymbol{\Lambda_0}\boldsymbol{\Lambda_{k-1}}}{1+\boldsymbol{\Lambda_1}+\boldsymbol{\Lambda_2}+\cdots+\boldsymbol{\Lambda_{k-1}}} \end{aligned}\right\} \tag{3.116}$$

As before, recursive application of the approximation yields an approximate inverse to S_{T+1}, again denoted R_{T+1}^{-1}, and this is used in the residual correction scheme (3.91). For ternary variables, in the bivariate case all but two components of $SR^{-1}(\mathbf{x})$ are mapped to themselves. However, only 15/27 are so mapped for $T = 3$, and just as with binary variables, R^{-1} is an increasingly worse approximation to the true inverse S^{-1} with increasing T.

Since no analytic solution to the bivariate system is given here when $k > 2$, an analogue of the 'quadratic correction' is not proposed.

From the simulations the extended SR algorithm seems to be much faster than Newton–Raphson, but less fast than the following extension to algorithm SQb.

The SQ algorithm extended

By analogy with equation (3.92) of Section 3.6, the last $k - 1$ sets of equations in (3.114) rearrange as

$$\boldsymbol{\pi_i} = S_T^{-1}(\boldsymbol{\Lambda_i} S_T(\boldsymbol{\pi_0})), \quad \mathbf{i = 1, 2, \ldots, k-1}. \tag{3.117}$$

The first set of equations then becomes

$$\boldsymbol{\pi}_0 + \sum_{\mathbf{i}} S_T^{-1}(\boldsymbol{\Lambda}_i S_T(\boldsymbol{\pi}_0)) = S_T^{-1}(\boldsymbol{\Lambda}_0). \tag{3.118}$$

Choosing $\mathbf{x_i}$ to satisfy

$$\boldsymbol{\Lambda}_i S_T(\boldsymbol{\pi}_0) = S_T(\mathbf{x_i}\boldsymbol{\pi}_0) \tag{3.119}$$

leads to a natural extension of the algorithm of Section 3.6. Specifically, given $\boldsymbol{\pi}_i^{(n)}$, $\mathbf{i} = \mathbf{0}, \ldots, \mathbf{k-1}$, let

$$\mathbf{x_i}^{(n=1)} = \boldsymbol{\pi}_i^{(n)}/\boldsymbol{\pi}_0^{(n)}, \quad \mathbf{i} = \mathbf{1}, \mathbf{2}, \ldots, \mathbf{k-1} \tag{3.120}$$

then

$$\boldsymbol{\pi}_0^{(n+1)} = \frac{S^{-1}(\boldsymbol{\Lambda}_0)}{\mathbf{1} + \mathbf{x_1}^{(n+1)} + \cdots + \mathbf{x_{k-1}}^{(n+1)}} \tag{3.121}$$

and

$$\boldsymbol{\pi}_i^{(n+1)} = S^{-1}(\boldsymbol{\Lambda}_i S(\boldsymbol{\pi}_0^{(n+1)})), \quad \mathbf{i} = \mathbf{1}, \mathbf{2}, \ldots, \mathbf{k-1}. \tag{3.122}$$

This is called recursively to evaluate the inverses, S^{-1}, to specified precision.

Since with binary variables SQb was seen to be considerably better than pure SQ or SQa, the same modifications are applied when fitting polytomous data. The potentially enormous improvement over Newton–Raphson is discussed in Section 3.8.6.

3.8.5 The derivatives of $s(\mathbf{p})$ for polytomous data

By direct analogy with Section 3.3.1 and with vector subscript notation as introduced above, for k-ary data define

$$\boldsymbol{\Delta}_{\mathbf{d}}^{\mathbf{c}} = \begin{pmatrix} \Delta_{\mathbf{d_0}}^{\mathbf{c_0}+\mathbf{c_1}+\cdots+\mathbf{c_{k-1}}} & \Delta_{\mathbf{d_1}}^{\mathbf{c_0}+\mathbf{c_1}+\cdots+\mathbf{c_{k-1}}} & \cdots & \Delta_{\mathbf{d_{k-1}}}^{\mathbf{c_0}+\mathbf{c_1}+\cdots+\mathbf{c_{k-1}}} \\ -\Delta_{\mathbf{d_0}}^{\mathbf{c_0}} & & & \\ \vdots & & \text{block diag}\left\{\Delta_{\mathbf{d_i}}^{\mathbf{c_i}}\right\}_{\mathbf{i}=\mathbf{1},\ldots,\mathbf{k-1}} & \\ -\Delta_{\mathbf{d_0}}^{\mathbf{c_0}} & & & \end{pmatrix}, \tag{3.123}$$

but for scalars a and b,

$$\Delta_b^a = \frac{b}{a}. \tag{3.124}$$

By analogous reasoning to that of Section 3.3.1, with this definition

$$\frac{\partial s_T(\mathbf{p})}{\partial \mathbf{p}} = \Delta_{\boldsymbol{\pi}}^{\boldsymbol{\pi}} \tag{3.125}$$

for any number, T, of k-ary observations.

The size of this matrix grows quickly with k as well as with T, leading to the poor performance of the Newton–Raphson technique reported in the following subsection.

3.8.6 Comparison of algorithms

Space and time constraints preclude study or discussion at the same level of detail as for binary data. Here attention is restricted to simulated ternary variables, at values of T and η shown to be of special interest in the binary studies.

Simulated values

Probability tables for multivariate ternary observations were obtained as in Section 3.7.1, guaranteeing the existence of a solution. The extremity index η as defined in (3.103) was again used to classify the tables thus obtained. It would appear that for small T, $\eta > 50,000$ again represents a 'very extreme' table, in that finding a solution becomes problematical under any algorithm.

Flop counts

The counts obtained in Section 3.7.2 are easily generalized.

For the SR algorithm, cf equation (3.105), flops per iteration are

$$(k+1)^T + (T+1)k^T \text{ div/multi}, \quad (k+1)^T + [T(k-1)-k]^{(T-1)} \text{ add/sub.} \tag{3.126}$$

A Newton–Raphson iteration, cf equation (3.107), requires

$$\left.\begin{array}{r} \frac{1}{3}(k^T)^3 + 2(k^T)^2 + \left(k_1 + k_2 - \frac{4}{3}\right)(k^T) + (k+1)^T \text{ div/multi,} \\ \frac{1}{3}(k^T)^3 + \left(\frac{1}{2} + \frac{1}{k(k-1)}\right)(k^T)^2 + \left(1 - \frac{1}{k(k-1)} - \frac{5}{6}\right)k^T + (k+1)^T \text{ add/sub.} \end{array}\right\} \quad (3.127)$$

The cost of an outer loop under SQ, cf equation (3.108), is

$$3^{(T-1)} + 2(k-1)k^{T-1} \text{ div/multi}, \quad (k+1)^{(T-1)} \text{ add/sub.} \quad (3.128)$$

Again here we need to maintain a running total when assessing the performance because counting only outer loops would ignore recursive calls.

Comparison of flop counts and robustness; summary

Table 3.8 details flop counts and failures for the limited number of simulations considered. Newton–Raphson is apparently not at all robust for very extreme tables, though this may be more due to poor starting values than to instability at the solution; compare discussion of Table 3.5 given a similar number of parameters but with binary data.

Algorithms SR and SQb do not fail fatally here, although in many cases convergence to required precision is not obtained with the control settings used for this simulation (maximum iterations 200 for SR; maximum iterations 50, $d_1 = 2$ and $d_2 = 5$ as in Section 3.6 for SQb). The binary study suggests that recalibrating these values for extreme tables could improve the performance.

The optimum strategy for choice of algorithm is the same as for binary data (page 144): in essence, use SQb by choice, Newton–Raphson only out of necessity, should SQb fail to reach desired precision. However, for very large T — for ternary data, I mean $T \geq 5$, giving vectors of length 243 or more — the use of Newton–Raphson is hardly an option because it is so slow at such extremes. Even for $T = 5$, a single Newton–Raphson iteration takes 4.9M multiplications (compared with median full convergence in 58K multiplications under SQb). This rises sharply to 130M multiplications per iteration for $T = 6$, compared with an average of 1.6M to convergence under SQb

Table 3.8: Multiplication counts to convergence of algorithms SR, SQb and Newton–Raphson, to 6 d.p., for ternary variables. Algorithms were fitted to the same simulated Λ values for each case illustrated, except that SR was not fitted at $T = 5$ where it was observed to be slower than SQb. There were no failures under any algorithm when $\eta < 50,000$.

3^T	Extremity	Summary	SR	SQb	NR
9	$\eta < 50,000$	min	120	78	4156
		med	**304**	**144**	**4156**
		max	1195	562	7273
	$\eta > 50,000$	min	163	210	5195
		med	**679**	**298**	**8312**
		max	4979	1046	24K
		soft	(13)	(17)	—
		fatal	—	—	(22)
27	$\eta < 500$	min	833	916	40K
		med	**1693**	**1387**	**50K**
		max	6853	2823	60K
	$\eta > 50,000$	min	3929	2011	50K
		med	**11K**	**4554**	**70K**
		max	30K	9516	170K
		soft	(46)	(49)	—
		fatal	—	—	(59)
243	$\eta < 50,000$	min	—	44K	24M
		med	—	**58K**	**24M**
		max	—	160K	30M

(sample of size 10 only; $\eta < 1 \times 10^6$). Restricting attention to $T = 5$ based on a sample of size 100, it is seen that SQb is on average over 400 times faster than Glonek and McCullagh's (1995) method, offering perhaps one model fit per day rather than one every 13 months.

In addition, SQb will often offer a full solution, or if not that, then an approximate one, while Newton–Raphson fails to do so due to numeric instability at or near the solution.

Chapter 4

Markov chain models

In this chapter we turn from marginal to transitional models, in the form of Markov chain models, introduced in Sections 4.1 and 4.2. I present new methodology, in Sections 4.3 and 4.4, for fitting logistic regression models to the set of transition probabilities at each timepoint simultaneously, allowing for parameters to be shared across timepoint models, in a more general setting than previously reported in the literature. Sections 4.3 and 4.4 cover unordered and ordered categories, respectively, in each case first examining the univariate models before applying them to multivariate outcomes. Both these types of data may occur in the same data set; the score equations are easily adapted to meet this situation, as shown in Section 4.5. Examples with discussion follow in Section 4.6. Further examples and the simultaneous modelling of dropout, which is very naturally handled within a Markov-chain framework, are taken up in Chapter 6.

In previous chapters I considered polytomous data as an extension to binary models, but in Markov chain models the factorized probability function is easier to specify, so that we may start with the assumption that the data are polytomous, and the binary case need not be considered separately.

4.1 Distributions defined by Markov chains

Transitional models have been briefly considered in Section 1.1; we now consider them in more detail. As in previous chapters, let $Y_1, Y_2, \ldots, Y_T$ be T random variables measuring some response on the same unit. In particular, assume for now that the responses are repeated measurements of an underlying variable Y^* at each of T timepoints, and that the responses are ordered naturally from time 1 to time T. The joint distribution may be factorized in time sequence as follows:

$$f(\mathbf{y}) = f(y_1)f(y_2 \mid y_1) \cdots f(y_T \mid y_1, y_2, \ldots, y_{T-1}). \tag{4.1}$$

The right-hand side of (4.1) represents a Markov process of order T. For useful models, we will probably wish to restrict the order to less than T, but in the following theoretical developments no such restriction is imposed.

A general class of multivariate distributions $f(\mathbf{y})$ may be defined according to (4.1) by letting each univariate, conditional distribution on the right-hand side be distributed as D. Restricted to D within the exponential dispersion family, and a process of first order only, such distributions have been discussed by Lindsey (1993). More generally, disregarding any consideration of an underlying stochastic process, D may be any univariate distribution. The resulting marginal distributions may not be D, nor may the joint distribution thus defined coincide with the standard definitions of multivariate analogues of D. For example, the multivariate conditionally-defined Poisson does not have marginal Poisson distributions, from timepoint 2 on, and the joint distribution is not the same as the symmetric multivariate Poisson as defined in, say, Lindsey (1993). Special cases of conditionally-defined Poisson distributions are considered by Lindsey (1993) and Diggle *et al.* (1994).

The implementation of the more general models is a subject for future research. Here I only consider multivariate polytomous data, for which the Markov-chain defined joint distribution is necessarily identical to that of the standard definition considered in Sections 1.4.1 and 2.2. Because of this restriction, I use here the discrete-data terminology *Markov chain* rather than the general *Markov process* and refer to the probabilities

defined by the univariate, conditional distributions as *transition probabilities.*

4.2 Markov chain models for polytomous data

For polytomous data, the canonical parameters, $\boldsymbol{\xi}$, for the joint distribution $f(\mathbf{y})$ in (4.1), considered as a member of the polytomous exponential family, are as described in Chapter 2. In transitional models, however, it is more natural to take as 'canonical' the canonical parameters of the univariate, conditional distributions that define the chain. To maintain a distinction, I refer to these latter as *conditional-canonical* and denote them $\boldsymbol{\phi}$, with $\boldsymbol{\phi}_t$ being the canonical parameters for the time-t conditional model.

A Markov chain model is taken here to mean a set of generalized linear models, one for each of the conditional distributions, assuming links of the form

$$g_t(\boldsymbol{\phi}_t) = \eta_t(\mathbf{h}_t, \mathbf{x}_t^+), \tag{4.2}$$

where

$$\mathbf{h}_t = (y_1, y_2, \ldots, y_{t-1})' \tag{4.3}$$

is as defined by Diggle and Kenward (1994) *history at time t*, and

$$\mathbf{x}_t^+ = (\mathbf{x}_1', \mathbf{x}_2', \ldots \mathbf{x}_t')' \tag{4.4}$$

is covariate history and current values combined. The distributional form (4.1) may now be written more fully as

$$\begin{aligned} f(\mathbf{y} \mid \boldsymbol{\phi}, \mathbf{x}_T^+) &= f(y_1 \mid \boldsymbol{\phi}_1, \mathbf{x}_1) f(y_2 \mid \boldsymbol{\phi}_2, \mathbf{h}_2, \mathbf{x}_2^+) \\ &\quad \times f(y_3 \mid \boldsymbol{\phi}_3, \mathbf{h}_3, \mathbf{x}_3^+) \cdots f(y_T \mid \boldsymbol{\phi}_T, \mathbf{h}_T, \mathbf{x}_T^+). \end{aligned} \tag{4.5}$$

Limiting the number of terms and/or interactions modelled in the linear predictors leads to models commonly given specific names in the literature. Limiting the number of modelled terms within $\mathbf{h}$, for example, reduces the order of the chain, and if

parameters are set to be common for different timepoints can lead to an ordinary, first-order Markov chain with stationary transition probabilities. If there are no interaction parameters for previous values, but only additive terms, the model is known as an *autoregression* (by, for example, Lindsey, 1993; others use the term more loosely). In autoregression, one generally also assumes stationarity; if this is not imposed, the result is an *ante-dependence* model.

In the implementation of models of the distribution (4.5) in this chapter, no restrictions are imposed. Indeed, users are free to specify a model of full order with every **h** interaction included, and include interactions between **h** and **x**, the latter to allow for different dependence structure between groups, or for different ages, for example. Practical models are unlikely to exploit every option, but it is the modeller, not the modelling mechanism developed here, that imposes restrictions.

In the following subsections, I introduce notation and discuss some issues common to both unordered and ordered categories.

4.2.1 Univariate polytomous data; notation

Consider a univariate random variable, Y, polytomous on k categories. Such variables fall into two broad types: either the classes are in some sense ordered, so that Y might be considered as a discretization of an underlying continuous variable, or else the class labels are arbitrary. In the latter case such variables are commonly described as *nominal*, but here I use the terms *ordered* and *unordered* to emphasise the contrast. Binary variables can be classified as either ordered or unordered.

Let the classes be labelled $0, 1, \ldots, (k-1)$, as in Section 2.2.2. If Y is *unordered*, assume that class 0 is the "baseline" category: we may relabel arbitrarily. Letting $y_i \in \{0, 1\}$ be an indicator for an observation falling in class i, a univariate observation may be represented as a vector (over $\boldsymbol{Z}_2^T$):

$$\mathbf{y}_\cdot = (y_1, y_2, \ldots, y_{k-1})'.$$

This is a vector of length $k-1$ with no more than one element non-zero, and with the

zero vector representing an observation of class 0 (that is, one for which $y_0 = 1$).

Another convenient way to describe an outcome is

$$\text{class}(y) = i \iff y_i = 1.$$

This form is preferred for concise representation of outcomes on computer.

I draw special attention to the fact that single subscripts of scalar y here do not refer to timepoints, as they have in all preceding discussion. To highlight the distinction, the general scalar subscript is i rather than t. Subscripts of *vectors* refer to timepoint, e.g. $\mathbf{y}_{t\cdot}$, with the single exception of subscript u, which denotes unit (or subject). When scalar y carries *two* subscripts, the *first* refers to timepoint: y_{3i} is the ith element, y_i, of $\mathbf{y}_{3\cdot}$.

For *ordered* categories, it is more appropriate to take the last class as baseline, because we will use cumulative probabilities in this case. In my notation this class is the $(k-1)$th, hence

$$\mathbf{y}_{\cdot\cdot} = (y_0, y_1, \ldots, y_{k-2})'.$$

Note the double dot subscript for ordered variables. The y_i and class(y) notation remain the same, except that for ordered categories I use the general subscript j rather than i, which is helpful when both types of data are under consideration simultaneously (as they are in an example in Chapter 6, Section 6.6).

In either case, the distribution of a single observation of Y is completely specified by the set of probabilities of Y being recorded in class k: only $k-1$ probabilities need to be specified as the sum must be unity. These probabilities are called *class* probabilities, distinguishing them from contingency-table *cell* probabilities. The distinction is made because a *multinomial* distribution is less easily specified and for this latter distribution inter-observation sampling must be independent with constant probabilities across observations. Note in particular that the familiar problem of multinomial over- or under-dispersion does not arise with single polytomous observations.

The univariate polytomous distribution may be written as

$$f(\mathbf{y}) = \pi_0^{y_0}\pi_1^{y_1}\pi_2^{y_2}\cdots\pi_{k-1}^{y_{k-1}} \tag{4.6}$$

where $\mathbf{y}$ is either $\mathbf{y}_{\cdot}$ or $\mathbf{y}_{:}$ and $\pi_i = \mathrm{P}(\mathrm{class}(y) = i)$, constrained either by $\pi_0 = 1 - (\pi_1 + \cdots + \pi_{k-1})$, with analogous constraint for y_0, for unordered categories, or by $\pi_{k-1} = 1 - (\pi_0 + \cdots + \pi_{k-2}$, with analogous constraint for y_{k-1}, for ordered categories.

4.2.2 Multivariate data

We now return to the joint distribution of polytomous variables

$$\mathbf{Y}_{\cdot} = (\mathbf{Y}'_{1\cdot}, \mathbf{Y}'_{2\cdot}, \ldots \mathbf{Y}'_{T\cdot})'.$$

Note that here $\mathbf{Y}_{\cdot}$ denotes a multivariate observation, but the indexed forms, e.g. $\mathbf{Y}_{t\cdot}$, represent univariate observations written in vector form following the conventions of Section 4.2.1 above. A colon replaces the dot if the categories are known to be ordered. For each timepoint t we model $f(\mathbf{y}_{t\cdot} \mid \boldsymbol{\phi}_t, \mathbf{h}_t, \mathbf{x}_t^+)$ for some appropriate link of the form (4.2). These conditional distributions are assumed to have variationally independent parameters $\boldsymbol{\phi}$, which will lead to fully general models, even though ultimately in practice $\boldsymbol{\phi} = g^{-1}(X\boldsymbol{\gamma})$ will be restricted by choice of design matrix. I illustrate this important point after developing the general form for the score equations, at the end of Section 4.3.2 (page 167), and consider it further in Section 4.5.

The conditional distributions $f(\mathbf{y}_{t\cdot} \mid \boldsymbol{\phi}_t, \mathbf{h}_t, \mathbf{x}_t^+)$ specify transition probabilities from previous state(s), $\mathbf{h}_t$ (possibly truncated), to value i at time t. For unordered categories, the conditional-canonical parameters $\boldsymbol{\phi}$ are the logits of such transitional probabilities, described in Section 4.3.1, so that an identity link is natural. For ordered categories, in Section 4.4, we prefer a cumulative link, which is a nonlinear function of $\boldsymbol{\phi}$.

In both Sections 4.3 and 4.4 below, we precede discussion of multivariate data with a detailed consideration of the univariate case, which is appropriate because Markov

chain models are essentially just a set of univariate models. This no doubt led Lindsey (1993) to say of Markov chains that "these may be studied using logistic or log-linear models; the theory and practice are standard and well known". For models with no parameters shared across timepoints, certainly, standard statistical packages may be used to fit separate models. Given, however, our preference to smooth rather than over-fit (Agresti, 1990), we will want to estimate several common parameters. For common parameter estimation, e.g. for stationary chains, Lindsey (1993) uses log-linear models, but these cannot incorporate continuous covariates. Diggle *et al.* (1994) show how ordinary logistic regression may be used to model stationary chains, but their methodology does not allow fitting of different explanatory variable models at different timepoints, nor fitting common baseline cutoff points for non-stationary chains, which might be called for if timepoints are not equally spaced, or if there are not many timepoints. In this chapter, I develop methodology for fitting the entire spectrum of models, from saturation to independence, using logistic regression for all analyses, within a unified framework.

4.3 Markov chain models for unordered categories

4.3.1 Conditional-canonical links

The univariate probability function (4.6) can be written as

$$f(\mathbf{y}_\cdot) = \pi_0 \left(\frac{\pi_1}{\pi_0}\right)^{y_1} \left(\frac{\pi_2}{\pi_0}\right)^{y_2} \cdots \left(\frac{\pi_{k-1}}{\pi_0}\right)^{y_{k-1}}, \tag{4.7}$$

giving the linear exponential-family canonical form

$$f(\mathbf{y}_\cdot) = \exp\left\{y_1\xi_1 + y_2\xi_2 + \cdots y_{k-1}\xi_{k-1} - C(\boldsymbol{\xi})\right\}, \tag{4.8}$$

where the ξ are multinomial-style logits with reference to class 0 as baseline: $\xi_i = \log\left(\frac{\pi_i}{\pi_0}\right)$. The normalizing constant is $C(\boldsymbol{\xi}) = -\log \pi_0$, as in equations (1.22) and (1.25) of Section 1.4.1.

The univariate models of interest in Markov chain modelling are conditional on pre-

vious values, and we denote their canonical parameters ϕ rather than ξ to emphasise that these are not the same canonical parameters as in Chapter 2. Using the standard form $\mathrm{P}(A \mid B) = \mathrm{P}(A \cap B)/\mathrm{P}(B)$ and denoting $\pi_{qr} = \mathrm{P}(Y_{1q} = 1, Y_{2r} = 1)$, etc., the conditional probabilities are seen to be of the form

$$\mathrm{P}(Y_{ti} = 1 \mid Y_{1q} = 1, Y_{2r} = 1, \ldots Y_{(t-1)s} = 1) = \frac{\pi_{qr\cdots s0}}{\pi_{qr\cdots s+}} \left(\frac{\pi_{qr\cdots s1}}{\pi_{qr\cdots s0}} \right)^i, \tag{4.9}$$

where $+$ denotes summation over an index. Note that while they are conditional in the ordinary sense, such probabilities are marginal with respect to future observations, showing that Markov chain models have some degree of reproducibility inherent.

Writing the set of probabilities (4.9), for $i = 1, 2, \ldots, k-1$, in the form of (4.7), shows that the conditional-canonical parameters are multinomial logits,

$$\phi_{ti} = \log \left(\frac{\pi_{qr\cdots si}}{\pi_{qr\cdots s0}} \right), \tag{4.10}$$

that is, logits with respect to baseline class 0 of the probabilities in the appropriate row of the transition matrix, given history.

An obvious (though not universally justifiable) choice of link is the identity function, since ϕ_t may lie anywhere on the real line. Thus, using α for 'intercepts' (which includes parameters setting different intercepts for different histories) and β for explanatory slopes, we get a sequence of links such as the following:

$$\begin{aligned} \phi_{1i} &= \alpha_{1i} + X^*_{1i}\beta_{1i} && (4.11) \\ \phi_{2j \mid y_{1i}=1} &= \alpha_{2j} + \alpha_{2y_{1i}} y_{1i} + X^*_{2j}\beta_{2j} && (4.12) \\ &\vdots \end{aligned}$$

where the X^* are design matrices. For the model to be ϕ-unconstrained, there is a separate parameter α_{1i} for *each* i at time 1, and a set of time-2 α guaranteeing a separate parameter $\phi_{2j \mid y_{1i}=1}$ for *each* time-2 class j given *each* time-1 class i, etc. Note that since at most one element of $\mathbf{Y}_{1\cdot}$ is nonzero this link is written more clearly

and concisely as

$$\phi_{2j\,|\,\mathbf{y}_{1\cdot}} = \alpha_{2j} + \boldsymbol{\alpha}'_{21j}\mathbf{y}_{1\cdot} + X^*_{2j}\boldsymbol{\beta}_{2j},$$

using the set of values $\mathbf{y}_{1\cdot}$ as dummy variables.

The time-3 link highlights the explosion of parameters for $\boldsymbol{\phi}$-unconstrained distributions:

$$\phi_{3k\,|\,\mathbf{y}_{1\cdot},\mathbf{y}_{2\cdot}} = \alpha_{3k} + \boldsymbol{\alpha}'_{31k}\mathbf{y}_{1\cdot} + \boldsymbol{\alpha}'_{32k}\mathbf{y}_{2\cdot} + \boldsymbol{\alpha}'_{3wk}\mathbf{w}_{12\cdot} + X^*_{3k}\boldsymbol{\beta}_{3k},$$

where $\mathbf{w}_{12\cdot}$ is a vector of time-1 and time-2 class indicator interactions.

Unless the sample size is large, there is danger of overspecification. An unconstrained intercept model, even if desirable, is frequently not feasible technically, and we will need to impose some model restrictions. This could be by specifying restrictions on the structure of the Markov chain through constraints on $\boldsymbol{\phi}$ using $\boldsymbol{\alpha}$, or by collapsing the original data to fewer classes. In certain circumstances it might be possible to treat *a priori* unordered classes as though ordered.

Despite the number of parameters, it is usually easier to interpret those relating to transitional probabilities than those relating to high-order marginal odds-ratios; see Section 4.6.2 for an example.

In the remainder of this section, concerned with estimation rather than interpretation, the distinction between intercepts, $\boldsymbol{\alpha}$, and slopes, $\boldsymbol{\beta}$, need not be made, and we simplify by writing the conditional-canonical links as

$$\phi_{tk\,|\,\mathbf{y}_{1\cdot},\mathbf{y}_{2\cdot},\ldots,\mathbf{y}_{(t-1)\cdot}} = X_{tk}\boldsymbol{\gamma}_{tk},$$

where

$$X_{tk} = ([I_{k_t} \quad H^*_{tk}] \quad X^*_{tk}) \qquad \text{and} \quad \boldsymbol{\gamma} = (\boldsymbol{\alpha}', \boldsymbol{\beta}')',$$

for H^* the design matrix for the historical observed values.

4.3.2 Maximum likelihood estimation of parameters

The univariate case

A univariate multinomial sample has likelihood proportional to that of a Poisson distribution (see for example Aitken *et al.*, 1989), so standard packages such as GLIM can be used for fitting. However, this method is impractical if there are many explanatory factors, and is not readily extended to the multivariate case. Thus I derive here direct maximum likelihood estimation techniques.

From equation (4.8), with ϕ replacing ξ for conditional logits, the contribution of the uth polytomous observation to the log likelihood is

$$\ell_u = y_{u1}\phi_{u1} + y_{u2}\phi_{u2} + \cdots + y_{u(k-1)}\phi_{u(k-1)} - C(\boldsymbol{\phi}_u), \tag{4.13}$$

where dependence of the parameters $\boldsymbol{\phi}_u$ on u is via the link

$$\mathbf{g}(\boldsymbol{\phi}_u) = X_u\boldsymbol{\gamma}.$$

The score contribution for the uth observation is obtained from

$$\frac{\partial \ell_u}{\partial \boldsymbol{\gamma}} = \frac{\partial \boldsymbol{\phi}_u'}{\partial \boldsymbol{\gamma}} \frac{\partial \ell_u}{\partial \boldsymbol{\phi}_u},$$

where

$$\frac{\partial \ell_u}{\partial \boldsymbol{\phi}_u} = \mathbf{y}_{u\cdot} - \boldsymbol{\pi}_{u\cdot}$$

where $\boldsymbol{\pi}_{u\cdot} = (\pi_{u1}, \pi_{u2}, \ldots, \pi_{u(k-1)})'$ is the vector of class probabilities determined by $\boldsymbol{\phi}_u$. Note this excludes $\pi_0 = 1 - \sum_{i=1}^{k-1} \pi_i$. We have used $\partial C(\boldsymbol{\phi}_u)/\partial \phi_{ui} = \pi_{ui}$, which is easily shown (as in equation 2.10 on page 45). In particular, for a canonical link, the score contribution is

$$\mathbf{U}_u(\boldsymbol{\gamma}) = X_u'(\mathbf{y}_{u\cdot} - \boldsymbol{\pi}_{u\cdot}). \tag{4.14}$$

In general the score contribution is

$$\mathbf{U}_u(\gamma) = \frac{\partial \phi_u'}{\partial \gamma}(\mathbf{y}_{u\cdot} - \boldsymbol{\pi}_{u\cdot}), \tag{4.15}$$

although for certain links it may be simpler not to write the likelihood and derivatives in terms of ϕ at all (see, for example, cumulative-logit links in the next section).

For the canonical link, the information matrix contribution from subject u is given by either taking the second derivative directly or by calculating the expected value of $\mathbf{U}_u\mathbf{U}_u'$; the direct approach yields

$$\mathcal{I}_u(\gamma) = X_u' \frac{\partial \boldsymbol{\pi}_u}{\partial \phi_u'} X_u, \tag{4.16}$$

where

$$\frac{\partial \boldsymbol{\pi}_u}{\partial \phi_u'} = \operatorname{diag}(\boldsymbol{\pi}_{u\cdot}) - \boldsymbol{\pi}_{u\cdot}\boldsymbol{\pi}_{u\cdot}' = \operatorname{var}(\mathbf{Y}_u), \tag{4.17}$$

easily derived by noting that $\pi_{ui} = \mathrm{e}^{\phi_{ui} - C(\phi_u)}$.

Estimates of γ are then obtained by the standard Newton–Raphson iterative scheme, here equivalent to Fisher scoring:

$$\gamma^{(s+1)} = \gamma^{(s)} - \left[\sum_u \mathcal{I}_u(\gamma^{(s)})\right]^{-1} \left[\sum_u \mathbf{U}_u(\gamma^{(s)})\right]. \tag{4.18}$$

The multivariate case

For a Markov-chain defined distribution, by (4.1) the contribution to the log likelihood of the uth multivariate observation is

$$\ell_u = \ell_{u1\cdot} + \ell_{u2\cdot} + \cdots + \ell_{uT\cdot} \tag{4.19}$$

where each of the univariate $\ell_{ui\cdot}$ is of the form given in the previous subsection (equation 4.13). For the canonical link, by concatenating

$$\mathbf{y}_{\cdot} = \begin{pmatrix} \mathbf{y}_{1\cdot} \\ \mathbf{y}_{2\cdot} \\ \vdots \\ \mathbf{y}_{T\cdot} \end{pmatrix}, \quad \boldsymbol{\pi}_{\cdot} = \begin{pmatrix} \boldsymbol{\pi}_{1\cdot} \\ \boldsymbol{\pi}_{2\cdot} \\ \vdots \\ \boldsymbol{\pi}_{T\cdot} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\phi} = \begin{pmatrix} \boldsymbol{\phi}_1 \\ \boldsymbol{\phi}_2 \\ \vdots \\ \boldsymbol{\phi}_T \end{pmatrix} = X\boldsymbol{\gamma}$$

where parameters $\boldsymbol{\gamma}$ are *not* assumed to partition into sets for each timepoint, we obtain from the previous results the score contribution of the uth subject:

$$\mathbf{U}_u(\boldsymbol{\gamma}) = X_u'(\mathbf{y}_{u\cdot} - \boldsymbol{\pi}_{u\cdot}). \tag{4.20}$$

This is identical to (4.14) except that the vectors have been extended. The information-matrix contribution is again

$$\mathcal{I}_u(\boldsymbol{\gamma}) = X_u' \frac{\partial \boldsymbol{\pi}_{u\cdot}}{\partial \boldsymbol{\phi}_u'} X_u, \tag{4.21}$$

but where now since the parameters $\boldsymbol{\phi}_t$, and hence the probabilities $\boldsymbol{\pi}_{t\cdot}$, of the individual conditional distributions are assumed to be distinct,

$$\frac{\partial \boldsymbol{\pi}_{u\cdot}}{\partial \boldsymbol{\phi}_u'} = \text{block diag}\left(\frac{\partial \boldsymbol{\pi}_{ui\cdot}}{\partial \boldsymbol{\phi}_{ui}'}\right). \tag{4.22}$$

The nonzero blocks are each as given for the univariate case (equation 4.17). A modification of Azzalini's approximate information matrix (Section 1.4.5) might be used; the appropriate modification is as discussed on page 175 for ordered categories. However, (4.21) is sufficiently simple to calculate that it is not obviously worth approximating.

If the elements of $\boldsymbol{\gamma}$ are distinct for every time point t, then clearly the system decomposes completely into T separate regression problems, and such parameters are orthogonal. An equivalent formulation of this limiting special case is that the design matrix X is block diagonal (with the obvious partitioning). It is assumed that this will not be the case for models of interest.

Since we have the link

$$\mathbf{g}(\boldsymbol{\phi}) = X\boldsymbol{\gamma},$$

the independence of the canonical parameters, $\boldsymbol{\phi}$, is lost as soon as we depart from γ orthogonality. Despite these constraints, we may still fit the model according to equations (4.20), (4.21) and (4.22) above; the necessary 'collapsing' over less-than-saturated γ is handled by multiplication by the appropriate model matrix, X, in (4.20) and (4.21). This is due to the general result in vector analysis that if $\ell = \ell(\phi_1(\gamma), \phi_2(\gamma), \ldots)$ we can obtain the partial derivatives with respect to γ either by the chain rule (as above) or by re-writing the likelihood in terms of γ and differentiating directly.

The importance of this assertion is that a single computer routine suffices to fit any of the possible restricted models, from the (probably meaningless) completely null model with one γ parameter, through to saturation; the required constraints are supplied by the user by specification of the model matrix. As already argued on page 47, in similar circumstances, for any particular model it could be of some computational advantage to re-write the likelihood function in terms of γ and to take derivatives directly, but the consequent loss of flexibility is a high price to pay; we usually fit several different models during any real data analysis exercise, and do not want to program each model separately.

4.4 Markov chain models for ordered categories

If there is a natural order to the classes $0, 1, \ldots, (k-1)$, then this should be reflected in the model. In particular, it is no longer appropriate to link to unconstrained, unordered multinomial logits, as in the previous section. A natural choice for ordered categories is a cumulative link, whether logit (for an underlying logistic), probit (for underlying normal) or complementary log-log (for underlying exponential distribution). Agresti (1990) gives a comprehensive introduction to the use of such link functions.

As in the previous section, an algorithm for fitting the univariate case is developed

and then its extension to multivariate data is described.

4.4.1 The univariate case

Notation and the cumulative link

In the present context it is easier to develop procedures for fitting models writing the distribution in the form (4.6) rather than the canonical (4.7). Thus the contribution of the uth unit to the log likelihood is

$$\ell_u = y_{u0}p_{u0} + y_{u1}p_{u1} + \cdots + y_{u(k-1)}p_{u(k-1)}, \tag{4.23}$$

where $p_{ui} = \log \pi_{ui}$ are the log class probabilities, which here assume a similar function to that of the canonical parameters for unordered data. Recall

$$\mathbf{y}_{u:} = (y_{u0}, y_{u1}, \ldots, y_{u(k-2)})',$$

where each $y_{ui} \in \{0, 1\}$, no more than one y_{ui} is nonzero, and the zero vector represents an observation of class $(k-1)$. Furthermore $y_{u(k-1)} = 1 - \sum_{i=0}^{k-2} y_{ui}$ and $\pi_{u(k-1)} = 1 - \sum_{i=0}^{k-2} \pi_{ui}$.

Unlike the previous section, the baseline class is now the last — the $(k-1)$th. The reason is simply because, in notation echoing that of Agresti, the cumulative link to the c.d.f., F, is

$$F_j(\mathbf{x}) = G_j(\alpha_j - \boldsymbol{\beta}_j'\mathbf{x}_j), \quad j = 0, 1, 2, \ldots, (k-2), \tag{4.24}$$

where j runs from 0 to $k-2$ only, since $F_{k-1}(\mathbf{x}) \equiv 1$. Note that $\boldsymbol{\beta}_j$ may vary with class, j; I do not impose a restriction to proportional odds models, as does, for example, Agresti (1990).

It is again convenient to define the common parameter vector $\boldsymbol{\gamma} = (\boldsymbol{\alpha}', \boldsymbol{\beta}')'$ and to write the vector of linear predictors as

$$\boldsymbol{\eta} = X\boldsymbol{\gamma}.$$

The cumulative link may be written as

$$\pi_0 + \pi_1 + \cdots + \pi_j = G_j(\eta_j). \tag{4.25}$$

Derivation and calculation of the score equations

Score equations for the general cumulative link are derived in the Appendix to the seminal paper of McCullagh (1980), although as Fienberg comments in the Discussion of that paper, "he skips rather blithely over the critical issues of computation". Also, the only models considered here in detail use a standard, logit link; nothing equivalent to McCullagh's scale parameter τ is introduced. Rather than leave the discussion at the point where the score is given only in terms of unexpanded derivatives (as in McCullagh, 1980), in the following an explicit form for the logit link is derived.

The log likelihood is naturally written in terms of log probabilities, $\mathbf{p}$, while the link is in terms of unlogged probabilities, $\boldsymbol{\pi} = \mathrm{e}^{\mathbf{p}}$. Thus, for the score equations, I use the chain-rule decomposition

$$\frac{\partial \ell_u}{\partial \boldsymbol{\gamma}} = \frac{\partial \boldsymbol{\pi}_u'}{\partial \boldsymbol{\gamma}} \frac{\partial \mathbf{p}_u'}{\partial \boldsymbol{\pi}_u} \frac{\partial \ell_u}{\partial \mathbf{p}_u}. \tag{4.26}$$

Dropping the subscript u for notational convenience, the rightmost term on the right-hand side of (4.26) is easily obtained from the derivatives for the components:

$$\frac{\partial \ell}{\partial p_j} = y_j + \left(y_{k-1} \frac{\partial p_{k-1}}{\partial p_j} \right), \quad j = 0, 1, \ldots, k-2.$$

Since at most one y_j is nonzero, $\partial \ell / \partial \mathbf{p} = \mathbf{e}_j$ is a vector of zeros except for the jth element, which is unity. Exceptionally, when $\mathbf{y}_\cdot = \mathbf{0}$, i.e. $y_{k-1} = 1$, all elements are equal to

$$\frac{\partial p_{k-1}}{\partial p_j} = \frac{\partial}{\partial p_j} \log\left(1 - \mathrm{e}^{p_0} - \mathrm{e}^{p_1} - \cdots - \mathrm{e}^{p_{k-2}}\right) = -\frac{\pi_j}{\pi_{k-1}}. \tag{4.27}$$

The second term on the right-hand side of (4.26) is

$$\frac{\partial \mathbf{p}'}{\partial \boldsymbol{\pi}} = \operatorname{diag}\left(\frac{\partial p_j}{\partial \pi_j} \right) = \operatorname{diag}(1/\pi_j), \quad j = 0, 1, \ldots, k-2. \tag{4.28}$$

Together, these results give the further computational simplification

$$\frac{\partial \ell}{\partial \boldsymbol{\pi}} = \frac{\partial \mathbf{p}'}{\partial \boldsymbol{\pi}} \frac{\partial \ell}{\partial \mathbf{p}} = \begin{cases} \frac{1}{\pi_j}\mathbf{e}_j & \text{if class}(y) = j < k-1, \\ -\frac{1}{\pi_{k-1}}\mathbf{1} & \text{if class}(y) = k-1. \end{cases} \tag{4.29}$$

where $\mathbf{1}$ is vector with all elements unity. This simplification hides the analytic form of the expression needed to find second derivatives. This form is

$$\frac{\partial \ell}{\partial \boldsymbol{\pi}} = \left(\left[\frac{y_0}{\pi_0} - \frac{y_{k-1}}{\pi_{k-1}} \right], \left[\frac{y_1}{\pi_1} - \frac{y_{k-1}}{\pi_{k-1}} \right], \ldots, \left[\frac{y_{k-2}}{\pi_{k-2}} - \frac{y_{k-1}}{\pi_{k-1}} \right] \right)'. \tag{4.30}$$

To derive the remaining term in (4.26), consider that the definition of the link function, equation (4.25), implies that for $j = 0, 1, \ldots, (k-2)$,

$$\pi_j = G_j - G_{j-1}. \tag{4.31}$$

Written in vector and matrix form, equations (4.31) become

$$\boldsymbol{\pi} = M\mathbf{G}$$

where

$$M = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ -1 & 1 & \ddots & 0 & \vdots \\ 0 & -1 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -1 & 1 \end{pmatrix}.$$

Then, by further application of the chain rule,

$$\frac{\partial \boldsymbol{\pi}'}{\partial \boldsymbol{\gamma}} = \frac{\partial \boldsymbol{\eta}'}{\partial \boldsymbol{\gamma}} \frac{\partial \mathbf{G}'}{\partial \boldsymbol{\eta}} \frac{\partial \boldsymbol{\pi}'}{\partial \mathbf{G}} = X' \Delta M', \tag{4.32}$$

where, if G_j^{-1} is the logit function, a general result already quoted (equation 2.11)

gives

$$\Delta = \text{diag}[G_j(1 - G_j)], \qquad j = 0, 1, \ldots, (k-2). \tag{4.33}$$

The general form for any cumulative link $\mathbf{G}$ is

$$\mathbf{U}(\boldsymbol{\gamma}) = \frac{\partial \mathbf{G}'}{\partial \boldsymbol{\gamma}} \mathbf{a}(\mathbf{y}, \boldsymbol{\pi}), \tag{4.34}$$

where

$$\begin{aligned} \mathbf{a}(\mathbf{y}, \boldsymbol{\pi}) &= \frac{\partial \ell}{\partial \mathbf{G}} = M' \frac{\partial \ell}{\partial \mathbf{p}} \\ &= \left(\left[\frac{y_0}{\pi_0} - \frac{y_1}{\pi_1} \right], \left[\frac{y_1}{\pi_1} - \frac{y_2}{\pi_2} \right], \ldots, \left[\frac{y_r}{\pi_r} - \frac{y_{r+1}}{\pi_{r+1}} \right], \ldots, \left[\frac{y_{k-2}}{\pi_{k-2}} - \frac{y_{k-1}}{\pi_{k-1}} \right] \right)'. \end{aligned} \tag{4.35}$$

For the logit link we may also simplify computation by calculating

$$\mathbf{U}(\boldsymbol{\gamma}) = X' \mathbf{b}(\mathbf{y}, \boldsymbol{\pi}) \tag{4.36}$$

where

$$\mathbf{b}(\mathbf{y}, \boldsymbol{\pi}) = \Delta \mathbf{a}(\mathbf{y}, \boldsymbol{\pi}).$$

A further simplification arises by considering that the definition of the cumulative link function (4.24) implies that

$$X\boldsymbol{\gamma} = I_{k-1}\boldsymbol{\alpha} - X^*\boldsymbol{\beta},$$

where I_{k-1} is an appropriately sized identity matrix. It follows that

$$\mathbf{U}(\boldsymbol{\gamma}) = \begin{pmatrix} \mathbf{b} \\ X^{*\prime}\mathbf{b} \end{pmatrix}.$$

The information matrix and model fitting

To fit the model by Fisher scoring we must also calculate the information matrix contribution

$$\mathcal{I}(\boldsymbol{\gamma}) = \mathrm{E}[\mathbf{U}\mathbf{U}'] = X'\Delta \mathrm{E}[\mathbf{a}\mathbf{a}']\Delta X$$

for each subject. The (i,j)th entries in $\mathbf{a}\mathbf{a}'$ are of the form

$$\frac{y_{i-1}y_{j-1}}{\pi_{i-1}\pi_{j-1}} - \frac{y_i y_{j-1}}{\pi_i \pi_{j-1}} - \frac{y_{i-1}y_j}{\pi_{i-1}\pi_j} + \frac{y_i y_j}{\pi_i \pi_j}, \tag{4.37}$$

which have expectation zero whenever there are no indices in common (since only one $\mathbf{y}_{\cdot}$ value may be nonzero). Hence $E = \mathrm{E}[\mathbf{a}\mathbf{a}']$ is tridiagonal with entries

$$E_{ii} = \frac{1}{\pi_{i-1}} + \frac{1}{\pi_i}, \tag{4.38}$$

$$E_{i(i-1)} = -\frac{1}{\pi_{i-1}}, \tag{4.39}$$

$$E_{i(i+1)} = -\frac{1}{\pi_i}. \tag{4.40}$$

Here i indexes matrix rows and columns and so runs from 1 to $k-1$ rather than from 0 to $k-2$, but the subscripts for $\boldsymbol{\pi}$ on the right-hand sides are in my standard notation. There is of course no entry E_{10} or $E_{(k-1)k}$.

Re-introducing the subject index u, we may then obtain estimates of $\boldsymbol{\gamma}$ by iterating according to

$$\boldsymbol{\gamma}^{(s+1)} = \boldsymbol{\gamma}^{(s)} + \left[\sum_u (X_u' \Delta_u E_u \Delta_u X_u) \Big|_{\boldsymbol{\gamma}^{(s)}} \right]^{-1} \left[\sum_u \mathbf{U}_u(\boldsymbol{\gamma}^{(s)}) \right]. \tag{4.41}$$

Simplifications using $X = (I \quad X^*)$ may be made here as described above. An even greater simplification is to use the approximate information $\mathcal{A} = \sum \mathbf{U}\mathbf{U}'$ for $\mathcal{I}$ (Azzalini, 1994; see Section 1.4.5). Simulations have shown that in the current context this approximation, which holds for large n, is effective for fairly small samples provided that the solution is not too close to the edge of the parameter space.

The true information matrix should always be calculated after fitting to obtain max-

imum likelihood estimates of the parameter standard errors. In practice, simulations have shown that as the sample size increases there is increasingly little to choose between $\mathcal{A}$, $\mathcal{I}$ and the "sandwich" estimator, $\mathcal{S}$ (Section 1.4.5). In the examples in Section 4.6, these matrices differ only in the fourth or fifth significant digit.

4.4.2 The multivariate case

The extension of the above results to the multivariate case closely parallels the development of the multivariate (unordered) analogue of the univariate (unordered) logit models given at the end of the previous section. Again the contribution of the uth multivariate observation to the log likelihood is the sum of the univariate contributions:

$$\ell_u = \ell_{u1:} + \ell_{u2:} + \cdots + \ell_{uT:} \tag{4.42}$$

We concatenate the vectors of observations, probabilities and linear predictors for the individual timepoints as

$$\mathbf{y}_{:} = \begin{pmatrix} \mathbf{y}_{1:} \\ \mathbf{y}_{2:} \\ \vdots \\ \mathbf{y}_{T:} \end{pmatrix}, \quad \boldsymbol{\pi}_{:} = \begin{pmatrix} \boldsymbol{\pi}_{1:} \\ \boldsymbol{\pi}_{2:} \\ \vdots \\ \boldsymbol{\pi}_{T:} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\eta} = \begin{pmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \\ \vdots \\ \boldsymbol{\eta}_T \end{pmatrix} = X\boldsymbol{\gamma},$$

as appropriate. Then the score contribution of the uth subject is

$$\mathbf{U}_u(\boldsymbol{\gamma}_u) = \frac{\partial \mathbf{G}_u'}{\partial \boldsymbol{\gamma}_u}\mathbf{a}_u, \tag{4.43}$$

which is identical in appearance and derivation to equation (4.34), but $\mathbf{a}$ is now the concatenation of univariate forms (4.35) and $\mathbf{G}$ is the concatenation of the T cumulative link functions $\mathbf{G}_{u1}, \ldots, \mathbf{G}_{uT}$. For all-logit links,

$$\mathbf{U}_u(\boldsymbol{\gamma}_u) = X_u'\Delta_u\mathbf{a}_u, \tag{4.44}$$

where

$$\Delta_u = \text{block diag}\,(\Delta_{u1}, \ldots, \Delta_{uT})\,,$$

with each δ_{ut} as in (4.33). The information matrix contribution is

$$X_u' \Delta_u E_u \Delta_u X_u \tag{4.45}$$

where

$$E_u = \text{block diag}\,(E_{u1}, \ldots, E_{uT})\,, \tag{4.46}$$

for the E_{ui} derived in (4.37)–(4.40). The block diagonal form of matrix $E_{u:}$ is obtained by considering the expectations of expressions like (4.37). Letting the first subscript denote a timepoint, while the second is the class indicator, the general cross-product in multivariate $\mathbf{aa}'$ (*cf* 4.37) becomes, before expansion,

$$\left(\frac{y_{s(i-1)}}{\pi_{s(i-1)}} - \frac{y_{si}}{\pi_{si}}\right)\left(\frac{y_{t(j-1)}}{\pi_{t(j-1)}} - \frac{y_{tj}}{\pi_{tj}}\right) = g_s(y_s) g_t(y_t) = h(y_s, y_t), \text{ say.} \tag{4.47}$$

Then

$$\begin{aligned} \mathrm{E}_{Y_s,Y_t}[h(Y_s,Y_t)] &= \mathrm{E}_{Y_s}\{\mathrm{E}_{Y_t \mid Y_s=y_s}[h(Y_s,Y_t)]\} \\ &= \mathrm{E}_{Y_s}\{g_s(Y_s)\,\mathrm{E}_{Y_t \mid Y_s=y_s}[g_t(Y_t)]\} \\ &= \mathrm{E}_{Y_s}\{g_s(Y_s)[1-1]\} \\ &= 0 \end{aligned}$$

for all $s \neq t$. (When $s = t$ we of course recover the univariate results already given.)

An approximation to $\mathcal{I}$ may be used in numerical work as for the univariate case. However, the simple form

$$\mathcal{A}^{(\text{raw})} = \sum_u \mathbf{U}_u \mathbf{U}_u'$$

is not in general close enough to $\mathcal{I}$ to ensure convergence. Instead, we need to ensure

that the "meat" in the "sandwich" $\mathrm{E}[X'\Delta\mathbf{a}\mathbf{a}'\Delta X]$ is set to be

$$\text{block diag}\left(\mathbf{a}_{u1}\mathbf{a}'_{u1}, \ldots, \mathbf{a}_{uT}\mathbf{a}'_{uT}\right),$$

rather than the raw, non-block-diagonal form $\mathbf{a}_u\mathbf{a}'_u$. I label this slight modification of Azzalini's approximation $\mathcal{A}^{(\text{mod})}$. The modification is clearly consistent with the considerations of block diagonality in (4.22) and (4.46). Substitution of $\mathcal{A}^{(\text{mod})}$ for $\mathcal{I}$ has been successful in simulation studies.

4.5 Models for multivariate chains

We will consider, in Chapter 6, data sets where at each timepoint two observations are made (one is the outcome of interest, Y, the other an indicator of whether it is observed, R). In the example considered in Section 6.6, Y is an ordered categorical variable, while R is unordered, ternary. By the device of assuming that R is observed 'before' Y at each timepoint, or vice versa, we can fit a 'Markov chain' model to each pair, modelling 'transition' probabilities conditional on all previous values of both Y and R:

$$f(\mathbf{y}_{t:}, \mathbf{r}_{t\cdot} \mid \mathbf{y}_{1:}, \ldots, \mathbf{y}_{(t-1):}, \mathbf{r}_{1\cdot}, \ldots, \mathbf{r}_{(t-1)\cdot}) =$$
$$f(\mathbf{r}_{t\cdot} \mid \mathbf{y}_{1:}, \ldots, \mathbf{y}_{(t-1):}, \mathbf{r}_{1\cdot}, \ldots, \mathbf{r}_{(t-1)\cdot}) f(\mathbf{y}_{t:} \mid \mathbf{y}_{1:}, \ldots, \mathbf{y}_{(t-1):}, \mathbf{r}_{1\cdot}, \ldots, \mathbf{r}_{t\cdot})$$

The methods developed above enable such models to be fitted with ease. Each univariate, conditional model with linear predictor $\boldsymbol{\eta}_s = X_s\boldsymbol{\gamma}_s$ leads to a score equation of the form

$$\mathbf{U}_s = X'_s \frac{\partial \ell_s}{\partial \boldsymbol{\eta}_s} = X'_s \mathbf{b}_s, \tag{4.48}$$

where $s = 1, 2, \ldots, S$ reflects the order chosen for the Markov chain style factorization (no longer a simple question of time sequence). Now as in both Sections 4.3 and 4.4,

the log likelihood for the whole series is

$$\ell = \ell_1 + \ell_2 + \cdots + \ell_S$$

and the $\boldsymbol{\eta}_s$ are assumed variationally independent both within and across conditional models. Thus writing $\boldsymbol{\eta} = (\boldsymbol{\eta}_1', \ldots, \boldsymbol{\eta}_S')'$ and $\mathbf{b} = (\mathbf{b}_1', \ldots, \mathbf{b}_S')'$, the contribution for the entire chain to the overall score equation is, for each subject,

$$\mathbf{U} = X' \frac{\partial \ell}{\partial \boldsymbol{\eta}} = X'\mathbf{b}. \tag{4.49}$$

By choice of design matrix X, different $\boldsymbol{\eta}_s$ may share some parameters $\boldsymbol{\gamma}$, but this does not affect the general form (4.49). Furthermore by the same reasoning the information matrix contribution is

$$\mathcal{I} = X'\mathrm{E}[\mathbf{b}\mathbf{b}']X, \tag{4.50}$$

where $\mathrm{E}[\mathbf{b}\mathbf{b}'] = -\partial \ell^2 / \partial \boldsymbol{\eta}' \partial \boldsymbol{\eta}$ is necessarily block diagonal.

Any mixture of data types could be fitted using (4.49) and (4.50) for Fisher scoring. One merely needs to be able to calculate the univariate elements $\partial \ell_s / \partial \boldsymbol{\eta}_s = \mathbf{b}_s$ and corresponding $\mathrm{E}[\mathbf{b}_s \mathbf{b}_s']$ for each pseudo-timepoint s according to the distributional form assumed for variable Y_s. Parameters common to different conditional models are specified by choice of X independently of the individual forms $\mathbf{b}_s$, which are the same for all models to be considered.

In this thesis I have derived $\mathbf{b}_s$ only for logistic regression models for ordered and unordered categorical data. Clearly, however, extensions to other types of data will be straightforward, and are worthy of further study.

Some new terminology would be useful: the conditional probabilities for Y given R, above, are not in any normal sense 'transitional' and the proposed factorizations of multivariate probability functions at each timepoint do not correspond to any underlying Markov process.

4.6 Examples and further discussion

We illustrate here with two simple examples where the data are binary, the object being to contrast the interpretation of Markov chain models with that already discussed for marginal models in Chapter 2. For a more demanding application, involving both ordered and unordered polytomous outcomes, see Chapter 6, Section 6.6.

4.6.1 Cerebrovascular deficiency revisited

This data set has been discussed in Section 2.3.3. For comparison consider now the Markov chain analysis. Such an analysis is not the obvious one for cross-over data; although there are clearly defined sequential and discrete timepoints, the measure is by design repeated under different conditions, and the focus of attention is the treatment and not the period.

The fully saturated Markov chain model is

$$\begin{aligned} \phi_1 &= \alpha_0 + \beta_1 x_1 \\ \phi_2 &= \alpha_0 + \alpha_2 + \alpha_{21} y_1 + \beta_2 x_2 + \beta_{21} x_2 y_1 \end{aligned}$$

where x_t is the treatment at time t ($x_t = 0$ for placebo), and there is no explicit period covariate since the period is included in the intercept terms. Equivalently, for the time 1 and 2 intercepts we could take $\alpha_1^* = \alpha_0$ and $\alpha_2^* = \alpha_0 + \alpha_2$. The model reflecting that chosen in the marginal literature is

$$\begin{aligned} \phi_1 &= \alpha_0 + \beta x_1 \\ \phi_2 &= \alpha_0 + \alpha_2 + \alpha_{21} y_1 + \beta x_2. \end{aligned}$$

This is not exactly equivalent to any presented before, because of the different mechanism for modelling dependency, but in essence there are both period (α_2) and common treatment effects (β) plus simplified allowance for dependence. This model fits insignificantly worse than the saturated one, the evaluated log likelihoods being -69.6572 and -69.7596, respectively. We cannot drop the time effect from this model as the log

likelihood reduces to -78.436, but as for the marginal analysis, we can drop β from the time-2 model (for a log likelihood of -70.98).

Again this has somewhat worrying consequences in interpretation; the drug appears to have no discernible effect at time 2 once previous outcome is allowed for, in contradiction to the original conclusions. The Markov chain model serves here to strengthen a controversial conclusion; we should not be surprised by failure to fit a common parameter for the drug effect at each timepoint. In this example, β_1 is the logs odds ratio increase due to active treatment, while β_2 is the ratio adjusted for both period and history. There is no reason to suppose that a model setting these equal would have any natural interpretation.

The nature of a cross-over trial introduces considerable confounding within a Markov chain approach. Here, for example, from the maximum likelihood estimate $\hat{\beta}_1 = -1.2$, $\mathrm{P}(Y_1 = 1)$ is elevated when a patient receives placebo at time 1, so that α_{21} is more likely to contribute to the time-2 linear predictor when drug is administered at that time. Because of such difficulties, I would not advocate Markov chain models for the analysis of such data.

4.6.2 6 cities revisited

A question that was not raised by Fitzmaurice and Laird (1993) in their analysis of the 6 cities data (see Section 2.3.3, Example 2) is whether the covariates age and maternal smoking are indeed significant if the dependency model is fully specified. The thrust of their presentation, and of my comparative fully marginal model, was the discussion of potential simplifications in dependency specification.

The brief answer is yes for age, but no for maternal smoking. A completely saturated model, for different smoking effects across timepoints, has evaluated log likelihood -790.8, whereas a saturated (ϕ-unconstrained) model for the table collapsed over smoking has log likelihood -792.9. Thus the deviance change is approx. 4 on 4 d.f., which is not significant. Both these models are saturated for age effect, which is here naturally aliased by period in the process of fitting a separate set of unconstrained intercept parameters for each timepoint (or one might model a linear trend through

Table 4.1: Intercept-unconstrained Markov chain model for the 6 cities data. Asterisks mark significance under a z-test. Standard deviations obtained from the true, approximate and sandwich estimators were identical to 4 d.p.

Parameter	Estimate	SD
α_1^*	-1.6433	0.1171
α_2^*	-2.0794	0.1500
α_3^*	-2.5489	0.1928
α_4^*	-3.0995	0.2556
α_{21}^*	1.9644	0.2620
α_{31}^*	1.1352	0.4187
α_{32}^*	2.1434	0.3471
α_{312}	-0.0730	0.6059
α_{41}^*	1.4573	0.5140
α_{42}	0.9023	0.6601
α_{412}	-0.1764	0.9923
α_{43}^*	1.7558	0.5248
α_{413}	-0.8067	0.9871
α_{423}	-0.1776	0.9304
α_{4123}	0.8380	1.4426

the non-history-dependent intercepts, α_t).

As already mentioned, not too much should be made of this finding clinically, because the data set is not particularly robust.

Parameter estimates and standard deviations for the intercept-unconstrained Markov chain model are given in Table 4.1. The asterisks mark parameters significant on univariate z-tests: not the most robust of analyses, but highly indicative nonetheless. In the marginal, canonical or mixed parametrizations of Chapter 2 the selected dependence structure was difficult to interpret, but here clearly we have a mixture of lag-one dependence and dependence on the initial value. Those who are wheezing by age 7 are likely to do so during the whole of the study period, and once contracted wheeze is likely to persist from one year to the next.

4.6.3 Further discussion

When they are appropriate, Markov chain models for longitudinal data are relatively easy to interpret, because dependency expressed as functions of previous values is

easier to conceptualize than quantities such as the high-order odds ratios of marginal-modelling approaches, and is in my view more plausible than the local independence assumption of random-effects models. Common-sense notions of dependency on previous values can be expressed as parameter constraints in an obvious way. Example 4.6.2 contrasting the two approaches illustrate these points. On the other hand, Example 4.6.1 stresses that Markov chain models are by no means always appropriate.

Another important consideration is that most longitudinal studies suffer from appreciable dropout, or attrition. In Chapter 6 the Markov chain approach is seen to be natural in this setting and also offers a concise proof that informative-missing data are unidentifiable by maximum likelihood. Unfortunately the Markov chain model does *not* deal particularly easily with missing data patterns other than the monotone pattern of dropout — but neither does any other type of model.

The generalized model discussed in Section 4.5 imposes no restrictions on the possible values of the vector $\mathbf{Y}$. The values may be continuous, discrete, or even mixed at each timepoint, and we need not even insist that each Y_t is a repeated measure on some Y^*. Such models could even be used for the analysis of clustered data, when there is only one 'timepoint', but one applies factorization (4.1) to some more or less arbitrary labelling of the Y_t. Of course one would then face problems of interpretation in terms of arbitrary "history" and choice of "baseline", Y_1, and in general then surely a marginal model would be more natural.

Chapter 5

Unbalanced data and multivariate models

5.1 Introduction

In the discussion of marginal models in Chapter 2, it was assumed that the outcome data were balanced: all the outcome vectors were the same size and there were no structural zeros. In this chapter the effect of violating each of these assumptions is considered.

Motivating applications include missing data and dropout problems when nonresponse is assumed to be either at random or completely at random (in the sense of Little and Rubin, 1987; defined here in Section 6.2 of the following chapter). Of course, one should never make such assumptions with regard to missing data without careful consideration.

In Section 5.2 we consider models for data subject to dropout, where the dropout mechanism (but not its presence) is ignored; in Section 5.3 we consider models for the probability of dropout, where the data model is ignored. We might fit both simultaneously to a suitable dataset.

5.2 Unbalanced data

Almost inevitably while collecting longitudinal data on people or animals, subjects will drop out of the study. In this chapter I assume that such dropout is *completely at random* or *at random* in the sense of Little and Rubin (1987); in particular, the observed likelihood is then adequate for valid inference. Another common situation arises when the outcome vectors represent data on siblings or litters, since then it is unlikely that all families will be of the same size. Or subjects might not all be followed up at every occasion, for reasons unconnected with the outcome measurement.

In any of these cases, suitably modified fully marginal models, as discussed in Section 5.2.1, may be preferable to other approaches, such as mixed or canonical parametrizations, because of their reproducibility (Section 1.4.3). However, when there are many timepoints, fully marginal models can be very difficult to fit (Chapter 3), so that other approaches are considered in Sections 5.2.2 to 5.2.5.

5.2.1 Fully marginal approach

As for balanced data, we model

$$\boldsymbol{\lambda} = X\boldsymbol{\gamma},$$

but now estimation is based only on those observations that are actually measured at each timepoint (i.e. we use all the *available* data). The 'full' model is that for the largest observed vector, although this may itself be considered a submodel for a potentially larger vector, by reproducibility.

This model is not inherently restricted to monotone missing-data patterns. When calculating the score contribution of a subject, we need only mask terms involving margins and/or interactions that are not present in the outcome vector concerned.

For incomplete observations, rows of X corresponding to unobserved marginal odds ratios and interactions may be set to zero, so that when the score function is evaluated, contributions may be added to the correct elements of the score vector. However, when calculating log probabilities $\mathbf{p}$ and $\partial s/\partial \mathbf{p}$ (Chapter 3), one should calculate an appropriate marginal probability table, *not* one of full size with unmodelled log odds

ratios set to zero. Practically, it is easiest to work with X set to the natural size of the sub-observation, and then after evaluating the corresponding raw score contribution, add the results to the correct entries in the full-size score vector.

5.2.2 Canonical parametrization

Referring to Section 1.4.3 we may write the canonical form of the distribution functions of sub-observations $\mathbf{Y}^{\mathcal{A}}$ in terms of $\xi_{\mathcal{B}}^{\mathcal{A}}$ and $z_{\mathcal{B}}^{\mathcal{A}}$, as $\mathcal{B}$ runs over the subsets of $\mathcal{A}$. To fit linear predictors to the canonical parameters, we must set, for the uth subject,

$$\xi_u^{\mathcal{A}} = X_u' \gamma^{\mathcal{A}}, \quad \mathcal{A} \subseteq \mathcal{T}. \tag{5.1}$$

That is, there is a different set of canonical parameters $\gamma^{\mathcal{A}}$ for every observation subset $\mathcal{A}$ occurring in the data. Hence it is not immediately sensible to use the same γ for all outcome vector sizes, because they parametrize different odds ratios. Nevertheless, this approach is followed here because it is the simplest algorithmically, and the consequences are considered.

5.2.3 The false identity link

Suppose we use the same γ for two or more different identity links:

$$\xi_{\mathcal{B}}^{\mathcal{T}} = X_{\mathcal{B}}' \gamma, \quad \xi_{\mathcal{B}}^{\mathcal{A}} = X_{\mathcal{B}}' \gamma, \qquad \mathcal{B} \subseteq \mathcal{A} \subset \mathcal{T}. \tag{5.2}$$

This makes the simplifying but unjustifiable assumption that the CORs of each sub-distribution are the same as for the full distribution — hence the terminology *false identity link.*

Since in general $\xi^{\mathcal{A}} \neq \xi^{\mathcal{B}}$ for $\mathcal{A} \neq \mathcal{B}$, in this case the fitted $\tilde{\gamma}$ from (5.2) will be a form of weighted average of the 'true' $\hat{\gamma}$ obtained from (5.1) to estimate the same-order interactions in the true models. Conceivably, if hierarchical analysis of deviance, or prediction, is the object of the exercise rather than parameter estimation or interpretation per se, it may be enough to fit according to (5.2). The precise nature of the

bias in the γ estimates and variances can only be established on a case-by-case basis, given the particular design matrix X to be used.

In certain circumstances, the CORs might actually be equal and the link then not be false. One such case is independence. For the bivariate case this is easily established, but for three or more variables in the full observation, the CORs might be equal in other circumstances too, as we now describe.

Addressing this issue indirectly, consider a closely related question: when are the CORs and their corresponding marginal odds ratios (MORs) equivalent? As previously noted, the 'saturating' (highest-order) $\text{COR}_{\mathcal{T}}$ and $\text{MOR}_{\mathcal{T}}$ are always equivalent. Of more current interest is when some lower-order CORs of the full distribution are equivalent to saturating CORs for corresponding subdistributions; these latter are also MORs by highest-order equivalence. Hence we can consider criteria for COR–MOR equivalence rather than $\text{COR}^{\mathcal{T}}$–$\text{COR}^{\mathcal{A}}$ equivalence.

Consider bivariate binary data, $\mathcal{T} = \{1, 2\}$, with probability distribution

$$\text{P}(\mathbf{Y}^{\mathcal{T}}) = \pi_{00} \left(\frac{\pi_{10}}{\pi_{00}}\right)^{y_1} \left(\frac{\pi_{01}}{\pi_{00}}\right)^{y_2} \left(\frac{\pi_{11}\pi_{00}}{\pi_{01}\pi_{10}}\right)^{y_1 y_2}, \tag{5.3}$$

where $\pi_{ij} = \text{P}(Y_1 = i, Y_2 = j)$. Summing over values of Y_2, we obtain the marginal for Y_1, which in current terminology is also the 'full' marginal distribution for a subobservation $\mathbf{Y}^{\mathcal{A}}$ with $\mathcal{A} = \{1\}$:

$$\begin{aligned} \text{P}(Y_1) &= \pi_{00} \left(\frac{\pi_{10}}{\pi_{00}}\right)^{y_1} + \pi_{01} \left(\frac{\pi_{11}}{\pi_{01}}\right)^{y_1} && (5.4) \\ &= \pi_{0+} \left(\frac{\pi_{1+}}{\pi_{0+}}\right)^{y_1} && (5.5) \end{aligned}$$

where subscript + denotes summation over the index in question.

The question of COR equivalences posed above becomes: when does the false-identity link, for say, COR_1,

$$\xi_1^{\mathcal{T}} = \xi_1^{\mathcal{A}} = X_1'\gamma$$

become a true equivalence? Explicitly here, when is the equality

$$\log \frac{\pi_{10}}{\pi_{00}} = \log \frac{\pi_{1+}}{\pi_{0+}}$$

satisfied? The answer is: when independence holds, and in the degenerate case when $P(Y_2 = 0) = 1$. To show this, consider the 2×2 table of specified odds ratios

1	χ_2
χ_1	$\chi_{12}\chi_1\chi_2$

where $\chi_{\mathcal{A}} = \mathrm{COR}_{\mathcal{A}} = \exp \xi_{\mathcal{A}}$, determined by the parameters $\boldsymbol{\xi}$ (see Section 2.4.5). From this table,

$$\mathrm{MOR}_1 = \frac{\chi_1(\chi_{12}\chi_2 + 1)}{\chi_2 + 1},$$

so that if (in fact iff, provided that $\chi_2 \neq 0$) $\mathrm{MOR}_1 = \mathrm{COR}_1^{\mathcal{A}} = \mathrm{COR}_1^{\mathcal{T}} = \chi_1$, then $\chi_{12} = 1$. QED.

Thus, given independence, the false-identity link becomes true identity. More importantly, using the false-identity link implies an independence assumption, with unpredictable effect on a parametrized ξ_{12}.

The simplicity of the bivariate case is lost for higher-order observations, and general rules as to the effect of a false-identity link are intractable. Immediately, for trivariate binary data, we may derive by direct analogy with the above

$$\begin{aligned} \mathrm{MOR}_{12} &= \frac{(\chi_{123}\chi_{12}\chi_{13}\chi_{23}\chi_1\chi_2\chi_3 + \chi_{12}\chi_1\chi_2)(\chi_3 + 1)}{(\chi_{23}\chi_2\chi_3 + \chi_2)(\chi_{13}\chi_1\chi_3 + \chi_1)} \\ &= \chi_{12}\frac{(\chi_{123}\chi_{13}\chi_{23}\chi_3 + 1)(\chi_3 + 1)}{(\chi_{23}\chi_3 + 1)(\chi_{13}\chi_3 + 1)}, \end{aligned}$$

so that $\mathrm{MOR}_{12} = \mathrm{COR}_{12}$ iff the ratio on the right-hand side is unity. One such case is when all the 2nd- and 3rd-order ratios are unity (independence); another is when $\chi_{13} = \chi_{123} = 1$ (a first-order Markov chain with not necessarily stationary transition probabilities). However, there are infinitely many other sets of values that meet the

criterion but that defy easy interpretation (e.g. if $\chi_3 = 1$ and $\chi_{123} = 1/\chi_{13}\chi_{23}$, then the ratio is unity for any $(1 + \chi_{13})(1 + \chi_{23}) = 4$). Note in particular that the quadratic exponential family assumption, here $\chi_{123} = 1$, is not sufficient to ensure that $\mathrm{MOR}_{12} = \mathrm{COR}_{12}$ although it does ensure that the zero-conditional ratios become ordinary conditional ratios (Section 2.4.3).

Analysis of deviance might be used to test whether a fuller model such as described in the next section might be reduced to a false identity fit without loss. However, since in so doing we lose almost all interpretability, this might not be a worthwhile strategy. The false identity link is proposed because it facilitates the compromise fit of a model when observation vectors are too large for any fuller alternative to be computationally viable. A false identity link might also be considered acceptable when only a small proportion of observations are less than full; however, it is not clear that this would be any more robust than a complete-case only analysis. Further work is needed here.

5.2.4 The corrected false identity link

In the previous subsection we rejected the idea of fitting the true model (5.1) because of the number of parameters involved. For short series, however, it may be feasible to do so, especially if we are prepared to compromise on explanatory-variable saturation. Thus we might fit

$$\boldsymbol{\xi}^{\mathcal{A}} = \boldsymbol{\alpha}^{\mathcal{A}} + X\boldsymbol{\beta}, \tag{5.6}$$

imposing considerably less untested assumptions than for a false identity link. If we were interested only in variable selection, say, or prediction perhaps, rather than interpretation of parameter estimates, this model is not restricted to the assumption that data are missing completely at random. Moreover, it is extremely quick to fit by comparison with the adaptations needed to the mixed model approach (Fitzmaurice and Laird, 1993, and Section 5.2.5).

Consider the following artificial example (adapted from Agresti, 1990) that is intended to illustrate the process. 696 students were asked, in 1979, whether or not they used drugs; the question was then asked of the same students the following year, by which

time 259 had left and so were not available for survey. Assuming that graduation is not related to drug abuse, it might seem a priori reasonable to assume these 259 were missing completely at random; this example will also show how wrong such an assumption can be.

The responses were

	Y_2		
Y_1	No	Yes	N/A
No	380	18	222
Yes	27	12	37

for which the false-identity model

$$\begin{aligned} \xi_1^* &= \alpha_1^* \\ \xi_2 &= \alpha_2 \\ \xi_{12} &= \alpha_{12} \end{aligned}$$

gives point estimates $\alpha_1^* = -2.24$, $\alpha_2 = -3.02$ and $\alpha_{12} = 1.84$. The corrected false identity model (which is here saturated) is

$$\begin{aligned} \xi_1 &= \alpha_1 + \alpha_c I_c \\ \xi_2 &= \alpha_2 \\ \xi_{12} &= \alpha_{12} \end{aligned}$$

where I_c is an indicator, being unity for observations of size one only, which has estimates $\alpha_1 = -2.64$, $\alpha_2 = -3.05$, $\alpha_{12} = 2.24$ and 'correction term' $\alpha_c = 0.85$. Although the value for α_2 is very similar for both models, in each case being estimated only from the fully observed (two-timepoint) part of the table, that for α_{12} is rather different between models. Again this latter parameter is only estimated from bivariate observations, but for the false identity link the calculation is biased by the bias induced in ξ_1^* evaluation.

The deviance change between these two models is approximately 10 on 1 d.f.; the false identity link is significantly worse than the saturated model. The false identity link implies, in particular, that dropout is missing completely at random; the saturated model does not. Thus we conclude that the data are at least missing at random, rather than completely at random, which is equivalent to asserting that the pattern of response amongst those who dropped out is significantly different to that of those who did not. In the present context, there is then evidence of at least a cohort effect; 1979 graduates used more drugs. (The data *might* also be missing informatively, but we cannot hope to test for this; see Chapter 6.)

5.2.5 Unbalanced data and mixed parametrizations

If we adopt the strategy of Fitzmaurice and Laird (1993) and model the means marginally but the odds ratios conditionally, that is,

$$\begin{aligned} \mathbf{g}(\boldsymbol{\mu}) &= X_{\mathrm{M}}\boldsymbol{\gamma}_{\mathrm{M}}, \\ \boldsymbol{\xi}^{\mathcal{A}}_{(\text{2nd+ order})} &= X_{\mathrm{C}}\boldsymbol{\gamma}_{\mathrm{C}}, \end{aligned}$$

where M denotes marginal-model and C zero-conditional, then at least the part of the model we are most interested in is reproducible. Unfortunately, as pointed out in the above reference, the higher-order part of the model is *not* reproducible; this led Fitzmaurice *et al.* (1994) to consider using an EM approach.

However, this complexity can in general be avoided by correcting the false identity link (for the higher-order ratios only) as in the previous section. Admittedly this assumes there are enough occurrences of each missing data pattern to make estimation feasible, and also it leaves the question of explanatory variable effect open. When interaction parameters are regarded as nuisance parameters, there will rarely be need for complicated models in the linear predictors, and an unconstrained intercept model should suffice.

5.3 Models for dropout

A commonly occurring situation, in which the cell probability table for multivariate binary data features *structural*, as distinct from merely *observed* zeros, is when that data set represents binary indicators for dropout, a topic discussed more fully in Chapter 6. If our attention is focussed on dropout itself, we are led naturally to the following models.

After introducing terminology and describing the relevant distribution function in Section 5.3.1, we consider two types of parametrization of the model. In Section 5.3.2 we look at marginally-linked parameters, which, however, lead to unidentifiable models in the presence of time-varying covariates and suffer from difficulties in constraining monotonicity of the survival function. To overcome these problems, in Section 5.3.3 we use a semi-canonical link that nevertheless offers a reproducible model. Interpretation of parameters is, however, not straightforward, as highlighted by an example data set studied in Section 5.3.4. Otherwise, the semi-canonical model has several advantages, discussed in Section 5.3.5.

Throughout this section I assume dropout is at worst missing at random in the covariate effects (Little, 1995), which is to say that unrecorded covariates at dropout time might affect the probability of dropout.

5.3.1 Odds ratios and canonical parameters

In this section the outcome vector is a set of binary indicator variables denoted $\mathbf{R} = (R_1, R_2, \ldots R_T)'$. Although for binary indicators the coding is not important, for consistency with the literature I will code $R_t = 0$ for dropout at time t. We assume that if there is dropout at time t, then $R_{t+1}, \ldots R_T$ are also zero (i.e. a subject does not re-enter a trial having once dropped out). The distribution of $\mathbf{R}$, a multivariate binary vector, must belong to the polynomial exponential family. However, simplifications can be made to the general formulation as a result of the structural zeros.

Consider first the case of bivariate binary data. Again using π to denote cell probabilities and adopting the convention that first subscript refers to first variable, etc.,

we have that

$$\begin{aligned}
\mathrm{P}(R_2 = 1 \mid R_1 = 0) &= 0 = \frac{\pi_{01}}{\pi_{0+}} \ (\Rightarrow \pi_{01} = 0); && (5.7)\\
\mathrm{P}(R_2 = 0 \mid R_1 = 0) &= 1 = \frac{\pi_{00}}{\pi_{0+}} \ \left(= \frac{\pi_{00}}{\pi_{00}} \text{ by } \pi_{01} = 0\right); && (5.8)\\
\mathrm{P}(R_2 = 0 \mid R_1 = 1) &= \frac{\pi_{10}}{\pi_{1+}}; && (5.9)\\
\mathrm{P}(R_2 = 1 \mid R_1 = 1) &= \frac{\pi_{11}}{\pi_{1+}} && (5.10)
\end{aligned}$$

(where + denotes summation over an index) so that the probability table

π_{00}	0
π_{10}	π_{11}

has two degrees of freedom, as elements must sum to unity. For multivariate data, the structural zeros are those of the form $\pi_{1\cdots10\cdots01}$ (where the first string of ones may be empty). Thus, in particular,

$$\pi_{1\cdots10\cdots0} \equiv \pi_{1\cdots10+}.$$

The probability table may be written in terms of the marginal means μ_t of R_t:

$1-\mu_1$	0
$\mu_1 - \mu_2$	μ_2

Note how easily this table is built, compared with the difficulty of completing a general 2×2 table from the means and the odds ratio (Section 3.2.1). The ease of construction continues into the general multivariate case.

Consider now the distribution function rather than its tabulation. The (marginal) univariate distribution of R_1 is by definition

$$\mathrm{P}(R_1 = r_1) = \pi_{0+}\left(\frac{\pi_{1+}}{\pi_{0+}}\right)^{r_1} \qquad (5.11)$$

(we can write π_{00} for π_{0+}), and combining (5.7)-(5.10) gives the conditional probability

$$\mathrm{P}(R_2 = r_2 \mid R_1 = r_1) = \left(\frac{\pi_{10}}{\pi_{1+}}\right)^{r_1} \left(\frac{\pi_{11}}{\pi_{10}}\right)^{r_1 r_2} c(r_1, r_2) \tag{5.12}$$

where the shape function $c(\cdot)$ is zero for the impossible observation $(0,1)$ and unity otherwise. This may be derived directly or by using the general formula for polynomial exponential family conditional distributions (equation 4.9 on page 162). Here

$$\mathrm{P}(R_2 = r_2 \mid R_1 = 1) = \frac{\pi_{10}}{\pi_{1+}} \left(\frac{\pi_{11}}{\pi_{10}}\right)^{r_2},$$

so that

$$\mathrm{P}(R_2 = r_2 \mid R_1 = r_1) = [\text{r.h.s. of above}]^{r_1} \times \text{shape}$$

because of the simplicity of the structural zero pattern.

Multiplying (5.12) by the marginal of R_1, (5.11), gives the joint frequency function

$$\mathrm{P}(R_1, R_2) = \pi_{0+} \left(\frac{\pi_{10}}{\pi_{0+}}\right)^{r_1} \left(\frac{\pi_{11}}{\pi_{10}}\right)^{r_1 r_2} c(r_1, r_2). \tag{5.13}$$

The derivation extends naturally to the trivariate case, for which

$$\mathrm{P}(R_3 = r_3 \mid R_2, R_1) = \left[\frac{\pi_{110}}{\pi_{11+}} \left(\frac{\pi_{111}}{\pi_{110}}\right)^{r_3}\right]^{r_1 r_2} \times \text{shape}$$

giving after multiplication by the joint frequency (5.13)

$$\mathrm{P}(R_1, R_2, R_3) = \pi_{0+} \left(\frac{\pi_{10+}}{\pi_{0+}}\right)^{r_1} \left(\frac{\pi_{110}}{\pi_{10+}}\right)^{r_1 r_2} \left(\frac{\pi_{111}}{\pi_{110}}\right)^{r_1 r_2 r_3} c(r_1, r_2, r_3).$$

Clearly this pattern will extend to the general T-variate case. On taking logs we see that the canonical parameters are log specified-conditional odds ratios (SCORs; see Section 2.2.1): these are the log odds ratios given that previous observations are unity and future values are zero. Moreover, these are ratios of first order only and there are no interaction terms. The nature of dropout allows the replacement of crossproduct terms, such as $r_1 r_2 \cdots r_t$, by just r_t; if $r_t = 1$ then $r_s = 1$ for all $s < t$, whereas if r_t is

zero then of course the product is zero. Thus the full joint distribution of $\mathbf{R}$ belongs to the *linear* exponential family:

$$\mathrm{P}(\mathbf{R}) = c(\mathbf{r}) \exp\{\boldsymbol{\xi}'\mathbf{r} - C(\boldsymbol{\xi})\}, \tag{5.14}$$

with $c(\mathbf{r})$ an indicator shape function and $C(\boldsymbol{\xi})$ the normalizing constant $(-\log \pi_{00\cdots0})$.

The SCORs, except that for the last timepoint in an observation, are reproducible, since they are functions of marginal expectations: for example,

$$\mathrm{P}(R_1, R_2, R_3) = (1-\mu_1)\left(\frac{\mu_1-\mu_2}{1-\mu_1}\right)^{r_1}\left(\frac{\mu_2-\mu_3}{\mu_1-\mu_2}\right)^{r_2}\left(\frac{\mu_3}{\mu_2-\mu_3}\right)^{r_3}. \tag{5.15}$$

Note that the dropout-time canonical parameter for a subject who drops out differs from that for the same time for a larger observation. For example, letting $\xi_i^{(t)}$ denote the canonical parameter for timepoint i given dropout at time t, an observation with dropout at $t = 3$ has

$$\xi_3^{(3)} = \log\left(\frac{\pi_{111+}}{\pi_{110+}}\right) = \log\left(\frac{\mu_3}{\mu_2-\mu_3}\right)$$

but an observation with dropout after time 3 has

$$\xi_3^{(t)} = \log\left(\frac{\pi_{1110+}}{\pi_{110+}}\right) = \log\left(\frac{\mu_4-\mu_3}{\mu_2-\mu_3}\right).$$

If we were considering a full identity-link model, in this example we would estimate $\xi_3^{(3)}$ only for subjects with observed dropout at time 3, and $\xi_3^{(t)}$ only for subjects with no dropout before or at time 3.

A quite different approach to dropout models stems from noting that the indicator vector $\mathbf{R}$ can be aliased by a random variable, D say, denoting dropout timepoint (i.e. the first timepoint with missing observation). This new variable D follows a geometric-type distribution, being a count of the number of 'successes' until a 'failure' of a Bernoulli process. Unlike the standard geometric distribution, successive trials do not have the same probability of success, and in addition the distribution will

be truncated to some maximum number of 'successes'. I let $D = 1, 2, \ldots, T$ denote dropout time and arbitrarily let $D = T + 1$ if there is no dropout. (In Sections 5.3.3 and 5.3.4 these values are all decreased by one.) It is, however, unnecessary to express the truncated success-varying geometric distribution explicitly in order to estimate the changing success probabilities, since these are more easily obtained as shown below; the same modelling process allows either interpretive framework.

5.3.2 Marginal/survival parametrization

Motivated by the widespread use in other settings of marginal models, consider linking to the marginal expectations. For dropout data this gives a discrete-time survival model, directly parametrizing the survival function (or in practice its logit). Since

$$\mathrm{P}(R_t = 1) = \mathrm{P}(R_1 = R_2 = \cdots = R_t = 1),$$

then

$$\mu_t = \mathrm{E}[R_t] = \mathrm{P}(R_t = 1),$$

is the probability of survival until *at least* time t.

Two different approaches to the marginal fitting/modelling process will now be considered. Both suffer from two serious problems that are addressed by moving away from the standard canonical parametrization in Section 5.3.3.

Marginal links — method 1

Because the distribution here belongs to the exponential family and is of the canonical form, we can most simply apply the score equations of Section 2.3.2, e.g. equation (2.20) on page 50, which are for the general polynomial exponential family, merely noting that there are no interaction parameters to model. The variance matrix here is

$$V = \begin{pmatrix} \mu_1(1-\mu_1) & \mu_2(1-\mu_1) & \mu_3(1-\mu_1) & \cdots & \mu_T(1-\mu_1) \\ & \mu_2(1-\mu_2) & \mu_3(1-\mu_2) & \cdots & \mu_T(1-\mu_2) \\ & & \mu_3(1-\mu_3) & \cdots & \mu_T(1-\mu_3) \\ & & & \ddots & \vdots \\ \text{symmetric} & & & & \mu_T(1-\mu_T) \end{pmatrix},$$

for dropout at time T. This is easily calculated, but must still be inverted for each subject when calculating the score and information contributions in fitting. Although inversion of V is feasible, we can avoid even having to evaluate it, by using the following method.

Marginal links — method 2

The score equations For this particular distribution, we can derive $\partial\ell/\partial\boldsymbol{\nu}$ directly (rather than as $\partial\boldsymbol{\xi}'/\partial\boldsymbol{\nu}\cdot\partial\ell/\partial\boldsymbol{\xi}$; *cf* equation 2.20), because here $\boldsymbol{\xi}$ can be written readily in terms of $\boldsymbol{\nu}$. In the following the more familiar $\boldsymbol{\mu}$ is used in preference to $\boldsymbol{\nu}$ because only first-order expectations need be considered.

For greater symmetry I introduce the notation

$$\xi_0 = -C(\boldsymbol{\xi})$$

and, with a change of previous notation write

$$\boldsymbol{\xi} = (\xi_0, \xi_1, \ldots, \xi_T)',$$

where the first element is indexed zero rather than one to avoid confusion with standard element names.

The distribution function can now be written as

$$\mathrm{P}(\mathbf{R}) = c(\mathbf{z}) \exp\{\boldsymbol{\xi}'\mathbf{z}\}, \tag{5.16}$$

a minor modification of (5.14), where

$$\mathbf{z} = (1, r_1, r_2, \ldots, r_T)',$$

indexed as z_0 to z_T. The probability constraint on $\boldsymbol{\xi}$ is now implicit.

In this notation we obtain a symmetric form for expressions of ξ_t in μ_1 to μ_T. Since $\pi_{110} = \pi_{11+} - \pi_{111}$, etc., and any number of subscript + can be added to all of these terms, generalizing (5.15) gives

$$\xi_t = \log\left(\frac{\mu_t - \mu_{t+1}}{\mu_{t-1} - \mu_t}\right), \quad t = 1, 2, \ldots, T \tag{5.17}$$

where $\mu_0 = 1$ and $\mu_{T+1} = 0$. These values for impossible timepoints are not entirely arbitrary. That $\mu_0 = 1$ follows from $\xi_1 = \log[(\mu_1 - \mu_2)/(1 - \mu_1)]$, read from the distribution function. The algebraic convention adopted here has no effect on inference. By contrast the highest-indexed SCOR (alone) dictates that μ_{T+1} should be zero if the general form is to hold for all SCORs. This supposition has certain consequences discussed shortly in deriving the derivative matrix $\partial\boldsymbol{\xi}'/\partial\boldsymbol{\mu}$.

For the uth subject, with log likelihood contribution

$$\ell = \boldsymbol{\xi}'\mathbf{z} \tag{5.18}$$

where subscript u is omitted for clarity, and links

$$\boldsymbol{\mu} = g^{-1}(X\boldsymbol{\gamma}),$$

the score contribution is

$$\mathbf{U}(\boldsymbol{\gamma}) = \frac{\partial \ell}{\partial \boldsymbol{\gamma}} = \frac{\partial \boldsymbol{\eta}'}{\partial \boldsymbol{\gamma}} \frac{\partial \boldsymbol{\mu}'}{\partial \boldsymbol{\eta}} \frac{\partial \boldsymbol{\xi}'}{\partial \boldsymbol{\mu}} \frac{\partial \ell}{\partial \boldsymbol{\xi}} = X'\Delta M\mathbf{z},$$

where $\boldsymbol{\eta}$ is the vector of linear predictors, and assuming a logit link,

$$\frac{\partial \boldsymbol{\mu}'}{\partial \boldsymbol{\eta}} = \Delta = \mathrm{diag}[\mu_t(1 - \mu_t)],$$

as derived in (2.11). The $T \times (T+1)$ derivative matrix M has nonzero entries only as follows: for $t = 1, 2, \ldots, T$

$$\begin{aligned} M_{tt} = \frac{\partial \xi_{t-1}}{\partial \mu_t} &= -\frac{1}{\mu_{t-1} - \mu_t}; \\ M_{t(t+1)} = \frac{\partial \xi_t}{\partial \mu_t} &= \frac{\mu_{t-1} - \mu_{t+1}}{(\mu_t - \mu_{t+1})(\mu_{t-1} - \mu_t)}; \\ M_{t(t+2)} = \frac{\partial \xi_{t+1}}{\partial \mu_t} &= -\frac{1}{\mu_t - \mu_{t+1}}; \end{aligned}$$

since these are the only SCORs with nonzero derivatives w.r.t. each μ_t (see equation 5.17). As there is no term $M_{T(T+2)}$, the last row of M has only two entries, compared with three nonzero terms in each of the other T rows.

The size of M is related to the definition $\mu_{T+1} = 0$ in (5.17). One could let M be $T \times (T+2)$, assuming an unobserved (indeed unobservable) z_{T+1} fixed at zero — this corresponds to explicitly treating a non-dropout subject as if dropped out at time $T+1$. But then μ_{T+1} must be fixed at zero too. With the extended scheme, we would need some arbitrary value for ξ_{T+1}. But there is no point in this; the added terms cannot contribute to the likelihood.

Information matrix In order to fit by Fisher scoring we need the sum of contributions of the form

$$\mathcal{I} = \mathrm{E}\left[X'\Delta M \mathbf{Z}\mathbf{Z}' M' \Delta X\right],$$

or else evaluate the second derivatives of the log likelihood and use Newton–Raphson. An extreme simplification follows from assuming that each subject contributes

$$\mathcal{I} = \mathbf{U}\mathbf{U}'.$$

The reason is perhaps not immediately apparent and E[$\mathbf{Z}\mathbf{Z}'$], the only part of $\mathcal{I}$ that might depend on $\mathbf{R}$, is derived by two different strategies.

First, suppose all $\mathbf{Z}$ vectors are of the same size, T: that is, we carry on 'observing' zeros after dropout until design timepoint T. If, in fact iff, there are no missing explanatory variables, all is well, because at each cycle of the fitting algorithm we can

calculate *all* the μ_t, and so we can evaluate

$$\mathrm{E}[\mathbf{Z}\mathbf{Z}'] = \begin{pmatrix} 1 & \mu_1 & \mu_2 & \cdots & \mu_T \\ \mu_1 & \mu_1 & \mu_2 & \cdots & \mu_T \\ \mu_2 & \mu_2 & \mu_2 & \cdots & \mu_T \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mu_T & \mu_T & \mu_T & \cdots & \mu_T \end{pmatrix} \tag{5.19}$$

But this is an unlikely scenario except in simple experiments with non-time-varying covariates only. Whenever there is a time-varying covariate, it can be assumed that it will not be observed at dropout time or subsequently. We already have a difficult enough problem with missing variables at dropout time especially if, as is indeed likely, the effect of such variables on the probability of dropout is of interest. We would not want to compound this problem by having to deal with unobserved (and unobservable) variables for several occasions after dropout.

As an alternative, consider models where $\mathbf{Z} = (1, \mathbf{R}')'$ is truncated after the first zero, if there is dropout. Then observations $\mathbf{z}$ are vectors of length $d+1$ if there is dropout at timepoint d, and $\mathbf{Z}$ has an induced marginal exponential distribution of the form (5.16), by reproducibility.

An immediate problem arising from different observation vector lengths is that sense must be made of the score contribution $X'\Delta M\mathbf{z}$ for different lengths of $\mathbf{z}$. For the score contribution itself, this is easily handled; form X so that the contributions for those $\boldsymbol{\beta}$ not inherently of interest, within each particular subject, are set to be zero, or in practice more efficiently use a smaller X for the nonzero contributions only, and pad the result with suitable zeros before adding to the cumulative sum. Either of these schemes is parallel to the standard marginal modelling approach exploiting reproducibility.

When there is dropout at time d, naively $\mathrm{E}[\mathbf{Z}\mathbf{Z}']$ might be taken to be a $(d+1)\times(d+1)$ version of that given in (5.19). But this is wrong if we attempt to use the 'smaller' X and $\mathrm{E}[\mathbf{Z}\mathbf{Z}']$ and pad the calculated result with zeros to make a matrix conforming

with the largest size, T. Reproducibility fails here because we have not taken account of the *structural* zeros; this procedure is also equivalent to assuming a false identity for the dropout-time parameters (e.g. $\xi_3^{(3)} = \xi_3^{(t)}$ in the example on page 192).

Instead, take a fully pattern-mixture approach (defined in Section 6.2) whereby $\mathbf{Z}$ is no longer a random variable, but is a constant, given dropout time. Hence $\mathrm{E}[\mathbf{Z}\mathbf{Z}'] = \mathbf{z}\mathbf{z}'$. Practically, this can be calculated for 'smaller' X and $\mathbf{z}$ up to dropout size only, or for the full version with design zeros explicitly given in X and with $\mathbf{z}$ padded with trailing zeros after dropout up to full time T.

Since the likelihoods with indicators $\mathbf{R}$ (or $\mathbf{Z}$) considered as outcome variable, and those with dropout time, D, as outcome, are necessarily equivalent, the information contributions must be the same, and substituting $\mathbf{z}\mathbf{z}'$ for $\mathrm{E}[\mathbf{Z}\mathbf{Z}']$ is always valid. This establishes the surprising result that (5.19) need never be evaluated, even though $\mathrm{E}[\mathbf{Z}\mathbf{Z}'] \neq \mathbf{z}\mathbf{z}'$ if $\mathbf{Z}$ is viewed as a random variable.

In simulations (not presented here), convergence speed under either scheme is roughly the same. Of course, different intermediate values for the information (and thus score, after the second step) occur before convergence to identical parameter and information estimates.

Only rather trivial examples (with time-constant explanatory variables) could be simulated for comparative analysis because of the problems of monotonicity of means and evaluation of dropout-time links. In the absence of constraints, the fitting algorithm is extremely sensitive to the starting values for the Fisher steps, and if there are any time-varying covariates, the linear predictor at dropout time cannot be evaluated. These problems are overcome in the following model.

5.3.3 A reproducible, semi-canonical model for dropout

Again we ignore the case when unobserved covariate values directly affect the probability of dropout — the case of informative dropout — and concentrate on so-called random dropout: the dropout probability depends here only on observed values of explanatory variables X.

We will again take dropout time, rather than the set of indicators $\mathbf{R}$ (or $\mathbf{Z}$), as the

random variable, although it is convenient to maintain the $\mathbf{z}$ notation as in the previous section.

Note especially that from now in this section, exceptionally, timepoints are labelled starting at zero, thus R_i indicates dropout at the ith follow-up; we also redefine T as the number of follow-ups, rather than number of timepoints, and in similar vein the dropout variable D now takes values starting at zero. These conventions are natural given the 'shifted' parametrization now considered.

We revert to using $\boldsymbol{\xi}$ only for the polynomial exponential family parameters and introduce a 'shifted' parameter set

$$\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_d)', \quad d \leq T$$

truncated after dropout at follow-up d (if there is dropout), where we let

$$\theta_i = \xi_i, \quad i = 1, 2, \ldots, d \tag{5.20}$$

where the subscripts of ξ are timepoints, as previously (i.e. follow-up plus one). We will let θ_0 be a free parameter to be modelled, with the role of normalization constant now taken by ξ_d^{d+1}. I will label this ξ_+ to avoid confusion in subscript numbering conventions.

From the canonical form

$$C(\boldsymbol{\xi}) = \log \sum_{\text{all } \mathbf{z}} c(\mathbf{z}) \exp\ \{\boldsymbol{\xi}'\mathbf{z}\} = \log \sum_{\text{possible } \mathbf{z}} \exp\ \{\boldsymbol{\xi}'\mathbf{z}\}\,.$$

Thus for a subject dropping out at follow-up d,

$$\begin{aligned} \theta_0 &= -C(\boldsymbol{\xi}) = -\log\left(1 + e^{\xi_1} + e^{\xi_1+\xi_2} + \cdots + e^{\xi_1+\cdots+\xi_+}\right) \\ &= -\log\left(1 + e^{\theta_1} + e^{\theta_1+\theta_2} + \cdots + e^{\theta_1+\cdots+\theta_d+\xi_+}\right). \end{aligned}$$

Rearranging expresses ξ_+ in terms of θ:

$$\begin{aligned} e^{\xi_+} &= \frac{e^{-\theta_0} - \left(1 + e^{\theta_1} + e^{\theta_1+\theta_2} + \cdots + e^{\theta_1+\cdots+\theta_d}\right)}{e^{\theta_1+\cdots+\theta_d}} \\ &= \frac{1 - \left(e^{\theta_0} + e^{\theta_0+\theta_1} + \cdots + e^{\theta_0+\cdots+\theta_d}\right)}{e^{\theta_0+\cdots+\theta_d}}, \end{aligned} \tag{5.21}$$

from which we can obtain derivatives $\partial\xi_+/\partial\boldsymbol{\theta}$. The numerator of (5.21) is the marginal mean for follow-up time d, μ_d, which gives us an explicit formula for converting from canonical to marginal estimates. Note that such means are constrained to monotonicity, $\mu_d \geq \mu_{d+1}$, overcoming a problem with the previous approach (Section 5.3.2). Alternatively, if we have already calculated ξ_+,

$$\mu_d = \exp\left(\theta_0 + \cdots + \theta_d + \xi_+\right).$$

Score equations

Assume for simplicity a full identity link:

$$\boldsymbol{\theta} = X'\boldsymbol{\gamma}.$$

In practice I use a complementary log-log link to the probability θ_0, but this introduces only trivial complications to the following theoretical development. We transform the likelihood contribution for the uth subject (subscript u omitted),

$$\ell = \boldsymbol{\xi}'\mathbf{z} - C(\boldsymbol{\xi}),$$

using the chain rule:

$$\mathbf{U}(\boldsymbol{\gamma}) = \tilde{X}'\frac{\partial\boldsymbol{\xi}'}{\partial\boldsymbol{\theta}}(\mathbf{z} - \boldsymbol{\mu}). \tag{5.22}$$

Here $\tilde{X}$ is the same as X except that the first row is multiplied by $\theta_0 = \partial\theta_0/\partial\alpha_0$ for the link $\theta_0 = -e^{\alpha_0}$. Also I have used the standard marginal-model result $\partial\ell/\partial\boldsymbol{\xi} = \mathbf{z} - \boldsymbol{\mu}$, for $\boldsymbol{\mu}$ the marginal means, and using (5.20) and (5.21)

$$\frac{\partial \boldsymbol{\xi}'}{\partial \boldsymbol{\theta}} = \begin{pmatrix} 0 & \cdots & 0 & f_1 \\ & & & \vdots \\ & I_{(d\times d)} & & \vdots \\ & & & f_{d+1} \end{pmatrix}$$

where I is an identity matrix and the $(d+1)$-vector $\mathbf{f}$ has jth entry

$$f_j = \frac{\partial \xi_+}{\partial \theta_{j-1}} = \frac{\left(e^{\theta_0} + e^{\theta_0+\theta_1} + \cdots + e^{\theta_0+\cdots+\theta_{j-2}}\right) - 1}{1 - \left(e^{\theta_0} + e^{\theta_0+\theta_1} + \cdots + e^{\theta_0+\cdots\theta_{d-1}}\right)}, \quad j = 1, 2, \ldots, d+1. \tag{5.23}$$

For $j = 1$, the numerator is -1, otherwise the numerator equals $-\mu_{j-2}$. If there is no dropout, when conventionally we set $d = d_{\max} + 1$, we set here instead $d = d_{\max}$.

If we again condition on observed dropout time D, rather than regarding $\mathbf{Z}$ as a random variable, the information matrix contribution for Fisher scoring is just

$$\mathcal{I} = \mathbf{U}\mathbf{U}'$$

as in the previous section.

Simplifications in fitting. Instead work from the observed likelihood (conditional on observed dropout time), expressed in terms of $\boldsymbol{\theta}$ and final design timepoint parameter ξ_+ only, i.e. equation (5.18) substituting $\boldsymbol{\xi} = (\boldsymbol{\theta}', \xi_+)'$. For subjects who drop out,

$$\frac{\partial \ell}{\partial \boldsymbol{\theta}} = \mathbf{z},$$

while for subjects who do not drop out,

$$\begin{aligned} \frac{\partial \ell}{\partial \boldsymbol{\theta}} &= \mathbf{z} + \frac{\partial \xi_+}{\partial \boldsymbol{\theta}} \\ &= \mathbf{z} + \mathbf{f} \end{aligned}$$

where $\mathbf{f}$ is given by equation (5.23).

5.3.4 Example

As part of on ongoing study into causes of renal transplant failure at the Royal Liverpool University Hospital, the relationship between graft failure and unusual levels of the treatment drug Cyclosporin A (CyA) was studied. If CyA level is too low then rejection might follow (the known overall effect of CyA being to repress the body's rejection of grafts), but high levels of CyA are toxic to the liver itself. It was of interest to determine if CyA as assessed by measuring the trough level, i.e. immediately before administration of the next dose, exhibited this expected behaviour.

Data were available for 150 transplant patients, with graft status and trough CyA (measured on a 3-point scale, -1 for low, 0 for normal, 1 for high) at 12, 24, 48, 60 and 72 months after the operation, hereafter labelled follow-up 0 to 4.

Graft failure is dropout in the terminology of this chapter, and the models of the preceding subsections may be used to study it. I will illustrate only the shifted-canonical model of Section 5.3.3; certain marginal-model inference can be drawn from this as shown below (page 205). The intercept-unconstrained parametrization for the null-explanatory model is

$$\begin{aligned} \theta_0 &= -\mathrm{e}^{\alpha_0} \\ \theta_i &= \alpha_i, \qquad i = 1, 2, 3, 4. \end{aligned}$$

This model is saturated for the observed marginal dropout pattern given in Table 5.1. Parameter estimates, standard errors and model deviance are given in Appendix A5.3.4; this is model A5.1 of that section.

The comparatively large standard errors for follow-ups 2 to 4 suggest that a common intercept for these timepoints would fit well, but this is not pursued here since the focus of attention is the explanatory model. If these intercepts were forced to be equal then the sample differences might incorrectly affect estimates for explanatory-model parameters.

Trough CyA levels prior to time 0 were not available in the database and θ_0 is modelled as a constant throughout. (In fact there are some CyA level values prior to time 0, at

Table 5.1: Renal graft failure, given as observed failures at each timepoint and as Kaplan–Meier proportions, i.e. failures/(number remaining); a constant dropout rate is unlikely.

Failure time					
0	1	2	3	4	No failure
25 (0.166)	9 (0.072)	15 (0.129)	13 (0.128)	26 (0.295)	62

6 months in the records, but these have not all been entered into the database.)

If low and high CyA trough levels affect the probability of graft failure uniformly we would expect (naively; see below) to find a significant improvement for the model

$$\theta_i = \alpha_i + \beta_{\text{low}} x_{\text{low}} + \beta_{\text{high}} x_{\text{high}}$$

where the x are dummy variables relating to follow-up $i - 1$ and the normal CyA level is the baseline. In fact, model (A5.2) is no better than the null model (deviance change 4.95 on 2 d.f.). There is a nominally significant improvement ($P = 0.047$) on the null model if only the effect of low CyA is added to it (A5.3), but this is not convincing.

To see why this model fits so poorly we can ascertain non-constant effects by using a separate β parameter pair at each timepoint. Although this model (A5.4) fits significantly better than the null model (deviance change 28.9 on 8 d.f.; $P < 0.001$), we find that a low trough CyA level appears to increase the failure rate at follow-up 2 but decrease it at follow-up 3; high CyA level appears to increase the probability of failure at follow-up 4 only. There are further discrepancies when considering the effects that are not significant on univariate z-tests. As this is clinically implausible, we infer this is an artifact.

Instead, let us consider a linear effect for the trough CyA level; model A5.5 shows the fit for a model not assuming such effects are constant over timepoints. This is also significant w.r.t. the null model but far more plausible, with high trough CyA

levels decreasing the probability of failure at follow-ups 1 and 2, but increasing it at follow-ups 3 and 4 (and vice versa for low CyA). A clinical interpretation is that trough CyA is a reasonably good surrogate for overall CyA level; high trough CyA in the early years does what CyA should do, prevent graft rejection, but later its toxic effect outweighs this advantage.

Suppose the drug regime were changed so that all patients had constant high trough values. Using the estimates of model A5.5, a sample of 150 would have pointwise expected failures (25, 8.8, 8.3, 12.5, 42.4). This is broadly the same as the observed sample excepting a large overestimate of failures at the last timepoint.

Interpretation of parameters

The linear predictors (except for time 0, for which θ_0 is just the log probability of immediate dropout) are, from equation (5.17),

$$e^{\theta_d} = \frac{P(D = d)}{P(D = d - 1)}. \tag{5.24}$$

This is *not* the usual conditional probability,

$$P(D = d \mid D \geq d - 1) = \frac{P(D = d)}{P(D \geq d - 1)};$$

its closest counterpart is the adjacent logit model (e.g. Agresti, 1990). The interpretation of the result that all the θ_d are approximately zero is that there is an even spread of observation vector lengths; the dropout rate itself would be increasing. A constant dropout rate would have $P(D = d + 1)/P(D = d) = k < 1$, so the θ_d should be constant but negative. Neither of these occurred in the CyA data set.

The parameters $\boldsymbol{\theta}$ have an interesting interpretation if the dropout rate is constant. Consider an artificial sample with $N = 1000$ with dropout vector $(100, 90, 81, \ldots)$, i.e. a constant dropout rate of 0.1. Then $P(D = d) = (0.1, 0.09, 0.081, \ldots)$ gives $e^{\boldsymbol{\theta}} = (0.9, 0.9, \ldots)$; that is, θ_d is the log probability of *not* dropping out at time d. This does not hold in general.

Naively, if an explanatory variable has a constant effect on the probability of dropout, it should have roughly equal numerical influence on the linear predictors, even if the baselines are not constant. But equality does not hold in this case, and a change in θ_d is more easily interpreted in terms of the marginal $\mathrm{P}(D = d-1)$, rather than an effect on the rate itself.

For this more marginal interpretation, noting $\mu_d = \mathrm{P}(D \geq d)$, we can write

$$\mathrm{e}^{\theta_0+\theta_1} = \mu_0 - \mu_1 = \mathrm{P}(D = 0).$$

In general

$$\theta_0 + \theta_1 + \cdots + \theta_{d+1} = \log\ \mathrm{P}(D = d).$$

If

$$\theta_{d+1} = \alpha_{d+1} + \beta_{d+1} x_d$$

then for two different values $x_d^{(1)}$ and $x_d^{(2)}$, assuming all other previous covariates are the same,

$$\beta_{d+1}(x_d^{(1)} - x_d^{(2)}) = \log\left(\frac{\mathrm{P}(D = d \,|\, x_d^{(1)})}{\mathrm{P}(D = d \,|\, x_d^{(2)})}\right).$$

This is easier to interpret than the effect on the adjacent-category ratio $\mathrm{e}^{\theta_{d+1}}$; namely $\exp\{\beta_{d+1}\}$ is the relative risk of dropout at follow-up d for a unit difference in covariate x at follow-up d.

5.3.5 Discussion

Choosing the 'shifted' parameter set $\boldsymbol{\theta}$ in Section 5.3.3 overcomes the problem that the canonical parameter ξ_d for a subject not dropped out by time d is different to ξ_d for a subject with dropout at time d. The $\boldsymbol{\theta}$ are identical for different observation sizes, so we gain full reproducibility. It is a curiosity that for non-degenerate distributions the marginal model is reproducible, whereas for this distribution for dropout, the (shifted) canonical parameters achieve this aim.

Another reason for introducing the new parameters is that θ_0, which corresponds exactly to the constant term, $-C(\boldsymbol{\xi})$, in the standard canonical form, is easily inter-

pretable as being the log probability of immediate dropout. It makes good sense to model this as a constant (in regression terms, an intercept), since in cases of immediate dropout we could never realistically expect to know the values of any explanatory variables.

In the shifted model, explanatory variables at dropout time are not explicitly modelled even if known, but their history determines $\boldsymbol{\theta}$ and hence ξ_+. The set of parameters $\boldsymbol{\theta}$ uniquely determines the full distribution function for the outcome vector, in the same way as a full set of $\boldsymbol{\xi}$ values determine the constant $C(\boldsymbol{\xi})$.

A further advantage of the new parametrization is that it ensures monotonicity of the derived marginal expectations without the need to build in explicit constraints.

Chapter 6

Data with possibly informative dropout

This chapter extends the discussion in Chapter 5 of the analysis of data when some values are missing to the case when those values are missing 'informatively', a concept defined in Sections 6.1 and 6.2. In Section 6.3, we consider the adaptation of the Markov chain model of Chapter 4 to deal with data of this type and find that it offers a very natural framework in which to specify a data model and the missing-data mechanism simultaneously, in a variety of new and established modelling frameworks.

In Section 6.4 we consider what I describe as *imputed maximum likelihood*, i.e. the computation of ordinary maximum likelihood estimators for data with imputed values substituted for missing values (analogous to the the M step of the classical EM algorithm). We consider both marginal and transitional modelling of such imputed probability tables, in Sections 6.4.2 and 6.4.3, respectively. In Section 6.4.4 we look at the process of estimation and the inter-relationship between specification of a completed table and pre-specification of otherwise non-identifiable parameters, and extend this discussion in Section 6.4.5 to show that pre-specifying such non-identifiable parameters is equivalent to the EM approach, since the missing data makes no contribution to the likelihood. This point is made again in Section 6.4.6 by noting the failure of a profile likelihood approach to determine the non-identifiable parameters.

The certain failure of maximum likelihood techniques to provide any evidence for or against data being missing informatively is demonstrated in Section 6.5.

The chapter concludes in Section 6.6 with an example analysis of a study of senile dementia. Here we demonstrate the use of the timepoint-wise pattern-mixture approach of Section 6.3.2, and reveal some perhaps unexpected difficulties in model selection, even after the simplifying decision not to model the missing data as informative. This example also illustrates the use of a ternary missing-value indicator and demonstrates how easily this is incorporated within a Markov chain approach.

6.1 Introduction

Almost inevitably, when measurements are repeated over time, some subjects will be lost to follow-up. If T observations are intended, a subject observed only up until timepoint $t-1$ is said to exhibit *dropout* at time t (with $t \leq T$). I do not consider here models for which some subjects return after being missing for a number of observations.

To illustrate, let variables Y_1 and Y_2 be observed sometimes ($\surd$) and sometimes not ($\times$), with each subject in one of the four patterns

Y_1	Y_2	
$\surd$	$\surd$	(a)
$\surd$	$\times$	(b)
$\times$	$\surd$	(c)
$\times$	$\times$	(d)

Data missing entirely (pattern d) is usually considered a question of sample selection bias, rather than as a missing data problem as such. I do not consider here in detail the possibly biasing effect of such entirely missing data, although such 'immediate dropout' has been discussed in Section 5.3.

Pattern (c) corresponds to *dropin*, which as mentioned above is not considered in detail, although the marginal approach (Section 6.4.2) is able to handle this pattern

too.

Dropout is the simplest class of missing data problem; in the tableau above, data are subject to dropout if there is a mixture of patterns (a) and (b). More generally dropout models apply to strictly monotone missing data patterns, with no missing values at time 1.

The term *dropout modelling* has come to be used ambiguously. Here *dropout model* means a model for the probability of dropout (as in Section 5.3). Otherwise when attention is focussed on the intended measurement, with dropout a nuisance, it is more correct to use *model for data subject to dropout*; the dropout model is only a part of the overall model and in many cases need not be parametrized.

6.2 Nomenclature

Let $\mathbf{Y}$ be the vector of observations, of length T if there is no dropout, with joint density $f(\mathbf{y})$. Missing observations are recorded by a vector of indicators $\mathbf{R}$, of length T, with

$$R_t = \begin{cases} 1, & Y_t \text{ observed} \\ 0, & Y_t \text{ unobserved.} \end{cases}$$

For inference, we model the joint density of $\mathbf{Y}$ and $\mathbf{R}$. This can be factorized in several ways (see for example Section 6.3) of which there are two extreme types: a *selection model* takes

$$f(\mathbf{y},\mathbf{r}) = f(\mathbf{y})f(\mathbf{r}\,|\,\mathbf{y}), \tag{6.1}$$

whereas a *pattern-mixture model* uses

$$f(\mathbf{y},\mathbf{r}) = f(\mathbf{r})f(\mathbf{y}\,|\,\mathbf{r}). \tag{6.2}$$

For a random sample of observation vectors $\mathbf{Y}_u$ with associated indicators $\mathbf{R}_u$, $u = 1 \ldots n$, the contribution of each to the overall likelihood is the observed marginal

$$\sum_{\text{missing } Y_{ut}} f(\mathbf{y}_u, \mathbf{r}_u). \tag{6.3}$$

In the following, the subscript u is dropped. Integration replaces summation for continuous variables.

It is natural to partition $\mathbf{Y}$ into observed and unobserved values and write

$$\mathbf{Y} = (\mathbf{Y}'_{\text{obs}}, \mathbf{Y}'_{\text{miss}})'.$$

Though this notation is quite general, for dropout patterns the order of the elements is unchanged. The likelihood is obtained from contributions $f(\mathbf{y}_{\text{obs}}, \mathbf{r})$ only.

Data are said to be missing

(a) completely at random (MCAR), if

$$f(\mathbf{r} \mid \mathbf{y}) = f(\mathbf{r}); \tag{6.4}$$

(b) at random (MAR), if

$$f(\mathbf{r} \mid \mathbf{y}) = f(\mathbf{r} \mid \mathbf{y}_{\text{obs}}); \tag{6.5}$$

(c) non-ignorably (NIGmiss) — also known as informatively, and as non-randomly — if

$$f(\mathbf{r} \mid \mathbf{y}) = f(\mathbf{r} \mid \mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{miss}});$$

Note that it is the presence of $\mathbf{y}_{\text{miss}}$ in the conditional distribution that determines NIGmiss, even if $f(\mathbf{r} \mid \mathbf{y})$ does not depend on $\mathbf{y}_{\text{obs}}$.

Under the MCAR assumption, data are missing at random in the obvious sense; under MAR the observed values determine the probability of data being missing. Although this terminology frequently causes confusion, it is so well established it is retained here. The conceptual problem stems from the fact that when data are *not* NIGmiss — hence either MCAR or MAR — it is common to refer to them as *non-informative* missing data (here denoted NINFmiss). But under MAR the data *are* informative in predicting dropout.

The NIGmiss assumption is that the *unobserved* data values are correlated with the indicators. For example, with binary outcomes, if subjects who would record a score

of 1 are less likely to turn up for interview than those who would score 0, then the data are NIGmiss. The observed mean is then a biased estimate of the population mean.

The NINFmiss definitions (6.4 and 6.5) can be written in conditional-independence operator notation as

$$\mathbf{R} \perp\!\!\!\perp \mathbf{Y}_{\text{miss}} \mid \mathbf{Y}_{\text{obs}},$$

Because of the symmetry of this form an equivalent defining condition for NINFmiss data is that

$$f(\mathbf{y}_{\text{miss}} \mid \mathbf{y}_{\text{obs}}, \mathbf{r}) = f(\mathbf{y}_{\text{miss}} \mid \mathbf{y}_{\text{obs}}). \tag{6.6}$$

Thus we could redefine

(c′) data are NINFmiss if the conditional distribution for the missing data only depends on observed data values, and not on dropout pattern; otherwise they are NIGmiss.

If NINFmiss holds we can impute missing values from those observed, which is fundamental to the EM algorithm for MAR or MCAR data (Dempster *et al.*, 1977). However, there is little reason to suppose NINFmiss holds, except for computational and conceptual simplicity. It is shown in Section 6.5 that it is always impossible to test for NINFmiss by maximum likelihood ratio tests alone.

6.2.1 Ignored vs ignorable missing values

For data with dropout, a distinction should be made between the *ignored* missing values, i.e. $Y_s \; \forall s > t$, and the *nuisance* missing value for Y_t at dropout time t.

In a seminal paper (Diggle and Kenward, 1994) the more general case of non-ignorably missing data is restricted to non-ignorable dropout (NIGdrop), which is defined as occurring if the (unobserved) value of the dropout-time variable, Y_t, affects the probability of dropout, $f(r_t \mid y_1, y_2, \ldots, y_t)$. The issue of possible dependence between post-dropout, Y_s values, and history, i.e. $(y_1, \ldots, y_u)'$ and $(r_1, \ldots, r_u)'$ for $t \leq u < s$, for each s, is not addressed, nor is that of allowing potential Y_s values to influence the

estimation of the mean for Y_s. No attempt at modelling the full joint distribution is made.

This strategy is reasonable, because the modelling of such variables would introduce many more nuisance missing values. However, it does introduce an implicit two-tier scheme of ignorability, under which

(a) Post-dropout Y_s values are always ignored, although no attempt is made to assess whether in terms of the true, full model such values are or are not actually NIGmiss ignorable; and

(b) Dropout-time Y_t values are not necessarily ignored but are called ignorable if they do not affect the true (or assumed) dropout model.

This same scheme is adopted here. One reason is that both dropout, Y_t, and post-dropout, Y_s, values can readily be ignored in the limited sense that such unobserved values cannot affect the likelihood as defined in (6.3). On the other hand, such values should not be ignored if we wish to allow for NIGmiss, although we cannot use likelihood-based inference to test for this. It seems sufficient to be concerned for now with NIGdrop only, given that a full NIGmiss model has greater problems than NIGdrop, which is itself unidentifiable (Section 6.5).

6.3 Markov chain models for data with dropout

The joint probability $f(\mathbf{y}, \mathbf{r})$ has many other possible factorizations than those given in equations (6.1) and (6.2). For longitudinal data is is natural to first factorize by timepoint,

$$f(\mathbf{y}, \mathbf{r}) = f(y_1, r_1) f(y_2, r_2 \,|\, y_1, r_1) \cdots f(y_T, r_T \,|\, \mathbf{h}_T, \mathbf{h}_{rT}) \tag{6.7}$$

where $\mathbf{h}_t$ is restrictively-defined history (Section 4.2) and $\mathbf{h}_{rt}$ is similarly-defined indicator history. This is a Markov model for bivariate observations. Because there is no dropout at time 1 in our models (Section 6.1), r_1 is redundant, and also indicator

history h_{rt} reduces to just r_{t-1} (*cf* Section 5.3). Thus equation (6.7) reduces to

$$f(\mathbf{y}, \mathbf{r}) = f(y_1)f(y_2, r_2 \mid y_1) \cdots f(y_T, r_T \mid \mathbf{h}_T, r_{T-1}). \tag{6.8}$$

After this factorization, a selection-model type approach is to take

$$f(y_t, r_t \mid \mathbf{h}_t, r_{t-1}) = f(y_t \mid \mathbf{h}_t, r_{t-1})f(r_t \mid y_t, \mathbf{h}_t, r_{t-1}), \tag{6.9}$$

whereas a pattern-mixture type model uses

$$f(y_t, r_t \mid \mathbf{h}_t, r_{t-1}) = f(r_t \mid \mathbf{h}_t, r_{t-1})f(y_t \mid r_t, \mathbf{h}_t). \tag{6.10}$$

These factorizations lead to new models, which I call *timepoint-wise* selection and pattern-mixture, respectively. Classically, a selection model specifies the full marginal joint distribution, $f(\mathbf{y})$, and a pattern-mixture model specifies the full joint conditional $f(\mathbf{y} \mid \mathbf{r})$. The models following from (6.9) and (6.10) are selection and pattern-mixture, respectively, at each timepoint, but not overall. Both of them contain elements of both selection and pattern-mixture approaches, as will be discussed further in Section 6.3.2.

6.3.1 Selection models

A full selection model can be developed by considering its timepoint-wise counterpart, as follows (Diggle and Kenward, 1994).

We wish to model the set of marginal expectations $\mathrm{E}[Y_t^*]$, where y_t^* is the actual value of a (possibly unobserved) measure at time t. Denote the *observed* value as Y_t; that is

$$Y_t = \begin{cases} Y_t^*, & t = 1, \ldots, D-1 \\ \mathrm{NA}, & t \geq D \end{cases}$$

where D is a random variable identifying dropout time. In terms of indicators $\mathbf{R} =$

$(R_1, R_2, \ldots, R_T)'$ this is

$$Y_t = \begin{cases} Y_t^*, & R_1, \ldots, R_{t-1} = 1,\ R_t \neq 0 \\ \text{NA}, & \text{at least one } R_i \neq 1,\ i = 1, \ldots, t \end{cases}$$

or in terms of observed values

$$y_t = \begin{cases} y_t^*, & y_t^* \in \{0, 1\} \text{ known} \\ \text{NA}, & y_t^* = \text{unknown} \end{cases}$$

The probability function of $\mathbf{Y}^*$ is given by the model for the underlying process, which here I assume to be polynomial exponential family. We need also a model for dropout; for binary indicators

$$P(R_t = r_t \mid y_t^*, \mathbf{h}_t) = \exp\{\phi_t r_t - C(\phi_t)\}, \quad r_t \in \{0, 1\} \tag{6.11}$$

for dropout at time t. Recall the convention here, and in the following, of using ϕ for the canonical parameters of the conditional distribution, reserving ξ for the canonical parameters in the joint polynomial exponential family form.

Assume a logit link for parameters of more immediate interest, $\boldsymbol{\delta}$:

$$\boldsymbol{\phi} = X_\delta \boldsymbol{\delta}$$

where X_δ is the dropout-model design matrix. This may share terms with the design matrix for the main model, and will probably include a time effect. Also, unless dropout is non-informative, X_δ must include values of y_t^* which are not observed for subjects who drop out. An important related point, not mentioned by Diggle and Kenward (1994), who simplified most of their presentation by excluding explanatory variables, is that X_δ must include only known terms and does not allow for missing explanatory-variable values.

Suppose first that dropout depends strongly on the (unobserved) $\mathbf{Y}^*$, but that Y_t^* itself depends strongly on $\mathbf{H}_t$; then a random dropout (MAR) model will probably be

quite adequate. Suppose instead that intra-subject association is very weak, but that $\mathbf{Y}$ is predicted quite accurately by explanatory variables, $\mathbf{X}^*$; then Y_t^* (and hence dropout) depends strongly on X_t^* rather than the observed H_t, and in most cases these critical predictors will not be observed.

The factorized distributions for $\mathbf{R}$ and $\mathbf{Y}^*$ together induce a conditional distribution for $\mathbf{Y}$; specifically

$$P(Y_t = y_t \mid \mathbf{h}_t) = P(Y_t^* = y_t^* \mid \mathbf{h}_t)P(R_t = r_t(y_t) \mid y_t^*, \mathbf{h}_t), \quad y_t \in \{0, 1, \text{NA}\}, \quad (6.12)$$

where

$$r(y) = \begin{cases} 1, & y \in \{0, 1\} \\ 0, & y = \text{NA} \end{cases}$$

In less cluttered notation, $P(\mathbf{y}) = P(\mathbf{y}^*, \mathbf{r}) = P(\mathbf{y}^*)P(\mathbf{r} \mid \mathbf{y}^*)$. By the law of total probability (no derivation is given in Diggle and Kenward, 1994) the probability of dropout at time t is then

$$\begin{aligned} P(R_t = 0 \mid \mathbf{h}_t) &= P(Y_t = \text{NA} \mid \mathbf{h}_t) \\ &= \sum_{y_t^* \in \{0,1\}} P(Y_t^* = y_t^* \mid \mathbf{h}_t)P(R_t = 0 \mid y_t^*, \mathbf{h}_t) \end{aligned} \quad (6.13)$$

provided always that $\mathbf{h}_t$ is known, that is, the observation unit has not already dropped out, since

$$P(Y_t = \text{NA} \mid y_{(t-1)} = \text{NA}) = 1$$

for permanent dropout.

This in turn induces a joint density for $\mathbf{Y}^{(d)}$, that is, for a subject dropping out at time d, as

$$P(\mathbf{y}^{(d)}) = P(\mathbf{y}^{*(d-1)}) \left(\prod_{t=2}^{d-1} P(R_t = 1 \mid y_t^*, \mathbf{h}_t) \right) P(Y_d = \text{NA} \mid \mathbf{h}_d), \quad (6.14)$$

with the final term on the right-hand side not present for a full observation. This is

expressed in terms of the unconditional joint density for $\mathbf{Y}^*$, parametrized as

$$\mathbf{g}(\mathrm{E}[\mathbf{Y}^*]) = X\gamma$$

and the conditionals for each R_t, parametrized as in (6.13).

The use of this form of the density is natural in the original setting of Gaussian data where the joint distribution is especially easily derived from the conditional distributions in canonical form. For categorical data, implicitly $\mathbf{Y}^*$ can be modelled marginally, so that the distributions for vectors of different sizes are easily obtainable (by reproducibility). However, for categorical observations there are other possibilities. Instead the super-observation $(\mathbf{Y}, \mathbf{R})$, which has straightforward polynomial exponential family distribution, can be modelled directly and the dependencies can be modelled by parametrizing the higher-order interactions, allowing for the structural zeros inherent in the dropout process. A simple example of this approach is given in Section 6.4.

Timepoint-wise selection models

Alternatively, one might parametrize the model directly in its Markov chain factorized form, here shown by example. Consider trivariate binary data with

$$f(\mathbf{y},\mathbf{r}) = \underbrace{f(y_1)}_{\alpha_1}\underbrace{f(y_2 \mid y_1)}_{\alpha_2,\alpha_{21}}\underbrace{f(r_2 \mid y_1, y_2)}_{\delta_2,\delta_{21},\delta_{22}}\underbrace{f(y_3 \mid y_1, y_2, r_2)}_{\alpha_3,\alpha_{32}}\underbrace{f(r_3 \mid y_1, y_2, r_2, r_3)}_{\delta_3,\delta_{21},\delta_{22}} \qquad (6.15)$$

where the parameters for each univariate conditional distribution are indicated under the braces. For simplicity assume there are no explanatory variables, although the possible danger of such simplification has been mentioned on page 214. For this example (with later discussion in Section 6.5.2) assume a non-saturated model with

$$\begin{aligned} \phi_{Y_1} &= \alpha_1 \\ \phi_{Y_2 \mid Y_1} &= \alpha_2 + \alpha_{21} y_1 \\ \phi_{R_2 \mid Y_1, Y_2} &= \delta_2 + \delta_{21} y_1 + \delta_{22} y_2 \end{aligned}$$

$$\phi_{Y_3 \mid Y_1,Y_2,R_2} = \alpha_3 + \alpha_{32} y_2$$

$$\phi_{R_3 \mid Y_1,Y_2,Y_3,R_2} = \delta_3 + \delta_{21} y_2 + \delta_{22} y_3$$

For binary data the expression $\phi_{R_2 \mid Y_1}$, for example, actually describes two separate, variationally independent (if unconstrained) parameters, $\phi_{R_2 \mid Y_1=1}$ and $\phi_{R_2 \mid Y_1=0}$. Similarly, $\phi_{Y_2 \mid Y_1,R_2}$ is shorthand for four canonical parameters.

The parameters $\boldsymbol{\alpha}$ are intercepts for the model of main interest while the $\boldsymbol{\delta}$ parametrize the dropout mechanism. Here the Y_t depend only on Y_{t-1}, but the dependency is not forced to be common. Dropout probabilities for time t are assumed to depend on Y_t and Y_{t-1} with the same parameters, but the intercept is not assumed common.

As with any Markov-type factorization, the time-1 model is of little intrinsic interest and could without loss be omitted. Also, the dropout predictors will in general not share parameters with the data model, so can be fitted separately (or not at all, if NINFmiss is assumed), but note that in the selection-type factorization possibly unobserved values enter the linear predictors for dropout indicators. This feature is shared with the Diggle and Kenward model above, where dropout is parametrized piecewise conditionally no matter what form is taken for P(Y^*).

6.3.2 Pattern mixture models

For a pattern-mixture model it is unnecessary to assume the existence of an underlying true variable $\mathbf{Y}^*$. With reference to the tableau on page 208, a selection approach models rows (a) and (b) combined, whereas a pattern-mixture approach has a separate model for (a) and (b).

In practice we might introduce simplifying parameter-led constraints while studying pattern-mixture models that would introduce elements of selection modelling; for example, we might assume certain common parameters for outcome vectors of varying size, and indeed this is the assumption in Section 6.6. Thus, the terminology is of less use in practical modelling than it is for theoretical developments. Pure selection models should be based on $f(\mathbf{y}^*)$, marginalized over $\mathbf{R}$, but many practitioners advocate that the number of observations should appear as a covariate; pure pattern-mixture

models are for $f(\mathbf{y} \mid \mathbf{r})$, but $\mathbf{r}$ need not appear as such in the linear predictors provided dropout is NINFmiss.

A full pattern-mixture model as defined by equation (6.2) models the marginal density of $\mathbf{R}$ and this can be reduced to a univariate variable, D, indicating dropout time (Section 5.3). Thus all the modelling can be based on $f(\mathbf{y} \mid d)$, which is fully observed, provided the dropout model itself is not of interest. A selection model such as above is biased for $\mathbf{Y}^*$ if dropout is misspecified, whereas the pattern-mixture $f(\mathbf{y} \mid d)$ is not. However, a selection model is often preferred because of ease of interpretation.

Timepoint-wise pattern-mixture models

The timepoint-wise pattern-mixture model, in common with timepoint-wise selection models, has elements of both selection and pattern-mixture approaches, since predictors for $\mathbf{Y}$ depend on $\mathbf{R}$ (or D) and predictors for $\mathbf{R}$ depend on $\mathbf{Y}$. The difference is that here I will factorize $f(y_t, r_t)$ as $f(y_t \mid r_t)f(r_t)$ where above we had $f(y_t)f(r_t \mid y_t)$, all dependent on history and probably covariates.

The analogue of the model at the end of the previous subsection, page 216, is

$$f(\mathbf{y}, \mathbf{r}) = \underbrace{f(y_1)}_{\alpha'_1} \underbrace{f(r_2 \mid y_1)}_{\delta'_2, \delta'_{21}} \underbrace{f(y_2 \mid y_1, r_2)}_{\alpha'_2, \alpha'_{21}, \delta'_{22}} \underbrace{f(r_3 \mid y_1, y_2, r_2)}_{\delta'_3, \delta'_{21}} \underbrace{f(y_3 \mid y_1, y_2, r_2, r_3)}_{\alpha'_3\, \alpha'_{32}, \delta'_{22}} \tag{6.16}$$

parametrized as

$$\begin{aligned}
\phi_{Y_1} &= \alpha_1 \\
\phi_{R_2 \mid Y_1} &= \delta'_2 + \delta'_{21} y_1 \\
\phi_{Y_2 \mid Y_1, R_2} &= \alpha'_2 + \alpha'_{21} y_1 + \delta'_{22} r_2 \\
\phi_{R_3 \mid Y_1, Y_2, R_2} &= \delta'_3 + \delta'_{21} y_2 \\
\phi_{Y_3 \mid Y_1, Y_2, R_2, R_3} &= \alpha'_3 + \alpha'_{32} y_2 + \delta'_{22} r_3
\end{aligned}$$

Primed parameters are broadly equivalent to their counterparts in the previous formulation; similarly named parameters have the same function as each other, but have different values, because they do not appear in the same linear predictors. Further

discussion and a comparison of this and the previous formulation are considered in Section 6.5.2.

6.4 Imputed maximum likelihood

Imputation (that is, assumption) lies at the heart of any strategy for estimation when data are missing. If the missing mechanism is MCAR or MAR, then the imputation is equivalent to the combined effects of the E-steps in the EM algorithm. Often only one such step will be needed for models near saturation. *Imputed maximum likelihood* (IML) represents the likelihood maximized in the M-step, after replacing the missing values by their E-step expectations.

When data are NIGmiss, there is insufficient information in the sample on which to base E-step expectations, and one approach is multiple imputation (Little and Rubin, 1987). IML estimates are in this context those obtained for each single imputation of the multiple process.

Throughout this section a simple example of missing data is considered (Section 6.4.1; discussed at some length in Little and Rubin, 1987). Marginal and Markov chain parametrizations for the example problem are set out in Sections 6.4.2 and 6.4.3, respectively. In Section 6.4.4 various approaches to imputing the missing values are introduced. Viewing the filled-in table as though it were a complete observation leads to the formulation of the log likelihood function described and analysed in Section 6.4.5, whereas a stricter definition of likelihood (as a function of the parameters given the *observed* data) leads to the analyses of Section 6.4.6. Both approaches are seen to fail to produce acceptable parameters: in the first case likelihood is maximized outside the parameter space and in the second, the likelihood maximum is not unique.

6.4.1 A simple example

The data in Table 6.1 are taken from Little and Rubin (1987). Data follow patterns (a) and (b) of Section 6.1, page 208, so we need a model for data subject to dropout, albeit the simplest possible case of this.

Table 6.1: Data from Little and Rubin (1987), Chapter 11.

Y_1	Y_2	R_2	Cell count	Numerical example
0	0	0	h_{000}	$h_{000} + h_{010} = 40$
1	0	0	h_{100}	$h_{100} + h_{110} = 60$
0	1	0	h_{010}	
1	1	0	h_{110}	
0	0	1	m_{001}	100
1	0	1	m_{101}	30
0	1	1	m_{011}	20
1	1	1	m_{111}	50

With no dropout at time 1, we require only one indicator variable, R_2, coded 0 for dropout. The data are assumed to be for a single group, with no covariates.

In Table 6.1, the cells denoted m are actually observed but the h-cells are not. We do, however, fully observe the (Y_1, R_2) margin, giving us the m-sums

$$m_{0+0} = h_{000} + h_{010}$$

and

$$m_{1+0} = h_{100} + h_{110}.$$

6.4.2 Marginal formulation

The trivariate distribution of the binary variables (Y_1, Y_2, R_2) can be written directly in polynomial exponential family form, with log likelihood contribution, for the uth subject,

$$\ell = \boldsymbol{\xi}'\mathbf{z} - C(\boldsymbol{\xi})$$

where

$$\mathbf{z} = (y_1, y_2, r_2, y_1y_2, y_1r_2, y_2r_2, y_1y_2r_2)'.$$

Despite the risk of introducing confusion, it is convenient to use subscript 3 to refer to R_2, as it is, the third variable listed, so that the above log likelihood contribution

can be written

$$\begin{aligned}\ell = \quad & \xi_1 y_1 + \xi_2 y_2 + \xi_3 r_2 + \xi_{12} y_1 y_2 + \xi_{13} y_1 r_2 \\ & + \xi_{23} y_2 r_2 + \xi_{123} y_1 y_2 r_2 - C(\boldsymbol{\xi}).\end{aligned}$$

The parameters $\boldsymbol{\xi}$ are zero-conditional log odds ratios; interest is more likely to focus on their marginal equivalents. Thus we might take a straightforward link for marginal saturation as

$$\begin{aligned}\text{logit}\,\mu_t = \lambda_t &= \alpha_t \quad t = 1, 2 \\ \log \text{MOR}_{12} = \lambda_{12} &= \alpha_{12} \\ \text{logit}\,\mu_{R_2} = \lambda_3 &= \alpha_3.\end{aligned}$$

The last term, λ_3, may be considered nuisance but is directly estimable from the observed table. The other odds ratios, λ_{13}, λ_{23} and λ_{123} will characterize the missing data mechanism. If two of these are zero then the missing data mechanism is MAR; if all three are zero, the missing data are MCAR.

6.4.3 Markov chain formulation

A Markov chain formulation is

$$\begin{aligned}p(y_1) &= \exp\{\phi_1 y_1 - C_1(\phi_1)\} \\ p(y_2 \mid y_1 = i) &= \exp\left\{\phi_{2|i} y_2 - C_{2|i}(\phi_{2|i})\right\}, \quad i = 0, 1 \\ p(r_2 \mid y_1 = i, y_2 = j) &= \exp\left\{\phi_{3|ij} r_2 - C_{3|ij}(\phi_{3|ij})\right\}, \quad i = 0, 1;\ j = 0, 1.\end{aligned}$$

In this simple case, the timepoint-wise selection model is a full selection model. A pattern-mixture type model for the probability of Y_2 given dropout pattern could equally well be applied but is omitted here.

In contrast to the previous section, we now assume that interest lies in the response-conditional means, which are the inverse logits of the parameters $\boldsymbol{\phi}$. The parameters of immediate interest, $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta})$, model the dependency between Y_1 and Y_2, and that

between the dropout indicator R_2 and the Y, as follows:

$$\begin{aligned} \text{logit}\,\mu_1 &= \alpha_0 \\ \text{logit}\,\mu_{2|Y_1=y_1} &= \beta_0 + \beta_1 y_1 \\ \text{logit}\,\mu_{3|Y_1=y_1,Y_2=y_2} &= \delta_0 + \delta_1 y_1 + \delta_2 y_2 + \delta_{12} y_1 y_2. \end{aligned}$$

The last of these links is a saturated dropout model as given. Whereas the MCAR assumption sets $\delta_1 = \delta_2 = \delta_{12} = 0$, the MAR model takes $\delta_2 = \delta_{12} = 0$. If (and only if) data are NIGmiss, $\delta_2 \neq 0$.

6.4.4 Estimation and imputation

ML and IML parameter estimation in general

Regarding the table of example data, Table 6.1, as an observation from a multinomial distribution with eight classes, immediately the maximum likelihood (ML) estimates of the cell probabilities are the sample proportions. Provided there are no observed zeros, the ML estimates of the odds ratios, which are one–one functions of the cell probabilities for any of the models discussed above, follow by invariance.

However, the h-cell totals are unknown. Following a joke suggestion in Crowder and Hand, one way round this problem is to fill in the missing cell counts while nobody's looking. Such a filled-in table would now have uniquely and clearly specified ML estimates, which are here called imputed maximum likelihood (IML) estimates (the terms pseudo- and quasi-ML are already in use for other purposes). Below, IML estimates are indicated with a tilde rather than a hat, which is reserved for true ML estimates.

ML estimates for incomplete multinomial counts

Even if two or more of the cell counts in a multinomial sample are unknown, the ML estimates for the *observed* cells are the observed cell counts. For example, consider a trinomial sample (a, b, c), with only c and the sample size, n, observed. Then this is

a fully observed binomial sample $(n-c, c)$, for which obviously $\hat{c} = c_{\text{obs}}$.
Thus for models that are saturated over the observed data, the observed cell estimators must be the observed cell counts themselves. More generally, any fully observed odds ratios are their own ML estimates.

Imputation by pre-specifying certain parameters

In a certain sense a set of true ML estimates actually *determines* the cell counts from which they were obtained, since the estimates could not be ML for any other observed values. Similarly, a set of IML estimates determines, uniquely, a filled-in table. The following are equivalent:

(a) filling-in the h-cell totals (and leaving the m-cells untouched);

(b) specifying IML cell probabilities $\tilde{\boldsymbol{\pi}}$ and sample size n;

(c) specifying IML odds ratios $\tilde{\boldsymbol{\xi}}$ (or $\tilde{\boldsymbol{\lambda}}$ or $\tilde{\boldsymbol{\phi}}$) and sample size n.

Consider for simplicity only the canonical parameters $\boldsymbol{\xi}$; analogous results will hold for any other set of parameters fully specifying the distribution. For the missing-data example introduced above, if we are prepared to specify

(a) *any two* $\tilde{\xi}$ parameters, and

(b) that the imputed table matches the observed table everywhere where known (i.e. the imputed table m-cells and counts match the observed m-cells and counts),

then all the other IML estimates will have been uniquely determined, since parameters cannot be IML for a table they would not determine.
The two parameters arbitrarily specified apparently have a different status to the others, which are truly IML, in the sense that IML estimates are functions of the data plus any constraints sufficient to complete the table. This in turn means that the arbitrarily-specified parameters must now become IML estimates for the odds ratios they themselves pre-specified. Hence the ordinary tilde notation is used for these.
By pre-specifying two parameters for the example data in Table 6.1, the observed cells can be taken as an observation on 6 categories, requiring only 5 parameters in

total (of the canonical 7) to fully specify. Since transformations to marginal $\boldsymbol{\lambda}$ and conditional $\boldsymbol{\phi}$ must be one–one, it is enough to prespecify two of these, although the choice of which two is no longer free if the table is to be saturated. Some details on the admissibility of pre-specified IML parameters are given in Appendix A6.4.4. It is seen there that one has considerable freedom in choosing parameters assuming a MAR or NIGmiss model that is saturated over the observed part of the table. The conceptually simpler MCAR model is harder to specify, unless it should so happen that the observed part of the table is consistent with this assumption.

6.4.5 EM and plug-in likelihood

At the kth step of the EM algorithm, given parameter estimates $\boldsymbol{\xi}^{(k)}$, the E step is to find the expected likelihood

$$\ell^{(k)} = \int \ell(\boldsymbol{\xi} \,|\, \mathbf{y}, \mathbf{r}) f(\mathbf{y}_{\text{miss}} \,|\, \mathbf{y}_{\text{obs}}, \mathbf{r}, \boldsymbol{\xi} = \boldsymbol{\xi}^{(k)}) \, \mathrm{d}\mathbf{y}_{\text{miss}}$$

(Little and Rubin, 1987). These authors note that there is no need to calculate this integral directly in their examples, because for exponential family distributions, the process is equivalent to substituting the expected sufficient-statistic values into the likelihood expressed in terms of the sufficient statistics, s say. The M-step maximizes the likelihood using the imputed values

$$s^{(k)} = \mathrm{E}[s(\mathbf{Y}) \,|\, \mathbf{Y}_{\text{obs}}, \mathbf{R}, \boldsymbol{\xi}^{(k)}].$$

This is equivalent to the following rather more naive approach.

Plugged-in values

Specifying the unknown CORs completes the probability table and so gives us an IML estimate of $\mathrm{E}[Y_2]$, denoted $\tilde{y}_2$, allowing for the (current) estimate of the missing datum. Suppose then that for those observations with Y_2 missing, $\tilde{y}_2$ is substituted,

then the (IML) plug-in likelihood would be

$$\ell(y_1, r_2, \tilde{y}_2) = \sum \boldsymbol{\xi}' \tilde{\mathbf{z}} - C(\boldsymbol{\xi}),$$

where $\tilde{\mathbf{z}} = (y_1, \tilde{y}_2, r_2, y_1\tilde{y}_2, y_1 r_2, \tilde{y}_2 r_2, y_1 \tilde{y}_2 r_2)'$. One idea would be to try various values for the unidentifiable CORs and so obtain a profile IML likelihood.

This does not work, because this profile is maximized for CORs zero or infinity. The likelihood function attains its maximum for a non-stochastic process — here for all the otherwise unassigned probability lying in only two of the four h-cells. In such cases the second-order COR becomes zero or infinite.

The overall process is still stochastic, but in some sense 'less so' for six possible cells than for seven or eight possible cells. More formally, for a multinomial distribution with a model that is cell-saturated where possible and marginally where not, the evaluated likelihood for *any* t-way table is greater than that for *any* s-way table with $s > t$. For example, the binomial model for cells $(n - c, c)$ introduced in Subsection 6.4.4 must have greater evaluated likelihood than the trinomial model allowing for an extra stochastic allocation of the $n - c$ marginal counts to one of cells a or b. That part of the log likelihood sum relating to fully-observed values, for which $\tilde{\mathbf{z}} = \mathbf{z}$, is constant, regardless of any choice of (admissible) unobserved CORs, even at the limit.

The value of the IML likelihood is finite as the parameters tend out of range. In the limit this value is exactly that of the log likelihood evaluated according to the methods of the following sections: the value is -499.8502 for data in Table 6.1. This is because the limiting case assesses the likelihood over only six cells, as do the following methods.

The imputed values for missing data, $\tilde{y}_2$, are non-integer, which is alarming for binary data. However, this is not the cause of the estimation problem, since the evaluated log likelihood is the same for non-integer values directly as for that which would be obtained if we assigned integer values randomized given $\tilde{\mu}_2$; these would, at least asymptotically, give the sample proportions again given the missing CORs.

Equivalence of EM and plug-in

Using the cell-count notation introduced in Table 6.1, for this data the sample log likelihood is

$$\begin{aligned}\ell(\boldsymbol{\xi} \mid h\ldots, m\ldots) \;=\; & (m_{101} + m_{111} + m_{1+0})\xi_1 \\ & +(m_{011} + m_{111} + h_{010} + h_{110})\xi_2 \\ & +(m_{111} + h_{110})\xi_{12} \\ & +m\xi_3 + (m_{101} + m_{111})\xi_{13} \\ & +(m_{011} + m_{111})\xi_{23} + m_{111}\xi_{123} \\ & -nC(\boldsymbol{\xi})\end{aligned}$$

where m is the sum of observed m-cells and n is the sample size. The given counts are sufficient, if the h-cells are known. As a check on the validity of this expression, note that if it is differentiated once with respect to $\boldsymbol{\xi}$ and set equal to vector zero, we get the familiar ML result

$$\text{counts} - n\hat{\mu} = 0.$$

Note that the above holds only for data in a single group: that is, there are no covariates in the model. For grouped data, similar results would hold within groups. With continuous covariates this breaks down, since the sufficient statistics become the data themselves.

Using $\tilde{y}$, the IML estimates, *always* gives the same observed m-cell counts, and so is equivalent to specifying sufficient statistics $s(\mathbf{Y})$, based on the current estimate of the missing mechanism. Thus, the IML plug-in values are equivalent to the EM algorithm here.

Apart from the likelihood attaining a maximum for values outside the parameter space, the EM algorithm fails to converge here because of the complete lack of information with respect to the h-cell counts required. Speed of convergence depends on the ratio missing/complete information, and the numerator here is zero unless a restricting model such as MAR is imposed.

6.4.6 Marginalized likelihood

Marginal formulation

Rather than look at the E-step (or plug-in) likelihood above, we can look at the observed likelihood. That is, when Y_2 is missing, the likelihood is taken from the joint marginal distribution of (Y_1, R_2). For this example the joint marginal distribution is

$$\begin{aligned} p(y_1, r_2) &= p(y_1, r_2, y_2 = 0) + p(y_1, r_2, y_2 = 1) \\ &= \exp\{\xi_1 y_1 + \xi_3 r_2 + \xi_{13} y_1 r_2 - C(\boldsymbol{\xi})\} \\ &\quad + \exp\{\xi_1 y_1 + \xi_3 r_2 + \xi_{13} y_1 r_2 + \xi_2 + \xi_{12} y_1 + \xi_{23} r_2 + \xi_{123} y_1 r_2 - C(\boldsymbol{\xi})\} \end{aligned}$$

in terms of the $\boldsymbol{\xi}$ parameters of the full trivariate distribution. This may also be written

$$p(y_1, r_2) = \exp\left\{\xi_1^{13} y_1 + \xi_3^{13} r_2 + \xi_{13}^{13} y_1 r_2 - C(\boldsymbol{\xi}^{13})\right\},$$

where $\boldsymbol{\xi}^{13}$ are the canonical parameters of the subdistribution for Y_1 and R_2. The transform from $\boldsymbol{\xi}$ to $\boldsymbol{\xi}^{13}$ is not straightforward, and may not be useful. It is the (Y_1, Y_2) marginal distribution that is likely to be of primary interest.

When Y_2 is missing, $R_2 = 0$, and the above simplifies:

$$\begin{aligned} p(Y_1 = y_1,\, R_2 = 0) &= \exp\{\xi_1 y_1 - C(\boldsymbol{\xi})\} \\ &\quad + \exp\{\xi_1 y_1 + \xi_2 + \xi_{12} y_1 - C(\boldsymbol{\xi})\}, \end{aligned}$$

which contributes

$$\xi_1 y_1 - C(\boldsymbol{\xi}) + \log[1 + \exp(\xi_2 + \xi_{12} y_1)]$$

to the log likelihood, while the frequency function in the absence of dropout is $p(y_1, y_2, R_2 = 1)$. Hence the observed log likelihood is

$$\begin{aligned} \ell(\boldsymbol{\xi} \mid m...) &= (m_{101} + m_{111} + m_{1+0})\xi_1 \\ &\quad + (m_{011} + m_{111})\xi_2 \end{aligned}$$

$$
\begin{aligned}
&+m_{111}\xi_{12} \\
&+m\xi_3 + (m_{101} + m_{111})\xi_{13} \\
&+(m_{011} + m_{111})\xi_{23} + m_{111}\xi_{123} \\
&-nC(\boldsymbol{\xi}) \\
&+m_{1+0}[\log(1 + \exp\{\xi_2 + \xi_{12}\})] \\
&+m_{0+0}[\log(1 + \exp\{\xi_2\})]
\end{aligned}
$$

Differentiating with respect to $\boldsymbol{\xi}$ gives the score equations

$$
\begin{aligned}
m_{1++} &= n\nu_1 \\
m_{+11} + \left(\frac{m_{1+0}}{1+\xi_2+\xi_{12}} + \frac{m_{0+0}}{1+\xi_2}\right)e^{\xi_2} &= n\nu_2 \\
m_{++1} &= n\nu_3 \\
m_{111} + \frac{m_{1+0}}{1+\xi_2+\xi_{12}}e^{\xi_{12}} &= n\nu_{12} \\
m_{+11} &= n\nu_{23} \\
m_{1+1} &= n\nu_{13} \\
m_{111} &= n\nu_{123}
\end{aligned}
$$

where $+$ denotes summation over an index and $\boldsymbol{\nu}$ are marginal expectations about the origin.

Clearly this is overparametrized, since ξ_2 and ξ_{12} must be specified to obtain marginal expectations ν_2 and ν_{12}, or vice versa, and none of these are directly identifiable from the observed data. One might consider proceeding by profile likelihood — discussed shortly.

Markov chain approach

When Y_2 is missing the likelihood contribution is from

$$
p(y_1)\sum_{Y_2} p(r_2 \,|\, y_1, y_2)p(y_2 \,|\, y_1)
$$

which is just

$$p(y_1)p(r_2 \mid y_1) = p(y_1, r_2)$$

exactly as above, except for the difference in parametrization.

The marginalized log likelihood is

$$\begin{aligned}
\ell(\phi \mid m\ldots) = & \\
& m_{111}\left[\phi_1 - C_1(\phi_1) + \phi_{2|1} - C_{2|1}(\phi_{2|1}) + \phi_{3|11} - C_{3|11}(\phi_{3|11})\right] \\
+ \; & m_{011}\left[-C_1(\phi_1) + \phi_{2|0} - C_{2|0}(\phi_{2|0}) + \phi_{3|01} - C_{3|01}(\phi_{3|01})\right] \\
+ \; & m_{101}\left[\phi_1 - C_1(\phi_1) - C_{2|1}(\phi_{2|1}) + \phi_{3|10} - C_{3|10}(\phi_{3|10})\right] \\
+ \; & m_{001}\left[-C_1(\phi_1) - C_{2|0}(\phi_{2|0}) + \phi_{3|00} - C_{3|00}(\phi_{3|00})\right] \\
+ \; & m_{0+0}\left[-C_1(\phi_1) + \log\left\{\exp\left(-C_{2|0}(\phi_{2|0}) - C_{3|00}(\phi_{3|00})\right)\right.\right. \\
& \qquad \left.\left. + \exp\left(\phi_{2|0} - C_{2|0}(\phi_{2|0}) - C_{3|01}(\phi_{3|01})\right)\right\}\right] \\
+ \; & m_{0+0}\left[\phi_1 - C_1(\phi_1) + \log\left\{\exp\left(-C_{2|1}(\phi_{2|1}) - C_{3|10}(\phi_{3|10})\right)\right.\right. \\
& \qquad \left.\left. + \exp\left(\phi_{2|1} - C_{2|1}(\phi_{2|1}) - C_{3|11}(\phi_{3|11})\right)\right\}\right]
\end{aligned}$$

This is the form evaluated for the following attempt at analysing the profile likelihood.

Failure of profile IML likelihood

Suppose we proceed as attempted earlier: that is, specify IML parameters $\tilde{\xi}_1$ and $\tilde{\xi}_2$ arbitrarily, insist that the predicted table match the observed where known, and thus determine all the other IML parameters. Alternatively we might specify any two $\phi_{3\,|\,ij}$ to give a set of IML conditional parameters.

The terminology profile IML likelihood represents the evaluation of the observed (marginalized) likelihood at the IML parameter estimate set obtained from the choice of the first two of this set.

No matter what values are assumed for ξ_1 and ξ_2, at least, provided they are admissible, the IML log likelihood has the same value (-499.85 for the example data). In other words, every possible completion of the table is equally likely. This result is shown more generally in the following section.

For the numerical example, I have observed that even quite large departures in numerical value from a set of valid IML parameters result in only a small decrease in evaluated likelihood. Non-saturated (hence non-IML, in the strict sense) models fit well or at least not substantially worse based on a likelihood ratio test. This is not a general claim; clearly sensitivity to departures from saturation is a function of the proportion of missing data, 33% in this example.

This observation prompts the idea that, in ascertaining the ratio of the likelihood of a non-saturated model to that of a saturated model, it suffices to assess the latter by the easily fitted MAR model, even if the simplified model is not strictly a submodel of this.

6.5 On the non-estimability of missing data

As already mentioned, a sample has no information on whether or not missing data are informative. In Section 6.5.1 a simple, but general, demonstration is given that in the presence of dropout the likelihood, the product of terms as in equation (6.3), does not have a unique maximum, for any random variables $\mathbf{Y}$. In Section 6.5.2 I reinforce this result with specific reference to the examples given at the end of Sections 6.3.1 and 6.3.2, showing that both models must be overparametrized with respect to estimation within incomplete observations. I conclude with further discussion of the issues raised.

6.5.1 General proof

Consider the pattern-mixture factorization of the joint probability density (or frequency) function given in equation (6.2). The right-hand term can be fully or partially expanded to give the equivalent forms (6.17), (6.18) and (6.19):

$$f(\mathbf{y},\mathbf{r}) = f(\mathbf{r})f(y_1 \mid \mathbf{r})f(y_2 \mid h_2,\mathbf{r})\cdots f(y_T \mid \mathbf{h}_T,\mathbf{r}), \tag{6.17}$$

which gives univariate conditional functions for given missing-data pattern $\mathbf{r}$;

$$f(\mathbf{y},\mathbf{r}) = f(\mathbf{r})f(y_1,\ldots,y_{t-1} \mid \mathbf{r})f(y_t,\ldots,y_T \mid \mathbf{h}_t,\mathbf{r}), \tag{6.18}$$

where the joint densities are obtained as the product of two or more conditional densities; and

$$\begin{aligned} f(\mathbf{y},\mathbf{r}) &= f(\mathbf{r})f(y_1,\ldots,y_{t-1}\,|\,\mathbf{r})f(y_t\,|\,\mathbf{h}_t,\mathbf{r}) \\ &\quad \times f(y_{t+1},\ldots,y_T\,|\,\mathbf{h}_{t+1},\mathbf{r}). \end{aligned} \tag{6.19}$$

where there is a specific time t of interest.

For a subject dropping out at time t, the observed likelihood contribution, according to (6.3), substituting factorization (6.18), is the product of contributions of the form

$$\begin{aligned} &\sum_{\text{missing } \mathbf{Y}} f(\mathbf{y},\mathbf{r}) = \\ &\quad f(\mathbf{r})f(y_1,\ldots,y_{t-1}\,|\,\mathbf{r}) \sum_{Y_t,\ldots,Y_T} f(y_t,\ldots,y_T\,|\,y_1,\ldots,y_{t-1},\mathbf{r}) \end{aligned} \tag{6.20}$$

with integration replacing summation if the variables are continuous. The first two terms in the expansion (6.18) are independent of Y_t and subsequent observations, and so can be written outside the summation. Then the right-hand summation in (6.20) is equivalent to unity because it is a sum over the range of a probability function.

Thus, the likelihood for subjects dropping out at time t reduces to contributions from

$$f(\mathbf{r})f(y_1,\ldots,y_{t-1}\,|\,\mathbf{r})$$

alone. This part of the joint density, then, is estimable; the last term in (6.18), i.e. $f(y_t,\ldots,y_T\,|\,\mathbf{h}_t,\mathbf{r})$, is not.

Parameters occurring in only the dropout and post-dropout conditional densities, that is in $f(y_t,\ldots,y_T\,|\,y_1,\ldots,y_{t-1},\mathbf{r})$, do not contribute to the score function.

Of course, the dropout time is not the same for all subjects, which might lead one to believe that, assuming there is at least one subject with no dropout at time T, all parameters would contribute to the overall likelihood. But, such contributions occur only over the *fully-observed* part of the data.

Suppose that the conditional density of Y_t (given history and given $r_i = 1$ for $i < t$ and

$r_i = 0$ thereafter) depends on some parameter(s) $\phi_{t|d_t}$, where d_t is an abbreviation for dropout at time t and history h_t. Now the conditional density at time t given *later* dropout depends on some other parameter(s) $\phi_{t|d_{t'}}$. From the unquestioned validity of the pattern-mixture decomposition, the densities for two different patterns, here dropout times, cannot in general be taken to be equal even if history up to time t or even t' are the same. Hence, they do not in general have the same parameters:

$$\phi_{t|d_t} \not\equiv \phi_{t|d_{t'}} \quad t < t'.$$

Inestimability now follows from noting that there is a nonzero contribution to the score only for timepoints before dropout; specifically for $s \geq t$, for dropout at time t, none of the $\phi_{s|d_t}$ are identifiable.

While almost any set of arbitrary constraints can be imposed to give a model that is not overparametrized, none of the parameters relating to portions of the data that are missing can affect the observed likelihood. Thus, the likelihood ratio between any two models which differ only in their specification of the unidentifiable parameters is unity, and so *none of the missing-data parameters are directly identifiable by maximum likelihood, and any set of constraints imposed to make them identifiable is untestable by likelihood-based inference.*

As a consequence, the model fitting all the observed values and marginal covariances and assuming that missing values come from exactly the same distribution as those observed (the 'saturated MAR model') yields the maximum likelihood; this holds only in the strict sense that no other model has *greater* likelihood, while infinitely many others have the *same* likelihood.

Note too that such a MAR fit is always attainable; we need only constrain

$$\phi_{t|d_t} = \phi_{t|d_{t'>t}},$$

that is, set the conditional parameter for dropout at time t equal to that estimated for timepoint t for subjects with no dropout by time t. We are always free to do this.

Since all other parametrizations of the likelihood are monotone transformations of ϕ, this result holds in full generality.

6.5.2 A specific example

Consider the model in Section 6.3.2, page 218. This model is not saturated for the observed likelihood, but despite the constraints the model remains overparametrized. Specifically, we can only estimate the two contrasts $\alpha'_2 + \delta'_{22}$ and $\alpha'_3 + \delta'_{22}$, which is insufficient to give a separate estimate of δ'_{22}. The underlying reason for this is given above, but for illustration, for those subjects who drop out later than time 2, the parameter that contributes to the score is

$$\phi_{Y_2 \mid Y_1=y_1, R_2=1} = \alpha'_2 + \alpha'_{21} y_1 + \delta'_{22},$$

while if there is dropout at time two (so that $r_2 = 0$), the canonical parameter

$$\phi_{Y_2 \mid Y_1=y_1, R_2=0} = \alpha'_2 + \alpha'_{21} y_1,$$

which would be needed to distinguish α'_2 from δ'_{22}, never contributes to the score. The same obviously holds at time 3, and for any outcome vector length.

Thus, we see that even under heavy constraints it is impossible to estimate, by maximum likelihood, the key parameter for informative dropout, here δ'_{22} (missing data is NINFdrop iff this is zero — see the equivalent definition for NIGmiss, equation 6.6). Similarly, in the formulation at the end of Section 6.3.1, page 216, we cannot estimate δ_{22}, which has the same interpretation for informative dropout as δ'_{22}. Less obviously for the selection-type formulation, this parameter is unidentifiable without further constraints; but by monotonicity of ML estimates, iff it were identifiable, then so would be δ'_{22}.

Again maximum likelihood is not a tool that can be used to assess whether missing data are, or are not, ignorable, nor can maximum likelihood be used to assess the quantitative effect of a wrong assumption.

6.5.3 Discussion

We cannot assess the pattern of missing data if we assume it is informatively missing (Little and Rubin, 1987). The NIGmiss assumption is that the missing data have a different distribution to that observed, and it is not surprising to find that the observed likelihood has nothing to contribute to the assessment in terms of estimation. Nevertheless if enough constraints are placed on the model for the observed data, estimates for the missing-data parameters are imputed since there are a fixed number of degrees of freedom. The selection model of Diggle and Kenward (1994), for example, assumes a simple form for the correlation structure, which enables a parameter for informative missingness, the equivalent of δ_{22} above, to be estimated. That approach fails for bivariate data and is unproven for categorical data; the unidentifiable model of Section 6.3.1 follows the same approach, except for its parametrization of dependency. Even if a model is sufficiently specified to estimate the equivalent of δ_{22}, it is deceptive to call this a 'maximum likelihood estimate of informative missingness', the implication being that such an estimate is determined by the data. Rather, it is the untestable assumptions that determine the parameter estimate.

6.6 Example: the Liverpool CHITC study

6.6.1 Data and dropout pattern

The Liverpool Continuing Health in the Community project was created to study the incidence and prevalence of dementia in the elderly. An original sample of 1070 people aged 65 or over were assessed by interview, questionnaire, cognitive tests and blood pressure measurements for dementia, with further assessments planned after three and six years. As would be expected in this population, there was considerable attrition due to subject death. Moreover there was a similar rate of refusal to complete the study. In all, of the 1063 non-excluded subjects at year 0, only 696 remained in the study by year 3 (179 had died, 188 refused to interview or were lost to follow-up); by year 6 the sample size reduced to 437 people (148 more had died; 111 more were

otherwise missing). Clearly a complete-case analysis of the 437 who completed the study would be inadequate.

The scores and measurements for these people were filtered through a computer diagnosis program, AGECAT (Copeland *et al.*, 1986), which outputs 8 ordered categorical scales known as "syndrome clusters", in addition to an overall diagnosis. In this example only the so-called organic cluster is considered as outcome variable; depression score (at baseline, thus time-invariant) is considered as a covariate. The original 6-point scale for organic cluster reduces naturally to a ternary outcome: original scores of 3 or above are *cases*, scores of 1 or 2 are called *subcases*, and scores of zero denote absence of syndrome. Despite the sample size, the data are too sparse to analyse on the finer scale. For example, there were no outcomes greater than 3 at year 0. In keeping with other analyses (e.g. Copeland *et al.*, 1992) only the three-point scale is considered further, relabelled from 0 (no syndrome) to 2 (case). Depression was only available as a binary variable (case or not at time 0).

The other available covariates of interest were age and sex.

The outcomes are shown in Table 6.2. This is presented in pattern-mixture style, with class 0, 1, and 2 totals for each pattern appearing before the dividing rules; the heavy rules delimit the three dropout patterns (outcome vector sizes). A further split into the two types of dropout is too cumbersome to present and the data may be conveniently represented in more conventional tables. For example, the association between year 0 organic score and year 3 dropout type is well illustrated by the following table:

	dropout		
organic cluster	0	1	2
0	620	126	154
1	41	35	19
2	35	18	15

where 0 indicates no dropout, 1 denotes death, 2 denotes otherwise missing. There are two important additions to the previous discussion of dropout indicators. Firstly, the de facto standard for binary indicators is 0 for dropout, but for computational convenience the opposite convention is adopted here: zero for no dropout, nonzero for

Table 6.2: CHITC outcome; 0 = no syndrome, 1 = subcase, 2 = case.

0	398	0	380	0	344
				1	18
				2	18
		1	8	0	3
				1	4
				2	1
		2	10	0	7
				1	1
				2	2
1	22	0	12	0	6
				1	1
				2	5
		1	6	0	1
				1	2
				2	3
		2	4	0	1
				1	1
				2	2
2	17	0	15	0	8
				1	3
				2	4
		1	1	0	0
				1	0
				2	1
		2	1	0	1
				1	0
				2	0
0	222	0	197		
		1	17		
		2	8		
1	19	0	6		
		1	10		
		2	3		
2	18	0	9		
		1	5		
		2	4		
0	280				
1	54				
2	33				

some type of dropout. Second, the use of a ternary dropout indicator is novel to this presentation.

A ternary indicator is called for here because it is important to maintain a distinction between dropouts due to death and dropouts due to refusal; the latter are believed by the researchers to be strongly informative, in that more refusees would be diagnosed as dementia cases than would be the case for participants. However, it must be recognised that the degree of informative missingness cannot be estimated here, as in general (Section 6.5), and there has been no follow-up study on a sample of withdrawals. Nevertheless distinguishing the two dropout types enables us to directly address important clinical questions such as the relationship between recorded dementia and subsequent death and/or subsequent withdrawal while fitting a single model. The relationship between dementia and subsequent *binary* dropout is of no direct interest.

A cross-tabulation of dropout status at years 3 and 6,

	dropout 6		
dropout 3	0	1	2
0	437	148	111
1	0	0	179
2	8	34	146

reveals that the data set includes missing patterns other than monotone; 8 subjects refused at year 3 but participated at year 6. This is such a small number that I simplify by treating these as monotone missing by year 3; there was no clear pattern to the response profile to suggest any appreciable bias would be introduced by this assumption. Also, 34 subjects were missing (refused) at time 2 but known to have died between years 3 and 6. Treating these as ordinary type-2 dropouts (which I do) might bias results for a lag-2 dropout model; a more comprehensive study than the current example should take this into account.

6.6.2 Analysis

Here I shall only consider the incidence of dementia, rather than its prevalence, because the former can conveniently be studied using Markov chain models. This approach facilitates the study of dropout models. Prevalence is better studied by a marginal model, which is how the initial sample at year 0 was studied by Copeland *et al.* (1987). The estimates are inefficient compared with those of a marginal model that takes into account the data from subsequent timepoints.

Data were analysed using a custom written program, `hcfit` (see Appendix A6.6). This program implements the timepoint-wise pattern-mixture model of Section 6.3.2, specifically equation (6.14), but the selection-type model (6.14) is also available although it is not analysed here. The `hcfit` program enables data and dropout models to be fitted simultaneously, assuming both are of interest. The models can also be fitted separately; indeed if data are not assumed informatively missing, only the data model need be fitted. The program-evaluated log likelihood is partitioned into contributions from each model, ℓ_Y and ℓ_R, to facilitate analysis of deviance when both data and dropout models are modified in the same step and the missing mechanism is assumed at worst MAR. If dropout is informative, the log likelihood partition is a mere artifact and only the sum of ℓ_Y and ℓ_R should be considered.

Cumulative logit models as described in Chapter 4 were fitted to the ordered data, $\mathbf{Y}$, while the nominal data, $\mathbf{R}$, were fitted via a canonical link to the multinomial logits, using $R_t = 0$ as baseline.

To study the effects of covariates a rule of thumb is to fit them to a intercept-unconstrained model. However, here 16 parameters would be included in the data model alone without considering potential interaction factors for the year 6 model, and it is extremely hard to find starting values that lead to convergence. I found that forward selection (from common intercept, to separate intercepts, to Markov lag one, etc.) was able to overcome this problem.

The modelling strategy adopted here is as follows:

1. Select a model for the intercept parameters (the 'intercept model' or 'depen-

dence model'). This features baseline cutoffs and history dependence, but excludes explanatory variables $\mathbf{x}$. I use analysis of deviance to compromise between adequacy of fit and reducing the number of parameters.

2. Forward selection of explanatory variables added to the intercept model. Each variable selection step consists of several substeps:

   - add each remaining effect in turn to the current model using the fullest possible submodel (i.e. separate parameters for each outcome value for each timepoint);
   - choose the effect that increases the log likelihood most significantly (if none, then stop);
   - simplify the model for the effect chosen;
   - repeat.

3. Reconsider simplifications of the intercept model now that covariates are included in the full model, and perhaps consider adding interactions between explanatory variables and previous outcomes.

It is important to fit full models for effects before simplifying them in step 2. A very simple model, such as a single parameter added to all links, might fail to detect an important effect altogether.

The overall strategy advocated, which is essentially block stepwise selection, does not consider every possible model and can be criticised on such grounds. However, all-subset regression is totally impractical for the CHITC data.

For the intercept ($\boldsymbol{\alpha}$-parameter) model, that is, the model ignoring covariates, I used forward stepwise selection from the simplest model that has any meaningful interpretation:

$$\begin{aligned} \eta_{Y_1} &= \alpha_{1j} \\ \eta_{R_2} = \eta_{R_3 i} &= \delta_i \\ \eta_{Y_2} = \eta_{Y_3 j} &= \alpha_{2j} \end{aligned}$$

where η are the linear predictors for a timepoint-wise pattern-mixture model, for conditional logits of the subscripted variables, j is 0 or 1 for the cumulative logit links (fitted to $\mathbf{Y}$), and i is 1 or 2 for the multinomial logits (fitted to $\mathbf{R}$). The history dependencies are suppressed from the η notation for simplicity (and indeed there is no conditioning on history in this model). The time 1 model is included primarily because `hcfit` expects one, though the baseline is always of some interest. This model should not include parameters in common with that for other timepoints, as this might bias the dependency model.

Labelled output from the `hcfit` program, showing parameter estimates and standard errors and evaluated log likelihoods for the models fitted here, is given in Appendix 7.5. Numbers in raised square brackets cross-reference the models; the above is [1]. Note that years 0, 3 and 6 become timepoints 1, 2 and 3 so that Y_3 is the outcome at year 6, not 3.

The above model is null in that it assumes no change over time. Another contender for the title of 'null model' is the independence model,[2]

$$\begin{aligned} \eta_{Y_1} &= \alpha_{1j} \\ \eta_{R_2} &= \delta_{2i} \\ \eta_{Y_2} &= \alpha_{2j} \\ \eta_{R_3} &= \delta_{3i} \\ \eta_{Y_3} &= \alpha_{3j} \end{aligned}$$

This has similar log likelihood to the null model, but has four more parameters. The time-1 and time-3 data-model parameters are very similar, suggesting either that the time-1 marginal analysis is adequate and that time-2 values are suspect, or that data are indeed missing informatively. Nevertheless, since we cannot test for this (Section 6.5), we proceed, for simplicity of illustration, assuming the data are at worst MAR. The first model of any intrinsic interest incorporates dependency on previous values y_{t-1}, that is, a lag-1 Markov model. History is assumed to have a linear effect for the

present and the time-2 and time-3 models are assumed to be the same, depending on y_{t-1}. That is, this simplest lag-1 Y_t model ($t = 2, 3$) is[3]

$$\eta_{Y_t j} = \alpha_{2j} + \alpha_h y_{t-1},$$

which assumes a common model across timepoints excepting the Y_1 model. This is a proportional odds model; the cutoff points α_{2j} are shifted equally by $\alpha_h y_{t-1}$, which corresponds to assuming the same distribution for each history except for a logit-linear change of location parameter. There is no significant improvement when the common-distribution assumption is dropped: that is, if two parameters α_{hj}, $j = 1, 2$, replace of α_h,[4] the deviance change is only 2.2 on 1 degree of freedom.

History can be fitted as a factor rather than as a scalar. This does more than depart from linearity as it avoids demanding a monotonic relationship; since here the scales are ordered, a non-monotone relationship would cast serious doubts on underlying assumptions. In practice, here we obtain a monotone fit to the factor model:[5]

$$\eta_{Y_t j} = \alpha_{2j} + \boldsymbol{\alpha}'_{h:}\mathbf{y}_{(t-1):}$$

where

$$\boldsymbol{\alpha}_{h:} = (0, \alpha_{h1}, \alpha_{h2})'$$

treats history 0 as baseline, GLIM style; the colon notation denoting ordered categories was introduced in Chapter 4. This is a significant improvement on the linear effect model.

Previously we saw a non-significant improvement when the common-distribution assumption was dropped. However for the nonlinear effect model there is a clear improvement when fitting[6]

$$\eta_{Y_t j} = \alpha_{2j} + \boldsymbol{\alpha}'_{hj:}\mathbf{y}_{(t-1):}$$

with separate parameters for $j = 0, 1$. This model is almost intercept unconstrained as far as the lag-1 relationship is concerned. If the α_{2j} were replaced by timepoint-varying

α_{tj} and the α_{hjk} by α_{thjk} we would have variationally independent $\boldsymbol{\phi}$. Equivalent to the first of these steps but more readily interpretable is the addition of a period effect α_{pj} to the time-3 model:[7]

$$\begin{aligned}\eta_{Y_2 j} &= \alpha_{2j} + \boldsymbol{\alpha}'_{hj:}\mathbf{y}_{(t-1):} \\ \eta_{Y_3 j} &= \alpha_{2j} + \alpha_{pj} + \boldsymbol{\alpha}'_{hj:}\mathbf{y}_{(t-1):}\end{aligned}$$

This gives a significant improvement over the common-intercepts model above, as perhaps suggested by the independence model earlier. The final step, to lag-1, non-stationary saturation,[8] offers no further improvement. Thus the lag-1 dependence structure is adequately modelled as stationary. This is encouraging for prediction, although the time-dependent baselines are a problem and this is addressed below. Also, we obtain time-invariant estimates of the effect of 3-year previous history on incidence, one of the main objectives of the study.

A lag-2 dependence model is less obviously desirable for a 3-wave study since it is based on only one pair of observations, and thus cannot be verified to be stationary. Moreover, when there is as much potentially informative dropout as here, the complete-case sample might not represent the study population. Nevertheless the effect of adding a simple lag-2 effect to the Y_3 model,[9]

$$\eta_{Y_3 j} = \alpha_{2j} + \alpha_{pj} + \boldsymbol{\alpha}'_{hj:}\mathbf{y}_{(t-1):} + \boldsymbol{\alpha}'_{hh:}\mathbf{y}_{(t-2):}$$

was highly significant and so could not be ignored (deviance 25.6 on 2 d.f.). Moreover, the period effect, α_{pj}, can be dropped from this model with almost no change in the log likelihood, and this gives a more appealing overall model[10] than one with different baselines for times 2 and 3. Substituting a non-proportional-odds parameter set $\boldsymbol{\alpha}_{hhj:}$ for $\boldsymbol{\alpha}_{hh:}$ gave no significant improvement.[11] Interactions between y_{t-1} and y_{t-2} were not then considered.

Estimates for the chosen intercept model, for the data model, are summarized in Table 6.3. Interpretation of these parameters follows in Section 6.6.3.

All the potential explanatory variables were found to be significant when added to this model. In general a separate parameter for each effect for each timepoint is required, as discussed on page 238ff., since the effects are adjusted for the time-varying circumstances of the intercept model. The full explanatory model could be reduced (without significant loss) to have a simple logit-linear effect for age (common across timepoints and proportional-odds), with depression and sex also significant but only for the time-2 model (given age). The individual fits for the stepwise selection procedure used are omitted from the Appendix for brevity; only the final model is shown there (and summarized in Table 6.3):[12]

$$\begin{aligned}
\eta_{Y_2 j} &= \alpha_{2j} + \boldsymbol{\alpha}'_{hj:}\mathbf{y}_{(t-1):} + \boldsymbol{\alpha}'_{hh:}\mathbf{y}_{(t-2):} + \beta_a(\text{age}) + \beta_{d2}(\text{depress}) + \beta_{s2}(\text{sex}) \\
\eta_{Y_3 j} &= \alpha_{2j} + \boldsymbol{\alpha}'_{hj:}\mathbf{y}_{(t-1):} + \boldsymbol{\alpha}'_{hh:}\mathbf{y}_{(t-2):} + \beta_a(\text{age})
\end{aligned}$$

Here 'age' is actual age minus 65 dived by 10 (to avoid magnitude problems for initial estimates of the logits), 'depress' is depression cluster at year 0 (0 if a clinical case, else 1; time-varying score was not available), and sex is 0 for females, 1 for males.

Once explanatory variables are incorporated into the overall model it might be possible to simplify the intercept model. Although not necessary for this analysis, if the chosen intercept model had been the period-effect model, then one would consider dropping the period effect from any model incorporating the simultaneously-varying covariate age. For these data, this approach gives a significantly worse fit than a model with both age and period (`hcfit` output omitted).

While fitting the successive data models, the dropout model can be developed simultaneously. The additions to the null model for $\mathbf{R}$ closely parallel those for the data model; intercept parameters (cutoffs δ_1 and δ_2 and history dependencies $\boldsymbol{\delta}_{hi:}$) could all be taken to be common for both timepoints, but one could not assume common parameters for the two types of dropout (as expected). Successive significant improvements were found for the addition of linear lag 1 (i.e. probability of dropout depending on previous Y outcome), then nonlinear lag, $\boldsymbol{\delta}_{hi:}$, analogous to $\boldsymbol{\alpha}_{hj:}$ in the data model.

Table 6.3: Estimates for the intercept (1) and final (2) models selected, for both data model (parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$) and dropout model (parameters $\boldsymbol{\delta}$). Standard errors (SE) of the estimates are those obtained from the information matrix. Approximate and sandwich estimates are given in Appendix A6.6.

Parameter	Estimate (1)	SE (1)	Estimate (2)	SE (2)
α_{10}	1.71	0.085	1.71	0.085
α_{11}	2.68	0.125	2.68	0.125
α_{20}	2.43	0.115	2.58	0.136
α_{21}	3.22	0.162	3.39	0.169
α_{h01}	-2.66	0.300	-2.46	0.313
α_{h02}	-1.51	0.324	-1.40	0.339
α_{h11}	-1.50	0.386	-1.37	0.404
α_{h12}	-1.42	0.377	-1.35	0.410
α_{hh1}	-2.22	0.400	-1.85	0.464
α_{hh2}	-2.00	0.477	-1.62	0.516
β_a			-0.62	0.157
β_{d2}			-0.92	0.270
β_{s2}			0.34	0.179
δ_1	-1.83	0.071	-2.05	0.113
δ_2	-1.33	0.065	-1.29	0.085
δ_{h11}	1.33	0.208	1.13	0.220
δ_{h12}	0.93	0.239	0.67	0.253
δ_{h21}	1.01	0.226	0.96	0.227
δ_{h22}	0.33	0.289	0.31	0.290
δ_{a1}			0.96	0.103
δ_{s1}			0.71	0.135
δ_{s2}			-0.40	0.142

There was no significant lag-2 effect.

For the explanatory model for dropout, after reduction, there is an effect for age on death, δ_{a1}, though not on refusal/missing, and a sex effect for both death, δ_{s1}, and refusal, δ_{s2}; all these are the same, i.e. not significantly different, for dropout times 2 and 3. Estimates and their SEs are given in Appendix A6.6, with those for the final model[12] in Table 6.3.

6.6.3 Comments

The chosen explanatory models indicate that age increases the probability of organic dementia and simultaneously increases the probability of death (this latter is hardly a surprising finding). Also previous history of dementia is strongly associated with increased probability of both dementia and death. Dementia subcases are associated with refusal to participate further (parameter δ_{h21} is significant) but, surprisingly, full cases are not significantly less likely to participate. Given the other results this may be simply because these people tend to die before having a chance to refuse.

It is hard to interpret the significance of depression score and sex at time 2 only, for the data model. The first can perhaps be explained by supposing depression has a lag 1 but not lag 2 effect; if we had the time-varying depression score our estimates might change. The sex effect may be confounded by age effect. Otherwise the estimates suggest that males are less likely to be cases at time 2, but according to the dropout model are more likely to die then (and at time 3). Also, the significance of δ_{s2} suggests that males are less likely to refuse. A plausible explanation for these joint findings is that females who refuse are less likely to be cases than those who do not refuse, that is, that dropout is informative. This hypothesis is however not testable by the available data.

Model selection is not an easy task here. There are far too many possible models for all-subsets regression to be applied blindly, but there is surely great danger of missing a good model if selection is not in some way systematic. Each model may take a considerable time to interpret as there are so many potential inter-relationships, but it is necessary to interpret each of them carefully when selecting on more intuitive

grounds; when I first started selection, I frequently found that adding an effect in a particular form would then suggest an unsuspected, non-nested simplification. The block stepwise selection algorithm given above was developed as a result of many earlier frustrated attempts. Even so, that algorithm might fail to find the best model, and still takes a considerable amount of time. Further guidelines for model selection are needed.

Chapter 7

Summary and conclusions

In longitudinal data analysis, we are concerned with an outcome variable, Y_t, measured at each of T timepoints, on each of N subjects or units. In this thesis, we have restricted attention to categorical outcomes, and to the case when the timepoints are fixed (and the same for all subjects). Then, writing

$$\mathbf{Y} = (Y_1, Y_2, \ldots, Y_T)',$$

the joint probability function, $f(\mathbf{y})$, is a member of the polynomial exponential family, with canonical form

$$f(\mathbf{y}) = \exp\{\boldsymbol{\xi}'\mathbf{z} - C(\boldsymbol{\xi})\}, \tag{7.1}$$

where

$$\mathbf{z} = (y_1, \ldots y_T,\ y_1y_2, \ldots, y_{T-1}y_T,\ \ldots,\ y_1y_2\cdots y_T)'.$$

In Chapter 2, reviewed in Section 7.1, we considered *multivariate models* for $f(\mathbf{y})$, of two main types, as now outlined. When a multivariate model directly parametrizes $\mathrm{E}[Y_t]$, the mean at each timepoint, it is called a *marginal model.* If the model further parametrizes all higher-order marginal expectations, from pairwise $\mathrm{E}[Y_iY_j]$ to full-order $\mathrm{E}[Y_1Y_2\cdots Y_T]$, it is called here a *fully marginal model.* This nomenclature extends to the case when marginal odds ratios, rather than expectations, are directly modelled. Fully marginal models are reviewed in Section 7.1.1.

If only first-order expectations are modelled marginally, with the remaining specification of the multivariate probability function being other than through marginal parameters, the model is here called *first-order marginal.* By extension, a *second-order marginal* model parametrizes each $\mathrm{E}[Y_t]$ and $\mathrm{E}[Y_iY_j]$ marginally, but completes specification of the distribution other than marginally. Collectively such models are called *partially marginal.* The only models considered here use a canonical link for the high-order interaction parameters. These and non-marginal models that use a canonical link for all parameters are reviewed in Section 7.1.2.

In another approach, the joint probability function may be factorized

$$f(\mathbf{y}) = f(y_1)f(y_2 \mid y_1) \cdots f(y_T \mid y_1, y_2, \ldots, y_{T-1}). \tag{7.2}$$

A model for the set of univariate, conditional distributions on the right-hand side of this expression is known in general as a *transitional* model. For categorical outcomes, this is a *Markov chain* model. Such models are reviewed in Section 7.2.

In a longitudinal study, very often not all subjects are available for measurement at every pre-determined timepoint. Commonly, a subject, once missing, is lost to all subsequent follow-up; this is known as *dropout.* Strategies for modelling data when some subjects drop out are reviewed in Section 7.3. A further summary of strategies available for a cross-classification of model type and dropout specification is given in Section 7.4.

I conclude in Section 7.5 with a brief summary of the novel contributions made in this thesis, and highlight some topics for future research.

7.1 Multivariate models

We consider here models for the joint, multivariate probability function $f(\mathbf{y})$. I review fully marginal models in Section 7.1.1, then in Section 7.1.2 consider models that are marginal only up to some given order of interaction, being canonically linked thereafter.

7.1.1 Fully marginal models

Fully marginal models were introduced by Molenberghs and Lesaffre (1994) and Glonek and McCullagh (1995), who studied marginal odds ratio models fitted to binary data or ordered categories (Glonek and McCullagh consider also unordered categories), and by Ekholm *et al.* (1995), who introduced dependence ratio models for binary data.

In Section 2.3.1 I have generalized the approach of Ekholm *et al.* (1995) in two ways. Firstly, I have derived score equations that do not impose constraints such as 'horizontal homogeneity' (i.e. constraining all ratios of like order to equality); I allow full freedom of choice of parametrization. Secondly, I have illustrated how the ratios and models can be extended to polytomous data, both ordered and nominal.

My main contribution to the field of marginal odds ratio models is computational. The new SQb algorithm (Section 3.6) makes it feasible to fit models on many more timepoints than hitherto practical, because it is much faster than Newton–Raphson iteration as adopted by Glonek and McCullagh (1995). The method employed by Molenberghs and Lesaffre (1994) has not been extended to more than three timepoints, as it requires analytic formulae for the coefficients of high-order polynomials, which are not currently available (as discussed in Section 3.2).

For many sets of odds ratios, at the extremes of the parameter space, the Newton–Raphson algorithm fails to find a solution for the probability table, π — even though one exists, and can be found readily using SQb. On the other hand, ratios also exist for which Newton–Raphson finds a solution but SQb cannot. Thus a robust computer routine for finding π must have both algorithms available in case one fails. A joint strategy is proposed (on page 144): briefly, attempt SQb first, but use Newton–Raphson if this fails. With an increasing number of timepoints an entire wasted run of SQb is increasingly quicker than a single Newton–Raphson step.

When all odds ratios lie between 0.1 and 10 — or are constrained to do so — the SQb algorithm has never been observed to fail to converge within reasonable time and can entirely replace Newton–Raphson. As an example of the speed advantage: if there are 7 timepoints, SQb is on average 400 times faster than Newton–Raphson.

Of the other algorithms studied, SMϕ_ϵ fails as seldom as does Newton–Raphson, but it is never very much quicker on average, and can be rather slower. The SR algorithm (modified to SRquad for less than seven timepoints), is often close to SQb in speed, but is occasionally much slower, and more often fails to find a solution within an acceptable amount of iterations. Because it is based on an approximate inverse for the logistic transform, the SR algorithm does, however, generally give very good approximations to within 2 or 3 decimal places within a remarkably small number of iterations.

Sometimes all algorithms fail. I have proposed a meta-algorithm whereby a less extreme problem is solved to give either an approximate solution, or better starting values for the original problem, but this has not yet been fully studied. It is more urgent to study whether the extreme conditions that lead to numerical instability are likely to occur in practical applications.

7.1.2 Partially marginal and zero-conditional models

Motivated by the considerable computational advantage of first-order marginal models as proposed by Fitzmaurice and Laird (1993), I have studied taking the process one stage further, whereby *all* the odds ratios are zero-conditional, giving a model with identity link to the canonical parameters, $\boldsymbol{\xi}$ in (7.1).

This model is computationally simpler than that of Fitzmaurice and Laird (1993). However, it is sensitive to mis-specification of the dependence structure, and in this it is considerably worse than the Fitzmaurice and Laird model, which has efficient estimators of marginal means even when high-order, zero-conditional ratios are poorly specified.

In the fully zero-conditional model, the probability table corresponding to a set of zero-conditional log odds ratios, $\boldsymbol{\xi}$, is easily expressed in closed form. This enables some analytical insight into the transform to marginal log odds ratios, $\boldsymbol{\lambda}$. The partially marginal set of ratios used in the Fitzmaurice and Laird model does not give a probability table expressed in closed form.

Canonical models are attractive because of the speed of the fit, and because for reasonably large numbers of observations the marginal algorithm may not be compu-

tationally feasible. These simpler models generate maximum likelihood estimates of the probability table, $\hat{\pi}$, from which marginal odds ratios, $\hat{\lambda}$, may be simply read. If one is ultimately interested in the parameters γ_M of a more or fully marginal model, one might use, say, least squares to fit them to $\hat{\lambda} = X\gamma_M$. We might obtain crude confidence intervals for such estimates by studying the range of values in the marginal interpretation as we vary the zero-conditional parameters between the extremes of their estimated confidence intervals.

This has not been studied; instead I have concentrated on extending the practical limit on the number of observations for which fully marginal parameters can be estimated directly (see previous subsection). Indirect techniques remain potentially useful, however, for data based on more than ten timepoints.

I have been perhaps too dismissive in Section 2.5.2 in claiming that mixed parametrizations with greater than first-order odds ratios modelled marginally are too difficult to fit. Although I have not yet found a recursive definition and/or algorithm as succinct as that for the fully marginal case, this would be a useful area for further study.

7.2 Markov chain models (and generalizations)

A series of discrete observations with conditional probability functions obeying

$$f(y_t \mid y_1, \ldots, y_{t-1}) = f(y_t \mid y_{t-m}, \ldots, y_{t-1})$$

is called a Markov chain of order m. A set of such probabilities for $t = 1, \ldots, T$ defines a multivariate distribution for $\mathbf{y} = (y_1, \ldots, y_T)'$ as in equation (7.2). In Section 4.5 we have seen how this is readily extended to the case when a vector of observations is taken at each timepoint, giving a multivariate Markov chain. Furthermore we have noted that decomposition (7.2) can be applied to any joint distribution even when the variables are not repeated measures; this is *not* a Markov chain model. In particular, in Sections 4.5 and 6.3 we have used the decomposition on the vector of observations at each timepoint of a multivariate chain.

This generalized model can be fitted using the same methodology as for any standard

Markov chain model, as indicated in Section 4.5. The log likelihood for a set of S conditional probability functions obeying the decomposition (7.2), whether they represent a Markov chain or not, is the sum

$$\ell = \ell_1 + \ell_2 + \cdots + \ell_S.$$

Assume that the linear predictors, $\boldsymbol{\eta}_s = X_s\boldsymbol{\gamma}$, are variationally independent both within and across conditional models; departure from this assumption is treated shortly below. Concatenating $\boldsymbol{\eta} = (\boldsymbol{\eta}_1', \ldots, \boldsymbol{\eta}_S')'$ and $\mathbf{b} = (\mathbf{b}_1', \ldots, \mathbf{b}_S')'$, where $\boldsymbol{b}_i = \partial\ell_i/\partial\boldsymbol{\eta}_i$, the contribution for the entire series to the overall score equation is, for each subject,

$$\mathbf{U} = X'\frac{\partial\ell}{\partial\boldsymbol{\eta}} = X'\mathbf{b}. \tag{7.3}$$

By choice of design matrix X, built from the set of X_s, different $\boldsymbol{\eta}_s$ may share some parameters $\boldsymbol{\gamma}$. But this does not affect the general form (7.3), because the results of applying the chain rule as above, or of re-writing the likelihood in terms of $\boldsymbol{\gamma}$ and differentiating directly, are necessarily identical. As a result, parameters common to different conditional models can be specified by choice of X independently of the individual forms $\mathbf{b}_s$, which are the same for all models to be considered. Furthermore by the same reasoning the information matrix contribution is

$$\mathcal{I} = X'\mathrm{E}[\mathbf{b}\mathbf{b}']X, \tag{7.4}$$

where $\mathrm{E}[\mathbf{b}\mathbf{b}'] = -\partial\ell^2/\partial\boldsymbol{\eta}'\partial\boldsymbol{\eta}$ is necessarily block diagonal.

In this thesis I have derived $\mathbf{b}_s$ only for logistic regression models — *not* assuming standard proportional odds restrictions — for ordered and unordered categorical data. I have used a logit link throughout; other standard links, such as probit or complementary log-log should also be considered. Extensions to other types of data are worthy of further study and can be easily incorporated into the above framework, given the relevant form of $\mathbf{b}_s$ and $\mathrm{E}[\mathbf{b}_s\mathbf{b}_s']$.

The use of the generalized score equations within standard Markov chain modelling

is novel and allows great flexibility. For example, we may fit different explanatory variable models at different timepoints while assuming stationarity with respect to previous outcomes, or fit common baseline probabilities within non-stationary chains, or allow for interaction between history and covariates in chains of arbitrary order. A single algorithm suffices to fit every model to be considered, which seems preferable to introducing a series of more tailored methods such as reviewed by Lindsey (1993) and Diggle *et al.* (1994). It also allows for easy forward or backward selection of nested models.

A large number of parameters is unavoidable, given the large number of potential interactions. As a rule of thumb, we should not set parameters common to timepoints adjusted for different histories. Parameters *can* safely be shared across timepoints when modelling stationary, fixed-order chains, for example. Otherwise, similar parameter values from an unrestricted set can suggest simplifications in the interests of parsimony, but care may be needed in subsequent interpretation.

I am still in search of a name for the generalized model. Restricted to longitudinal data, 'timepoint-wise factorized Markov chain' is adequate, though clumsy. Whatever it is called, the 'Markov chain' approach offers an intuitively appealing model when there is dropout, as seen in Chapter 6 and discussed in the following section.

7.3 Dropout

As mentioned in the introduction to this chapter, *dropout* occurs when a subject becomes and remains unavailable for further measurement before the final timepoint of the planned study. A standard approach to modelling when there is dropout follows the more general method of Little and Rubin (1987): we introduce a vector of indicator variables for occurrence of dropout, $\mathbf{R}$, and model the joint probability $f(\mathbf{y}, \mathbf{r})$. In the nomenclature of Little and Rubin, dropout is *non-random* if the probability of dropout depends on the value of the outcome variable that would have been observed had the subject not dropped out. Dropout is said to be *completely at random* (MCAR) if the probability of dropout is independent of all observed data values, and *at random*

(MAR) if dropout probability depends on observation history.

The joint distribution $f(\mathbf{y}, \mathbf{r})$ may be factorized

$$f(\mathbf{y}, \mathbf{r}) = f(\mathbf{y}) f(\mathbf{r} \mid \mathbf{y}), \tag{7.5}$$

which gives a *selection model*, or otherwise as

$$f(\mathbf{y}, \mathbf{r}) = f(\mathbf{r}) f(\mathbf{y} \mid \mathbf{r}), \tag{7.6}$$

which is called a *pattern-mixture model.* In a selection model, the observed data are assumed to be a subset of an underlying complete observation $\mathbf{y}^*$; this assumption is not required in pattern-mixture models, because $f(\mathbf{y} \mid \mathbf{r})$ is fully observed. A different factorization is introduced in Section 6.3 and discussed below in Section 7.3.2.

I use the terminology *data model* for $f(\mathbf{y})$ or $f(\mathbf{y} \mid \mathbf{r})$, and *dropout model* for $f(\mathbf{r})$ or $f(\mathbf{r} \mid \mathbf{y})$. Unless dropout is non-random, data and dropout models can be fitted separately. Thus in this case one could model both $f(\mathbf{y} \mid \mathbf{r})$ and $f(\mathbf{r} \mid \mathbf{y})$ even though these do not combine directly to give $f(\mathbf{y}, \mathbf{r})$.

In the following I split discussion of marginal and transitional models into separate subsections, though it is possible and sometimes desirable to mix the two approaches, using, say, a marginal data model with transitional dropout model, as in Molenberghs *et al.* (1997).

First, though, I re-iterate an important general point spelled out in Section 6.5: we cannot assess the pattern of missing data if we assume they are informatively missing. If enough constraints are placed on the model for the observed data, however, estimates for the missing-data parameters, and hence the missing data, are imputed, since there are a fixed number of degrees of freedom. I argue though that even if a model is sufficiently specified to estimate such parameters, their significance should not be construed as a 'test for informative missingness', the implication being that the outcome of such a test is determined by the data. Rather, it is the untestable assumptions that determine the outcome of the test.

7.3.1 Multivariate models

In Chapter 5 we considered models based on either full selection or full pattern mixture approaches, that is, we considered $f(\mathbf{y})$ or $f(\mathbf{y} \mid \mathbf{r})$ in Section 5.2, and $f(\mathbf{r})$ or $f(\mathbf{r} \mid \mathbf{y})$ in Section 5.3.

For the data models, a fully marginal model naturally follows a selection approach, while for partially marginal, and zero-conditional models, inherent lack of reproducibility demands imputation of missing values before a selection model can be fitted (even assuming MCAR dropout). I have shown in Section 5.2.4 how the introduction of a correction term — a separate intercept parameter for each dropout pattern — enables a quasi-pattern-mixture model to be fitted without recourse to imputation. Parameters other than the intercept may be the same for all dropout patterns, which violates the definition of full pattern-mixture modelling. A full pattern-mixture model is easily fitted without imputation; observations of different lengths are modelled separately. This approach has not been favoured in the literature.

For dropout models, I have considered ordinary marginal models adapted to take account of the structural zeros in the vector of $\mathbf{R}$ values, but encountered problems both in fitting and in modelling time-varying covariates.

To overcome such problems, a semi-canonical model has been proposed in Section 5.3.3. Advantages are that this is very easy to fit, that successive means are inherently constrained to be monotone decreasing, as they need to be, and that the linear predictors are identical for different observation sizes, so we gain reproducibility (or more correctly a type of reproducibility; such that the model for timepoints 1 to d is a subset of that for timepoints 1 to $d+1$, etc.).

In this model, explanatory variables at dropout time are not explicitly modelled even if known, but their history determines the canonical parameters and hence uniquely determines the distribution function for the outcome vector. A loss is that time-varying explanatory variables at the final timepoint are not incorporated.

The vector of indicators $\mathbf{R}$ can be aliased by a variable D, denoting dropout time, which follows a truncated geometric-type distribution where probabilities of 'success'

vary in successive trials. However, there is no obviously simpler way to model the time-varying probabilities than by using the above model.

Even for a trial with primary outcomes measured at discrete timepoints, dropout time may be known more fully, such as when dropout is due to death and this is followed up and recorded. In such cases, unless dropout is non-random, any suitable, standard survival model can be fitted.

7.3.2 Markov chain models

After factorizing the joint probability $f(\mathbf{y}, \mathbf{r})$ according to (7.5) or (7.6), we may use standard Markov chain models for the data and/or dropout models. Rather than pursue this directly here I have studied instead factorizing first by timepoint, giving a multivariate Markov chain:

$$f(\mathbf{y}, \mathbf{r}) = f(y_1, r_1) f(y_2, r_2 \mid y_1, r_1) \cdots f(y_T, r_T \mid \mathbf{h}_T, r_1, \ldots, r_{T-1}).$$

Assuming that we do not attempt to model timepoints beyond dropout time, the conditional model for (y_t, r_t) depends only on previous outcomes and r_{t-1}.

After this factorization, a selection-model style approach is to take

$$f(y_t, r_t \mid \mathbf{h}_t, r_{t-1}) = f(y_t \mid \mathbf{h}_t, r_{t-1}) f(r_t \mid y_t, \mathbf{h}_t, r_{t-1}),$$

whereas a pattern-mixture style model uses

$$f(y_t, r_t \mid \mathbf{h}_t, r_{t-1}) = f(r_t \mid \mathbf{h}_t, r_{t-1}) f(y_t \mid r_t, \mathbf{h}_t).$$

These are called here *timepoint-wise selection* and *timepoint-wise pattern-mixture* models, respectively. Both combine elements of both ordinary selection and pattern-mixture models: in general the conditional probabilities for $\mathbf{y}$ depend on $\mathbf{r}$ and those for $\mathbf{r}$ depend on $\mathbf{y}$.

Model selection can be difficult even for data on a small number of timepoints. There are far too many possible models for all-subsets regression to be applied blindly, but

there is surely great danger of missing a good model if selection is not in some way systematic. The block stepwise selection algorithm proposed in Section 6.6 is a first attempt at tackling this problem, but certainly might fail to find the best model. More work is needed here.

Parameters will not be shared between data and dropout models, so that each can be optimized separately, unless dropout is non-random. If one is simultaneously imputing values according to various non-random dropout assumptions, however, the models do not separate and model selection becomes even more difficult.

Setting common parameters within a timepoint-wise pattern-mixture model, for each timepoint model for each dropout pattern, as in the example in Section 6.6, is very similar to a selection model approach. The assumption made is that all the observed values up to timepoint d are realizations of a common distribution $f(\mathbf{y}^{(d)} \mid D > d)$, which is itself nested within a common distribution $f(\mathbf{y}^{(d+1)} \mid D > d+1)$, and so on. These distributions are neither the selection $f(\mathbf{y}^*)$ nor the pattern-mixture $f(\mathbf{y} \mid d)$, but are much closer to the former in concept and interpretation. Given a marked preference in the literature for selection models, the proposed mixed approach should prove appealing.

7.4 Complete and incomplete observations — summary

I here summarize considerations for the four main types of model studied, i.e. the fully marginal, first-order marginal, canonical (zero-conditional) and Markov parametrizations, according to missing data presence or assumption, i.e. complete data, MCAR, MAR and non-random missing or dropout.

The marginal Koch and Stram models, the subject-specific Korn & Whittemore, beta-binomial and random-effects models, and split-plot ANOVA methods, discussed in Chapter 1, are not included in the following. For GEE1, see 'first-order marginal'. Comments would be similar for the GEE2 and other partially marginal models not studied here in depth.

Fully marginal, complete data. These models are discussed extensively in Chapter 2, Section 2.3, and in Section 7.1.1 above, q.v.

Fully marginal, MCAR data. As discussed in Section 5.2.1, in this case ordinary likelihood methods are applied to the observed data. It suffices to mask out score and information matrix contributions relating to variables not observed for a particular subject. Reproducibility (Section 1.4.3) makes fully marginal models very attractive in this setting. Such models easily deal with any pattern of missing data (except entirely missing), not just the monotone pattern of dropout.

Fully marginal, MAR data. The comments of the previous paragraph apply here too, if one is concerned only with the data model. This gives a selection model. A full pattern-mixture model would ignore reproducibility, raising doubt as to whether one should ever attempt to use marginal models in a pattern-mixture setting. Parameters might be common for each of the odds ratios, as in selection models, but with observation-vector size introduced as a further explanatory variable. If there is interest in the missing-data mechanism, it needs to be modelled separately, perhaps using one of the methods proposed in Section 5.3.

Fully marginal, informative missing. I have considered this case in Section 6.4, specifically for binary data, in Subsections 6.4.2 and 6.4.6. One must be prepared to make untestable assumptions about the missing mechanism in order to obtain an identifiable model. Recently Molenberghs *et al.* (1997) have proposed a selection model for ordinal data, with a logistic regression model for the missing-data mechanism, which is restricted to dropout. With informative missing data, selection models *require* imputation of missing $\mathbf{Y}$ values (Molenberghs *et al.* use the EM algorithm).

First-order marginal, complete data. These models were proposed by Fitzmaurice and Laird (1993) and are discussed in Section 2.5. Orthogonality between the marginal (first-order) and zero-conditional (higher-order) odds ratio estimates makes these models particularly attractive when only the univariate marginal odds are of

interest, since estimators of these are efficient even if the dependency model is very poorly specified. GEE1 serves a similar purpose but fails to take account of high-order dependence and gives less efficient marginal estimates.

First-order marginal, MCAR and MAR data. Even in the relatively simple case of MCAR data, the lack of reproducibility of the zero-conditional (higher-order) odds ratios presents a problem. I have highlighted the problems of assuming common zero-conditional odds ratios for outcomes with a different number of observations in Section 5.2.3, and suggested a pattern-mixture style correction for this (the 'corrected false-identity link') in Section 5.2.4, reviewed in Section 5.2.5. Fitzmaurice *et al.* (1994) proposed to use the EM algorithm in an alternative, selection-model approach. Another option is to use weighted GEE methods (Robins *et al.*, 1995).

First-order marginal, informative missing. The difficulties mentioned in the preceding paragraph apply here too, with the added problem of non-identifiability of the missing-data mechanism. I have not studied this particular combination in any detail, though it is clear that a combination of corrected identity links and multiple imputation would offer a feasible modelling strategy.

Canonical link, complete data. Canonical link models, proposed and discussed in Section 2.4, are extremely easy to fit but difficult to interpret. Their use as a stepping stone to marginal inference is considered in Section 2.4.4. More work is needed on this particular aspect, but it is clear that for a particular set of covariate values, a canonically-linked model that is not too simplified gives a good estimate of the joint probability table, from which marginal and/or ordinary conditional odds ratios can be read, as desired. A disadvantage is sensitivity to variance mis-specification.

Canonical link, MCAR and MAR data. The false identity link and its correction have already been discussed in the paragraph for first-order marginal models with MCAR or MAR data, above. This offers a pattern-mixture type model. I have not considered selection models in this context, though one could adapt the EM approach

of Fitzmaurice *et al.* (1994) here.

Canonical link, informative missing. The issue of non-identifiability of such models is detailed in Section 6.4.5. As for first-order marginal models, a combination of corrected identity links and multiple imputation would offer a feasible modelling strategy.

Markov, complete data. Discussed extensively in Chapter 4 and Section 7.2 above, q.v.

Markov, MCAR data. Provided that the missing-data pattern is monotone, i.e. the only missing data is due to dropout, fitting is identical to that for complete data except that one truncates the Markov chain immediately preceding dropout time for each incomplete observation. Other missing-data patterns present a more difficult problem, which has not been addressed here.

Markov, MAR dropout. In timepoint-wise models, both pattern-mixture and selection, data are MAR if previous y values enter the linear predictors for the dropout model. The data model itself is identical to that under the simpler MCAR assumption.

Markov, informative missing. I have considered this case primarily in illustrating inherent non-identifiability of informative missing data, in Section 6.5. Practically, one must be prepared to make some strong assumptions to gain identifiability, perhaps as part of an exercise in multiple imputation.

7.5 Conclusions

The algorithms introduced for numerical solution of the inverse logistic transform, especially algorithm SQb, make feasible the fitting of fully marginal models for data observed over more timepoints, and with more extreme dependence structure, than hitherto practical. Algorithm SQb could be improved still further by automating judicious choice of control parameters; the important gain would be in further increasing

the probability of obtaining a solution, though there can be improvement in speed also.

Algorithms SR and SRquad are chiefly useful in obtaining approximate solutions quickly; otherwise they compare unfavourably with SQb. Algorithm SMϕ_ϵ is optimum only in exceptional circumstances (a very large number of timepoints with odds ratios close to machine infinity or zero). In common with all other algorithms studied, Newton–Raphson iteration does not always offer a solution. Otherwise it has excellent convergence properties, though at the cost of being prohibitively slow for large problems. A combined strategy is proposed exploiting the best features of all four algorithms.

A generalization of the dependence-ratio model of Ekholm *et al.* (1995) is given, imposing fewer constraints, and dependence ratios have been defined for polytomous data. Score equations have been derived for these extended models.

The model of Fitzmaurice and Laird (1993) has been extended to cater for unbalanced data without the need for imputation. Another extension has been to models with all the canonical parameters parametrized directly using identity links. This is the computationally fastest to fit of this family of multivariate models, but it is sensitive to mis-specification of the dependency structure.

Score equations have been derived for generalized Markov chain models, allowing a flexible framework for fitting a broad class of models regardless of whether or not the factorization exploited corresponds to an underlying stochastic process. The methodology, specified for categorical outcomes, is amenable to extension to other data distributions.

Timepoint-wise factorizations of multivariate Markov chains introduce models for dropout that are neither purely selection nor purely pattern-mixture, but which may lean strongly towards one or the other extreme as the modeller prefers. For polytomous data, and ordinal data not assuming proportional odds models, there are many potential parameters to consider, and though I have suggested heuristic guidelines, more work is needed on subset selection.

Three broad areas for further work relate to both multivariate and transitional models.

Firstly, a useful addition would be the ability to incorporate random effects (*not* necessarily including a local independence assumption), enabling one to allow for heterogeneity between subjects not accounted for by existing models. Secondly, the techniques considered here for discrete timepoints need to be extended to the case of unequally spaced observations. Finally, more work is needed on software development, in that programs need to be both more efficient and more user-friendly.

Appendices

The section number of these appendices reflects that of the chapter and section to which they relate.

A2.2 PEF canonical parameters for polytomous data

Consider a bivariate distribution of ternary variables with the following cell probabilities:

	Y_2		
Y_1	0	1	2
0	π_{00}	π_{01}	π_{02}
1	π_{10}	π_{11}	π_{12}
2	π_{20}	π_{21}	π_{22}

Let ξ_i^r be the zero-based, zero-conditional univariate log odds ratios for variable $i = 1, 2$:

$$\xi_1^r = \log \frac{\pi_{r0}}{\pi_{00}} \tag{A1}$$

$$\xi_2^r = \log \frac{\pi_{0r}}{\pi_{00}} \tag{A2}$$

where $r = 1, 2$; the convention is subscript for variables, superscript for value. The bivariate ratios are based on the 'anchor' cell π_{00} giving

$$\xi_{12}^{rs} = \log \frac{\pi_{rs}\pi_{00}}{\pi_{r0}\pi_{0s}}. \tag{A3}$$

These are canonical parameters for the polynomial exponential family form (1.24), as may be readily verified directly; for example

$$\begin{aligned} \mathrm{P}(Y_1 = 2, Y_2 = 1) &= \exp\left\{\xi_1^2 + \xi_2^1 + \xi_{12}^{21} - C(\boldsymbol{\xi})\right\} \\ &= \chi_1^2 \chi_2^1 \chi_{12}^{21} \pi_{00} \\ &= \left(\frac{\pi_{20}}{\pi_{00}}\right)\left(\frac{\pi_{01}}{\pi_{00}}\right)\left(\frac{\pi_{21}\pi_{00}}{\pi_{20}\pi_{01}}\right)\pi_{00} \\ &= \pi_{21}, \end{aligned}$$

where $C(\boldsymbol{\xi}) = -\log \pi_{00}$ (by definition) and $\boldsymbol{\chi} = \mathrm{e}^{\boldsymbol{\xi}}$ as for binary variables with the same sub/superscript conventions as for $\boldsymbol{\xi}$.

It is easy (but tedious) to verify that other choices of minimal sets of ratios, such as adjacent-cell ratios (Agresti, 1990, equation 2.9) cannot be the canonical set. Starting with $C(\boldsymbol{\xi}) = -\log \pi_{00}$, there is a lack of freedom of subsequent choice by the very definition of the polyno-

mial exponential family form (1.24).

For three ternary variables, the 27-cell probability table can be considered as a $3 \times 3 \times 3$ cube rather than as three tables side by side; the direction of the levels of Y_3 is orthogonal to the plane of (Y_1, Y_2). With four or more variables we consider the equivalent tables as hypercubes. For three variables, the 'bottom layer' of the cube, $Y_3 = 0$, can be categorized by the univariate and bivariate ratios $\boldsymbol{\xi}$ with the modification that each of the $\boldsymbol{\pi}$ terms has an extra subscript zero. This is precisely what we define a zero-conditional odds ratio to be, conditional on all other cells being zero. That is, the CORs of the bivariate system or 'face' (Y_1, Y_2) — implicitly taking $Y_3 = 0$ — are the CORs for (Y_1, Y_2) in the trivariate case — explicitly taking $Y_3 = 0$ — and the former set have been shown already to be the canonical parameters. This argument readily extends to more dimensions; in general, we regard the $\boldsymbol{\pi}$ terms as having implicit trailing zeros.

By symmetry, considering the face of the (hyper)cube where $Y_1 = 0$ for the CORs of (Y_2, Y_3), the zero-based first and second order CORs, for all pairs, in all dimensions, are the canonical parameters.

Having verified that the zero-based trivariate log CORs are canonical parameters for (Y_1, Y_2, Y_3), the argument can be extended to arbitrary dimension by similar reasoning. Unfortunately it is tedious to verify the three-dimensional case, and it becomes increasingly tedious for more dimensions, and mere verification is not an inductive proof. The general verification procedure reduces to showing that the system of equations

$$\pi_{i_1 i_2 \cdots i_T} = \chi_1^{i_1} \chi_2^{i_2} \cdots \chi_{12\cdots T}^{i_1 i_2 \cdots i_T} \times \pi_{00\cdots 0} \tag{A4}$$

has a unique solution, which it will have if and only if the matrix of the log of this system is non-singular.

A proof of this general result is not attempted here, but by following an example using the geometric approach outlined above, it becomes increasingly obvious that the result must hold. Although no formal proof of this fundamental assertion has been offered, the result has been tacitly assumed by both Liang and Zeger (1986) and Glonek and McCullagh (1995) in their key papers.

A2.3.3 Six Cities data: fully marginal model

The fit for the intercept-unconstrained model is given in the main text (Section 2.3.3). Full output for the fit for the model dropping the 4-way interaction term is as follows:

```
$out.table:
            estimated SE.robust    SE.info
  alpha_0 -1.90667422 0.1190109 0.11835835
   beta_1 -0.16336173 0.0558646 0.05686523
   beta_2  0.30731250 0.1879929 0.18890689
   beta_3  0.08454771 0.0878113 0.08855784
 alpha_12  2.00308953 0.2625182 0.26096447
 alpha_13  1.74965922 0.2696587 0.26729904
 alpha_14  2.07416785 0.2697889 0.27920543
 alpha_23  2.47007633 0.2888587 0.27944056
 alpha_24  2.05603879 0.2800217 0.28211751
 alpha_34  0.09611362 0.6201641 0.62699526
alpha_123 -0.28213826 0.6145637 0.60676994
alpha_124  2.08680648 0.2876003 0.28863899
alpha_134 -0.23136507 0.6177819 0.62222742
alpha_234  0.10020013 0.6604717 0.66207244

$ell:
[1] -793.1531

$G2:
[1] 7.910195

$df:
[1] 16

$P:
[1] 0.951495

$ptables:
$ptables[[1]]:
 [1] 0.678534427 0.067149364 0.043844147 0.018709618 0.036618872
0.010879501
 [7] 0.017821337 0.014411863 0.031894926 0.014161739 0.004836081
0.011527087
[13] 0.006278528 0.005602062 0.009380985 0.028363648

$ptables[[2]]:
 [1] 0.628071562 0.061889351 0.044638676 0.017497885 0.041657968
0.011947229
 [7] 0.021399113 0.015573369 0.040585507 0.017711151 0.006713393
0.014833823
[13] 0.010099881 0.008668500 0.015568736 0.043158510


$est.fit:
$est.fit[[1]]:
 [1] 237.487049  23.502277  15.345452   6.548366  12.816605
3.807825
 [7]   6.237468   5.044152  11.163224   4.956609   1.692628
```

```
4.034481
[13]   2.197485   1.960722   3.283345   9.927277

$est.fit[[2]]:
 [1] 117.449382  11.573309   8.347432   3.272104   7.790040
2.234132
 [7]   4.001634   2.912220   7.589490   3.311985   1.255405
2.773925
[13]   1.888678   1.621009   2.911354   8.070641
```

Abbreviated output for the model dropping all 3-way interactions (note cells for which the fit is very poor):

```
$out.table:
          estimated  SE.robust    SE.info
 alpha_0 -1.9118070 0.11833014 0.10280651
  beta_1 -0.1470229 0.05793643 0.05817628
  beta_2  0.2989133 0.18762564 0.16353474
  beta_3  0.0968851 0.09155967 0.09159680
alpha_12  1.7653004 0.26295761 0.26169888
alpha_13  1.1472224 0.27383779 0.26446182
alpha_14  1.5925187 0.25982558 0.27247016
alpha_23  1.9736280 0.26821286 0.26900619
alpha_24  1.2521553 0.25738166 0.26935413
alpha_34 -0.1837351 0.59884757 0.57862228

$ell:
[1] -818.4471

$G2:
[1] 58.49811

$est.fit[[1]]:
 [1] 225.6267515  23.0727700  15.5782397   6.1473181  19.5290349
 4.0269188
 [7]   8.6796025   7.7368034  17.5435445   7.7345094   3.5216582
 5.7073917
[13]   0.7200952   0.5835083   0.8755088   2.9214883

$est.fit[[2]]:
 [1] 109.7247226  10.7137111   7.9331377   2.7060803  12.5209205
 2.5440744
 [7]   6.0846692   4.9754357  12.5846938   5.3397017   2.8282184
 4.0949216
[13]   0.6529852   0.5233501   0.9097640   2.8662653
```

For the lag-1 Markov model:

```
$out.table:
           estimated  SE.robust    SE.info
 alpha_0 -1.93180775 0.11801410 0.09443219
  beta_1 -0.14605059 0.05856323 0.07103304
  beta_2  0.30349280 0.18730100 0.14875857
  beta_3  0.09902214 0.09137343 0.11376171
alpha_ij  1.44270373 0.22175107 0.15785077
```

```
$ell:
[1] -858.0214

$G2:
[1] 137.6469
```

For verification, the independence model as fitted by Fitzmaurice and Laird (1993):

```
$out.table:
          estimated  SE.robust    SE.info
alpha_0 -1.90345085 0.11941216 0.08912967
 beta_1 -0.16404626 0.05601693 0.06937986
 beta_2  0.31656227 0.18804971 0.13968567
 beta_3  0.09363723 0.08686099 0.11063898

$ell:
[1] -912.6204

$G2:
[1] 246.8449
```

The S-PLUS code used to fit the above models follows. `fmm.fit` is the top level program; this loops through `fmm.loop` (which calls `fmm.vars`) until convergence, after which it calls `fmma.vars`, an extended version of `fmm.vars` that also evaluates the log likelihood and G^2 and returns probability and expected tables for diagnosis. For brevity `fmm.vars` is omitted from the following and only `fmma.vars` is included. The S-PLUS function `algor.c`, which calculates probabilities from odds ratios, is a wrapper calling the C version of algorithm SQb given in Appendix A3.6.

```
function(m, counts, X, sizes, gammas, gamma.names, ptables, etables)
{
#
############################ fmm.fit ############################
#
# Get SEs (robust and info) and ell, G^2 value for fmm (ALL-MOR) model
#
# Arguments: as in fmm.loop; additionally
#       ptables         a list as long as X (garbage on call, but
#                       names will be retained if you want pretty)
#       etables         ditto
#       gamma.names     (character data) for pretty array print
#
#################################################
#
        num.params <- length(gammas)
        uut <- array(0, c(num.params, num.params))
        info <- uut
        U <- rep(0, num.params)
        ell <- 0
        G2 <- 0
        stop.difference <- 1
        iters <- 0
        gammas.new <- gammas
        while(stop.difference >= 0.0001) {
                gammas.old <- gammas.new
                gammas.new <- fmm.loop(m, counts, X, sizes, gammas.old)
                iters < iters + 1
                stop.difference <- infty.norm(gammas.old - gammas.new)
        }
        for(i in 1:length(X)) {
                vars <- fmma.vars(X[[i]], m[[i]], counts[[i]], gammas.new,
                        sizes$num.obs, sizes$msize, sizes$vmsize, U, uut, info,
                        ell)
                U <- vars$U
                uut <- vars$uut
                info <- vars$info
                ell <- vars$ell
                ptables[[i]] <- vars$ptable
                counts.sum <- sum(counts[[i]])
                etables[[i]] <- vars$ptable * counts.sum
                G2 <- G2 + 2 * sum(counts[[i]] * log(counts[[i]]/etables[[i]]))
        }
        inv.info <- solve(info)
        out.table <- cbind(gammas.new, sqrt(diag(inv.info %*% uut %*% inv.info)
                ), sqrt(diag(inv.info)))
        dimnames(out.table) <- list(gamma.names, c("estimated", "SE.robust",
                "SE.info"))     #
        df <- 2 * (2^sizes$num.obs - 1) - length(gammas)
        P <- 1 - pchisq(G2, df) #
        list(out.table = out.table, ell = ell, G2 = G2, df = df, P = P, ptables
                 = ptables, est.fit = etables)
}
```

```
function(m, counts, X, sizes, gammas)
{
############################ fmm.loop ############################
#
# One loop towards iterative solution of MOR score equations
#
# Needs to have
#
#       m       a list of observed cluster indices (plus 1), ie
#               index 1 is all obs zero, index 2 is Y1=1 only, etc
#               Each vector of observed indices in the list has
#               the same X matrix.
#
#       counts  a list of vectors of observed 'cell counts' (each in list
#               corresponds to counts for the same list for m
#
#       X       a list of design matrices X_i, to give X_i%*%gammas;
#               each in list applies to all m observations at
#               this level
#
#       sizes   a list (with named components) as follows:
#
#       sizes$num.obs   = (common) cluster size
#       sizes$msize     = 2^num.obs (size of cell probability table)
#       sizes$vmsize    = msize - 1 (size of variance matrix including zeros)
#
#       gammas  vector of current parameter estimates
#
###############################################
#
        num.params <- length(gammas)
        uut <- array(0, c(num.params, num.params))
        info <- uut
        U <- rep(0, num.params) #
#
        for(i in 1:length(X)) {
                vars <- fmm.vars(X[[i]], m[[i]], counts[[i]], gammas, sizes$
                        num.obs, sizes$msize, sizes$vmsize, U, info)
                U <- vars$U
                info <- vars$info
        }
        newgammas <- gammas + solve(info, U)
        newgammas
}
```

```
function(X, m, counts, gammas, num.obs, msize, vmsize, U, uut, info, ell)
{
############################ fmma.vars ##############################
#
# calculate the variance matrix, marginal expectations, and ptable
# for design matrix X and params gammas
#
# --- FULLY MARGINAL MODEL ----
#
# ptable is msize = 2^T long (it is given as vector, not array, but
# in the sequence it would have as an array (ie first subscript changing
# fastest, etc)
#
#
#################################################
#
# first find (all) lambda values:
        lambdas.wrapped <- c(0, X %*% gammas)   #
        X <- rbind(0, X)        # put in the constraining row for lambda_0 = 0
# go find p table:
        algorans <- algor.c(lambdas.wrapped)
        p <- exp(algorans$ans)  #
        dmat <- array(.C("sdiff",
                as.double(p),
                as.integer(msize),
                as.double(1:(msize * msize)))[[3]], c(msize, msize))
        for(i in 1:length(counts)) {
#
# invert it and times by m-observed (somewhat inefficient for now)
                mi <- rep(0, msize)
                mi[m[i]] <- 1
                partans <- solve(dmat, mi)      # finally u is X' times this:
                contrib <- t(X) %*% partans
                U <- U + counts[i] * contrib
                uut <- uut + counts[i] * contrib %*% t(contrib)
                ell <- ell + counts[i] * algorans$ans[m[i]]
        }
        info <- info + sum(counts) * t(X) %*% solve(dmat) %*% diag(p) %*% solve(
                t(dmat)) %*% X
        list(U = U, uut = uut, info = info, ell = ell, ptable = p)
}
```

A2.4.5 Calculating probabilites from conditional odds ratios

We need to establish that algorithm steps A1 and A2 on page 70 give the correct probability table for all problem sizes T. It is trivial to check that the algorithm is correct for $T = 1, 2$. Then using induction, suppose that the algorithm yields a valid table for $T - 1$ variables, with odds ratios $\chi_1, \chi_2, \ldots \chi_{12\cdots(T-1)}$. Let us call the Tth variable the 'new' variable, since in the T-variate case we have a set of odds ratios labelled the same as the $(T-1)$-variate case, the 'old' values, plus a new set of the same length as the old that has the same indices except for the addition of a further subscript indicating the new variable, Y_T, together with the new univariate ratio χ_T.

First, fill in the half of the table corresponding to $Y_T = 0$ as for the $(T-1)$-variate case, but do not yet divide by the sum, step A2; by the induction hypothesis this 'old' half of the table has the required odds ratios for these cells. No matter what values go into the 'new' half of the table, where $Y_T = 1$, these odds ratios will be unaffected since by definition they are based entirely on probabilities with $Y_T = 0$.

The odds ratios required in the 'new' half of the table, where throughout $Y_T = 1$, are not the 'new' ratios themselves, but rather the *conditional* ratios $\chi_{\mathcal{B}\,|\,y_T=1}$ for all subsets $\mathcal{B}$ of $\mathcal{T}^* = \{1, 2, \ldots (T-1)\}$. Using the algorithm to fill this $(T-1)$-variate subtable with such conditional ratios, the induction hypothesis ensures that we obtain a subtable having these ratios, as required. Subsequent division of the entire T-variate table by the sum of all its entries then necessarily converts it to the required probability table.

However, the algorithm does not explicitly fill the table in two halves as just proposed, and so it remains to show that the two schemes are equivalent. By definition the 'new' ratios are of the form

$$\chi_{\mathcal{B}T} = \frac{\chi_{\mathcal{B}\,|\,y_t=1}}{\chi_{\mathcal{B}\,|\,y_t=0}} \tag{A5}$$

where $\chi_{\mathcal{B}\,|\,y_t=0}$ are the ordinary $\chi_{\mathcal{B}}$, and $\mathcal{B} \subseteq \mathcal{T}^*$; the subscript $\mathcal{B}T$ is an abbreviation for $\mathcal{B} \cup \{T\}$. Thus

$$\chi_{\mathcal{B}\,|\,y_t=1} = \chi_{\mathcal{B}T}\chi_{\mathcal{B}}, \tag{A6}$$

where the right-hand side is in terms of zero-conditional ratios. When $\mathcal{B}$ is the empty set, $\chi_{\emptyset\,|\,y_t=1} = \chi_T$ since by convention $\chi_0 = 1$. The two-half algorithm fills the 'old' half of the table as

$$c_{s(\mathcal{B})0} = \prod_{\mathcal{C}\subseteq\mathcal{B}} \chi_{\mathcal{C}}, \tag{A7}$$

while over the 'new' half

$$c_{s(\mathcal{B})1} = \prod_{\mathcal{C}\subseteq\mathcal{B}} \chi_{\mathcal{B}\,|\,y_t=1} = \prod_{\mathcal{C}\subseteq\mathcal{B}} \chi_{\mathcal{C}T}\chi_{\mathcal{C}} = \prod_{\mathcal{D}\subseteq\{\mathcal{B}\cup\{T\}\}} \chi_{\mathcal{D}}, \tag{A8}$$

as in each case $\mathcal{B}$ runs through all the subsets of $\mathcal{T}^*$. But consider what happens as $\mathcal{B}$ runs through all subsets of $\mathcal{T}$, as in algorithm step A1. When variable Y_T is not selected, we obtain cells according to (A7); when Y_T is selected we obtain the remaining cells, equation (A8). Thus the 'two-half' version, which is directly validated by induction, is equivalent to the stated version (steps A1 and A2). Division of the the cells by the sum to obtain probabilities is not controversial provided that the log odds ratios are real, or equivalently that the odds ratios are strictly positive, and clearly yields a unique solution.

A3.1.2 The tildeplus operator

If we hypothesise a complete ordered field $T(\not+, +)$ order-isomorphic to $R(+, \times)$ and if we let $+$ in T be the *exact same* numeric operator as $+$ in R, the definition of $\not+$ as given in the main text must follow. The isomorphism from R to T is that of taking logarithms, and T is an extension of R to include (unique) logarithms of the negative numbers. Put rather less formally: $\not+$ is the operation we would have to perform on the logs of two numbers in order to obtain, by taking antilogs of the result of the operation, the sum of the original numbers.
Some other observations are that the 'one' of T is 0; the 'zero' I write as @. This @ represents negative infinity, and under addition (and hence multiplication) it behaves formally as one would naively expect: $@ + x = @$, $n@ = @ \quad \forall x, n \in R$. Given that the isomorphism from R to T is the logarithmic function this is clear, since we must map $\log 0 \mapsto @$, where $@ \in \tilde{R}$, one of the set of new 'tilde numbers'. Also, tildeplus has a natural inverse denoted $\sim$ (tildeminus). Tildeminus and tildeplus are also unary operators leading to the idea of 'tilde-sign'. Denoting $@ \sim x = \tilde{x}$ is another way of deriving the unique set of 'log negatives'.

A3.2 Darroch's conjecture

Darroch (1962) conjectured that, in problems of all dimensions, if there is a valid solution (in my terminology) at all, then it is unique. This conjecture was shown in that paper to hold good for bivariate and trivariate binary outcomes (and also a $3 \times 2 \times 2$ table).
The conjecture is claimed to hold good more generally by Lemma 1 in Molenberghs and Lesaffre (1994). The steps of the proof are identical to those of Darroch (1962), the essential extension being that of establishing that the highest-order interaction equation is always of the form, in notation closer to that of Darroch,

$$\Lambda_{12\ldots T} = \frac{\prod_{i=1}^{2^{(T-1)}} (\pi_{1\ldots 1} - a_i)}{\prod_{i=1}^{2^{(T-1)}} (b_i - \pi_{1\ldots 1})}, \tag{A9}$$

where a_i and b_i are functions of the lower-order odds ratios. A valid solution must satisfy

$$a = \max(a_i) \leq \pi_{1\ldots 1} \leq \min(b_i) = b, \tag{A10}$$

with strict inequalities for nondegenerate distributions; the unique valid solution lies between a and b (since the function is continuous and monotone increasing, and $\Lambda_{12\ldots T} = 0$ at a and $\to \infty$ as $\pi_{1\ldots 1} \to b$; Darroch, 1962).
The justification for the form (A9) — the critical step — is not explicit in Molenberghs and Lesaffre (1994). It is a consequence of their equation (4.6), but this itself is not formally proved. It is not immediately obvious from the defining equations, nor by considering the hierarchical approach outlined in Section 3.2.4, the controversial issue being whether the successive substitutions do indeed give rise to terms with the same signs as in (A9), for all dimensions.
As an extension, I note on page 119 that the conjecture also appears to hold when the 'probabilities' sum to some arbitrary value $\Lambda_0 > 0$. Such odds ratios occur in practice when the logistic transform is applied to a subset of a full probability table — as in the derivation of (3.21), for example.

A3.3.1 Evaluating the derivative $\partial s(\mathbf{p})/\partial \mathbf{p}$

The top-level call is to `sdiff`, which is a wrapper to `sdiff_internal` to avoid the S-PLUS calling routine having to pass too many arguments.

```
/*
   sdiff(super,sub,vlen,nowlen,row_offset,col_offset,ans)

   Find diff S / diff p, initial call with super=sub= pvector
*/

void sdiff_internal(super,sub,vlen,nowlen,row_offset,col_offset,ans)
     double *super, *sub, *ans;
     long vlen, nowlen, row_offset, col_offset;
{
  long i,halflen;
  double newsuper[nowlen/2];
  double negsuper1[nowlen/2];

  if (nowlen == 2) {
    *(ans + row_offset*vlen + col_offset) =
      sub[0]/(super[0]+super[1]);
    *(ans + row_offset*vlen + col_offset+1) =
      sub[1]/(super[0]+super[1]);
    *(ans + (row_offset+1)*vlen + col_offset) =
      - sub[0]/super[0];
    *(ans + (row_offset+1)*vlen + col_offset+1) =
      sub[1]/super[1];
  } else {
    halflen = nowlen/2;
    for (i = 0; i < halflen; i++) {
      newsuper[i] = *(super+i) + *(super+i+halflen);
      negsuper1[i] = - *(super+i);
    }
    sdiff_internal(newsuper, sub, vlen, halflen,
                   row_offset, col_offset, ans);
    sdiff_internal(newsuper, sub+halflen, vlen, halflen,
                   row_offset, col_offset+halflen, ans);
    sdiff_internal(negsuper1, sub, vlen, halflen,
                   row_offset+halflen, col_offset, ans);
    sdiff_internal(super+halflen, sub+halflen, vlen, halflen,
                   row_offset+halflen, col_offset+halflen, ans);
  }
}

void sdiff(pvector,plen,ans)
     double *pvector, *ans; long* plen;
{
  double halflen;
  sdiff_internal(pvector,pvector,*plen,*plen,0,0,ans);
}
```

A3.4 The SM algorithms

Comparison of algorithm SM and its modifications

A comparison of the number of iterations until **p** convergence to 6 d.p. is shown in Table A1, for the SM algorithms discussed in Section 3.4. Only sets of $\mathbf{\Lambda}$ values up to the first simulated set that contained any failure to converge are presented here. For brevity $2^T = 16$ is omitted entirely, and the second-quartile η set for $2^T = 128$, where failure was common, is not shown. Such omission is redressed in the tables in the main text.

The failures reported in Table A1 were all due to reaching a preset maximum number of allowed iterations. Values not far from the true solution were obtained. By increasing the allowed number of iterations, convergence to desired precision would be obtained. This would of course increase the median and other flop counts reported.

Exceptionally algorithm SMϕ! can fail fatally (for ratios more extreme than those shown): either successive iterate values oscillate between two fixed sets of values, or a numeric error occurs when, for example, the logarithm of (machine) zero aborts the algorithm.

Full conversion to flop counts is not made in Table A1 as the count per iteration is roughly the same for all four algorithms, ignoring the small overhead of the accelerator steps. Flop counts rather than iterations are given below when comparing with other types of algorithm. The accelerated algorithms consistently offer fewer flop counts (including the cases not shown in the table) but are less robust. With few exceptions, SMϕ! is quicker than SMϕ_ϵ not only on average (as shown) but also for each set of odds ratios fitted. However the situation is different for the more extreme sets of ratios not shown here, when SMϕ! fails to converge more frequently than SMϕ_ϵ — and the latter's failures are all non-fatal, except for very extreme η, when all algorithms fail. Given that any speed advantage is minor, only the more robust SMϕ_ϵ is considered further.

Extremity index ϵ is a fair, but not excellent, indicator of algorithm performance. The maximum flop counts shown in Table A1 were not necessarily for those tables with the most extreme η within the given range, though η was in the upper half of the range.

The SM algorithm: implementation

```
/*
   Function infty_norm finds infinity norm of its input vector
*/

double infty_norm(v,vlen)
     double *v; long *vlen;
{
  double max;
  long i;
  max = 0;
  for (i = 0; i < *vlen; i++){
    if (max < fabs(v[i]) ) max = fabs(v[i]);
  }
  return max;
}
```

Table A1: Iterations to convergence to 6 d.p. of algorithms SM, SMϕ, SMϕ! and SMϕ_ϵ, for sets of 105 simulated odds ratios for each range of η indicated. All four algorithms were fitted to the same values. For each T, tabulation stops after the first set of ratios for which there was any failure. Minima (rows 'min') and medians (rows 'med') are close, so that lower quartiles are omitted for brevity. The upper quartile is abbreviated 'upper', maximum as 'max'. Summary counts are calculated only on those runs where there was convergence.

2^T	$\eta(\Lambda)$		SM	SMϕ	SMϕ!	SMϕ_ϵ
8	< 3.4	min	95	23	7	7
		med	95	38	10	10
		upper	95	41	11	13
		max	96	53	14	17
	$[3.4, 10.9)$	min	95	32	9	9
		med	95	49	14	17
		upper	96	56	17	20
		max	101	83	27	37
	$[10.9, 276)$	min	95	38	15	16
		med	106	97	33	46
		upper	129	127	48	64
		max	399	398	180	330
		fails	—	—	(1)	—
32	< 19.8	min	367	85	11	12
		med	367	140	18	18
		upper	367	162	22	23
		max	378	307	46	67
	$[19.8, 591)$	min	367	114	14	15
		med	369	228	38	46
		upper	425	320	82	96
		max	1949	998	738	935
		fails	(4)	(11)	(4)	(8)
64	< 93	min	705	159	12	13
		med	706	264	22	23
		upper	707	320	32	33
		max	742	660	75	206
	[93,11K)	min	706	178	17	19
		med	708	387	47	59
		upper	725	546	95	110
		max	1933	962	859	631
		fails	(16)	(24)	(6)	(20)
128	< 896	min	1346	305	17	17
		med	1347	468	27	29
		upper	1348	557	36	41
		max	1460	957	109	505

```
/*
   minvmulti(v, vlen, ans)
   Premultiply by M-inverse (not full efficiently here)
*/

void minvmulti(v,vlen,ans)
     double *v; double *ans; long *vlen;
{
  double *top, *bot, term1[*vlen], term2[*vlen];
  long halflen, termlen=*vlen/2, i, j;

  if (*vlen == 1) {
    *ans = *v;
  } else {
    top = v;
    halflen = *vlen/2;
    bot = v + halflen;
    minvmulti(top,&halflen,&term1[0]);
    minvmulti(bot,&halflen,&term2[0]);
    for ( i = 0, j = halflen; i < halflen; i++, j++) {
      ans[i] = (term1[i] - term2[i])/2.0;
      ans[j] = (term1[i] + term2[i])/2.0;
    }
  }
}


/*
   Seval(v,vlen,ans)
   Evaluate S(v) (unlogged version)
*/

void Seval(v,vlen,ans)
     double *v; long *vlen; double *ans;
{
  double *top, *bot, term1[*vlen/2], term2[*vlen/2], term3[*vlen/2];
  long i, halflen;

  if (*vlen == 1) {
    *ans = *v;
  } else {
    top = v;
    halflen = *vlen/2;
    bot = v + halflen;
    for (i = 0; i < halflen; i++)
      term1[i] = top[i] + bot[i];
    Seval(term1,&halflen,ans);
    Seval(bot,&halflen,term2);
    Seval(top,&halflen,term3);
    for (i = 0; i < halflen; i++)
      ans[i+halflen] = term2[i] / term3[i];
  }
}
```

The function seval called in smphi below is omitted for brevity; it simply passes the exponential of its argument vector to Seval above, and returns the logarithm of that answer.

```
/*
   smphi(lambda,vlen,ans,status)

   The SMphi_epsilon algorithm, all in C

*/

void smphi(lambda,vlen,epsi,maxiters,iters,ans,status,
           phi_epsi,phi_multi,multis, multicount)
     double *lambda; double *ans; double *epsi;
     double *phi_epsi; double *phi_multi; double *multis;
     long *vlen; long *maxiters; long *iters; long *status; long *multicount;
{
  double p_old[*vlen], c_old[*vlen], c_new[*vlen];
  double phi_old[*vlen], phi_new[*vlen];
  double workspace1[*vlen], landc[*vlen];
  double stop_difference;
  double rvec[*vlen], multidenoms[*vlen];
  long   i, changed;

  *multicount = 0;

  for (i = 0; i < *vlen; i++) {
    rvec[i] = 1 - (1/multis[i]);
    multidenoms[i] = 1 / (1 - rvec[i]);
  }

  minvmulti(lambda,vlen,ans); /* ans (p_new) holds M-1 lambda */
  seval(ans,vlen,workspace1); /* hold s(this) */
  for (i = 0; i < *vlen; i++) {
    c_new[i] = lambda[i] - workspace1[i];
    phi_new[i] = c_new[i];
    landc[i] = lambda[i] + c_new[i];
  }
  minvmulti(landc,vlen,ans);
  stop_difference = 1;
  *iters = 1;   /* in fact first iteration is cheaper */
  *status = 0;

  while ( (stop_difference >= *epsi) && (*iters < *maxiters) ) {
    for (i = 0; i < *vlen; i++) {
      p_old[i] = ans[i];
      c_old[i] = c_new[i];
      phi_old[i] = phi_new[i];
    }
    seval(ans,vlen,workspace1);
    for (i = 0; i < *vlen; i++) {
      c_new[i] = landc[i] - workspace1[i];
      phi_new[i] = c_new[i] - rvec[i]*c_old[i];
      if (fabs(phi_new[i] - phi_old[i]) <= phi_epsi[i]) {
        phi_epsi[i] = phi_epsi[i] * phi_multi[i];
```

```
        c_new[i] = phi_new[i] * multidenoms[i];
        *multicount = *multicount + 2;
      }
      landc[i] = lambda[i] + c_new[i];
    }
    minvmulti(landc,vlen,ans);
    for (i = 0; i < *vlen; i++)
      workspace1[i] = ans[i] - p_old[i];
    stop_difference = infty_norm(workspace1,vlen);
    *iters = *iters + 1;
  }
  if (*iters == *maxiters)
    *status = 1;
}
```

A3.5 The SR algorithms

Comparison of algorithms SR and SRquad

The SRquad algorithm of Section 3.5 seldom outperforms the raw SR algorithm for any of the simulations shown in Table A2, not only on average as tabulated, but also for most individual simulations. However, for extreme ratios, provided $2^T \leq 64$, SRquad converges more frequently. On the other hand, at $2^T = 128$, SRquad very rarely converges at all, while SR is comparatively robust. For $2^T \geq 256$ (discussed below) I have never observed SRquad to reach a second iteration before generating a fatal error.

The problem is numerical: some of the SRquad correction terms are calculated as machine zero, though this is not the analytic solution. The quadratic equation could here be solved by a more robust formula (Burden and Faires, 1985, p. 15). But as SRquad offers no speed gain over SR, this is not considered further.

Certain modifications to SRquad offer better speed performance in particular cases, especially for small T and non-extreme λ, but are far less robust in other circumstances. They are therefore omitted from discussion.

The failures shown in Table A2 are all non-fatal, in that the algorithms were stopped after 1000 iterations. Fatal failures become more common than non-fatal with increasing η (not shown in Table A2; but see main text).

As with the SM variants, since the flop counts are very similar, in later comparisons I use the most robust: the SR algorithm for $T \geq 7$, otherwise SRquad.

Table A2: Flop counts and number of iterations to convergence of algorithms SR and SRquad, to 6 d.p., for the same non-extreme odds ratios (and other considerations) as in Table A1, excepting $2^T = 128$ which is discussed in the text. Columns 'multi', 'adds' and 'iters' count multiplications and divisions, additions and subtractions, and iterations, respectively.

			SR			SRquad		
2^T	$\eta(\mathbf{\Lambda})$		multis	adds	iters	multis	adds	iters
8	< 3.4	min	287	195	5	319	147	3
		med	464	312	8	537	245	5
		upper	464	312	8	646	294	6
		max	641	429	11	755	343	7
	$[3.4, 10.9)$	min	346	234	6	428	196	4
		med	641	429	11	755	343	7
		upper	759	507	13	864	392	8
		max	1349	897	23	1191	539	11
	$[10.9, 276)$	min	582	390	10	537	245	5
		med	1290	858	22	1300	588	12
		upper	1939	1287	33	1945	882	18
		max	15863	10491	269	11982	5390	110
		fails			(1)			—
32	< 19.8	min	5188	3876	12	6318	3630	10
		med	6928	5168	16	8223	4719	13
		upper	8668	6460	20	9493	5445	15
		max	19534	14535	45	17113	9801	27
	$[19.8, 591)$	min	6058	4522	14	7588	4356	12
		med	13888	10336	32	14573	8349	23
		upper	26503	19703	61	36798	21054	58
		max	353K	262K	811	521K	298K	821
		fails			(11)			—
64	< 93	min	17591	13815	15	18860	12012	12
		med	27007	21183	23	31476	20020	20
		upper	37600	29472	32	39361	25025	25
		max	343K	268K	291	130K	83K	83
		fails			(2)			—

The SR algorithm: implementation

Both the SR and SQ algorithms call the `infty_norm` and `Seval` functions already listed in Appendix A3.4. SR is given rather than SRquad, for brevity.

```
/*
   Rinv(v,vlen,ans)

   Evaluate R-inverse (unlogged version)
*/

#include <S.h>
#include <math.h>

void  Rinv(v,vlen,ans)
     double *v; long *vlen; double *ans;
{
  double denom[*vlen/2];
  double vtop[*vlen/2];
  double vbot[*vlen/2];
  long halflen;
  long i;

  if (*vlen == 2) {
    denom[0] = 1 + v[1];
    ans[0] = v[0]/denom[0];
    ans[1] = ans[0]*v[1];
  } else {
    halflen = *vlen/2;
    for (i = 0; i < halflen; i++) {
      denom[i] = 1 + v[halflen + i];
      vtop[i] = v[i] / denom[i];
      vbot[i] = vtop[i] * v[halflen + i];
    }
    Rinv(vtop,&halflen,ans);
    Rinv(vbot,&halflen,ans+halflen);
  }
}


/*
        sralgor(Lambda,vlen,ans,status)

        The SR algorithm, unlogged, all in c

*/

void sralgor(Lambda,vlen,epsi,maxiters,iters,ans,status)
     double *Lambda; double *ans; double *epsi;
     long *vlen; long *maxiters; long *iters; long *status;
{
  double C_new[*vlen], C_old[*vlen];
  double workspace1[*vlen], workspace2[*vlen], stop_difference;
  long  i;
```

```
  Rinv(Lambda,vlen,workspace1);
  Seval(workspace1,vlen,workspace2);
  for (i = 0; i < *vlen; i++)
    C_new[i] = Lambda[i] / workspace2[i];
  stop_difference = 1;
  *iters = 1;
  *status = 0;

  while ( (stop_difference >= *epsi) && (*iters < *maxiters) ) {
    for (i = 0; i < *vlen; i++) {
      C_old[i] = C_new[i];
      C_new[i] = C_new[i] * Lambda[i]; /* note: overload use of C_new */
    }
    Rinv(C_new,vlen,workspace1);
    Seval(workspace1,vlen,workspace2);
    for (i = 0; i < *vlen; i++) {
      C_new[i] = C_new[i] / workspace2[i];
      workspace1[i] = C_new[i] - C_old[i];
    }
    stop_difference = infty_norm(workspace1,vlen);
    *iters = *iters + 1;
  }
  for(i = 0; i < *vlen; i++)
    workspace1[i] = C_new[i] * Lambda[i];
  Rinv(workspace1,vlen,ans);
  if (*iters == *maxiters)
    *status = 1;
}
```

SR algorithm: proof of convergence for $T = 2$

We show that the sequences of correction terms starting from $\mathbf{C}^{(0)} = \mathbf{1}$ are either constant or monotone increasing and bounded, and hence converge by an elementary proposition in mathematical analysis.
Iterations take the form

$$\mathbf{C}^{(n+1)} = \mathbf{g}(\mathbf{C}^{(n)}) \tag{A11}$$

and for such schemes the standard sufficient condition for convergence is that the infinity-norm of the Jacobian, i.e. $\sum_j \mid \partial g_i/\partial C_j \mid < 1$, for all i, in some interval about the root (Gerald and Wheatley, 1984). However, here I am unable to prove this relationship holds, and hence turn to the following direct proof of necessity of convergence.

Existence of a unique solution. Expansion of the bivariate system $SR^{-1}(\mathbf{\Lambda})$ shows — after some algebraic manipulation — that the elements Λ_0, Λ_2 and Λ_{12} are mapped to themselves, that is, the residual correction term for these components is simply unity. In other words, here $\mathbf{C} = (1, C_1, 1, 1)'$ and we need only consider the second element of $SR^{-1}(\mathbf{\Lambda C})$, denoted $[SR^{-1}(\mathbf{\Lambda C})]_1$ for iterations

$$C_1^{(n+1)} = \Lambda_1 C_1 / [SR^{-1}(\mathbf{\Lambda C})]_1 = \frac{1}{f_1(\mathbf{\Lambda C})},$$

where

$$f_1(\mathbf{\Lambda C}) = \left(\frac{a_1 \Lambda_1 C_1 + a_2}{a_3 \Lambda_1 C_1 + a_4}\right), \tag{A12}$$

where

$$\begin{aligned} a_1 &= (1 + \Lambda_2)\,\Lambda_{12}, \\ a_2 &= (1 + \Lambda_{12})\,(1 + \Lambda_2 \Lambda_{12}), \\ a_3 &= (1 + \Lambda_{12})\,(\Lambda_2 + \Lambda_{12}), \\ a_4 &= (1 + \Lambda_{12})^2\,(1 + \Lambda_2). \end{aligned}$$

The last of these is strictly redundant: $a_4 = a_2 + a_3$. At convergence, when $C_1 = C_1^\infty$ say, $C_1^\infty f_1(\mathbf{\Lambda C}^\infty) - 1 = 0$. Substituting expression (A12) into this equation and collecting terms gives the following quadratic in C_1:

$$a_1 \Lambda_1 C_1^2 + (a_2 - a_3 \Lambda_1) C_1 - a_4 = 0. \tag{A13}$$

This equation could of course be solved directly, giving us the answer to the problem without recourse to iteration. However we must first establish here that a real solution exists, and that such a solution gives a valid probability table.
The left-hand side of (A13) is negative when $C_1 = 0$, and as the leading coefficient of (A13) is positive, there must be one real positive root and one real negative root. The negative solution always yields an invalid solution to the system, since $R^{-1}(\mathbf{\Lambda})$ is not positive in all components when Λ_1 (or here $\Lambda_1 C_1$) is negative and the other components are positive. To show this, consider

$$R_2^{-1}(\mathbf{\Lambda C}) = \begin{pmatrix} \frac{\Lambda_0\,(1+\Lambda_{12})}{(1+\Lambda_2)(1+\Lambda_{12}+\Lambda_1 C_1)} \\ \frac{\Lambda_0\,\Lambda_1}{(1+\Lambda_2)(1+\Lambda_{12}+\Lambda_1 C_1)} \\ \frac{\Lambda_0\,\Lambda_2\,(1+\Lambda_{12})}{(1+\Lambda_2)(1+\Lambda_{12}+\Lambda_1 C_1 \Lambda_{12})} \\ \frac{\Lambda_0\,\Lambda_1 C_1\,\Lambda_2\,\Lambda_{12}}{(1+\Lambda_2)(1+\Lambda_{12}+\Lambda_1 C_1 \Lambda_{12})} \end{pmatrix}. \tag{A14}$$

If $1 + \Lambda_{12} + \Lambda_1 C_1$ is negative, the first and third components are negative, otherwise the second and fourth components must be negative. Thus, the positive solution of (A13) for $C_1 = C_1^\infty$ gives a unique solution to the constrained system.

Importantly for what follows, the solution C_1^∞ is not only positive, but also greater than unity; the left-hand side of (A13), evaluated at $C_1 = 1$, can be written

$$-(\Lambda_{12}^2 + \Lambda_2)(\Lambda_1 + 1) - \Lambda_{12}(\Lambda_2 + 1), \tag{A15}$$

which is negative for all $\Lambda > 0$.
Having established the existence of a valid solution, we now show that the sequence $\{C_1^{(n)}\}$ converges to it.

Magnitude of correction terms. *Starting at $C_1^{(0)} > 0$, consecutive C_1 terms generated by the algorithm are greater than unity.* This follows because $f_1()$ lies between zero and unity, for any positive $\boldsymbol{\Lambda}$ and C_1; the difference between denominator and numerator, after factorization, is the positive quantity

$$(\Lambda_2 + \Lambda_{12}^2)\Lambda_1 C_1 + \Lambda_2(1 + \Lambda_{12}) + \Lambda_{12}(1 + \Lambda_{12}). \tag{A16}$$

Thus each new correction term $C_1^{(n+1)} = 1/f_1(\boldsymbol{\Lambda}\mathbf{C}^{(n)})$ is greater than unity.

Monotonicity. *If $1 < C^{(n)} < C^\infty$ then $C^{(n)} < 1/f_1(\boldsymbol{\Lambda}\mathbf{C}^{(n)}) = C^{(n+1)}$.* As shown by the discussions of equations (A13) and (A15), here $C_1^{(n)} f_1(\boldsymbol{\Lambda}\mathbf{C}^{(n)}) - 1 < 0$ and monotonicity follows immediately.

Boundedness. *If $1 < C^{(n)} < C^\infty$ then $1 < C^{(n+1)} < C^\infty$.* Consider the expression

$$C^{(n+1)} f_1(\boldsymbol{\Lambda}\mathbf{C}^{(n+1)}) - 1 = \frac{1}{f_1(\boldsymbol{\Lambda}\mathbf{C}^{(n)})} f_1\left(\frac{\boldsymbol{\Lambda}}{f_1(\boldsymbol{\Lambda}\mathbf{C}^{(n)})}\right) - 1 = 0. \tag{A17}$$

Expanding $f_1(\boldsymbol{\Lambda}\mathbf{C}^{(n)})$ according to (A12) and collecting terms, this is

$$\Lambda_1^2 a_1(a_2a_3 - a_1a_4)C_1^2 + \Lambda_1(a_2a_3 - a_1a_4)(a_2 - a_3\Lambda_1)C_1 - \Lambda_1 a_4(a_2a_3 - a_1a_4) = 0 \tag{A18}$$

But this is precisely equation (A13), on division by the common factors Λ_1 and $(a_2a_3 - a_1a_4)$. This second factor can be written

$$(a_2a_3 - a_1a_4) = \lambda_2(\Lambda_{12}^2 - 1)^2. \tag{A19}$$

For independent variables, $\Lambda_{12} = 1$, (A18) is identically zero, and the division is invalid. However, in this special case convergence is achieved at $C_1^\infty = 1/f_1(\Lambda)$ after only one step, whenever $C_1^{(0)} = 1$, which is the usual starting value here.
Since (A18) and (A13) have the same roots, and the common factor, by (A19), is positive except in the special case just discussed, both functions change sign at the same values of C_1. Thus $C_1^{(n+1)} f_1(\boldsymbol{\Lambda}\mathbf{C}^{(n+1)}) - 1$, the left-hand side of (A18), is negative, which happens iff $C_1^{(n+1)} < C_1^\infty$, which completes the proof.
It can be shown similarly that starting from $C_1^{(0)} > C_1^\infty$, the sequence is monotone decreasing to its limit. Thus we obtain convergence starting at any positive choice of $C_1^{(0)}$.

Table A3: Flop counts for algorithms SQ, SQa and SQb until convergence of π components to 8 d.p. (though elsewhere convergence is only to 6 d.p.), for the same non-extreme odds ratios as in Table A1.

			SQ		SQa		SQb	
2^T	$\eta(\Lambda)$		multis	adds	multis	adds	multis	adds
8	< 3.4	min	307	116	294	183	240	148
		med	469	182	450	291	370	240
		upper	523	204	514	335	424	276
		max	631	248	622	407	516	340
	$[3.4, 10.9)$	min	415	160	440	283	305	194
		med	577	226	669	441	525	346
		upper	631	248	779	517	572	380
		max	1009	402	1378	923	865	582
	$[10.9, 276)$	min	523	204	542	355	361	234
		med	1009	402	1325	889	925	625
		upper	1333	534	1888	1272	1160	783
		max	2683	1084	4970	3331	2669	1826
		fails	(2)		(2)		(3)	
32	< 19.8	min	39330	16910	15290	10920	8071	5918
		med	75760	32370	26700	19040	11180	8200
		upper	105K	45K	35K	25K	12K	9K
		max	254K	108K	73K	51K	21K	16K
64	< 93	min	403K	175K	91K	66K	36K	27K
		med	1.0M	446K	180K	131K	51K	38K
		upper	1.7M	713K	306K	220K	63K	47K
		max	15.2M	6.4M	1.1M	801K	106K	78K
		fails	(18)		(4)		(0)	
128	< 896	min	1.6M	705K	556K	415K	86K	69K
		med	4.4M	1.9M	1.1M	786K	129K	103K
		upper	10.0M	4.5M	1.9M	1.4M	168K	133K
		max	110M	47M	6.6M	4.8M	425K	336K
		fails	(15)		(8)		(0)	

A3.6 The SQ algorithms

Comparison of the SQ algorithms

Referring to Table A3, for $2^T = 8$ it might seem surprising that algorithms SQa and SQb do not necessarily outperform unmodified SQ, since both modified algorithms call themselves recursively with start values based on previous iterations while SQ starts afresh from arbitrary values. However, for $T = 3$, the two halves of the system are solved exactly using SQ at each iteration, so that start values are not required. As T increases, the advantage of SQb becomes clear both in speed and robustness.

It is perhaps a surprise that SQb should be as superior to SQa as demonstrated in Table A3. The speed gain applies to almost every individual run, and is not merely on average as tabulated here. A trace of intermediate values shows that less 'outer' iterations are needed under SQb than SQa, where one might have expected the only speed gain to be in the computation of $S^{-1}(\mathbf{\Lambda} S(\boldsymbol{\pi_0}))$ to less precision.

A further gain for which full details are omitted is found in running the algorithms to greater precision. Attempting to run SQa to 15 d.p. generally fails, with successive iterate values oscillating between two sets, each accurate to within perhaps 13 d.p. of the solution. But SQb seldom if ever suffers from this, at least for non-extreme odds ratios.

The failures to converge reported in Table A3 are non-fatal, the error flag being raised on reaching the maximum allowed number of iterations (set at 50 for these simulations) in the outermost loop. But convergence was in fact very nearly obtained (to no worse than 5 d.p. precision).

All SQ variants can crash fatally in generating zero values and then attempting to divide by them. The relative frequencies of these types of failure (for SQb) are shown in the main text. Because it is up to 40 times quicker than SQ, 10 times quicker than SQa, and far more robust, only variant SQb is considered further below and in the main text.

The SQb algorithm: implementation

Code for the simpler SQ and SQa variants, a subset of the following, is omitted for brevity. The unlisted function `Rinvq` returns the exact solution for the bivariate case.

```
/*
   Algorithm SQb in C

   with cumulative flop counts

*/

void sqbalgor(lambda, starts, ans, vlen, epsi, divisor1, divisor2,
              starts_flag, upper_flag, maxiters, status, anyfail, multis,
              adds, trace)
     double *lambda, *starts, *ans, *epsi, *divisor1, *divisor2;
     long *vlen, *starts_flag, *upper_flag, *maxiters,
     *status, *anyfail, *multis, *adds, *trace;
/* all args are pointers, saves S-PLUS interface overhead
   ans = solution on completion
   vlen = length of Lambda vector
   epsi = required precision
   divisor1 = precision for sinvlam0, epsi/divisor1
   divisor2 = precision for S^{-1}(Lambda_2 S(\pi_0)), stop_difference/divisor2
   starts_flag = 1 if starts mean anything, eg if self-called;
   upper_flag = 1 if inverting constant lambda_top to full precision
   maxiters =
      max allowed iterations through loop (auto reduced to prevent 'hang')
   status = 0 if nominally ok (but this includes when NaN are returned),
            1 if maxiters reached/exceeded;
            from this and ans it is possible to distinguish other failure types
   anyfail = 1 set on return if any iters failure at any time
     note: status on return refers only to outer loop, overwriting
     any previous setting; anyfail remains set if ever set.
   multis = multiplications and divisions
   adds = additions and subtractions
   trace = flag, set if you want a screen trace
*/

{

/* in the following, 'a' is pi_0 in write-up, 'b' is pi_1
the newXXXXXX variables are needed to created new instances to be pointed
to for recursive calls
*/
  double xnew; /* not actually a vector, only need one component at a time */
  double stop_difference, newepsi, a_first;
  double tmp[*vlen/2], a_new[*vlen/2], a_old[*vlen/2], a_diffs[*vlen/2];
  double b_new[*vlen/2], b_old[*vlen/2];
  double sinvlam0[*vlen/2], newstarts[*vlen/2];
  long iters, i, capt, threet, halflen, newflag;
  long newmaxiters, nextmaxiters, newupper;
```

```
/* FOR 2-VECTORS, JUST FIND ANALYTIC SOLUTION:
*/

  if (*vlen == 2) {
    ans[0] = lambda[0] / (1 + lambda[1]);
    ans[1] = ans[0] * lambda[1];
    *multis = *multis + 2;
    *adds = *adds + 1;
    *status = 0;

/* FOR 4-VECTORS, FIND ANALYTIC ONLY IF LAMBDA_0, FIXED:
*/

  } else if (*upper_flag && *vlen==4) {

    Rinvq(lambda, vlen, ans);
    *multis = *multis + 37;
    *adds = *adds + 9;
    *status = 0;

/* ALL OTHER CASES:
*/

  } else {

    halflen = *vlen/2;
    if (*upper_flag) {
      newepsi = *epsi/(*divisor1);
    } else {
      newepsi = *epsi;
    }
    newflag = *starts_flag;
    newmaxiters = *maxiters;
    newupper = *upper_flag;

/* FIND INVERSE S^{-1){Lambda_0} (called sinvlam0)
*/

    if (newflag)
      for (i = 0; i < halflen; i++)
        newstarts[i] = starts[i] + starts[halflen+i];
    sqbalgor(lambda, newstarts, sinvlam0, &halflen, &newepsi,
             divisor1, divisor2, &newflag, &newupper, &newmaxiters,
             status, anyfail, multis, adds, trace);
    if (*status && newupper)
      printf("\n **** Warning: sinv(lam0) failure, halflen = %3d ****\n",
             halflen);

/* BEGIN TO SET UP FOR LOOPING
*/

    capt = log(halflen)/log(2); /* ie number of variables */
    threet = pow(3,capt); /* ie 3^T, for flop count */
```

```
    if (newflag) {
      for (i = 0; i < halflen; i++) {
        a_new[i] = starts[i];
        b_old[i] = starts[halflen+i];
      }
      newepsi = *epsi;
    } else {
      for (i = 0; i < halflen; i++) {
        a_new[i] = sinvlam0[i] / (1 + lambda[halflen]);
        b_old[i] = a_new[i] * lambda[halflen];
      }
      *multis = *multis + halflen + halflen;
      *adds = *adds + 1;
      newepsi = 0.01; /* only find 2 dp, just to kick off */
    }

/* DO THE FIRST ITERATION OUTSIDE THE LOOP TO SET IT UP
*/

    Seval(a_new, &halflen, tmp);
    for (i = 0; i < halflen; i++) {
      tmp[i] = lambda[halflen+i] * tmp[i];
    }
    newupper = 1; /* this will flag that we dn't want analytic for 4-vector */
    newflag = 1; /* use b_old from above as start values */
    sqbalgor(tmp, b_old, b_new, &halflen, &newepsi,
             divisor1, divisor2, &newflag, &newupper,
             &newmaxiters, status, anyfail, multis, adds, trace);
    *adds = *adds + threet;
    *multis = *multis + threet;
    iters = 1;
    stop_difference = 1;

/* NOW ENTER MAIN LOOP:
*/

    while (1) {

    /* we break out on stop_difference < epsi or iters =newmaxiters */

/* save old values, find a_new, and keep count */

      for (i = 0; i < halflen; i++) {
        a_old[i] = a_new[i];
        b_old[i] = b_new[i];
        xnew = b_old[i]/a_old[i];
        a_new[i] = sinvlam0[i] / (1 + xnew);
      }
      *adds = *adds + halflen;
      *multis = *multis + halflen + halflen;

/* now find S^{-1}(Lambda_2 * S(a)) */

      Seval(a_new, &halflen, tmp);
```

```
      for (i = 0; i < halflen; i++) {
        tmp[i] = lambda[halflen+i] * tmp[i];
        a_diffs[i] = a_new[i] - a_old[i];
      }
      stop_difference = infty_norm(a_diffs, &halflen);
      if (stop_difference < *epsi) break;
      newepsi = stop_difference / (*divisor2);
      if (newepsi >= 0.1) newepsi = 0.1;
      if (newepsi < *epsi/(*divisor1)) newepsi = *epsi/(*divisor1);
      nextmaxiters = newmaxiters - iters;
      sqbalgor(tmp, b_old, b_new, &halflen, &newepsi,
               divisor1, divisor2, &newflag, &newupper,
               &nextmaxiters, status, anyfail, multis, adds, trace);
      *adds = *adds + threet;
      *multis = *multis + threet;
      iters++;
      if (iters >= newmaxiters) break;

/* screen trace if required */

      if (*trace) {
        for (i=0; i < halflen; i++) {
          printf("%10.8f",a_new[i]);
        }
        printf("\n");
      }
    }}

/* END OF LOOP. TEST FOR CONVERGENCE STATUS THEN WRITE TO ANS:
*/
    if (iters >= newmaxiters) {
      *status = 1;
      *anyfail = 1;
    } else {
      *status = 0; /* but leave anyfail alone */
    }
    for (i = 0; i < halflen; i++) {
      ans[i] = a_new[i];
      ans[halflen+i] = b_new[i];
    }
  }
```

A5.3.3 Trough CyA example

The following fits were obtained for the models described in the main text.

```
===== MODEL 5A.1 =====

         param     SE
alpha_0  0.5831981 0.1018966
alpha_1 -1.0216512 0.3887301
alpha_2  0.5108256 0.4216370
alpha_3 -0.1431008 0.3789324
alpha_4  0.6931472 0.3396831

$deviance:
[1] 473.5809

===== MODEL 5A.2 =====

         param     SE
alpha_0  0.5831981 0.1018966
alpha_1 -1.1499672 0.3995642
alpha_2  0.4237940 0.4270334
alpha_3 -0.2726461 0.3901164
alpha_4  0.5696920 0.3464134
beta_1   0.2367043 0.1040708  # effect for low trough
beta_2   0.1182576 0.1178449  # effect for high trough

$deviance:
[1] 468.6298

===== MODEL 5A.3 =====

         param     SE
alpha_0  0.5831981 0.10189659
alpha_1 -1.0831102 0.39331738
alpha_2  0.4889126 0.42321617
alpha_3 -0.2137026 0.37989183
alpha_4  0.6260172 0.34373365
beta_1   0.1731713 0.08182806  # effect for low trough

$deviance:
[1] 469.6489

===== MODEL 5A.4 =====

         param      SE
alpha_0  0.58319807 0.1018966
alpha_1 -1.19694076 0.4529151
alpha_2  0.17218295 0.4907894
alpha_3  0.39053238 0.4425680
alpha_4  0.01626171 0.5333484
beta_11  0.31032651 0.2627729  # time-varying effect for low trough
beta_12  1.25381389 0.3109085
beta_13 -1.21880124 0.4136122
beta_14  0.61865843 0.5644120
```

```
beta_h1  0.16264428 0.2648334  # time-varying effect for high trough
beta_h2 -0.23638602 0.2729881
beta_h3  0.01783771 0.3221271
beta_h4  1.20297315 0.4664767

$deviance:
[1] 444.6433

===== MODEL 5A.5 =====

          param      SE
alpha_0  0.58319808 0.1018966
alpha_1 -1.02250470 0.3928144
alpha_2  0.55147614 0.4244982
alpha_3 -0.09838013 0.3989115
alpha_4  0.65990276 0.3479203
beta_1  -0.12503234 0.1108798  # time-varying linear trough effect
beta_2  -0.64415769 0.1500666
beta_3   0.36885414 0.1686123
beta_4   0.36935033 0.1972701

$deviance:
[1] 457.6116
```

A6.4.4 Imputation by parameter pre-specification

Cell counts h and m refer to unobserved and observed observations, respectively, in the example data presented in main text Table 6.1.

Pre-specification of PEF canonical parameters

Suppose we specify, arbitrarily, the IML parameters $\tilde{\xi}_1$ and $\tilde{\xi}_2$. Then it follows that

$$\begin{aligned} h_{000} &= \frac{m_{0+0}}{1+\tilde{\chi}_2} \\ h_{010} &= h_{000}\tilde{\chi}_2 \\ h_{100} &= h_{000}\tilde{\chi}_1 \\ h_{110} &= m_{1+0} - h_{100} \end{aligned}$$

where $\tilde{\chi} = \exp\{\tilde{\xi}\}$ are the unlogged CORs, this being the solution to the systems

$$\begin{aligned} h_{000} + h_{010} &= m_{0+0} \\ h_{010}/h_{000} &= \tilde{\chi}_2 \end{aligned}$$

and

$$\begin{aligned} h_{100} + h_{110} &= m_{1+0} \\ h_{100}/h_{000} &= \tilde{\chi}_1. \end{aligned}$$

Notice that $\tilde{\chi}_2$ determines h_{000}, so that $\tilde{\chi}_1$ is restricted by the choice of $\tilde{\chi}_2$, since we need

$$0 < \chi_1 < m_{1+0}/h_{000} = (1+\chi_2)\frac{m_{1+0}}{m_{0+0}}.$$

This example shows that even pre-specification of canonical parameters is subject to constraints; the curvature induced in the distribution by specification of the first chosen value is subtle.

Selection-model constraints

In the presentation of Diggle and Kenward (1994), the dropout interaction parameter, δ_{12} here, is always zero. Equivalently, in this context, $\xi_{123} = 0$ in the canonical parametrization. The derivation of this result is as follows. Dropping the tilde throughout and setting $\delta_{12} = 0$,

$$e^{\phi_{3|11}} = e^{\delta_0+\delta_1+\delta_2} = e^{\phi_{3|10}}e^{\delta_2} = e^{\phi_{3|10}}e^{\phi_{3|01}-\delta_0} = e^{\phi_{3|10}}\frac{e^{\phi_{3|01}}}{e^{\phi_{3|00}}}.$$

Hence

$$e^{\phi_{3|11}}e^{\phi_{3|00}} = e^{\phi_{3|10}}e^{\phi_{3|01}}, \tag{A20}$$

i.e. $\chi_{123} = 1$.

As a corollary: equation (A20) becomes, in terms of cell counts, for any admissible pre-specified filled-in values (here assumed to be the two cells h_{000} and h_{100}),

$$\frac{m_{111}}{m_{1+0}-h_{100}}\frac{m_{001}}{h_{000}} = \frac{m_{101}}{h_{100}}\frac{m_{011}}{m_{0+0}-h_{000}}.$$

Rearranging,

$$h_{000} = \frac{\chi^1_{12}m_{0+0}h_{100}}{m_{1+0}-(1-\chi^1_{12})h_{100}},$$

or

$$h_{100} = \frac{m_{1+0}h_{000}}{\chi^1_{12}m_{0+0} + (1-\chi^1_{12})h_{000}},$$

where

$$\chi^1_{12} = \frac{m_{111}m_{001}}{m_{011}m_{101}}$$

is the odds ratio of the observed part of the table, that is, the one-conditional rather than zero-conditional ratio.

Having restricted once by choosing $\tilde{\delta}_{12} = 0$, there is effectively only one degree of freedom remaining. The choice of an admissible h_{000} forces a unique value of h_{100} and vice versa. Provided

$$0 < h_{000} < m_{0+0},$$

h_{000} is admissible. Substituting from above and rearranging,

$$\begin{aligned} \chi^1_{12}m_{0+0}h_{100} &< m_{0+0}[m_{1+0} - (1-\chi^1_{12})h_{100}] \\ \Rightarrow h_{100} &< m_{1+0} \end{aligned}$$

which is the condition that h_{100} must satisfy in order to be a priori admissible.

In other words, admissibility of the one chosen cell, be it h_{000} or h_{100}, is sufficient to ensure admissibility of the rest of the table. This also means that there are still infinitely many ways of completing the table given only the Diggle and Kenward restriction $\delta_{12} = 0$, with uniqueness following only after further specification of any one cell.

As already mentioned, the assumption $\delta_2 = 0$ corresponds to a MAR model. The joint assumptions $\delta_{12} = \delta_1 = 0$ also give a unique table, this time with possibly informative dropout since δ_2 is now fixed by the pre-conditions and the data to some probably nonzero value. Little and Rubin (1987) highlight this model as being the only one of the hierarchical models from $\{Y_1Y_2R_2\}$ to $\{Y_1, Y_2, R_2\}$ with both informative dropout and identifiable parameters. Equivalently this is an arbitrary pre-specification, and the pre-conditions are not jointly verifiable in any way. The likelihood ratio between the full $\{Y_1Y_2R_2\}$ and this $\{Y_1Y_2, Y_2R_2\}$ is unity, as is the likelihood ratio between either of these and the much more simply estimated MAR assumption.

The joint assumptions $\delta_1 = \delta_2 = \delta_{12} = 0$ give a non-saturated model implying MCAR data.

Imputing MCAR and MAR tables

If the data satisfied the MCAR or MAR assumptions, R_2 would be independent of Y_2, conditional on Y_1 for MAR, and unconditionally for MCAR. In both cases, the EM imputed values for the missing y_2 would be given by the regression of the observed y_2 on the observed y_1 and, more generally, on other covariates, although there are none in this example.

In the current context, under MCAR, this is equivalent to taking the unobserved odds ratios equal to those in the observed m-cells, from which we estimate ξ_1, ξ_2 and ξ_{12}, which immediately gives IML estimates $\tilde{\xi}_{13} = \tilde{\xi}_{23} = \tilde{\xi}_{123} = 0$. These are in fact ordinary ML estimates for the MCAR model: IML implies saturation over observed cells but ML does not. Such IML estimates do not exist in general; only in special cases is it possible to specify CORs that match the observed half of the table *and* the observed marginal constraints. The IML approach is here attempting to impose three constraints given only two degrees of freedom. Of course, ordinary ML estimates always exist for the MCAR model.

Under the weaker MAR assumption

$$Y_2 \perp\!\!\!\perp R_2 \,|\, Y_1,$$

which becomes MCAR if also $Y_1 \perp\!\!\!\perp R_2$ marginally, IML forces

$$m_{011}/m_{001} = h_{010}/h_{000} \quad = \quad \chi_2 \text{ (observed)}$$

$$h_{010} + h_{000} = m_{0+0}$$

which uniquely determines these unobserved cell counts. Also

$$m_{111}/m_{101} = h_{110}/h_{100} = \text{observed}$$
$$h_{110} + h_{100} = m_{1+0}$$

determines the remaining unknown cells. Thus all the IML $\tilde{\pi}$, and hence $\tilde{\xi}$, estimates are fixed simply by assuming the data to be MAR. The table is unique.

Conversely — and importantly — we can thus *always* impute a MAR table perfectly fitting the observed data. This is true not only for this example but in full generality (Section 6.5).

Note that here $\tilde{\chi}_{23} = 1$ and $\tilde{\chi}_{123} = 1$, but $\tilde{\chi}_1$ is a function of the marginal totals m_{0+0} and m_{1+0}, and therefore not necessarily equal either to unity or to the odds ratio of the observed data; nor is the marginal odds ratio Λ_{23} necessarily equal to χ_{23} (i.e. unity); nor are the conditional $\tilde{\chi}_{12}$ or marginal Λ_{12} necessarily unity, nor need they be equal.

With the MCAR assumption we do not need to specify the trivariate distribution, since all inference follows from analysing the fully-observed data. The induced ξ (log χ) values are given here to indicate those values which must be excluded if the model is forced to incorporate informative missingness: that is, at least one of the above χ values must not hold.

A6.6 Timepoint-wise models for data with dropout

Labelled output for models fitted to the CHITC data set (Section 6.6) follows. In the printouts, `se(A)`, `se(I)` and `se(S)` are the square roots of the diagonals of the estimated ($\mathcal{A} = UU'$), observed expected ($\mathcal{I}$) and sandwich ($\mathcal{S}$) variance matrix estimators (Section 1.4.5).
Model numbering reflects the order in which the data models were introduced in the main text. Dropout model selection is in practice simultaneous so that in the following sometimes a modified dropout model appears before being discussed in the text. Since the parameters of the two models are always independent and the likelihoods separate, this should present no difficulty.
Matlab code used to fit these models follows.

```
========== Model 1 (null) ==========

    param     se(A)     se(I)     se(S)
    1.7086    0.0851    0.0851    0.0851 % alpha_10
    2.6832    0.1253    0.1253    0.1253 % alpha_11
   -1.2427    0.0628    0.0628    0.0628 % delta_21 = delta_31
   -1.3322    0.0650    0.0650    0.0650 % delta_22 = delta_32
    1.9349    0.0895    0.0895    0.0895 % alpha_20 = alpha_30
    2.7830    0.1268    0.1268    0.1268 % alpha_21 = alpha_31

elly = -1.0944e+03   ellr =  -1.5784e+03

========== Model 2 (independence) ==========

    param     se(A)     se(I)     se(S)
    1.7086    0.0851    0.0851    0.0851 % alpha_10
    2.6832    0.1253    0.1253    0.1253 % alpha_11
   -1.3580    0.0838    0.0838    0.0838 % delta_21
   -1.3089    0.0822    0.0822    0.0822 % delta_22
    2.0843    0.1208    0.1208    0.1208 % alpha_20
    3.0658    0.1837    0.1837    0.1837 % alpha_21
   -1.0827    0.0951    0.0951    0.0951 % delta_31
   -1.3704    0.1063    0.1063    0.1063 % delta_32
    1.7265    0.1336    0.1336    0.1336 % alpha_30
    2.4411    0.1762    0.1762    0.1762 % alpha_31

elly = -1.0913e+03  ellr = -1.5756e+03

========== Model 3 (simple lag 1) ==========

    param     se(A)     se(I)     se(S)
    1.7086    0.0851    0.0851    0.0851 % alpha_10
    2.6832    0.1253    0.1253    0.1253 % alpha_11
   -1.3922    0.0700    0.0695    0.0690 % delta_1
   -1.4016    0.0703    0.0699    0.0696 % delta_2
    2.2117    0.1069    0.1033    0.1001 % alpha_20
    3.1142    0.1383    0.1422    0.1466 % alpha_21
    0.6618    0.1059    0.1075    0.1095 % delta_h1
    0.3807    0.1238    0.1230    0.1226 % delta_h2
   -1.1467    0.1336    0.1357    0.1395 % alpha_h

elly = -1.0630e+03   ellr = -1.5594e+03
```

```
========== Model 4 (less simple linear lag 1) ===========

    param      se(A)      se(I)      se(S)
    1.7086     0.0851     0.0851     0.0851 % alpha_10
    2.6832     0.1253     0.1253     0.1253 % alpha_11
   -1.3922     0.0700     0.0695     0.0690 % delta_1
   -1.4016     0.0703     0.0699     0.0696 % delta_2
    2.2165     0.1071     0.1035     0.1003 % alpha_20
    3.0172     0.1494     0.1460     0.1429 % alpha_21
    0.6618     0.1059     0.1075     0.1095 % delta_h1
    0.3807     0.1238     0.1230     0.1226 % delta_h2
   -1.1964     0.1333     0.1410     0.1506 % alpha_h1
   -0.9591     0.1813     0.1721     0.1650 % alpha_h2

elly = -1.0619e+03   ellr = -1.5594e+03

========== Model 5 (lag 1 history as factor) ===========

    param      se(A)      se(I)      se(S)
    1.7086     0.0851     0.0851     0.0851 % alpha_10
    2.6832     0.1253     0.1253     0.1253 % alpha_11
   -1.4291     0.0710     0.0710     0.0710 % delta_1
   -1.4291     0.0710     0.0710     0.0710 % delta_2
    2.1608     0.0995     0.0999     0.1003 % alpha_20
    2.7830     0.1268     0.1268     0.1268 % alpha_21
    1.3338     0.2079     0.2079     0.2079 % delta_h11
    1.0053     0.2258     0.2258     0.2258 % delta_h21
   -2.2775     0.2988     0.2777     0.2585 % alpha_h1
    0.9312     0.2385     0.2385     0.2385 % delta_h12
    0.3305     0.2889     0.2889     0.2889 % delta_h22
   -1.0192     0.3278     0.3045     0.2832 % alpha_h2

elly = -1.0573e+03   ellr = -1.5502e+03

========== Model 6 (lag 1 history as factor) ===========

    param      se(A)      se(I)      se(S)
    1.7086     0.0851     0.0851     0.0851 % alpha_10
    2.6832     0.1253     0.1253     0.1253 % alpha_11
   -1.4291     0.0710     0.0710     0.0710 % delta_1
   -1.4291     0.0710     0.0710     0.0710 % delta_2
    2.3188     0.1093     0.1093     0.1093 % alpha_20
    3.0829     0.1524     0.1524     0.1524 % alpha_21
    1.3338     0.2079     0.2079     0.2079 % delta_h11
    1.0053     0.2258     0.2258     0.2258 % delta_h21
   -2.7242     0.2961     0.2961     0.2961 % alpha_h01
   -1.6966     0.3700     0.3700     0.3700 % alpha_h11
    0.9312     0.2385     0.2385     0.2385 % delta_h12
    0.3305     0.2889     0.2889     0.2889 % delta_h22
   -1.7126     0.3127     0.3127     0.3127 % alpha_h02
   -1.6719     0.3842     0.3842     0.3842 % alpha_h12

elly = -1.0431e+03   ellr = -1.5502e+03
```

```
========== Model 7 (lag 1 history, plus period) ==========

    param      se(A)      se(I)      se(S)
    1.7086     0.0851     0.0851     0.0851 % alpha_10
    2.6832     0.1253     0.1253     0.1253 % alpha_11
   -1.4291     0.0710     0.0710     0.0710 % delta_1
   -1.4291     0.0710     0.0710     0.0710 % delta_2
    2.5885     0.1494     0.1489     0.1484 % alpha_21
    3.4610     0.2189     0.2147     0.2106 % alpha_22
    1.3338     0.2079     0.2079     0.2079 % delta_h11
    1.0053     0.2258     0.2258     0.2258 % delta_h21
   -2.8485     0.3042     0.3030     0.3019 % alpha_h01
   -1.8343     0.3775     0.3779     0.3785 % alpha_h11
    0.9312     0.2385     0.2385     0.2385 % delta_h12
    0.3305     0.2889     0.2889     0.2889 % delta_h22
   -1.7829     0.3174     0.3171     0.3171 % alpha_h02
   -1.7536     0.3861     0.3889     0.3922 % alpha_h12
   -0.5927     0.1959     0.1954     0.1951 % alpha_p0
   -0.7851     0.2654     0.2637     0.2622 % alpha_p1

elly = -1.0377e+0    ellr = -1.5502e+03

========== Model 8 (lag 1 data saturated) ==========

    param      se(A)      se(I)      se(S)
    1.7086     0.0851     0.0851     0.0851 % alpha_10
    2.6832     0.1253     0.1253     0.1253 % alpha_11
   -1.4291     0.0710     0.0710     0.0710 % delta_1
   -1.4291     0.0710     0.0710     0.0710 % delta_2
    2.5966     0.1581     0.1581     0.1581 % alpha_20
    3.5099     0.2392     0.2392     0.2392 % alpha_21
    1.3338     0.2079     0.2079     0.2079 % delta_h11
    1.0053     0.2258     0.2258     0.2258 % delta_h21
   -2.8418     0.3522     0.3522     0.3522 % alpha_h201
   -1.9294     0.4790     0.4790     0.4790 % alpha_h211
    0.9312     0.2385     0.2385     0.2385 % delta_h12
    0.3305     0.2889     0.2889     0.2889 % delta_h22
   -1.8165     0.3969     0.3969     0.3969 % alpha_h202
   -1.9343     0.5083     0.5083     0.5083 % alpha_h212
   -0.6079     0.2195     0.2195     0.2195 % alpha_p0
   -0.8656     0.3113     0.3113     0.3113 % alpha_p1
   -2.9050     0.6109     0.6109     0.6109 % alpha_h301
   -1.7280     0.6242     0.6242     0.6242 % alpha_h311
   -1.7374     0.5265     0.5265     0.5265 % alpha_h302
   -1.5457     0.6107     0.6107     0.6107 % alpha_h312

elly = -1.0375e+03   ellr = -1.5502e+03

========== Model 9 (lag 2 with period) ==========

    param      se(A)      se(I)      se(S)
    1.7086     0.0851     0.0851     0.0851 % alpha_10
    2.6832     0.1253     0.1253     0.1253 % alpha_11
   -1.4291     0.0710     0.0710     0.0710 % delta_1
```

```
   -1.4291    0.0710    0.0710    0.0710 % delta_2
    2.5310    0.1500    0.1467    0.1436 % alpha_20
    3.3704    0.2200    0.2096    0.1997 % alpha_21
    1.3338    0.2079    0.2079    0.2079 % delta_h11
    1.0053    0.2258    0.2258    0.2258 % delta_h21
   -2.7164    0.3071    0.3085    0.3103 % alpha_h01
   -1.5704    0.3937    0.4013    0.4109 % alpha_h11
    0.9312    0.2385    0.2385    0.2385 % delta_h12
    0.3305    0.2889    0.2889    0.2889 % delta_h22
   -1.5534    0.3265    0.3326    0.3406 % alpha_h02
   -1.4732    0.3776    0.4116    0.4515 % alpha_h12
   -2.0079    0.4315    0.4661    0.5062 % alpha_p0
   -1.8248    0.4928    0.5107    0.5294 % alpha_p1
   -0.2626    0.2156    0.2135    0.2118 % alpha_hh1
   -0.3510    0.2913    0.2861    0.2819 % alpha_hh2

elly = -1.0249e+03   ellr = -1.5502e+03

========== Model 10 (lag 2, no period) ==========

    param     se(A)     se(I)     se(S)
    1.7086    0.0851    0.0851    0.0851 % alpha_10
    2.6832    0.1253    0.1253    0.1253 % alpha_11
   -1.4291    0.0710    0.0710    0.0710 % delta_1
   -1.4291    0.0710    0.0710    0.0710 % delta_2
    2.4263    0.1157    0.1145    0.1134 % alpha_20
    3.2205    0.1623    0.1584    0.1546 % alpha_21
    1.3338    0.2079    0.2079    0.2079 % delta_h11
    1.0053    0.2258    0.2258    0.2258 % delta_h21
   -2.6591    0.3004    0.3031    0.3060 % alpha_h01
   -1.5006    0.3851    0.3943    0.4041 % alpha_h11
    0.9312    0.2385    0.2385    0.2385 % delta_h12
    0.3305    0.2889    0.2889    0.2889 % delta_h22
   -1.5111    0.3236    0.3299    0.3375 % alpha_h02
   -1.4230    0.3767    0.4084    0.4447 % alpha_h12
   -2.2174    0.4002    0.4381    0.4813 % alpha_hh1
   -2.0033    0.4746    0.4923    0.5109 % alpha_hh2

elly = -1.0258e+03   ellr = -1.5502e+03

========== Model 11 (lag 2, non proportional) ==========

    param     se(A)     se(I)     se(S)
    1.7086    0.0851    0.0851    0.0851 % alpha_10
    2.6832    0.1253    0.1253    0.1253 % alpha_11
   -1.4291    0.0710    0.0710    0.0710 % delta_1
   -1.4291    0.0710    0.0710    0.0710 % delta_2
    2.4252    0.1158    0.1146    0.1134 % alpha_20
    3.2168    0.1655    0.1610    0.1567 % alpha_21
    1.3338    0.2079    0.2079    0.2079 % delta_h11
    1.0053    0.2258    0.2258    0.2258 % delta_h21
   -2.6608    0.3013    0.3030    0.3051 % alpha_h01
   -1.4668    0.3842    0.3978    0.4126 % alpha_h11
    0.9312    0.2385    0.2385    0.2385 % delta_h12
```

```
 0.3305    0.2889    0.2889    0.2889 % delta_h22
-1.5143    0.3250    0.3296    0.3366 % alpha_h02
-1.3967    0.3761    0.4100    0.4504 % alpha_h12
-2.0620    0.4721    0.4813    0.4966 % alpha_hh01
-2.3471    0.4410    0.4838    0.5320 % alpha_hh11
-2.0827    0.4951    0.5091    0.5257 % alpha_hh02
-1.8330    0.5547    0.6051    0.6615 % alpha_hh12

elly = -1.0253e+03   ellr = -1.5502e+03

========== Model 12 (final model) ==========

   param     se(A)     se(I)     se(S)
 1.7086    0.0851    0.0851    0.0851 % alpha_10
 2.6832    0.1253    0.1253    0.1253 % alpha_11
-2.0480    0.1130    0.1126    0.1123 % delta_1
-1.2880    0.0845    0.0846    0.0848 % delta_2
 2.5122    0.1441    0.1395    0.1360 % alpha_20
 3.3261    0.1766    0.1793    0.1843 % alpha_21
 1.1295    0.2187    0.2195    0.2203 % delta_h11
 0.9580    0.2278    0.2268    0.2260 % delta_h21
-2.4604    0.3330    0.3169    0.3038 % alpha_h01
-1.1970    0.3903    0.4068    0.4263 % alpha_h11
 0.6686    0.2453    0.2527    0.2609 % delta_h12
 0.3125    0.2890    0.2896    0.2905 % delta_h22
-1.3775    0.3409    0.3441    0.3494 % alpha_h02
-1.2660    0.3878    0.4182    0.4555 % alpha_h12
-1.9708    0.4100    0.4482    0.4920 % alpha_hh1
-1.7026    0.5047    0.5030    0.5033 % alpha_hh2
-0.6114    0.1581    0.1604    0.1648 % beta_a
-0.9458    0.3037    0.2908    0.2850 % beta_d2
 0.7459    0.2740    0.2858    0.2996 % beta_s2
 0.9559    0.1010    0.1030    0.1052 % delta_a1
 0.7146    0.1357    0.1353    0.1351 % delta_s1
-0.4019    0.1422    0.1420    0.1420 % delta_s2

elly = -1.0090e+03   ellr = -1.4887e+03
```

The above models were fitted in Matlab using the following program. Matlab was chosen over Splus because of its speed advantage when running native code; even so models took up to 10 minutes to fit (the program is not, however, by any means optimized as yet).
The routine `hcevals` is not given here as it is the same as `hceval` except that it is stripped of comments and does not calculate the likelihood, which is not needed until convergence. The routine `alogit` returns the inverse logit.
The user needs to write the program with name passed to `eta_eval` according to the desired model. An example is given below. This part of the program suite is very inefficient and not at all user friendly.
The Y and R data, in columns ordered by time as in the main text, need to assigned globally to a variable named `y`. Explanatory variables also need to be assigned globally (as `x`) for my sample logit evaluation routine.

```
function [table,elly,ellr] = ...
hcfit(startgamma,N,k,indicators,T,links,eta_eval)
%
```

```
%   HCFIT --- LOOP THROUGH HCEVALS AND HCEVAL UNTIL CONVERGENCE (OR NOT)

oldgamma=zeros(size(startgamma));
iters=0;
okflag=1;

while sum(abs(startgamma - oldgamma)) > 0.001
    difference = sum(abs(startgamma - oldgamma))
    [u,inform] = ...
        hcevals(startgamma,N,k,indicators,T,links,eta_eval);
    oldgamma = startgamma;
    startgamma = startgamma + inform\u';
    iters=iters+1;
    if iters==20
        ell = NaN;
        okflag = 0;
        break;
    end
end

if okflag==1
    [u,uut,inform,elly,ellr] = ...
        hceval(startgamma,N,k,indicators,T,links,eta_eval);
    table = [ startgamma sqrt(diag(inv(uut))) sqrt(diag(inv(inform))) ...
        sqrt(diag(inform\uut/inform)) ];
else
    table = [];
end


function [u,uut,inform,elly,ellr] = ...
hceval(gamma,N,k,indicators,T,links,eta_eval)
%
% HCEVAL
%
% evaluate multivariate score etc
% for the mixed/cumulative link model, and sum of uu' for pseudo info;
% and true information matrix.
%
% log likelihood returned split into y and r models contributions
%
% gamma is vector of parameters at which to evaluate ell
% N is sample size
% k is a vector giving number of classes for timepoints 1,2,...,T
% indicators is a vector coded 0 for no dropout, 1 for dropout-indicator,
%   for each of the T variables
%   (ie stop processing loop after including this term if > 0)
% T is number of timepoints (for a full cluster; constant)
% links is a character vector, holding 'c' for cumulative link,
%   else anything
% eta_eval is a user-written function which must return a vector
%   of linear predictors and the matrix Xi for cluster i, called by
%   'eta_eval(i,gamma)'
%
```

```
% observed data and covariates must be pre-assigned global

global y

gammalength = length(gamma);
u = zeros(1,gammalength);   %eventually holds the score function
uut = u*u';     %eventually holds the pseudo-info
inform = uut;   %eventually holds the true info
elly = 0;        %eventually holds the likelihood evaluation
ellr = 0;

for i = 1:N

    [Xi,eta] = eval([eta_eval '(i, gamma)']);   % user evaluate all this
                                                % cluster's linear predictors


    % general housekeeping ----------------
    b_build = zeros(size(eta))';    % eventually to hold b' (see notes)
    mark = 1;       % initialize for vector-index subscripts
    uutmp = zeros(length(eta),length(eta)); % temp space to build this
    inftmp = uutmp;             % clusters uu and info matrices
    dropout = 0;        % no dropout time zero
    % -------------------------------------

    for h = 1:T
        kh = k(h);
        k_less_one = kh - 1;
        k_less_two = kh - 2;
        indices = mark:(mark+k_less_two);
        mark = mark + k_less_one;

        b = zeros(k_less_one,1)';           %will contribute to b_build
        e = zeros(k_less_one,k_less_one);   %and inftmp

        j = y(i,h);     % read the y value, coded in j notation

        eta_h = eta(indices); % extract relevant bits

        if links(h) == 'c'

        % ***************** CUMULATIVE LOGIT LINKS ***************

            gvec = alogit(eta_h);   % find G(eta) for time h
            delta = gvec .* (1 - gvec);

            if j == 0
                b(1) = (1 - gvec(1));
            elseif j == k_less_one
                b(k_less_one) = - gvec(k_less_one);
            else
                p = gvec(j+1) - gvec(j);
                b(j) = - 1 ./p.*gvec(j).*(1 - gvec(j));
                b(j+1) = 1 ./p.*gvec(j+1).*(1-gvec(j+1));
```

```
        end

        ph = gvec - [0 gvec(1:(k_less_two))']'; % NB p above redundant
        ph = [ph' 1-gvec(k_less_one)]';     % vector of p values
        pis = log(ph);                      % and their logs (pi values)
        iph = 1 ./ ph;                      % and their inverses

        e(1,1) = iph(1) + iph(2);
        if kh > 2
            e(1,2) = - iph(2);
            e(k_less_one,k_less_two) = - iph(k_less_one);
            e(k_less_one,k_less_one) = iph(k_less_one) + iph(kh);
            for hi = 2:(k_less_two)
                e(hi,hi-1) = - iph(hi);
                e(hi,hi) = iph(hi) + iph(hi+1);
                e(hi,hi+1) = - iph(hi+1);
            end
        end

        inftmp(indices,indices) = e .* (delta*delta');

    else

% *************** UNORDERED MULTINOMIAL LOGIT LINKS *********

        % convert from eta (equiv \xi) to p
        A = log( 1 + sum(exp(eta_h)) );
        p = exp( eta_h - A )'; % note this has p_1 as p(1)
        pis = [-A log(p)]';      % note this contains pi_0 as pis(1)

        % find b = (y - p)
        b = -p;
        if j > 0
            b(j) = b(j) + 1;
        end

        % the innards of the info contribution
        inftmp(indices,indices) = diag(p) - p'*p;

    end

% ****************** COMMON TO EITHER TYPE OF LINK ***********

    uutmp(indices,indices) = b' * b;
    b_build(indices) = b;

    ell = ell + pis(j+1);

    if indicators(h) == 1
        ellr = ellr + pis(j+1);
        if j > 0
            break
        end
    else
```

```
            elly = elly + pis(j+1);
        end

    end

    evalu = b_build * Xi;
    u = u + evalu;
    uut = uut + Xi' * uutmp * Xi;
    inform = inform + Xi' * inftmp * Xi;
end

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function [Xi, eta] = etas47(i,gamma)
% final model matrix

global y x

Xi = zeros(10,22);
Xi(1:6,1:6)=eye(6);
Xi(7:8,3:4)=eye(2);
Xi(9:10,5:6)=eye(2);

if y(i,1)==1
    Xi(3:6,7:10)=eye(4);
    Xi(9:10,15)=ones(2,1);
elseif y(i,1)==2
    Xi(3:6,11:14)=eye(4);
    Xi(9:10,16)=ones(2,1);
end

if y(i,3)==1
    Xi(7:10,7:10)=eye(4);
elseif y(i,3)==2
    Xi(7:10,11:14)=eye(4);
end

Xi(5:6,17)=ones(2,1).*(x(i,1)-2)./10;
Xi(9:10,17)=ones(2,1).*(x(i,1)+1)./10;
Xi(5:6,18)=ones(2,1).*x(i,3);
Xi(5:6,19)=ones(2,1).*(x(i,2)-1);
Xi(3,20)=(x(i,1)-2)./10;
Xi(7,20)=(x(i,1)+1)./10;
Xi(3:4,21:22)=eye(2).*(x(i,2)-1);
Xi(7:8,21:22)=eye(2).*(x(i,2)-1);

eta = Xi*gamma;

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

References

Agresti, A. (1989). A survey of models for repeated ordered categorical response data. *Statistics in Medicine*, **8,** 1209–1224.

Agresti, A. (1990). *Categorical data analysis.* Wiley, New York.

Agresti, A. (1993). Distribution-free fitting of logit models with random effects for repeated categorical responses. *Statistics in Medicine*, **12,** 1969–1987.

Agresti, A. and Lang, J.B. (1993). A proportional odds model with subject-specific effects for repeated ordered categorical responses. *Biometrika*, **80,** 527–534.

Aitken, M., Anderson, D., Francis, B. and Hinde, J. (1989). *Statistical Modelling in GLIM.* Clarendon Press, Oxford.

Anderson, D.A. and Aitkin, M. (1985). Variance component models with binary response: interviewer variability. *Journal of the Royal Statistical Society* B, **85,** 203–210.

Azzalini, A. (1994). Logistic regression for autocorrelated data with application to repeated measures. *Biometrika*, **81,** 767–775.

Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W. (1975). *Discrete Multivariate Analysis.* M.I.T. Press, Cambridge, Mass.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88,** 9–25.

Buist, A.S. and Vollmer, W.M. (1988). The use of lung function tests in identifying factors that affect lung growth and aging. *Statistics in Medicine*, **7,** 11–18.

Burden, R.L. and Faires, J.D. (1985). *Numerical analysis*, 3rd edn. Prindle, Weber & Schmidt, Boston.

Conaway, M.R. (1989). Analysis of repeated categorical measurements with conditional likelihood methods. *Journal of the American Statistical Association*, **84**, 53–62.

Conaway, M.R. (1990). A random effects model for binary data. *Biometrics*, **46**, 317–328.

Conaway, M.R. (1992). The analysis of repeated categorical measurements subject to nonignorable nonresponse. *Journal of the American Statistical Association*, **87**, 817–824.

Copeland, J.R.M., Davidson, I.A., Dewey, M.E., Gilmore, C., Larkin, B.A., McWilliam, C., Scott, A., Sharma, V. and Sullivan, C. (1992). Alzheimer's disease, other dementias, depression and pseudo-dementia: prevalence, incidence and three-year outcome in Liverpool. *British Journal of Psychiatry*, **161**, 230–239.

Copeland, J.R.M., Dewey, M.E. and Griffiths-Jones, H.M. (1986). Psychiatric case nomenclature and a computerised diagnostic system for elderly subjects: GMS and AGECAT. *Psychological Medicine*, **16**, 89–99.

Copeland, J.R.M., Dewey, M.E., Wood, N., Searle, R., Davidson, I.A. and McWilliam, C. (1987). Range of mental illness amongst the elderly in the community: prevalence in Liverpool using the GMS-ASGECAT package. *British Journal of Psychiatry*, **150**, 815–823.

Crowder, M.J. (1978). Beta-binomial ANOVA for proportions. *Journal of the Royal Statistical Society* C, **27**, 34–37.

Crowder, M.J. (1985). Gaussian estimation for correlated binomial data. *Journal of the Royal Statistical Society* B, **47**, 229–237.

Crowder, M.J. and Hand, D.J. (1990). *Analysis of Repeated Measures.* Chapman & Hall, London.

Darroch, J.N. (1962). Interaction in multifactor contingency tables. *Journal of the Royal Statistical Society* B, **24,** 251–263.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* B, **39,** 1–38.

Diem, J.E. and Liukkonen, J.R. (1988). A comparative study of three methods for analysing longitudinal pulmonary function data. *Statistics in Medicine*, **7,** 19–28.

Diggle, P. and Kenward, M.G. (1994). Informative drop-out in longitudinal data analysis (with discussion). *Journal of the Royal Statistical Society* C, **43,** 49–93.

Diggle, P.J., Liang, K.-L. and Zeger, S.L. (1994). *Analysis of longitudinal data.* Oxford Statistical Science Series 13. Clarendon Press, Oxford.

Dobbs, D.E. (1980). *A modern course on the theory of equations.* Polygonal Publishing, Passaic, New Jersey.

Ekholm, A. and Green, M. (1994). Fitting nonlinear models in GLIM4 using numerical derivatives. *GLIM Newsletter*, **23,** 12–20.

Ekholm, A., Smith, P.W.F. and McDonald, J.W. (1995). Marginal regression analysis of a multivariate binary response. *Biometrika*, **82,** 847–854.

Fitzmaurice, G.M. and Laird, N.M. (1993). A likelihood-based method for analysing longitudinal binary repsonses. *Biometrika*, **80,** 141–151.

Fitzmaurice, G.M., Laird, N.M. and Lipsitz, S.R. (1994). Analyzing incomplete longitudinal binary responses — a likeliohood-based approach. *Biometrics*, **50,** 601–612.

Fitzmaurice, G.M., Laird, N.M. and Rotnitzky, A.G. (1993). Regression models for discrete longitudinal responses. *Statistical Science*, **8,** 284–309.

Gerald, C.F. and Wheatley, P.O. (1984). *Applied numerical analysis*, 3rd edn. Addison–Wesley, Reading, MA.

Gilmour, A.R., Anderson, R.D. and Rae, A.L. (1985). The analysis of binomial data by a generalized linear mixed model. *Biometrika*, **72,** 593–599.

Glonek, G.F.V. and McCullagh, P. (1994). Multivariate logistic models. Technical Report. School of Information Science and Technology, The Flinders University of South Australia, G.P.O. Box 2100, Adelaide, SA, 5001.

Glonek, G.F.V. and McCullagh, P. (1995). Multivariate logistic models. *Journal of the Royal Statistical Society* B, **57,** 533–546.

Gourieroux, C., Monfort, A. and Trognon, A. (1984). Pseudomaximum likelihood methods: theory. *Econometrica*, **52,** 681–700.

Griffiths, D.A. (1973). Maximum likelihood esitation for the beta-binomial distribution and an application to the household distribution of the total number of cases of disease. *Biometrics*, **29,** 637–648.

Grizzle, J.E., Starmer, C.F. and Koch, G.G. (1969). Analysis of categorical data by linear models. *Biometrics*, **25,** 489–504.

Hanfelt, J.J. and Liang, K.-Y. (1995). Approximate likelihood ratios for general estimating functions. *Biometrika*, **82,** 461–477.

Jones, B. and Kenward, M. (1989). *Design and Analysis of Cross-over Trials.* Chapman & Hall, London.

Koch, G.G., Landis, J.R., Freeman, J.L. and Freeman, D.H. Jr (1977). A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics*, **33,** 133–158.

Korn, E.L. and Whittemore, A.S. (1979). Methods for analyzing panel studies of acute health effects of air pollution. *Biometrics*, **35,** 795–802.

Laird, N.M. and Ware, J.H. (1982). Random-effects models for longitudinal data. *Biometrics*, **38,** 963–974.

Landis, J.R., Miller, M.E., Davis, C.S. and Koch, G.G. (1988). Some general methods for the analysis of categorical data in longitudinal studies. *Statistics in Medicine*, **7,** 109–137.

Lee, Y. and Nelder, J.A. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society* B, **58,** 619–678.

Li, B. (1993). A deviance function for the quasi-likelihood method. *Biometrika*, **80,** 741–753.

Li, B. and McCullagh, P. (1994). Potential functions and conservative estimating functions. *Annals of Statistics*, **22,** 340–356.

Liang, K-L. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73,** 13–22.

Liang, K.Y., Zeger, S.L. and Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society* B, **54,** 3–40.

Lindsey, J.K (1993). *Models for Repeated Measurements.* Clarendon Press, Oxford.

Lipsitz, S.R., Laird, N.M. and Harrington, D.P. (1991). Generalized estimating equations for correlated binary data — using the odds ratio as a measure of association. *Biometrika*, **91,** 153–160.

Little, R.J.A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, **90,** 1112–1121.

Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data.* Wiley, New York.

Louis, T.A. (1988). General methods for analysing repeated measures. *Statistics in Medicine*, **7,** 29–45.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society* B, **42,** 109–142.

McCullagh, P. (1989). Models for discrete multivariate responses. *Bulletin of the International Statistical Association*, **53**, 407–418.

McCullagh, P. (1994). Exponential mixtures and quadratic exponential families. *Biometrika*, **81**, 721–729.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*, 2nd edn. Chapman & Hall, London.

McLeish, D.L. and Small, C.G. (1992). A projected likelihood function for semiparametric models. *Biometrika*, **79**, 93–102.

Molenberghs, G., Kenward, M.G. and Lesaffre, E. (1997). The analysis of longitudinal ordinal data with nonrandom dropout. *Biometrika*, **84**, 33–44.

Molenberghs, G. and Lesaffre, E. (1994). Marginal modelling of correlated ordinal data using a multivariate plackett distribution. *Journal of the American Statistical Association*, **89**, 633–644.

Neuhaus, J.M. (1992). Statistical methods for longitudinal and clustered designs with binary responses. *Statistical Methods in Medical Research*, **1**, 249–273.

Neuhaus, J.M., Kalbfleisch, J.D. and Hauck, W.W. (1991). A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review*, **59**, 23–35.

Owen, A.B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**, 237–249.

Owen, A.B. (1991). Empirical likelihood for linear models. *Annals of Statistics*, **19**, 1725–1747.

Prentice, R.L. (1986). Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. *Journal of the American Statistical Association*, **81**, 321–327.

Prentice, R.L. and Zhao, L.P. (1991). Estimating equations for parameters in means and covriances of multivariate discrete and continuous responses. *Biometrics*, **47,** 825–839.

Qu, Y., Piedmonte, M.R. and Medendorp, S.V. (1995). Latent variable models for clustered ordinal data. *Biometrics*, **51,** 268–275.

Qu, Y., Williams, G.W., Beck, G.J. and Medendorp, S.V. (1992). Latent variable models for clustered dichotomous data with multiple subclusters. *Biometrics*, **48,** 1095–1102.

Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, **90,** 106–121.

Rochon, J. (1996). Accounting for covariates observed post randomization for discrete and continuous repeated measures data. *Journal of the Royal Statistical Society* B, **58,** 205–219.

Rosner, B. (1984). Multivariate methods in opthalmology with application to other paired data situations. *Biometrics*, **40,** 1025–1035.

Rosner, B. (1989). Multivariate methods for clustered binary data with more than one level of nesting. *Journal of the American Statistical Association*, **84,** 373–380.

Rosner, B. (1992a). Multivariate methods for binary longitudinal data with heterogeneous correlation over time. *Statistics in Medicine*, **11,** 1915–1928.

Rosner, B. (1992b). Multivariate methods for clustered bibary data with multiple subclasses, with application to binary longitudinal data. *Biometrics*, **48,** 721–731.

Rosner, B. (1992c). Multivariate methods for clustered binary data with multiple subclasses, with application to binary longitudinal data. *Biometrics*, **48,** 721–731.

Skellam, J.G. (1948). A probability distribution derived from the binomial distribution

by regarding the probability of success as variable between the sets of trials. *Journal of the Royal Statistical Society* B, **10,** 257–261.

Stiratelli, R., Laird, N. and Ware, J.H. (1984). Random-effects models for serial observations with binary response. *Biometrics*, **40,** 961–971.

Stram, D.O., Wei, L.J. and Ware, J.H. (1988). Analysis of repeated ordered categorical outcomes with possibly missing observations and time-dependent covariates. *Journal of the American Statistical Association*, **83,** 631–637.

Tsou, T.-S. and Royall, R.M. (1995). Robust likelihoods. *Journal of the American Statistical Association*, **90,** 316–320.

Ware, J.H., Lipsitz, S. and Speizer, F.E. (1988). Issues in the analysis of repeated categorical outcomes. *Statistics in Medicine*, **7,** 95–107.

Wei, L.J. and Stram, D.O. (1988). Analysing repeated measurements with possibly missing observations by modelling marginal distributions. *Statistics in Medicine*, **7,** 139–148.

Wild, C.J. and Yee, T.W. (1996). Additive extensions to generalized estimating equation methods. *Journal of the Royal Statistical Society* B, **58,** in the press.

Williams, D.A. (1975). The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics*, **31,** 949–952.

Yee, T.W. and Wild, C.J. (1996). Vector generalized additive models. *Journal of the Royal Statistical Society* B, **58,** 481–493.

Zeger, S.L. (1988). Commentary (on the session on repeated categorical response). *Statistics in Medicine*, **7,** 161–168.

Zeger, S.L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, **42,** 121–130.

Zeger, S.L., Liang, K.-Y. and Albert, P.S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, **44,** 1049–1060.

Zeger, S.L. and Liang, K.Y. (1992). An overview of methods for the analysis of longitudinal data. *Statistics in Medicine*, **11**, 1825–1839.

Zhao, L.P. and Prentice, R.L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika*, **77**, 642–648.

Zhao, L.P. and Prentice, R.L. (1991). Use of quadratic exponential model to generate estimating equations for means, variances, and covariances. In *Estimating Functions* (ed. V.P. Godambe), pp. 103–117. Clarendon Press, Oxford.