1998

LATENT VARIABLE GENERALIZED LINEAR MODELS

CREAGH-OSBORNE, JANE

LATENT VARIABLE GENERALIZED LINEAR MODELS

by

JANE CREAGH-OSBORNE

A thesis submitted to the University of Plymouth

in partial fulfilment for the degree of

DOCTOR OF PHILOSOPHY

Department of Mathematics and Statistics

Faculty of Technology

March 1998

Jane Creagh-Osborne

Latent Variable Generalized Linear Models

ABSTRACT

Generalized Linear Models (GLMs) (McCullagh and Nelder, 1989) provide a unified framework for fixed effect models where response data arise from exponential family distributions. Much recent research has attempted to extend the framework to include random effects in the linear predictors. Different methodologies have been employed to solve different motivating problems, for example Generalized Linear Mixed Models (Clayton, 1994) and Multilevel Models (Goldstein, 1995). A thorough review and classification of this and related material is presented. In Item Response Theory (IRT) subjects are tested using banks of pre-calibrated test items. A useful model is based on the logistic function with a binary response dependent on the unknown ability of the subject. Item parameters contribute to the probability of a correct response. Within the framework of the GLM, a latent variable, the unknown ability, is introduced as a new component of the linear predictor. This approach affords the opportunity to structure intercept and slope parameters so that item characteristics are represented. A methodology for fitting such GLMs with latent variables, based on the EM algorithm (Dempster, Laird and Rubin, 1977) and using standard Generalized Linear Model fitting software GLIM (Payne, 1987) to perform the expectation step, is developed and applied to a model for binary response data. Accurate numerical integration to evaluate the likelihood functions is a vital part of the computational process. A study of the comparative benefits of two different integration strategies is undertaken and leads to the adoption, unusually, of Gauss-Legendre rules. It is shown how the fitting algorithms are implemented with GLIM programs which incorporate FORTRAN subroutines. Examples from IRT are given. A simulation study is undertaken to investigate the sampling distributions of the estimators and the effect of certain numerical attributes of the computational process. Finally a generalized latent variable model is developed for responses from any exponential family distribution.

LIST OF CONTENTS

Page

COPYRIGHT STATEMENT

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENT

The supervisor for this project was Dr David Wright, Department of Mathematics and Statistics, University of Plymouth. I would like to thank him for his contribution and assistance.

AUTHOR'S DECLARATION

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award.

This study was financed for the first two years and one term with the aid of bursaries from the Department of Mathematics and Statistics and Department of Psychology, University of Plymouth. The remainder of the study was undertaken on a part-time basis.

A course on Generalized Linear Models with Random Effects at the University of Lancaster was attended in September 1995.

Conference attended: 9th International Workshop on Statistical Modelling. Exeter. 1994

Signed.....

Date.....

CHAPTER 1. INTRODUCTION.

1.1. LATENT VARIABLES AND THEIR APPLICATIONS.

Latent variables have been described as "random variables which cannot be measured directly, but which play essential roles in the description of observable quantities." (Brillinger and Preisler, 1982). In a latent variable model an attempt is made to explain measured observable (or *manifest*) data by including the effect of at least one covariate that cannot be measured. For example, in social survey data, it might be thought that the answers to questions on political outlook are determined by a variable which could be labelled 'conservatism', and which individuals could be assumed to possess in varying degrees. An individual's conservatism, if such a quantity exists, cannot of course be measured in any direct way. Instead it manifests itself through attitude and behaviour. Attitude and behaviour can be investigated by questionnaires or recorded in other ways. A latent variable model for the data gathered in such a way would then include an unknown covariate to represent the effect of this underlying 'conservatism'. 'Quality of life' is another latent variable which belongs to the social sciences.

A latent variable is therefore hidden or in some sense hypothetical. Probably the first application considered arose from the work of Spearman (e.g. Spearman, 1904) in the early 20th century. He was interested in studying human abilities and introduced the concept of general intelligence which could not be directly measured but which appeared to influence the results of various different types of tests. Work on measuring IQ and other types of 'latent traits', as they are termed, continues in the field of Item Response Theory (IRT) where the responses to test questions are modelled as dependent on the unknown abilities.

Economics is another field in which latent variable models are used. A random variable which could be called 'business confidence' probably contributes to such things as the level of prices on the stock exchange and the value of international currencies. The exact

nature of 'business confidence' is debatable since it arises from a variety of differing opinions and attitudes, but there is little doubt that the result has a tangible effect on the economy and that it is useful to be able to account for it in economic modelling (Bartholomew, 1987). Latent variable models are also found in engineering. For example, in optical signal estimation an unobserved random signal is of interest. The signal is associated with the absorption of photons which can be observed and measured (Brillinger and Preisler, 1982).

Sometimes, although a latent variable may be measurable in principle, it is often too difficult to record an accurate measurement in practice. For example, Brillinger and Preisler (1982) write about a latent variable model in medicine where red blood cell counts depend on the volume of a blood sample which cannot be accurately recorded. In economics, personal wealth comes into this category. In Down's Syndrome screening the date of the last menstrual period and therefore foetal age is a covariate that is often measured with error. Accident counts are predicted by traffic flow rates which contain errors (Wright and Barnett, 1991) and there are latent variables models for count data which have applications in Hematology and Cardiology (Barnett and Wright, 1992).

1.2. ANALYTICAL FRAMEWORKS FOR LATENT VARIABLE MODELS.

In the past latent variable models have been developed within several different statistical frameworks. For example there are random effects models where one or more unknown random variables are assumed to contribute towards the observed data. A distribution is usually assumed for each of these variables and interest is centred on the variances of their distributions and thereby their contributions to the overall variability of the data. In some models each individual observation depends on a different realisation of a random covariate; in others a single realisation influences a group of observations, which are correlated as a result. In the latter case a clustering or nesting effect is produced in the data. If several random variables are included in a model there may be several

corresponding layers of nesting. Alternatively, the random effects may produce a crossed design where realisations of the random variables influence the observations in different combinations but without implying a hierarchy. More complicated models can include both hierarchical and crossed effects.

The classic latent variable model is the factor analysis model (Bartholemew, 1987). Starting with a correlation or covariance matrix for continuous manifest variables, the analyst tries to discover an unknown number of underlying continuous latent variables (factors) which account for the relationships amongst the observations. Also to be determined are the slope coefficients (factor loadings) on each unknown factor. These are directly related to the variances of the factors, so estimating variances in random effects models and estimating factor loadings in factor analysis models are essentially equivalent means of estimating the effects of latent covariates. Latent trait and latent class analysis (Andersen, 1990) are two extensions of factor analysis designed to deal with discrete manifest variables; the former is appropriate when the latent variables are continuous and the latter when they are categorical.

In the examples discussed in Section 1.1 the response data may take many different forms. It may be normally distributed continuous data or, as in intelligence tests or social survey questionnaires, it may be sets of dichotomous or polytomous responses. Alternatively in the transport and medical applications mentioned the data is in the form of Poisson counts. Linear random effects models for continuous data are well developed. When observations are discrete much attention has been given to the problem of incorporating random effects into various well-known linear fixed effects models. For example for binary data McCulloch (1994) used a probit model with random effects and Drum and McCullagh (1993) fitted a logistic model with crossed random effects; Tsutakawa (1988) and Hagenaars (1993) have considered mixed log-linear models for count data with a Poisson distribution.

Generalized linear models (GLMs) (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989) provide a unified framework for fixed effect models where the response data can arise from a variety of different probability distributions within the exponential family. Several attempts have been made in recent years to develop methodology to deal with random effects within the GLM framework. Various researchers have explored different approaches to these models, often referred to as generalized linear mixed models (GLMMs). For example, Liang and Zeger (1986) and Zeger, Liang and Albert (1988) developed 'generalized estimating equations' (GEEs). Later, Zeger and Karim (1991) and Karim and Zeger (1992) used Gibbs sampling (Gelfand and Smith, 1990) within a Bayesian framework to find parameter and variance component estimates.

A principal area of past and current research related to the GLMM is maximum likelihood (ML) estimation using the Expectation-Maximization (EM) algorithm (Dempster, Laird and Rubin, 1977). References for this work include Anderson and Aitkin (1988) and Aitkin and Francis (1996). Another growing research area which is attracting perhaps the majority of current interest focuses on the related techniques of penalised (or predictive) quasi-likelihood (PQL) and marginal quasi-likelihood (MQL) estimation which depend on the linearization of non-linear models (Breslow and Clayton, 1993; McGilchrist, 1994; Lee and Nelder, 1996).

Multilevel modelling (Goldstein, 1995) grew out of a need to provide a general methodology for analysing a variety of data whose structure depended on one or more random effects. Originally developed for mixed linear models (Goldstein, 1986) multilevel procedures have been extended to a variety of non-linear models (including GLMs) where the random effects may be nested or crossed. Multilevel modelling involves MQL and more recently PQL procedures which depend upon a linearization of the model. Amongst widespread applications, Pickles, Pickering and Taylor (1996) use multilevel modelling software to fit a mixed generalized linear model with random effects.

The analysis of longitudinal or repeated measures data (e.g. Diggle, Liang and Zeger, 1994) is a wide field where GLMMs or other latent variable models may be applicable.

Latent variable models have been widely exploited in IRT in recent years and a separate body of research has developed in this area. Many of the approaches that have been applied to mainstream random effects models have also been applied to IRT models. The EM algorithm has been used to obtain maximum likelihood estimates of item parameters (Bock and Aitkin, 1981) and for joint estimation of item parameters and ability covariates (Mislevy, 1989). A Bayesian framework has also been used (Swaminathan and Gifford, 1986) and this has been combined with Gibbs sampling (Albert, 1992). More recently attention has turned to multidimensional models, that is models with more than one latent variable (Segall, 1996). Using a factor analysis framework Meng and Schilling (1996) have developed a Monte Carlo Expectation Maximization (MCEM) algorithm using the Gibbs sampler to fit a multidimensional IRT model.

Finally latent variables play a large part in errors-in-variables modelling (Fuller, 1987). The observed covariates in models of this type are considered to consist of an unobserved latent covariate and an error component. Estimation of the variation of this error component and of the parameters of the distributions of the latent variables is again of interest in modelling the variability of the response data.

1.3. A MOTIVATING EXAMPLE.

One of the major concerns of this thesis is the use of latent variable GLMs in the analysis of item response data. A specific example of an IRT application where these models have been employed is a timed item test of mental arithmetic described by Wright *et al* (1994). By applying latent variable GLM methodology to the response data obtained from this computerised test the researchers were able to model the relationships between the

parameters and the characteristics of the test items. The subjects were presented with a series of mathematical equalities to which they were asked to respond 'true' or 'false'. For example the correct response to the equality 12-17+9=4 is 'true' and to 17+19-23=15 the correct response is 'false'. Considerable attention was given to the design of items at five different levels of difficulty with strict rules defining each expression type.

An additional feature of the test design was the control of response time for individual items. Tests of this type are often subject to 'strategy' on the part of the subjects who have to choose between speed and accuracy. Because the overall time is limited a subject may decide to devote it to completing a few items as accurately as possible, or alternatively he or she may rush through the test answering all the questions with little better than a guess. A scoring system which can effectively compare the abilities of subjects adopting these opposing strategies has yet to be devised. Furthermore it can be argued that different skills are in fact being employed in the two cases. To overcome the confounding effect of strategy the mental arithmetic test was presented in a way which controlled the response time of the candidates. Each equality was shown on the computer screen for a set period of 4, 6 or 8 seconds. At the end of this period the subject was told to 'respond now' and given 1.5 seconds to press the right or left mouse button to indicate his or her answer. In this way both the lower and upper limits of the time allowed are fixed and become a characteristic of the test item.

The ten basic expression types, i.e. true and false at each of the five difficulty levels, were each presented for each of the three time periods to give a block of 30 different item types. A 60 item test was then constructed from two such blocks and given to 293 subjects. The results were analysed using the latent variable generalized linear modelling software developed by the author at the Human Assessment Laboratory at the University of Plymouth and described later in this thesis. The IRT framework requires that a logistic function known as the item response curve should be fitted to each item. This curve, which

maps ability to probability of success on a given item, is defined by three parameters, the guessing parameter or lower asymptote, the difficulty or location parameter and the discrimination or slope parameter. The guessing parameter is set at 0.5 for all the items. The difficulty and discrimination parameters are modelled as various functions of the expression type and/or response time. The ability covariates appear in the model as unknown random effects. Using the latent variable GLM software several different models with item parameters structured in this way can be fitted to the data.

The results of this analysis suggested that this particular data set could be adequately described by a model with a constant discrimination parameter and a difficulty parameter determined by the item time and the item difficulty level. The item response curves of 60 items were therefore defined by eight parameters. These results can be generalized to predict the difficulties of new items from their expression type and permitted response time and used for the construction of new item banks and new tests. In this way latent variable GLMs provide a formal methodology for modelling item parameters in terms of the structural characteristics of the items.

1.4. GUIDE TO THE THESIS.

This thesis consists of ten chapters the first of these being a short introduction to the topic of latent variables.

The subject of the thesis is latent variable generalized linear models, an immensely wide topic. One of the objectives of the research has been to identify and bring together many of the diverse models which may be classified under this heading and to trace their common characteristics. As a result the methodology developed within this thesis is placed in its context and its relationship to the many contributions which have been made in the field is defined. Chapter 2 consists of a review of some related linear models which include unknown covariates. In Chapter 3 the theory of generalized linear models and its extension

to GLMs with random effects is outlined. This then leads to a review of the published literature in the field. Much of the research referred to earlier in this introduction is discussed here in greater detail. In view of the major interest of this thesis in modelling in the IRT field, a more detailed examination of some of the models used for item response data is presented in Chapter 4 and the relationship of IRT models to latent variable GLMs is clarified.

Chapter 5 is concerned with the methodology for ML estimation that has been developed for latent variable GLMs. It contains a discussion of the EM algorithm and the generalized linear modelling software package GLIM (Payne, 1987). These two elements are combined to produce a general fitting algorithm for models in this class. A shortcoming of the procedure is the lack of a convenient means of calculation for the standard errors of the resulting parameter estimates so some possible solutions to this problem are explored. Chapter 6 moves from the general to the specific. A binary response model for IRT applications is discussed at some length and it is shown how, by considering the 'expected complete data log likelihood function', the general fitting algorithm developed in Chapter 5 can be applied to this model. The computational techniques used to fit the model require the application of a method of numerical integration. In Chapter 7 the influence of the choice of integration strategy on parameter estimation is investigated. Previous researchers (e.g. Bock and Aitkin, 1981) have favoured Gauss-Hermite integration. This alternative method is contrasted with and compared to the Gauss-Legendre method of approximation which was adopted in the methodology presented in this thesis.

Chapter 8 describes the implementation of the general fitting algorithm using EM and GLIM for fitting the binary response model developed in Chapter 6. There is a detailed description of software written for the analysis of data from the timed mental arithmetic test referred to in Section 1.3. The software is also used to analyse a second example, a timed transitive inference test. The contents of Chapter 9 evolved from a pilot simulation study. It

became apparent during this study that an analysis of the effect of the variables required by the computation process at run-time (such as starting values) was needed. In this chapter several issues of this nature are discussed.

Chapter 10 is the final chapter of the thesis. In this chapter the latent variable model for binary responses is extended to Poisson and normal data. It is then shown how the methodology can be extended to all response data from the exponential family in order to arrive at a truly generalized latent variable linear model.

CHAPTER 2. LINEAR MODELS WITH RANDOM EFFECTS.

2.1. INTRODUCTION.

This chapter consists of a review of some of the linear models whose development has led up to the latent variable GLM. Emphasis is placed on the variance/covariance structure of the models and the additional components of dispersion introduced by the inclusion of latent covariates. In order to establish a fixed reference point Section 2.2 begins with the general linear model with fixed effects; this model has a single dispersion parameter, the error variance. The general linear model has been extended to a general mixed model in order to incorporate components of variation attributable to random effects (or latent covariates). Extra variation may also arise from grouping or nesting which leads to non-zero covariances between the responses. In Section 2.3 it is shown that the same model results whether random effects are assumed or whether components of the covariance matrix produced by clustering are modelled directly (variance components models).

The same distinctions between fixed and random effects are found in factor analysis models (Section 2.4). Here responses are modelled as linear combinations of small numbers of unknown latent variables, plus an independent error. At the start of analysis the number of unknown factors is itself usually unknown and the problem is to find the smallest number of underlying variables which will explain the correlations between the responses. In the factor analysis model as in other random effects models the covariance structure depends on the parameters of the distribution of the latent variable (hyperparameters).

2.2. THE GENERAL LINEAR MODEL.

The general linear model,

$$\underline{y} = X\underline{\beta} + \underline{e} \qquad (2.1)$$

where $\underline{y}$ is a realisation of $\underline{Y}$, an n-vector random variable, is one of the most widely used models in applied statistics. Models of this form include simple and multiple regression, analysis of variance and analysis of covariance (see, for example, Draper and Smith (1981) and Hocking (1985)). In this model $\underline{Y}$ consists of a systematic and a random component. The systematic component is the vector $X\underline{\beta}$ formed from X, a known n x p design matrix, and $\underline{\beta}$, an unknown p-vector parameter. The random n-vector $\underline{E}$ of which $\underline{e}$ is a realisation has $E(\underline{E}) = \underline{0}$ and $Var(\underline{E}) = V = \sigma^2 I_n$, where I_n is the n x n identity matrix. The basic normal-theory model requires in addition that the errors are independently normally distributed (with mean zero and constant variance σ^2).

The response variable has mean vector

$$E(\underline{Y}) = \underline{\mu} = X\underline{\beta}$$

so the expected value of each response is a linear combination of parameters representing treatment effects and/or regression covariates. Both are considered fixed mathematical quantities which do not contribute any extra random variation to the model.

In addition the dispersion matrix V has a very simple structure:

$$Var(\underline{Y}) = V = \sigma^2 I_n$$

This is because the responses are independent, the variance is assumed constant over the observations, and all the variance is attributed to a single random component.

2.3. VARIANCE COMPONENT AND RANDOM EFFECTS MODELS.

2.3.1. Variance Components.

Often situations arise when the assumptions of independence and constant variance for $\underline{Y}$ are violated. Typically this occurs when the observations are nested in some way. Suppose there are I units or clusters on which observations are made. These may be, for

example, human or animal subjects, or natural groups such as family units, classes of students or fields of wheat plants. Suppose a series of J observations is made on each of the main units, to give I x J = n responses. There may be different treatments or covariates to distinguish between the units and/or the subunits. It would be reasonable to surmise that a response on a particular unit is more closely related to another response on the same unit than to a response on a different unit. The n observations can no longer be considered mutually independent and the general linear model (equation 2.1) with its single dispersion component is inadequate to represent the data.

A variance components model is one in which the stochastic dependence amongst the data is directly modelled in the covariance matrix Σ where $\underline{Y} \sim MVN(\underline{\mu}, \Sigma)$ (Lindsey, 1993). In a simple example, the data consist of n observations, as described above, with J measurements from each of I clusters. A possible assumption is that the covariance of any two responses from the same cluster is a constant, say τ, and that this value applies to all the clusters. Responses from different clusters remain independent however. This is a constant covariance model. The variance of each observation has two components: the within cluster variability, σ^2, which is again assumed constant for all clusters, and the between cluster variability which is assumed equal to the within unit covariance τ.

The J x J dispersion matrix for the observations in cluster i, V_i, is therefore the same for all i, i = 1,2,...,I :

$$V_i = \begin{bmatrix} \sigma^2+\tau & \tau & \cdots & \tau \\ \tau & \sigma^2+\tau & \cdots & \tau \\ \vdots & \vdots & \ddots & \tau \\ \tau & \tau & \cdots & \sigma^2+\tau \end{bmatrix}$$

Therefore the n x n dispersion matrix Σ is block diagonal with the I identical matrices V_i, forming the blocks on the diagonal and all other elements zero:

$$\Sigma = \begin{bmatrix} V_1 & 0 & \cdots & 0 \\ 0 & V_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & V_I \end{bmatrix}$$

The covariance structure can therefore be specified as

$$\Sigma = I_n \otimes V_i \text{ where } V_i = \tau 1_J + \sigma^2 I_J \qquad (2.2)$$

where I_J is the J x J identity matrix and 1_J is the J x J matrix consisting entirely of ones.

In this model the component τ may be negative, indicating greater variability within the units than between them. This is acceptable since τ is a covariance and a component of total variance, not a variance by itself (Lindsay,1993). In more complex variance component models, the variance may be broken down into more than two components corresponding to further levels of nesting. In a 2-way design (Rao,1973; Ch.4) there are c responses in each of pq cells arranged as, say, p rows and q columns. Components of variance can be defined to model the correlation between observations. A pair in the same cell are assumed to have a common covariance equal to the sum of the following: the covariance between observations in the same row, the covariance between observations in the same column, plus any covariance attributable to an interaction effect between rows and columns. Assuming all other pairs of responses are independent, the variance of each observation is a sum of these three components plus the common between-responses variance resulting from the independent error term.

2.3.2. Fixed Effects, Random Effects and Variance Components.

The differences between fixed and random effects are fully described in Searle (1971; Ch.9). Fixed effects are not subject to a sampling process. An experiment may be designed to estimate the effects of different levels of a factor. The treatment levels are pre-

chosen and there is no interest in any other levels or in the parameters of any general population of effects. Inferences drawn from the data concern only the chosen factor levels. Similarly covariates in a regression model are pre-determined quantities. In contrast, a random effect in a statistical model corresponds to an independent variable which can be considered to be drawn at random from a larger population of similar variables. The value of the realisation of the variable is often unknown and seldom of direct interest. What is of interest is the variation in the data which is attributable to the random effect. Inferences drawn from the data therefore concern the whole population. For example, when a social survey is conducted by different interviewers, there is a measurable interviewer effect on the data (Anderson and Aitkin, 1985: Anderson, 1988). The contribution of an individual interviewer to the responses is not important . However, the variability amongst the general population of interviewers adds to the variability of the responses and as such becomes a component of variance in the model.

In the literature the terms 'variance components model' and 'random effects model' are frequently used interchangeably. Whereas the variance components model above emphasises the homogeneity found within the main units, the stress in the random effects model is on the extra variation across these units. Although the philosophy behind them may be different, resulting models can be identical. In a random effects model the total variance in the model is again partitioned into components. Because all the components are defined as true variances there is a restriction on them to be positive. The random effects model, if so defined, is therefore less general than the direct modelling of the covariance matrix described in Section 2.3.1.

In this section a model described as a variance components model and a model used for random effects are examined and found to be identical.

2.3.2.1. A General Model for Variance Components.

A general variance components model (Rao, 1973; Ch.4, and, for example, Jenrich and Sampson, 1976) can be written

$$\underline{y} = X\underline{\beta} + Z_1\underline{\gamma}_1 + Z_2\underline{\gamma}_2 + \cdots + Z_C\underline{\gamma}_C + \underline{e} \qquad (2.3)$$

Here the response vector $\underline{y}$ is a realisation of $\underline{Y}$, an n-vector random variable with a multivariate normal distribution i.e. $\underline{Y} \sim MVN(\underline{\mu}, \Sigma)$ where $\underline{\mu}$ and Σ are determined by the components of the model (2.3). $X\underline{\beta}$ is the n-vector of systematic effects seen in the general linear model and $\underline{e}$ is an n-vector of independent random error terms with each term a realisation of the random variable $E_i \sim N(0, \sigma_e^2)$. The Z_c s are known n x q_c design matrices, where $c = 1,2,\ldots,C$. These matrices consist of dummy variables which indicate the clusters to which the response variables belong. Each Z_c corresponds to a level of clustering. The $\underline{\gamma}_c$ s are unknown q_c-vectors of random values with zero mean vector and variance matrix $\sigma_c^2 I_{q_c}$. Each of these vectors is associated with a level of nesting and the unknown components of the vector can be considered to represent the effect of the clusters at that level in the hierarchy. Because of their common variance all the units at a particular level contribute the same component of variance to the model. The $\underline{\gamma}_c$'s and $\underline{e}$ are assumed independent of each other.

It follows that

$$E(\underline{Y}) = \underline{\mu} = X\underline{\beta}$$

and

$$Var(\underline{Y}) = \Sigma = \sigma_1^2 Z_1 Z_1^T + \sigma_2^2 Z_2 Z_2^T + \ldots + \sigma_C^2 Z_C Z_C^T + \sigma_e^2 I_n$$

The parameters to be estimated are the fixed effects $\underline{\beta}$, and the variance components $\sigma_1^2, \sigma_2^2, \ldots, \sigma_C^2, \sigma_e^2$.

The following example will clarify the relationship between the model in (2.3) and the model directly specified in the example in Section 2.3.1 both having the same covariance structure. For the sake of simplicity let the data consist of 4 observations, 2 on each of two subjects which means $I = J = 2$ and $n = 4$. The response y_{ij} refers to the j th observation on the i th subject. There is only one level of clustering so $C = 1$. The structure of the mean vector $\underline{\mu}$ will not be considered. The model can be written

$$\underline{y} = X\underline{\beta} + Z_1\underline{\gamma}_1 + \underline{e}$$

or, more fully,

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \end{bmatrix} = \underline{\mu} + \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ e_{21} \\ e_{22} \end{bmatrix}$$

where $\gamma_i \sim N\left(0, \sigma_1^2\right)$ represents the effect of the ith subject.

The covariance matrix is:

$$\Sigma = \sigma_1^2 Z_1 Z_1^T + \sigma_e^2 I_4$$

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_1^2 & 0 & 0 \\ \sigma_1^2 & \sigma_1^2 & 0 & 0 \\ 0 & 0 & \sigma_1^2 & \sigma_1^2 \\ 0 & 0 & \sigma_1^2 & \sigma_1^2 \end{bmatrix} + \begin{bmatrix} \sigma_e^2 & 0 & 0 & 0 \\ 0 & \sigma_e^2 & 0 & 0 \\ 0 & 0 & \sigma_e^2 & 0 \\ 0 & 0 & 0 & \sigma_e^2 \end{bmatrix} \qquad (2.4)$$

By putting $\sigma_1^2 = \tau$, the within subject covariance and between subject variance, and $\sigma_e^2 = \sigma^2$, the within subject variance, the dispersion matrix in equation (2.4) can be seen to have the same structure as that in equation (2.2). However in this model all the variance components are defined as true variances and therefore must all be positive. In this sense the model is less general than that described in section 2.3.1 because it does not allow for

negative correlations between observations on the same unit. The situation where there is greater variability within the clusters than between them is not allowed for.

2.3.2.2. The General Linear Mixed Model.

The general linear model can be extended to a mixed model which includes the possibility of both fixed and random effects (Searle, 1971; Ch.10).

$$\underline{y} = X\underline{\beta} + Z\underline{\gamma} + \underline{e} \qquad (2.5)$$

Again $\underline{y}$ is the response n-vector and $X\underline{\beta}$ is the systematic component vector which appears in (2.1) and (2.3). The error vector $\underline{e}$ is from a normal distribution with mean $\underline{0}$ and variance $R = \sigma_e^2 I_n$. $Z\underline{\gamma}$ is an additional random component vector that does not appear in (2.1). Z is a known n x q design matrix and $\underline{\gamma}$ is an unknown q-vector of mutually independent random effects with expected means $\underline{0}$ and variance matrix D. D is a q x q diagonal matrix with elements $\sigma_1^2,\sigma_1^2,\ldots,\sigma_1^2,\sigma_2^2,\sigma_2^2,\ldots,\sigma_2^2,\ldots,\sigma_C^2,\sigma_C^2,\ldots,\sigma_C^2$, where the q_c components of $\underline{\gamma}$ associated with the cth level of nesting ($c = 1,2,\ldots,C$; $q_1 + q_2 + \cdots + q_C = q$) share a common variance σ_c^2, which therefore appears on the diagonal of D q_c times in succession. No assumption about the distribution of $\underline{\gamma}$ is made at this stage. Element i of $Z\underline{\gamma}$ is a linear combination of the random effects associated with response y_i. $\underline{\gamma}$ and $\underline{e}$ are assumed independent of each other.

This formulation is exactly the same as (2.3) except that the design matrices $Z_1, Z_2, \ldots, Z_C$ and the vectors $\underline{\gamma}_1, \underline{\gamma}_2, \ldots, \underline{\gamma}_C$ have been combined in one matrix Z and one vector $\underline{\gamma}$ (Harville, 1977). That is

$$Z = [Z_1 \vdots Z_2 \vdots \cdots \vdots Z_C]$$

$$\underline{\gamma} = \left(\underline{\gamma}_1^T, \underline{\gamma}_2^T, \ldots, \underline{\gamma}_C^T\right)^T$$

Therefore

$$E(\underline{Y}) = \underline{\mu} = X\underline{\beta}$$

and

$$Var(\underline{Y}) = \Sigma = ZDZ^T + R \text{ where } R = \sigma_e^2 I_n$$

Since D and R are diagonal and have diagonal elements equal to the variances of the random effects and error terms, the variances and covariances that appear in Σ are sums of these variances. In fact the structure of Σ is again exactly the same as in equations (2.2) and (2.4).

In their study of the effect of teaching styles on pupil achievement, Aitkin *et al* (1981) used a mixed model of this type. Here, a child's score is dependent on a covariate (pre-test score), a fixed effect (teaching style) and a random effect (teacher ability).

2.4. FACTOR ANALYSIS MODELS.

The origins of factor analysis (Anderson, 1984; Bartholomew, 1987) lie in the first decade of this century. The concepts, models and methods were first devised to suit the needs of psychologists in order to assist the study and testing of mental abilities. The subject also has applications in other social sciences and to economic data.

It is supposed that a vector of observed responses $\underline{y}$ can be explained by dividing each observation into two parts, as in all the models examined in this chapter. The first is a "predictive" component which is a linear combination of a small *unknown* number of unobservable underlying (i.e. latent) factors, each one of which might be influencing a subset of the observations. These subsets might not be mutually exclusive. The second part is an independent error term peculiar to a particular observation. This assumption is a necessary consequence of the requirement that the mean of response y_i, conditional on the latent factors that enter its predictive component, is independent. Factor analysis is

concerned with estimating the number and nature of the latent factors and the parameters of the equations that give the conditional means of the responses.

As an example, suppose the response vector consists of a series of scores obtained by one individual on a set of test questions. It might be that a subset of the questions test spatial ability, another subset tests reasoning ability and there are also some questions that test both. It is hoped that the result of fitting a factor analysis model reveals that two factors can explain the responses, with the responses to those items requiring good spatial ability for success depending only on the spatial ability factor in the predictive component, the responses to those items requiring reasoning skills depending on the factor representing reasoning ability, and the responses to those items requiring both abilities having a predictive component consisting of a linear combination of both factors.

The coefficients of the factors in the predictive component are termed the factor loadings and are equivalent to the slopes on the covariates in a regression model. The factors can be treated as fixed parameters if interest is centred on the particular subjects in the investigation or experiment. More commonly the factors are assumed to be random variables drawn from a wider population. This choice of analyses mirrors that found in fixed and random effect general linear models.

2.4.1. The Linear Factor Model.

The linear factor model is

$$\underline{y} = \underline{\mu} + Z\underline{f} + \underline{e}$$

where $\underline{y}$ is the n-vector of observed responses with mean vector $\underline{\mu}$ (c.f. fixed effects vector $X\underline{\beta}$ in general linear model) and covariance matrix Σ. $\underline{f}$ is the q-vector of factors $(q \leq n)$ and Z is the n x q matrix of factor loadings. If $\underline{f}$ is fixed then the model is equivalent to the general linear model (equation 2.1), with $Z\underline{f}$ incorporated into $X\underline{\beta}$. If $\underline{f}$ is a realisation of

random variable $\underline{F}$, it is equivalent to the general linear mixed model (equation 2.5). As usual, $\underline{e}$ is the n-vector of independent random errors with $E(\underline{E}) = \underline{0}$ and $Var(\underline{E}) = R$ (diagonal).

One of the objectives of factor analysis is to determine the least possible q, the number of factors, so that the conditional means of the y_i are independent. If a subset of responses depends on one or more common factors then there is a correlation amongst those responses. If, by determining and conditioning on those common factors, the correlation is eliminated then it is assumed that there are no other factors influencing the response. The conditional means are written

$$E(\underline{Y}|\underline{f}) = \underline{\mu} + Z\underline{f}$$

and so by implication

$$Var(\underline{Y}|\underline{f}) = R$$

Therefore in the conditional distribution of $\underline{Y}|\underline{f}$ only the mean depends on $\underline{f}$.

If the factors are random variables then it is assumed that $E(\underline{F}) = \underline{0}$, in order that $E(\underline{Y}) = \underline{\mu}$, and $Var(\underline{F}) = D$. If D is diagonal the factors are 'orthogonal' (i.e. independent if $\underline{F}$ is distributed normally); if not they are termed 'oblique'. The covariance structure of the model is therefore given by

$$Var(\underline{Y}) = \Sigma = ZDZ^T + R$$

This is identical to the covariance matrix previously seen in the general linear mixed model in Section 2.3.2.2. and in the variance components models in Sections 2.3.1 and 2.3.2.1. The assumptions that the random errors $\underline{E}$ and the factors $\underline{F}$ have normal distributions are needed to ensure the normality of the distribution of the responses.

By assuming standard normal distributions for the factors, D can be replaced by the identity matrix, which means that

$$\Sigma = ZZ^T + R$$

Here the variances that enter the structure of Σ are made up of additive sums of squares of the factor loadings plus an error variance. The covariances are sums of products of the common factor loadings. If D is not the identity matrix but is diagonal then each component of variance (apart from the error variance) is multiplied by the variance of the associated factor and each component of covariance is multiplied by the variance of the associated common factor. To illustrate this consider a model with 3 observations and 2 factors. The first factor enters into the model for the first observation, the second factor enters into the model for the second observation and both factors enter into the third model. The matrix of factor loadings is therefore of the form

$$Z = \begin{bmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \\ \alpha_3 & \alpha_4 \end{bmatrix}$$

Let the dispersion matrix of $\underline{F} = (F_1, F_2)^T$ be

$$D = \begin{bmatrix} d_1^2 & 0 \\ 0 & d_2^2 \end{bmatrix}$$

and let the error variance matrix be

$$R = \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix}$$

This results in the following covariance structure

$$\Sigma = \begin{bmatrix} \alpha_1^2 d_1^2 + \sigma^2 & 0 & \alpha_1 \alpha_3 d_1^2 \\ 0 & \alpha_2^2 d_2^2 + \sigma^2 & \alpha_2 \alpha_4 d_2^2 \\ \alpha_1 \alpha_3 d_1^2 & \alpha_2 \alpha_4 d_2^2 & \alpha_3^2 d_1^2 + \alpha_4^2 d_2^2 + \sigma^2 \end{bmatrix}$$

Thus by assuming identical standard normal distributions for the factors (i.e. by putting $d_1 = d_2 = 1$) the variances are merely absorbed into the factor loadings.

2.4.2. Other Factor Models.

An important area of recent research in factor analysis deals with categorical responses variables which depend on unknown normally distributed factors. Included in this field, also known as latent trait analysis, are many of the models used in Item Response Theory (see Chapter 4).

CHAPTER 3. GENERALIZED LINEAR MODELS AND GENERALIZED LINEAR MODELS WITH RANDOM EFFECTS.

3.1. INTRODUCTION.

Generalized linear models, the basics of which are outlined in Section 3.2, were first introduced in the 1970s. Due to the ready availability of software to implement the associated fitting algorithms they have proved an invaluable and widely-used tool. Formulated to deal with several types of non-normal response including binomial and Poisson data, the GLM provides a generalisation of the normal-theory general linear fixed effects model. Under the GLM, observations are independent and the variance of each is still attributable only to the error component. More recently the term generalized linear mixed model (GLMM) has been used to describe GLMs which include one or more random effects. GLMMs have components of variance and covariance in excess of the dispersion due to random error and they are briefly examined in Section 3.3. This section includes an extensive review of the statistical literature that has resulted from research into these models over the past 15 years.

3.2. THE GENERALIZED LINEAR MODEL.

The models described in Section 2.3 are normal-theory models: the distributions of the responses are assumed to be normal. In addition, in the general linear model the expected value of the response variable is predicted by a linear combination of the explanatory effects. Nelder and Wedderburn (1972) introduced an important generalisation of the general linear model. Included within the same theoretical framework were well-known existing models for responses with non-normal distributions and expected values which are non-linear functions of the predictor variables; for example, logistic regression

and probit analysis models which link binary responses to continuous covariates, and log-linear models used in the analysis of contingency tables where the observations are sets of Poisson counts dependent on categorical effects.

The GLM has a structure shared by the models mentioned above and many others. In this new class of model the response variable is assumed to come from a member of the exponential family of distributions (which includes the normal, binomial, Poisson and gamma distributions), and the non-random part of the model is expressed as a transformation of a linear combination of effects. In addition the theory provides a general fitting algorithm (see Section 5.3.1.). The purpose of this section is to review the principal ideas behind GLMs. The theory of the GLM is developed and expanded in the book 'Generalized Linear Models' (McCullagh and Nelder, 1989).

3.2.1. The components of a GLM.

Assume a vector $\underline{Y}$ of n independent random variables with expected values $E(\underline{Y}) = \underline{\mu}$. Let the response data $\underline{y}$ be a realisation of $\underline{Y}$. In the general linear normal-theory model, $\underline{Y}$ is assumed normally distributed and $\underline{\mu}$ is equated to the systematic component $X\underline{\beta}$. Under the GLM these assumptions are extended to include responses from certain non-normal distributions and situations where $E(\underline{Y})$ is a non-linear function of the systematic component.

There are three components of a GLM: these are (1) the error distribution, (2) the linear predictor and (3) the link function. These are outlined briefly below.

3.2.1.1. The Error Distribution.

Under the GLM it is assumed that the response y_i is a realisation of a random variable Y_i which has a distribution from the exponential family of distributions. This means

that the probability density function. (or probability mass function) of Y_i can be written in the following form:

$$f_Y(y_i;\theta_i,\phi) = \exp\{(y_i\theta_i - b(\theta_i))/a(\phi) + c(y_i,\phi)\} \qquad (3.1)$$

where $a(.)$, $b(.)$ and $c(.)$ are specific functions, θ_i is known as the canonical parameter, and ϕ is a known scale parameter constant over observation $\underline{y}$.

In Appendix A, where the reader is referred for more details, it is shown that

$$E(Y_i) = \mu_i = b'(\theta_i)$$

and

$$\mathrm{Var}(Y_i) = b''(\theta_i)a(\phi)$$

The function $b''(\theta_i)$ is known as the variance function and is dependent upon the mean μ_i; it is also expressed as $V(\mu_i)$. $a(\phi)$ is usually of the form $\frac{\phi}{w_i}$ where the w_i are known prior weights.

	NORMAL	BINOMIAL	POISSON
θ_i	μ_i	$\ln\left(\frac{\pi_i}{1-\pi_i}\right)$	$\ln\lambda_i$
$E(Y_i)$	θ_i	$\frac{1}{1+\exp(-\theta_i)}$	$\exp(\theta_i)$
$\mathrm{Var}(Y_i)$	σ^2	$\frac{n_i\exp(\theta_i)}{(1+\exp(\theta_i))^2}$	$\exp(\theta_i)$
$V(\mu_i)$	1	$\pi_i(1-\pi_i)$	λ_i

TABLE 1. Canonical parameters, means, variances and variance functions of normal, binomial and Poisson distributions.

Table 1 shows the canonical parameters of the normal, binomial (where π_i represents the expected proportion of successes) and Poisson distributions as functions of the respective mean values. The expected values and variances are expressed in terms of the canonical parameters and the variance functions as functions of the means.

3.2.1.2. The Linear Predictor.

Associated with each response vector $\underline{y}$ is a vector $\underline{\eta}$ of linear predictors where

$$\underline{\eta} = X\underline{\beta} \tag{3.2}$$

X is the n x p design matrix for the model, the elements of which are 0s and 1s or values of known covariates. The vector $\underline{\beta}$ is a p-vector of fixed effect parameters. The 0s and 1s correspond to the fixed effect parameters which are included in the model for each response and the covariates have slope parameters to be estimated. The linear predictor can thus be a highly structured combination of parameters.

3.2.1.3. The Link Function.

The systematic component $\underline{\mu}$ of a GLM is connected to the linear predictor by a link function, usually the same one for each response. That is

$$\eta_i = g(\mu_i)$$

where $g(.)$ is monotonic and differentiable. The linear predictor and the link function together describe how the location of the distribution of Y_i is explained by the covariates. (The mean value fixes the position of the distribution on the numeric scale whereas the variation helps define its shape).

For each member of the exponential family there is a canonical link function which transforms the location parameter to the canonical parameter of the given distribution. That is,

$$\eta_i = g(\mu_i) = \theta_i$$

For a normally distributed variable with mean μ_i the canonical link function is the identity function. That is,

$$\eta_i = g(\mu_i) = \mu_i = \theta_i$$

In this case the GLM is the usual normal-theory general linear model. When the error distribution is binomial with $\mu_i = \pi_i$, the expected proportion of successes, a logit transform is used to give

$$\eta_i = g(\pi_i) = \ln\frac{\pi_i}{1-\pi_i} = \theta_i$$

This GLM therefore reduces to the familiar logistic regression model for binomial responses. For Poisson variables where $\mu_i = \lambda_i$, the expected value of the ith count

$$\eta_i = g(\lambda_i) = \ln\lambda_i = \theta_i$$

Here the canonical link function is the log transform used in traditional log linear models.

The theory of GLMs is therefore a generalized framework through which various normal and non-normal linear and non-linear models can be analysed as one. This methodology was originally developed only for fixed effects in the linear predictor; with independent responses all variation is accounted for by the error distribution (Section 3.2.1.1.).

3.3. THE GENERALIZED LINEAR MIXED MODEL.

3.3.1. The Model.

The GLM as described above is formulated for fixed effects in the linear predictor and allows for a single component of variance. As in normal theory models it is sometimes necessary to include extra sources of variation in GLMs: this has led to the development of the generalized linear mixed model (GLMM), defined as a GLM that includes at least one random effect (Clayton, 1994). Under the general linear mixed model (equation 2.5) the

means are modelled as the sum of fixed and random effects. There is an obvious generalisation to non-normal models: the random effects are added to the linear predictors. The n-vector of linear predictors associated through link function $g(.)$ with response vector $\underline{y}$ therefore becomes

$$\underline{g(\mu) = \eta = X\beta + Z\gamma} \tag{3.3}$$

where $\underline{\beta}$ is a p-vector of unknown parameters representing fixed effects associated with n x p design matrix X and $\underline{\gamma}$ is a q-vector of random effects associated with n x q design matrix Z which may be partitioned as in equation (2.3). The distribution from which the random effects are sampled is usually assumed to be multivariate normal with mean vector $\underline{0}$ and q x q dispersion matrix D where $D = D(\underline{\omega})$ and $\underline{\omega} = (\sigma_1^2, \sigma_2^2, ..., \sigma_C^2)$, the vector of components of variance in the model attributable to the random effects. D is also often taken to be diagonal but assumptions about the random effects may vary.

Under the GLMM responses are conditionally independent with means

$$E\left(Y_i|\underline{\gamma}\right) = \mu_i^{\gamma} = g^{-1}(\underline{x}_i^T \underline{\beta} + \underline{z}_i^T \underline{\gamma})$$

and variances

$$Var\left(Y_i|\underline{\gamma}\right) = a(\phi)V\left(\mu_i^{\gamma}\right)$$

where ϕ is the dispersion parameter and $V(.)$ is the known variance function dependent on the conditional mean.

Suppose the linear predictor for observation y_i with a single random effect γ_1, a realisation of random variable Γ_1, is

$$\eta_i = \underline{x}_i^T \underline{\beta} + \gamma_1 \text{ where } \Gamma_1 \sim N(0, \sigma_1^2) \tag{3.4}$$

If the random effect γ_2 is a realisation from a standard normal distribution, it is easily seen that $\sigma_1\Gamma_2 \sim N(0, \sigma_1^2)$. The equation for the linear predictor (3.2) can be re-written

$$\eta_i = \underline{x}_i^T \underline{\beta} + \sigma_1 \gamma_2 \text{ where } \Gamma_2 \sim N(0,1) \qquad (3.5)$$

In matrix notation the model is written as in (3.3) but in the design matrix Z, the components which were ones (1s) are replaced by the standard deviations $\sigma_1, \sigma_2, \ldots, \sigma_C$ and $\underline{\omega}$ becomes a vector of 1s. Equation (3.5) is therefore equivalent to equations (3.4) and (3.3). In other words the variances of the random effects are absorbed into the design matrix Z where they can be estimated as slope parameters on the random effect (c.f. linear factor model in Section 2.4.1).

In a normal error model the covariance structure is independent of the means, allowing extra variation to be easily accommodated. In GLMs with non-normal error distributions the variance is a fixed function of the mean. Sometimes when all possible explanatory variables have been fitted the amount of residual variation is greater than the variance function for the given error structure allows. If no other explanation can be found this extra variation is termed overdispersion which may be modelled by the addition of a random effect to the linear predictor (Aitkin, 1994; Anderson and Hinde, 1988). Overdispersion can lead to underestimation of the standard errors of the fixed effect parameters of a GLM since the extra uncertainty is not included in the likelihood function and hence the information matrix (see section 5.3.1.). Overdispersion in specific GLMs such as the Poisson model has been examined by Hinde (1982) and in binomial GLMs by Anderson (1988), and Czado (1994).

3.3.2. Estimation in GLMMs.

Over the last decade and a half much research interest has been focused on the problem of finding effective fitting algorithms for non-linear models with random effects, including GLMs with latent variables. As time has moved forward there has been a greater degree of generalisation. The methodologies have basic elements in common: in order to fit

a GLMM a likelihood function for the unknown parameters including the variance components and possibly the random effects is formulated. From this likelihood estimating equations are derived and must be solved. Models with various different likelihood functions have been devised by different researchers, some requiring strong assumptions about the distributions of the data and the random effects. Other more general approaches have been based on quasi-likelihood models (McCullagh and Nelder, 1989). These models have the very weak assumptions that (i) $E(Y_i) = \mu_i(\underline{\beta})$ and (ii) $Var(Y_i) = a(\phi)V(\mu_i)$ as in a GLM but without the requirement of an exponential family distribution for Y_i. Quasi-likelihood estimating equations have the same properties as ML estimating equations for a GLM and the same asymptotic theory can be applied to the parameter estimates. Likelihood methods used for exponential family models can therefore be applied to much broader models and vice-versa. As the estimating equations are invariably non-linear an iterative algorithm based on the Newton-Raphson procedure or Fisher's Method of Scoring (see Section 5.3.1) is usually employed.

If full distributional assumptions can be made, it is possible to specify the joint distribution of the conditional data and the random effects. If estimates of the random effects themselves are not required they can be integrated out of the joint distribution to obtain a marginal distribution of the data dependent on the fixed parameters and the variance components. Maximum likelihood estimates can be obtained from this likelihood with the use of the EM algorithm (Dempster, Laird and Rubin, 1977). The development of this area of research is described in Section 3.3.2.1 below. Section 3.3.2.2 describes the progress of the other major technique which has been applied to the analysis of GLMMs, the basis of which is an approximation of the non-linear GLMM by a linear model. The resulting likelihood function then allows the application of repeated normal theory techniques. The procedures for the analysis of multilevel models (Goldstein, 1995) of which GLMMs are a special case, can be implemented with widely available software. They are

also based upon linearization methods and are discussed in Section 3.3.2.3. Some authors have concentrated on incorporating random effects into one particular GLM such as the probit or logistic model, sometimes with severe restrictions on design matrices X and Z and narrow distributional assumptions for the random effects. Some of these miscellaneous models are discussed in Section 3.3.2.4.

Although an attempt has been made to review most of the relevant published material in this wide field it should be noted that some omissions have had to be made. Research of specific relevance to IRT is reviewed in Section 4.3 of the next chapter.

3.3.2.1. ML Estimation with the EM Algorithm.

This approach leads to ML estimators of the fixed parameters and variance components. The random effects are not estimated and can only appear in the model in a nested, not crossed, design. The use of the marginal distribution in the likelihood function results in a difficult integration which becomes more difficult as the levels of nesting increase. For this reason it is necessary to introduce an approximation to the marginal distribution of the data. This distribution is an integral obtained by integrating the joint distribution of the data and the random effects with respect to the random effects. It is approximated by using Gaussian quadrature the effects of which on the estimates are for the most part unknown. Perhaps the greatest contribution to this methodology was made by Bock and Aitkin (1981) who proposed the use of the EM algorithm (Dempster *et al*, 1977) for ML estimation of item parameters in item response models with a latent ability covariate. Following Bock and Lieberman (1970), Bock and Aitkin (1981) used a normal cumulative distribution function for the conditional probability of a correct response and obtained an unconditional likelihood for each possible response pattern by using Gauss-Hermite quadrature to approximate the integral over the ability distribution in the marginal distribution. The likelihood equations resulting from this model were reformulated and

shown to be the likelihood equations for a probit analysis in which the independent variables are the quadrature points (nodes) and the data are (i) the expected frequencies of correct responses to each item at each ability level and (ii) the expected size of the sample responding to each item at each ability level.

The authors also showed how the same expressions for (i) and (ii) can be derived by borrowing from the principles of missing data used in the EM algorithm. They replaced the missing data (abilities) in the log likelihood equations with their expectation conditional on the observed data and current parameter estimates. This, as they point out, is not quite the same as the approach outlined by Dempster *et al* (1977) which in its most general form computes the expected value of the log likelihood of the complete data conditional on the observed data and current parameter estimates.

Bock and Aitkin (1981) also showed that it is unnecessary to make any assumption about the distribution of the ability variable. Discrete posterior densities conditional on the data for each ability node can be calculated and used as weights in the corresponding probit analysis.

The two-step EM algorithm is employed iteratively as follows: the first step is the expectation step which results in the computation of (i) and (ii) above, given working estimates of the item parameters and the second is the maximisation step where the probit model is fitted to this data in order to update the estimates. Bock and Aitkin do not reveal details of their software but report slow convergence of the algorithm and the lack of a readily available inverse information matrix to provide standard errors as disadvantages of the methodology.

Hinde (1982) adopted a similar approach using GLIM software (Payne, 1987) for fitting GLMs, to help with the problem of over-dispersion in Poisson data. It is assumed that the extra variability in the data can be attributed to some unknown random effect, just as the variability in item response data is in a similar way attributed to a random latent

ability covariate. In this application there is a resulting expansion of the data as K copies of each response are created (K is the number of quadrature points) each with a different weight derived from the approximate distribution of the latent variable. In the Bock and Aitkin application however there is a reduction of the data. This is because the data consists of all the different item score patterns observed and the numbers of subjects recording each possible pattern. Later a further summation of subjects at each ability node occurs. This reduction in the length of the vectors being processed by the computer is advantageous when handling large data sets.

The ML methodology using the EM algorithm was extended by Brillinger and Preisler (1983) to a wider class of latent variable models where the responses are conditionally independent depending on parameter $\underline{\beta}$ and the latent variables have independent distributions depending on parameter $\underline{\alpha}$. They applied it specifically to a problem with Poisson counts. They were followed by Anderson and Aitkin (1985) who considered a logistic model with random effects to describe binomial responses to a social survey where interviewer variability was thought to influence the data. They used the EM algorithm with GENSTAT software (Alvey, 1977) rather than GLIM to enable them to accommodate more than one level of nesting and different sized clusters, but reported problems with limitations on data space for large data sets made even larger as a result of the expansion required.

Anderson (1988) compared this same logistic model with random effects to other models that might explain overdispersion in binomial data. She found it an appropriate model when the extra variation could be attributed to clustering and a distribution for the random effect could be assumed. She listed slow convergence of the EM algorithm, computational intensity and the necessity for numerical methods to approximate the integral in the marginal likelihood as disadvantages of the procedure.

Later, Anderson and Hinde (1988) generalized the EM methodology to all GLMs with random effects by adding the random component to the linear predictor. They suggested that extensions to more than one level of nested random effects could be easily incorporated within the general algorithm. Problems of implementation including the suitability and accuracy of any particular method of Gaussian quadrature are not dealt with in the paper (or elsewhere).

Aitkin and Francis (1996) produced GLIM4 (Francis *et al*, 1993) macros which implemented the methodology described by Hinde (1982) and Anderson and Hinde (1988) and used them to solve several apparently different types of problem. These include both overdispersion in binomial and Poisson models and maximum likelihood estimation of the unknown parameters of the component distributions in finite mixture problems. A random effect included in the linear predictor of a GLM is assumed to be normally distributed. When the integration of the marginal distribution of the data is approximated using Gaussian quadrature the resulting likelihood function is the same as the likelihood of a finite mixture of exponential family distributions. The location parameters of the underlying distributions are the known quadrature nodes and the proportions attributable to each component are weights. During the expectation step of the EM algorithm these weights, which depend on the current parameter values, are computed. This likelihood is then maximised during the M-step to obtain better parameter estimates.

As before (Hinde, 1982; Anderson and Hinde, 1988) the data must be expanded to K (number of quadrature points) times its original length. As Aitkin himself suggests that $K > 20$ is necessary for a reasonably accurate approximation, it appears that large data sets might strain the data space limitations of GLIM4. This is borne out in practice. The software was supplied with Gauss-Hermite nodes and weights although these were easy to change.

The algorithm implemented in these macros can also be extended to non-parametric estimation of the distribution of the random effect. In this case the nodes and weights (or mixture proportions) are unknown and are estimated along with the usual model parameters for a given K which is increased from $K=1$ until the likelihood is maximised.

The authors have adapted the non-parametric procedure to produce further macros for variance component estimation in models where the data is nested in a two-level structure. Here weights are computed during the E-step at the higher level of nesting. These are then used to weight the individual responses during the M-step.

Since Bock and Aitkin wrote their original paper in 1981 ML estimation and the EM algorithm have been widely used statistical techniques both in the mainstream field of GLMMs and in item response modelling. Although other methodologies have been developed, some with a great deal of success, to deal with the problem of random effects in various non-linear models, the ML-EM procedures still offer a valid alternative and research to improve upon them continues today. Recently, Meng and Schilling (1996) have attempted to eliminate the error due to the numerical integration required in this methodology by using Gibbs sampling to compute the E-step. Meng and Schilling assert, with reference to the use of Gauss-Hermite quadrature, that "*the predictive* (i.e. posterior) *distributions for the individual latent abilities become more peaked as the number of items increases, leading to 'lumpy' observed-data likelihood for the model parameters, but the reliability of the fixed point Gauss-Hermite quadrature method relies on the smoothness of the integrand*" (see Section 7.4). Although criticisms of this method of integral approximation and therefore of the entire ML methodology abound in the literature it appears that there is no published work which attempts ML estimation using the EM algorithm with any alternative numerical method. One of the objects of this thesis is to contribute towards the ML-EM methodology by exploring the implementation of the algorithm with alternative quadrature rules (see Chapter 7).

3.3.2.2. 'Linearization' Methods.

This section covers a variety of closely related models and methods. Some models include assumptions about the conditional distributions of the data and some of these also have distributions for the random effects. More general results have been obtained using quasi-likelihood models. One approach has been to replace the non-linear part of the model by a linear function and another to approximate the likelihood function instead. The score equations derived from the likelihood are solved by iterative processes the form of which may differ slightly from author to author. Similarly theoretically different estimators are frequently used to obtained the same estimates. A common advantage of 'linearization' methods is their applicability to all kinds of crossed and nested models. However estimates have been found to exhibit bias and the use of an approximating linear model must be in doubt in many situations where the data is far from normally distributed. One of these doubtful cases is that of binary response data.

Schall (1991) was one of the first researchers to suggest an algorithm for estimating fixed effects, random effects and variance components in GLMs with random effects. It was based on a proposal (Fellner, 1986, 1987) for the iterative computation of maximum likelihood estimates of variance components in the normal linear model. The link function $g(.)$ of the GLM is linearized using Taylor's first order approximation. This results in a linear random effects model for the 'adjusted dependent variable', $\underline{z}$, (see McCullagh and Nelder, 1989) where

$$\underline{z} = g(\underline{y}) = X\underline{\beta} + Z\underline{\gamma} + \underline{e}g'(\underline{\mu})$$

Schall's fitting algorithm consists of two-steps: the first step provides least-squares estimates of the fixed and random effects given current estimates for the dispersion parameters; the second step updates the estimates for the dispersion parameters given current values for the fixed and random effects. The estimates obtained from this procedure

can be taken as approximate ML estimates when the conditional distribution of the data and the prior distribution of the random effects are from the exponential family. A variation of the second step gives approximate restricted maximum likelihood (REML) estimates, which take account of the loss of degrees of freedom due to the estimation of the fixed effects (Patterson and Thompson, 1971). The algorithm can be programmed fairly simply using GLIM and is applicable to designs incorporating both nested and crossed random effects. Although a covariance matrix is given it is dependent upon the estimates of the random effects and the extra variability due to these estimates is not taken into account.

Very similar algorithms to Schall were derived by Engel and Keen (1992) who dispensed with the need for full distributional assumptions for either the data or the random effects. Their approach is based on a combination of quasi-likelihood methods and minimum norm quadratic unbiased estimation (MINQUE) (Rao, 1973).

Schall (1991) and Engel and Keen (1992) both describe special cases of penalised quasi-likelihood (PQL) estimation in GLMMs. This methodology is generalized further by Breslow and Clayton (1993). They worked with a quasi-likelihood function for the data and a multivariate normal distribution for the random effects. They obtained a marginal quasi-likelihood by integrating the exponent of the sum of the two likelihoods over the random effects. They approximated this marginal quasi-likelihood using Laplace's method for integral approximation (Tierney and Kadane, 1986) and arrived, after several simplifying assumptions, at a likelihood function for the fixed and random effects equivalent to the PQL used by Green (1987). The resulting estimating equations are solved iteratively. The fixed and random parameter estimates are substituted in the approximation to the marginal quasi-likelihood to give an approximate quasi-likelihood function for the variance components. This is then adjusted to obtain REML estimates. The new estimates for the variance components are used to obtain improved parameter estimates and so on until convergence. The estimating equations are recognisable as those derived by Harville (1977) for the

normal linear mixed model, in which case they give best linear unbiased prediction (BLUP) (McGilchrist, 1994; see below) estimates in the case of the fixed and random effects and REML estimates in the case of the variance components. Like Schall (1991), Breslow and Clayton, (1993) found the fixed and random effect estimators to be approximate marginal ML estimators. The assumptions and approximations made in order to arrive at these equations suggest that the more normally the data are distributed the greater the validity of the model.

Breslow and Clayton, (1993) compare PQL with another similar approach to inference in GLMMs which they call marginal quasi-likelihood (MQL) and is the procedure proposed by Goldstein (1991) (see Section 3.3.2.3). In MQL an approximate marginal mean is specified. This does not include the random effects which are therefore not included in the linear predictor. For given components of dispersion the fixed effects only are computed iteratively with Fisher scoring. The resulting estimates are then used in the same REML equations as in the PQL approach to calculate updated variance parameters, although for this model the equations are derived by applying the method of pseudolikelihood (Carroll and Ruppert, 1982). In the MQL version the random effects are not estimated until convergence has occurred.

Laplace's integral approximation was independently applied to marginal distributions of the data in non-linear mixed models by Wolfinger (1993). In the case of the GLMM his resulting estimating equations are equivalent to those of Schall (1991) assuming normality for the random effects.

McGilchrist (1994), following McGilchrist and Aisbett (1991), adapted a method known as best linear unbiased prediction (BLUP) (e.g. Henderson, 1975) for linear models with fixed and random effects and applied the theory to GLMs with random effects. They showed how BLUP estimators can be adjusted to find approximate ML and REML estimators (Harville, 1977) for fixed parameters, random effects and variance components

in a GLMM and other non-linear models where the random effects are assumed to be normally distributed. The iterative BLUP procedure maximises the joint 'log-likelihood' of the fixed and random parameters. This likelihood can, under certain assumptions, be approximated by a quadratic expression. If the likelihood derived from the approximate asymptotic distribution of the ML estimators of the fixed and random effects (McGilchrist and Aisbett, 1991) is substituted into this joint 'log-likelihood' the same quadratic expression results. This reasoning justifies the use of the approximate asymptotic likelihood. Together with a normal prior, this allows restrictions on the random effects to be incorporated into the model. When this theory is applied to the GLMM, it is reduced to a normal linear mixed model with an adjusted dependent variable. This approach is again effectively equivalent to Schall (1991). The resulting estimators, although an improvement on straight BLUP procedures, have however been shown in simulation studies to be biased particularly in the case of the variance components (Kuk, 1995). Kuk proposes an iterative Monte Carlo method to correct the bias shown in initial estimates obtained by BLUP or similar estimation.

The term hierarchical generalized linear model (HGLM) was defined by Lee and Nelder (1996). These are GLMs with linear predictors that include random variables whose distributions are not confined to the normal. (In this paper, the term 'GLMM' is restricted to those HGLMs with normally distributed random effects). The authors bring together, generalise and extend the work of many of the researchers in this area.

The approach has much in common with McGilchrist (1994) being based on the joint likelihood derived from the conditional distribution of the data and the distribution of the random effects (also Henderson, 1975). This is called the h-likelihood and the estimates which maximise it the maximum h-likelihood estimates (MHLE). The resulting score equations for the MHLEs are those derived by Schall (1991), Engel and Keen (1992) Breslow and Clayton (1993), Wolfinger (1993), and McGilchrist (1994). For the estimators

of the dispersion components Lee and Nelder, (1995) define the adjusted profile *h*-likelihood (APHL) and the maximum adjusted profile *h*-likelihood estimators (MAPHLEs). It is shown how these estimators lead to the REML estimators also derived by the previous researchers. Analysis using HGLMs is simplified when the model is a GLMM and also when the distribution of the random effects is conjugate to that of the data. For example the Poisson-gamma model, the binomial-beta model and the gamma-inverse gamma are all conjugate HGLMs. However the problem of bias, which can be particularly serious in the variance components associated with binary responses, is not addressed by this paper. In addition, the authors make no mention of the multilevel modelling software (Goldstein, 1995) which is extensively used to fit models of this type.

3.3.2.3. Multilevel Models.

Multilevel models (Goldstein, 1995) and the general-purpose software for their application ML3 (Prosser *et al*, 1991) and Ml*n* (Rasbach *et al*, 1995), were developed as tools for the systematic analysis of data with a hierarchical structure. Data is grouped in levels corresponding to the clustering mechanisms present. For example, in IRT terms, the binary item responses are level one units. These units are grouped by subject, the level two units. If the subjects were clustered further such as by age groupings these would become level three units and so on. Fixed and random effects may appear at any level. More complex structures where the data is cross-classified can be encompassed in the general framework. In multilevel modelling terms the GLM is a single-level model.

The methodology, originally developed for continuous data (Goldstein,1986), has been generalized to non-linear models for discrete data (Goldstein, 1991). The procedure is the MQL approach described by Breslow and Clayton (1993) (see previous section) and can be implemented with the software package ML3. The general multilevel model is expressed as the sum of two parts, one linear and one non-linear. Random variables can belong to

either component. Both the fixed and the random parts of the non-linear function are then linearized using first and possibly second order terms from a Taylor's expansion. The result is a standard multilevel linear model to which the linear estimation procedure can be applied. This involves the use of an iterative generalized least squares (IGLS) algorithm where the estimating equations are based upon quasi-likelihoods rather than full distributions. Multivariate normality is however assumed in the random effects in order to compute a weight matrix. Approximate maximum likelihood estimates of the fixed parameters and the variance components are obtained. These may be biased in the case of the variance component estimates even in the linear model and REML modifications can be applied to correct the bias. The addition of the quadratic terms from Taylor's approximation may in some circumstances produce substantially improved estimates but may not in others. Convergence of the algorithm is not always guaranteed.

MQL methods for analysing hierarchical data are also implemented in the program VARCL (Longford, 1988). This software produced identical estimates to ML3 using 10 simulated data sets in a comparative study of the two packages carried out by Rodriguez and Goldman (1995). They found that both Goldstein and Longford model discrete data using the same linear approximation. The algorithms used to produce the parameter estimates vary slightly in that Longford uses Fisher scoring rather than GLS for the variance components but this does not effect the results. Rodriguez and Goldman used the packages to fit multilevel models for simulated binary response data with two- and three-level structures. They found the estimates to be severely downwardly biased particularly in the case of the variance components when the random effects were large or when the number of units within a level were small. Including quadratic terms in the approximation improved matters only slightly.

In response to this criticism, Goldstein and Rasbash (1996) suggest adopting a slightly different procedure during the computation. This corresponds to the PQL (penalised

or 'predictive' quasi-likelihood) approach (Breslow and Clayton, 1993). In the PQL approach current estimates of the random effects are included in the linear part of the expansion of the non-linear function; in the MQL approach they are not. Thus at each iteration a current estimate of the random effects is included in the estimating equations. The PQL modification is incorporated in the updated software package M*ln* and appears to improve the estimates considerably when used in conjunction with the second order approximation. In a simulation study with binary data (Goldstein, 1995, p99) the best results were again produced using the PQL version with second-order terms included in the model.

Bias in the variance component estimates has not however been completely eliminated and further research is needed in this area to assess which methods (PQL or MQL with or without second order approximations) should be used in which situations.

Pickles, Pickering and Taylor (1996) used M*ln* software with first-order PQL estimation to fit a mixed generalized linear model with random effects.

3.3.2.4. Miscellaneous examples of extensions of random effects models to GLMs.

One of the first researchers to develop methodology to include latent variables within a particular GLM was Williams (1982). He used GLIM for ML estimation in a logistic model incorporating extra-binomial variation associated with unobserved random variables. In this model the response, conditional on the random effect, is binomially distributed. The relationship between the mean and the variance of the random effect is specified and this leads to the relationship between the unconditional expectation and variance of the response variable. Estimation is therefore based upon the quasi-likelihood and does not include estimation of the random effects. A restriction of the model is that the covariates cannot vary within a unit (or cluster).

Stiratelli, Laird and Ware (1984) presented a more flexible logistic model for serial binary observations from a panel of subjects with general covariates and normal random effects at the subject level. ML and Bayesian estimation techniques were combined with the EM algorithm. Later, using a probit model with normal random effects for binomial data, Gilmour, Anderson and Rae (1985) described a 'joint-maximisation' method. The estimating equations derived for both these models are special cases of those developed later by Schall (1991).

An early attempt to model data with extra components of dispersion within the GLM framework was an analysis of longitudinal data with time-dependent covariates by Liang and Zeger (1986). They derived 'generalized estimating equations' (GEEs) based on maximum quasi-likelihood estimation. These equations give consistent estimates of the fixed parameters under weak assumptions about the joint distribution of the repeated measurements. A working correlation matrix is estimated to model the dependency between observations on the same subject but the focus of this method was essentially on estimation of the fixed parameters. Following this, Zeger, Liang, and Albert (1988) distinguished between subject-specific and population averaged models and applied GEEs to both. Subject-specific models are those in which each subject's individual response is of interest rather than that of the population as a whole. In these situations variation across subjects is explicitly modelled as in the GLMM. Moment estimates for the variance components and the fixed parameter solutions to the GEEs are calculated simultaneously within an iterative procedure. The authors found that convergence may not be achieved when the data is extremely non-normal.

Im and Gianola (1988) used two different maximisation methods for computing ML estimates in mixed probit and logistic models for binomial data on lamb mortality. They preferred the simplex method (Nelder and Mead, 1965) to the EM algorithm because it could be adapted to produce an asymptotic covariance matrix. Conaway (1990) took an

unusual approach by modelling binary responses with the 'log-log' function as an alternative to the logit or probit link. In addition he suggested a log-gamma distribution for the random effects. This allowed the marginal likelihood to be computed without the need for numerical integration.

Zeger and Karim (1991) applied a Bayesian framework to the GLMM and used the Gibbs sampler to address the computational problems posed by the complex numerical integration that can occur in the likelihood function. Samples are drawn repeatedly from the conditional distributions of the fixed, random and then variance parameters in turn, given the most recently sampled values of the other parameters. After a sufficiently large number of sampling iterations the process converges to the joint distribution of the parameters. More values can then be generated to simulate the empirical joint distribution from which inferences can be made. The method can be applied to both nested and crossed effects models and can accommodate different assumptions about the random effects. Although computationally intensive the method is easy to implement. The Bayesian/Gibbs approach was applied to the well-known salamander mating data (McCullagh and Nelder, 1989) by Karim and Zeger (1992). This data has a complicated structure with crossed random effects and has been the subject of several analyses including that of Drum and McCullagh (1993) who adapted restricted maximum likelihood (REML) estimation to logistic models with crossed random effects in the linear predictor. They compared their results favourably with estimation using linearization methods (Schall, 1991).

McCulloch (1994) considered a probit model for binary data with normally distributed random effects and used the EM algorithm for ML and REML estimation of the variance components. For crossed effects EM is combined with Gibbs sampling to avoid a complicated integration.

CHAPTER 4. GENERALIZED LINEAR MODELS IN ITEM RESPONSE THEORY.

4.1. INTRODUCTION.

The previous two chapters have been concerned with presenting a general review of some of the ways that latent variables have been incorporated into linear models (Chapter 2) and non-linear models with particular reference to GLMs (Chapter 3). Chapter 4, although still forming part of a review of a wide field, is much more detailed than the previous presentation because its subject matter is Item Response Theory (IRT) (Hambleton and Swaminathan, 1985; Hambleton, Swaminathan and Rogers, 1991) which is the chief application area of the methodologies described in this thesis. IRT is a branch of psychometrics which is concerned with measurement in the field of psychology. By examining this single application area in greater depth this chapter will demonstrate that IRT is a rich source of opportunities for the application of latent variable GLMs for dichotomous responses. Demands for better modelling tools within this area has therefore led to the development of the modelling software which will be described in later chapters.

One of the objectives of IRT is the development of tests to measure latent traits in human subjects (Lord and Novick, 1968). A latent trait is usually some kind of underlying ability or aptitude such as general intelligence, suitability for a certain career, talent for a particular task, etc. Test questions or 'items' are designed to measure a particular trait. A bank of test items is created and the subject is given a test consisting of a subset of items selected from the bank. Each item has its own properties such as type, difficulty or time allowed for completion. Statistical models (Birnbaum,1962 and 1969) relate the probability of a correct answer to the item parameters and a subject's latent ability. In Section 4.2 some of the most

common models used to interpret item response data are described. It is shown how these particular models can be brought within the framework of the GLM. Several different procedures which have been developed for fitting the models are also discussed (Section 4.3).

In a typical IRT situation there is a vector of binary response data which consists of the results of testing I subjects on J items. The response of each subject to each item depends on the item parameters and the ability (the latent trait) of the subject. This ability is unknown and cannot be found by any direct methods of measurement. The problem is to fit the model, estimating the item parameters, without knowledge of the latent covariate. Having been calibrated in this manner the items can then be used for ability estimation for another set of subjects at some future time. The following discussion is restricted to test items where a dichotomous response variable is recorded; that is, the response is either a 1 if the answer is correct, or 0 if it is incorrect. In addition, the assumption of unidimensionality is adopted; in other words, it is proposed that the responses can be explained by a single latent trait. In theory the application of GLMs can be extended to polytomous responses and multidimensional latent variables.

4.2. THE MODELS.

The response y_{ij} of subject i to item j is modelled as the additive combination of its expected value and an error component, where the expected response depends on the parameters of the item, $\underline{\beta}_j$, and the latent ability of the subject, γ_i. That is

$$y_{ij} = \mu_{ij}(\underline{\beta}_j, \gamma_i) + e_{ij} \qquad (4.1)$$

Since the responses are dichotomous it can be assumed that their distribution in equation (4.1) is binomial. When the response variable is from the exponential family, as in this case, and the function which gives the expected value can be written in the form

$$\mu_{ij} = g^{-1}(\eta_{ij})$$

with

$$\eta_{ij} = \underline{x}_{ij}^{T} \underline{\beta}_{j} + z_{ij}\gamma_{i} \tag{4.2}$$

i.e. with a link function g and a linear predictor η_{ij} in this linear form, then the model is a GLMM as described in Section 3.3 (see equation 3.5).

The probability, π_{ij}, of a correct response by subject i to item j is equal to the expected value of the response. In IRT the function relating probability of success to latent ability is known as the Item Characteristic Function (or Curve), assumed to be monotonically increasing between the limiting values of 0 and 1. It has been modelled at various times both by the ogive curve of the cumulative normal distribution and by a logistic regression function. For a logistic regression curve

$$\pi_{ij} = \mu_{ij} = \frac{1}{1+e^{-\eta_{ij}}} = g^{-1}(\eta_{ij}) \tag{4.3}$$

This is the inverse of the logit link function

$$\eta_{ij} = \ln\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) \tag{4.4}$$

which is the canonical link function for the binomial distribution. In IRT the ability variable $\underline{\gamma}$ is not known and, since subjects are presumably selected at random for purposes of item calibration and are of no interest themselves, it is reasonable to treat the abilities as random effects added to the linear predictor. (If the γ_i are known fixed effects then these models are

ordinary GLMs.) Therefore binary response IRT models with logit link functions and linear predictors in the form of equation (4.2) are latent variable GLMs.

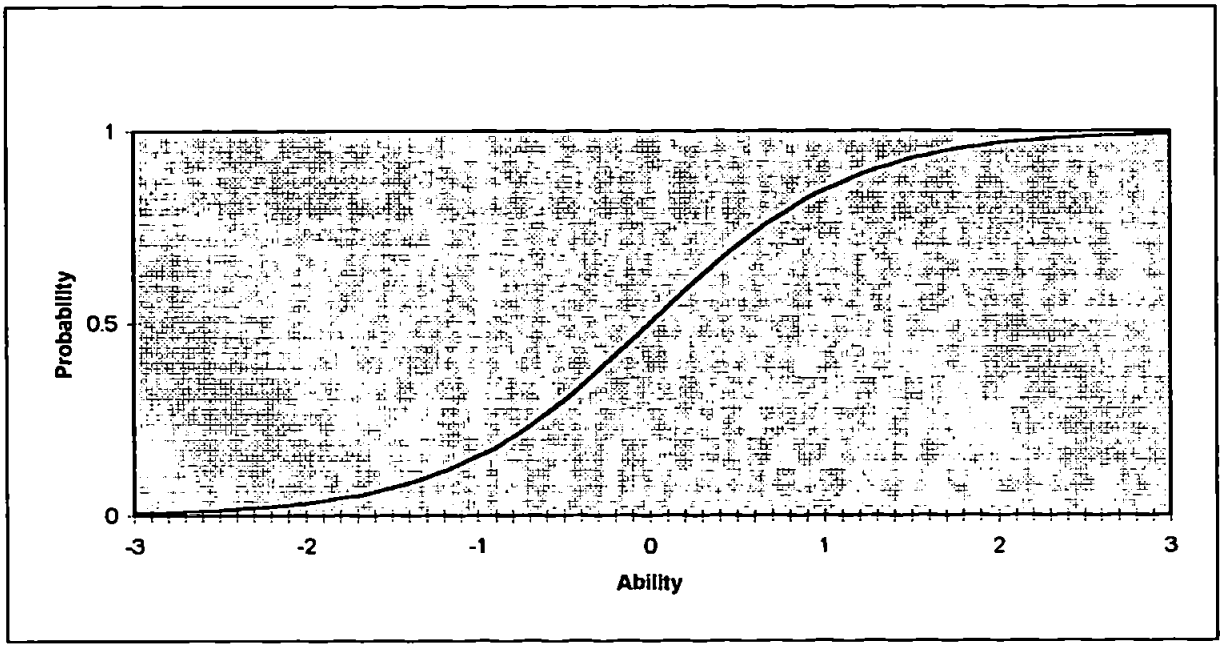


FIGURE 1. A Typical Item Characteristic Function.

The typical S-shaped logistic curve is shown in Figure 1. The parameters of the item alter the exact shape and location of the curve and IRT models are distinguished by the number of parameters items are assumed to possess. The three principal models are distinguished by having either one, two or three parameters per item.

4.2.1. Item Parameters.

IRT models incorporate up to three item parameters which represent specific properties of the items in a test bank. The one-parameter model, which is also known as the Rasch model, includes a difficulty parameter, denoted b. Specifically, in the one-parameter model, b is the

ability level of a subject with 0.5 probability of success. This parameter determines the position of the logistic curve in relation to the ability scale (Figure 2). Large positive values of b indicate very difficult items where the curve is at the right-hand end of the scale. Large negative values are associated with easy items where the curve is situated towards the left-hand end of the ability axis.

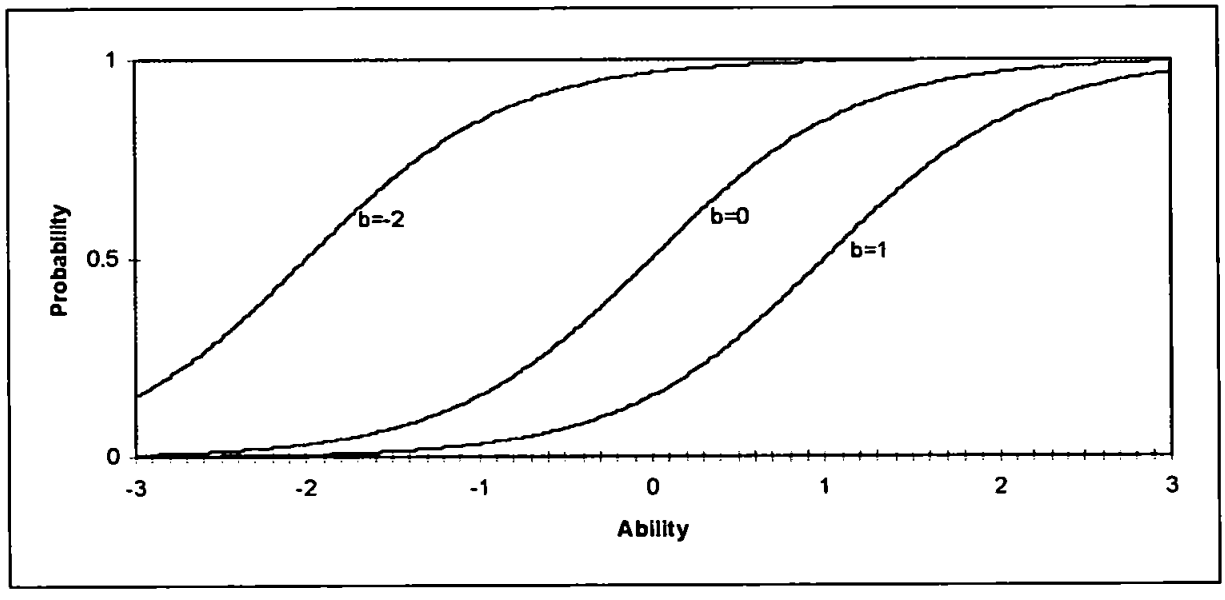


FIGURE 2. Item Characteristic Functions showing difficulty parameter.

The two-parameter model includes a discrimination parameter, denoted a. The discrimination parameter is equivalent to the slope on γ_i at the point on the curve where π_{ij} is equal to 0.5. Consider two subjects with abilities differing by one unit and whose probabilities of success are neither unusually high nor unusually low. If the curve slopes steeply then their probabilities of success will differ widely. The same two subjects will have much closer probabilities of success on an item associated with a curve with a shallow slope (Figure 3). Therefore, the greater the value of a, the more discriminating the item.

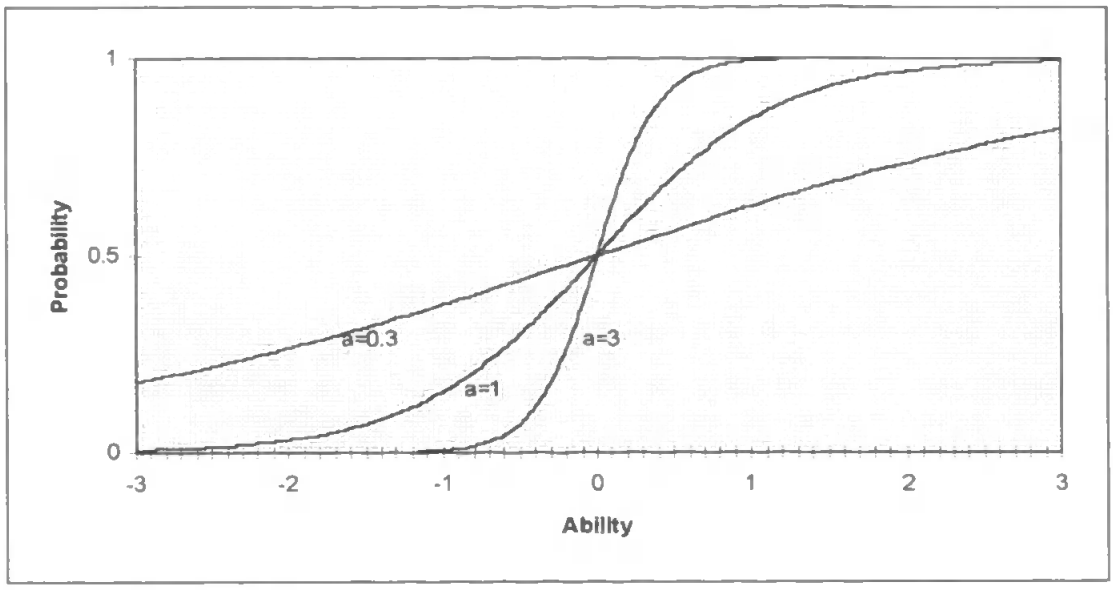


FIGURE 3. Item Characteristic Functions showing discrimination parameter.

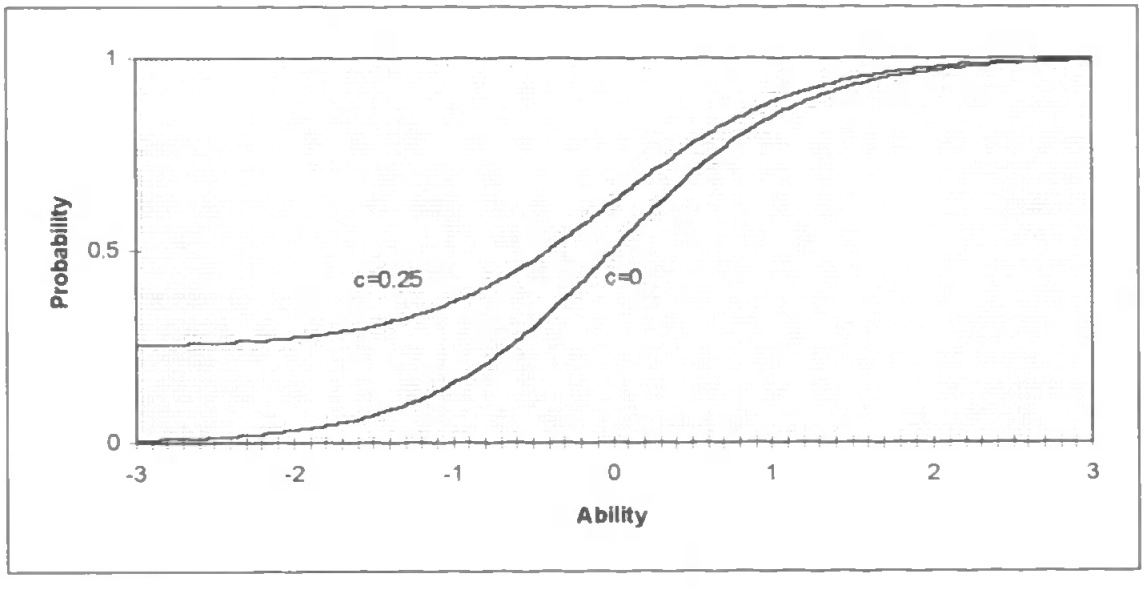


FIGURE 4. Item Characteristic Functions showing guessing parameter.

In the three-parameter model a third parameter, denoted c, is used. This is sometimes called the guessing parameter. This parameter represents a lower asymptote of the function

(Figure 4), or equivalently, the probability that a subject with minimum ability has of answering the test item correctly. Such a subject's response is determined only by chance. Therefore the guessing parameter represents the probability of guessing the correct answer.

Yen (1981) compared the performances of one-, two- and three parameter models in explaining data obtained from eight different achievement tests which were given to students aged approximately 12 to 14. The tests were in Maths and English and all consisted of multiple-choice items with four options each. The results of this study indicated that the three-parameter model was the most appropriate model for all the eight data sets and might well be the best choice for all data from multiple-choice tests. The three-parameter model does however require larger sample sizes than the one- and two-parameter models to estimate parameters to a given level of accuracy.

4.2.2. One-, Two-, and Three-Parameter Models.

When all three parameters are included in the model it is known as the three-parameter logistic model. In this model the probability of subject i responding correctly to item j is

$$\pi_{ij} = c_j + \frac{1-c_j}{1+e^{-\eta_{ij}}} \text{ where } c_j \leq \pi_{ij} \tag{4.5}$$

where

$$\eta_{ij} = a_j\left(\gamma_i - b_j\right) \tag{4.6}$$

The two-parameter model is obtained by setting c_j to zero and the one-parameter model by further setting a_j to 1. The linear predictor, equation (4.6), is therefore always of the form shown in equation (4.2) and the link function is the logit link, equation (4.4), in the one- and two-parameter models. The three-parameter model has an unknown parameter c_j which is not part of the linear predictor and the model is not a GLM. However if this parameter is

known the model becomes a GLM although the link function is no longer the canonical link function. The link function is obtained from equation (4.5) as shown in Appendix B.

$$\eta_{ij} = \ln \frac{\pi_{ij} - c_j}{1 - \pi_{ij}} \text{ where } c_j \leq \pi_{ij}$$

In Section 3.3.1 (equations (3.4) and (3.5)) it was shown that the variances of the random effects are equivalent to the square of the slope parameters on the random effects. If it is assumed that the components of the random effects vector are sampled from independent and identical normal distributions with zero means and a common variance, then this is equivalent to assuming that all items have the same power of discrimination. The standard deviation is equivalent to a discrimination parameter. Increasing the discrimination of the items has the same effect as spreading out the distribution of ability. In this case the linear predictors are of the form

$$\eta_{ij} = \beta_j + z\gamma_i$$

where γ_i is a realisation of random variable Γ_i and $\Gamma_i \sim N(0,1)$

If the discrimination parameter is allowed to vary between items then it must be indexed by j. Then the linear predictors are of the form

$$\eta_{ij} = \beta_j + z_j\gamma_i$$

where $\Gamma_i \sim N(0,1)$

Both versions of the linear predictor conform to the GLM.

4.2.3. Likelihood Functions for the One-, Two- and Three-Parameter Models.

If $\underline{y}$ is the response pattern of I subjects attempting J items, then the log likelihood function for ability vector $\underline{\gamma}$ and item parameters $\underline{\beta}$ conditional on response pattern $\underline{y}$ is

$$\ln L(\underline{\gamma}, \underline{\beta}|\underline{y}) = \sum_{i=1}^{I}\sum_{j=1}^{J}\left[y_{ij}\ln\pi_{ij} + (1-y_{ij})\ln(1-\pi_{ij})\right]$$

where π_{ij}, which is dependent on item parameter $\underline{\beta}_j$ and ability γ_i, is the probability of subject i responding correctly to item j.

In the one-parameter model where $\underline{\beta}_j = (b_j)$

$$\pi_{ij} = \frac{1}{1+e^{-\eta_{ij}}} \text{ where } \eta_{ij} = \gamma_i - b_j$$

Hence,

$$\ln L(\underline{\gamma}, \underline{\beta}|\underline{y}) = \sum_{i=1}^{I}\sum_{j=1}^{J}\left[y_{ij}\ln\frac{1}{1+e^{-(\gamma_i-b_j)}} + (1-y_{ij})\ln\left(1-\frac{1}{1+e^{-(\gamma_i-b_j)}}\right)\right]$$

In the two-parameter model where $\underline{\beta}_j^{\,T} = (b_j, a_j)^T$

$$\pi_{ij} = \frac{1}{1+e^{-\eta_{ij}}} \text{ where } \eta_{ij} = a_j(\gamma_i - b_j).$$

Hence,

$$\ln L(\underline{\gamma}, \underline{\beta}|\underline{y}) = \sum_{i=1}^{I}\sum_{j=1}^{J}\left[y_{ij}\ln\frac{1}{1+e^{-a_j(\gamma_i-b_j)}} + (1-y_{ij})\ln\left(1-\frac{1}{1+e^{-a_j(\gamma_i-b_j)}}\right)\right]$$

In the three-parameter model where $\underline{\beta}_j^{\,T} = (b_j, a_j, c_j)^T$.

$$\pi_{ij} = c_j + \frac{1-c_j}{1+e^{-\eta_{ij}}} = \frac{1+c_j e^{-\eta_{ij}}}{1+e^{-\eta_{ij}}} \text{ where } \eta_{ij} = a_j(\gamma_i - b_j).$$

Hence,

$$\ln L(\underline{\gamma}, \underline{\beta}|\underline{y}) = \sum_{i=1}^{I}\sum_{j=1}^{J}\left[y_{ij}\ln\frac{1+c_j e^{-a_j(\gamma_i-b_j)}}{1+e^{-a_j(\gamma_i-b_j)}} + (1-y_{ij})\ln\left(1-\frac{1+c_j e^{-a_j(\gamma_i-b_j)}}{1+e^{-a_j(\gamma_i-b_j)}}\right)\right]$$

4.3. ESTIMATION.

4.3.1. Ability Estimation.

When the item parameters of all the test questions answered by a subject are known, the maximum likelihood estimates of the subjects' latent abilities can be obtained by standard methods of maximum likelihood estimation. Let $\underline{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iJ})^T$ represent the responses of subject *i* to *J* test items. Assuming that, conditional on latent trait γ_i, the responses of subject *i* are independent, then the log likelihood for response pattern $\underline{y}_i$ is

$$\ln L(\gamma_i | \underline{y}_i) = \sum_{j=1}^{J} \left[y_{ij} \ln \pi_{ij} + (1 - y_{ij}) \ln(1 - \pi_{ij}) \right]$$

where π_{ij} is a function (either equation (4.3) or (4.5)) of the item parameters and the subject's ability. Differentiating with respect to γ_i and equating the result to zero gives a set of non-linear ML equations which are usually solved iteratively by the Newton-Raphson method. Mislevy (1984, 1985) used the EM algorithm (Dempster *et al,* 1977) to compute estimates of the parameters of the ability distribution when the item parameters of the item response model are known.

Problems with this procedure arise when zero or perfect scores are recorded. When $\underline{y}_i = \underline{0}$, i.e. all responses are incorrect, the likelihood equation is satisfied only when $\gamma_i = -\infty$. When $\underline{y}_i = \underline{1}$, i.e. all responses are correct, the likelihood equation is satisfied only when $\gamma_i = \infty$. There are therefore no maximum likelihood estimates for ability in these cases. In addition, convergence to a local rather than a global maximum of the function can take place in some situations. However this is unlikely to happen if there are more than 20 items in a test (Hambleton and Swaminathan, 1985).

4.3.2. Item Parameter Estimation.

If the latent variables, $\underline{\gamma}$, were known the model parameters could be estimated using the maximum likelihood methods associated with either logistic regression or, where appropriate, the fitting algorithms for GLMs provided by GLIM (Payne, 1987) or other software packages. Since the values of $\underline{\gamma}$ cannot be known, model fitting in IRT has been problematic. In a fixed effects model the abilities appear in the likelihood function as nuisance parameters, the number of which increases with the number of subjects and for this reason it is often not possible to apply asymptotic theory to the estimators of the item parameters. The most widely used procedure is joint maximum likelihood estimation where both item and ability parameters are estimated simultaneously. Conditional maximum likelihood is a method which applies only to the one-parameter model. A method which involves eliminating $\underline{\gamma}$ from the likelihood equations is commonly known as 'marginal' maximum likelihood. All these methods have problems associated with them and parameter estimation for these models is a subject of current research.

4.3.2.1. Conditional Maximum Likelihood Estimation.

Estimation of item parameters is easier if the ability parameters are not present in the likelihood function. Conditional maximum likelihood (CML) (Anderson, 1970 and 1972) estimation is a method of achieving this in the one-parameter (Rasch) model. The total number of items answered correctly by subject i, r_i, is a sufficient statistic for $\underline{\gamma}$. By conditioning on r_i the likelihood can be expressed in terms of $\underline{r}$ instead of $\underline{\gamma}$. However there are no similar sufficient statistics to enable the two- and three-parameter models to be fitted by this method.

4.3.2.2. Marginal Maximum Likelihood Estimation.

This method is so-called because the likelihood function is derived from the marginal distribution of the data. However the method also relies on expressing the likelihood without reference to the ability parameters. If $\underline{y}$ is the response pattern of I subjects attempting J items, then the probability of $\underline{y}$ conditional on ability vector $\underline{\gamma}$ and item parameters $\underline{\beta}$ is

$$P\left(\underline{y}|\underline{\gamma},\underline{\beta}\right)=\prod_{i=1}^{I}\prod_{j=1}^{J}\left[\pi_{ij}^{\,y_{ij}}\left(1-\pi_{ij}\right)^{\left(1-y_{ij}\right)}\right]$$

The joint probability of $\underline{y}$ and $\underline{\gamma}$ is

$$P\left(\underline{y}|\underline{\gamma},\underline{\beta}\right)=\prod_{i=1}^{I}\prod_{j=1}^{J}\left[\pi_{ij}^{\,y_{ij}}\left(1-\pi_{ij}\right)^{\left(1-y_{ij}\right)}\right]f\left(\underline{\gamma}\right)$$

where $f_{\Gamma}\left(\underline{\gamma}\right)$ is the probability distribution of $\underline{\gamma}$ which may be taken to be standard normal. When this joint probability is integrated with respect to the ability parameters the result can be interpreted as the likelihood of $\underline{\beta}$ given $\underline{y}$.

$$L\left(\underline{y}|\underline{\beta}\right)=\int_{-\infty}^{\infty}\prod_{i=1}^{I}\prod_{j=1}^{J}\left[\pi_{ij}^{\,y_{ij}}\left(1-\pi_{ij}\right)^{\left(1-y_{ij}\right)}\right]f_{\Gamma}\left(\underline{\gamma}\right)\delta\underline{\gamma}$$

Bock and Lieberman (1970) originally developed 'marginal' maximum likelihood (MML) estimation for item response models. They used a two-parameter normal ogive curve to model the probability of a correct response. However, because of the high computational requirements their algorithm for maximising the likelihood with respect to the parameters was impractical for use with more than 10-12 item tests.

Bock and Aitkin (1981) (see Section 3.3.1.1) considerably advanced the computation of ML item parameter estimates in the two-parameter item response model with latent ability covariates. They proposed the use of the EM algorithm and approximated the continuous ability distribution of the subjects by a discrete distribution with a finite number of ability

points. The E-step of EM calculates the expected number of correct responses and expected sample size at each ability level. The M-step produces updated item parameter estimates using probit analysis, but does not make use of the general methodology for fitting GLMs.

Thissen (1982) applied Bock and Aitkin's algorithm to the one-parameter logistic model and compared it to CML estimation. The results indicated that the estimates given by Bock and Aitkin's marginal likelihood method were as reliable as the CML estimates and the procedure was easier to implement than CML. In addition CML is not applicable to models with more than one parameter. Mislevy and Bock (1984) produced computer software, BILOG to estimate up to three parameters in logistic item response models using the MML procedures developed and refined by Bock and Lieberman (1970) and Bock and Aitkin (1981). This software included the facility to specify prior distributions for the item parameters and produced optional Bayes' estimates. Although widely used, it could be expensive to run on mainframe computers. A version for PCs was introduced later (PC-BILOG: Mislevy, 1989).

4.3.2.3. Joint Maximum Likelihood Estimation.

This procedure can be used to fit one-, two- or three-parameter models. By assuming that subjects with the same response patterns or equal total scores have the same ability, the $\underline{\gamma}$ can be treated as a finite number of fixed effects and estimated simultaneously with the other model parameters $\underline{\beta}$. The log likelihood function for response pattern $\underline{y} = \left(\underline{y}_1, \underline{y}_2, \ldots, \underline{y}_I\right)^T$ is

$$\ln L\left(\underline{\gamma}, \underline{\beta} | \underline{y}\right) = \sum_{i=1}^{I} \sum_{j=1}^{J} \left[y_{ij} \ln \pi_{ij} + \left(1 - y_{ij}\right) \ln\left(1 - \pi_{ij}\right) \right]$$

where π_{ij} is a function of $\underline{\beta}_j$ and γ_i. If indeterminacy in the model is eliminated the maximum likelihood equations obtained by equating the first derivatives to zero can be solved iteratively.

First, a suitable starting value for $\underline{\beta}$ is obtained. The first ability estimates can then be found by solving the set of N non-linear equations, as described in the previous section. Having obtained an initial set of ability estimates, the item parameters are estimated by solving another system of non-linear equations. This may be accomplished by Newton-Raphson iteration or by the method of scoring (see Chapter 5). New ability estimates can then be calculated. This procedure hopefully results in convergence at the maximum likelihood estimates of $\underline{\beta}$ and $\underline{\gamma}$.

Unfortunately convergence is not always rapid and sometimes does not occur at all, particularly when there are items which have been answered correctly or incorrectly by all subjects, or subjects with zero or perfect scores. Further, it is not always clear whether the iterative procedure has converged to a local or a global maximum of the joint likelihood function. The assumption of a finite number of fixed abilities is difficult to justify when the subjects themselves are not specifically of interest, even though it may lead to asymptotically unbiased and consistent estimators, particularly in the case of the Rasch model.

Rigdon and Tsutakawa (1983) investigated the application of the EM algorithm to joint maximum likelihood estimation of item parameters and ability estimates from the same data. They used a more general version of EM than Bock and Aitkin to find item parameters that maximised the expected log likelihood given the data and estimates from the previous iteration. They assumed a normal distribution for the ability variable. Point estimates are obtained from the posterior ability probability distribution function. using a semi-Bayesian approach. That is, a prior probability distribution is used to obtain Bayesian estimates for the ability parameters but for the item parameters marginal maximum likelihood estimates are calculated using EM. As in all these applications Gauss-Hermite quadrature is employed for approximating integrals on the grounds that the exponent from the normal distribution appears as a factor in the integrands. The full procedure and a second modified version were applied to the one-

parameter logistic model. Tsutakawa (1984) afterwards applied the full version of this algorithm to the two-parameter logistic model.

4.3.2.4. Bayesian Estimation.

Bayesian methods in which prior distributions for item and ability parameters are specified and incorporated into the likelihood function have also been used in item response modelling (O'Hagan, 1976; Sun *et al*, 1996). Both the item parameters and the ability estimates are considered to be random variables. Prior densities for these variables can express knowledge about the difficulty and discrimination power, for example, of the test items. Alternatively, vague priors can be used. A joint posterior distribution for the parameters is obtained by combining the prior information with the conditional distribution of the data and possibly integrating out any nuisance parameters. Finally Bayesian modal estimates are obtained by maximising the joint posterior density function with respect to each parameter.

Swaminathan and Gifford (1982, 1985, 1986) applied a Bayesian approach to the problem of joint item parameter and ability estimation in the logistic model. Three separate papers dealt with the one-, two- and three-parameter cases respectively. In the two-parameter model problems of inadmissible estimates of the discrimination parameter had been experienced with other joint estimation procedures and although the MML methods had produced an improvement these problems still occurred. By specifying a prior distribution for this and the other item parameters as well as for the ability estimates, the authors kept the discrimination estimate from going out of range and found that all the parameters were estimated with increased accuracy. The authors reported similar success when they applied Bayesian methods to joint estimation in the three-parameter logistic model. Specification of priors ensured that the discrimination and chance level parameters stayed within range and improved estimates were

obtained. Mislevy (1986) devised a more general Bayesian framework for logistic item response models with up to three parameters.

In general methods in which both ability and item parameters are estimated from the same data are open to question. It will be demonstrated in this thesis that by integrating out the ability parameters a distinct advantage can be gained.

4.3.2.5. Recent Developments and New Directions.

More recent developments in item parameter estimation focus on adaptations of Monte-Carlo simulation. Albert (1992) introduced Gibbs sampling (Gelfand and Smith, 1990) to IRT. Using a Bayesian model he first specified a joint posterior density for the item and ability variables derived from a two-parameter probit model for the data, a normal density for the random effects and a vague prior for the item parameters. In maximum likelihood estimation the ability parameters are integrated out of the joint posterior density to obtain the marginal posterior density function which can be maximised with respect to the parameters. Albert questioned the use of ML methods such as the EM algorithm which approximate the joint distribution of the parameters with a multivariate normal function. This may not be valid unless samples are large. Instead Albert suggested using the Gibbs sampler to simulate a sample from the joint posterior distribution of the item and ability parameters. From this sample posterior means, modes and standard errors can be calculated for the parameter estimates.

Meng and Schilling (1996) also used Gibbs sampling, this time applying the method to item response models with high dimensional latent variables (i.e. more than one latent ability is assumed to influence the test results of each subject). Meng and Schilling retain the use of the EM algorithm for ML estimation of the item parameters in a two-parameter probit model in the manner of Bock and Aitkin (1981). In the multidimensional case the integration required to

obtain the marginal density of the data becomes increasingly complex as the number of dimensions increases. The authors raise many doubts about the accuracy and usefulness of the usual numerical method of approximation based on Gauss-Hermite quadrature. The Gibbs sampler is employed therefore to simulate the expected complete data log likelihood function (see Section 6.2.4.) whose calculation is normally the task of the E-step of the EM algorithm. By this method the computations required for numerical integration are avoided and the expected frequencies and sample sizes for use by the M-step are calculated from the simulated Gibbs sample. A Newton-Raphson method is suggested for the maximisation routine. No reference is made to GLM methodology. Meng and Schilling (1996) call their procedure the Monte-Carlo Expectation-Maximisation (MCEM) algorithm.

In IRT a subject currently attracting a great deal of interest is multidimensional adaptive testing (MAT) (Segall, 1996). Candidates for testing sit at computers and are presented with an individual selection of test items based on their on-going responses to items in the current test. Algorithms based on ML and Bayesian techniques are required for the simultaneous estimation of multidimensional ability vectors and the selection of items to be presented to the subjects. These tests require large banks of items whose response functions are dependent on several latent variables. Developing methodology for the fitting of response curves in the multidimensional context is a topic for future research.

CHAPTER 5. MAXIMUM LIKELIHOOD ESTIMATION USING THE EM ALGORITHM AND GLIM

5.1. INTRODUCTION.

The EM algorithm (Dempster, Laird and Rubin, 1977) is a general iterative method for obtaining maximum likelihood (ML) estimates of model parameters in situations where the observed data is in some way incomplete. It is called the EM algorithm because it combines two procedures at each iteration, an expectation phase (E-step) and a maximisation phase (M-step). During the E-step the expected complete data log likelihood is computed using estimates of the unknown parameters. This likelihood is then maximised during the M-step to give new parameter estimates. This chapter describes how the EM algorithm can be used in conjunction with the model fitting software package GLIM (Payne, 1987) to obtain maximum likelihood parameter estimates for latent variable GLMs.

Section 5.2 of this chapter looks at the definition of an EM algorithm in its most general form and gives a theoretical description of the two steps. A simple example which has been much quoted in the literature is used to illustrate the main points of the theory and several areas of application are listed. This is followed by a note on aspects of convergence, an area which appears to be not yet fully understood.

Since 1977 the literature on EM has grown considerably. Many publications describe applications of the algorithm. For example, Kimura (1992) applies it to a functional calibration model in which there is an observed data vector $\underline{y}$ which approximates an unknown vector of true measurements $\underline{x}$. The model is assumed to have a normal error distribution. The E-step of EM computes $\hat{\underline{x}}$ and the M-step estimates the model parameters $\underline{\beta}^{(i)}$.

Attempts have been made to improve upon the basic algorithm. The slow convergence of EM in certain situations has led to the development of methods for speeding up the process. The lack of a covariance matrix for the parameter estimates has prompted several proposals for incorporating its computation into the EM iterations; see Louis (1982), Meilijson (1989) and Meng and Rubin (1991). In certain situations the complete data likelihood, which is maximised during the M-step, may not be as computationally simple as it is in most of the cases where the algorithm is an obvious choice. For these situations an ECM algorithm has been proposed (Meng and Rubin, 1993) where the complete data likelihood function is conditional on some function of the parameters. The ML estimates based on the conditional likelihood are simpler to compute in a series of CM-steps than they would be in an equivalent M-step.

GLIM (Payne, 1987) is a computer software package designed principally to fit generalized linear models (see Section 3.2). It does this by an iterative estimation procedure called Iterative Re-weighted Least Squares (IRLS) which is described in Section 5.3. It is shown how GLIM can be extended and adapted to fit GLMs with latent variables by using the package to perform an EM algorithm. A major drawback of the methodology is the lack of easily calculated standard errors for the parameter estimates. Some possible methods of dealing with this problem are discussed in section 5.4.

5.2. THE EM ALGORITHM.

The EM algorithm was first described by Dempster, Laird and Rubin (1977). Prior to this date various forms of the algorithm were in existence, each version written for a particular application or problem. By presenting a unified framework Dempster, Laird and Rubin provided a widely applicable tool. They were able to define the incomplete data situation mathematically and to describe a generalized EM algorithm (GEM) for computing ML estimates in appropriate circumstances.

5.2.1. The Generalized EM Algorithm.

Let there be a data set $\underline{x} = (x_1, x_2, \ldots, x_n) \in \mathrm{X}$ and let its probability density function be $f_{\underline{X}}\left(\underline{x}|\underline{\beta}\right)$. Usually, finding the ML estimate of $\underline{\beta}$ by maximising the log likelihood function $l_{\underline{X}}\left(\underline{\beta}|\underline{x}\right)$ would not be a problem. Suppose for one or more of a variety of possible reasons the observations $\underline{x}$ cannot be recorded. Instead $\underline{y} = \underline{y}(\underline{x}) \in \mathrm{Y}$ is observed. That is, the complete data $\underline{x}$ is only observed *indirectly* through $\underline{y}$ and $\underline{\beta}$ must be estimated from $\underline{y}$, using knowledge of $f_{\underline{X}}\left(\underline{x}|\underline{\beta}\right)$.

Let the probability density (or mass) of the observed data be $g_{\underline{Y}}\left(\underline{y}|\underline{\beta}\right)$.

Then

$$g_{\underline{Y}}\left(\underline{y}|\underline{\beta}\right) = \int_R f_{\underline{X}}\left(\underline{x}|\underline{\beta}\right) d\underline{x} \qquad (5.1)$$

where $R = [\underline{x} : \underline{y} = \underline{y}(\underline{x})]$.

Each iteration of the EM algorithm consists of two steps: the expectation step followed by the maximisation step. The m^{th} iteration begins with the m^{th} estimate of the parameter $\underline{\beta}$, $\hat{\underline{\beta}}^{(m)}$. During the m^{th} expectation step $\hat{\underline{\beta}}^{(m)}$ and the observed data $\underline{y}$ are used to compute an estimate $t^{(m)}\left(\underline{x}\right)$ of a sufficient statistic of the complete data $\underline{x}$. This is used in the maximisation step to estimate new values of $\underline{\beta}$ which maximise the expected log likelihood function $l_{t(\underline{X})}\left[\underline{\beta}|t^{(m)}\left(\underline{x}\right)\right]$. The updated estimates $\hat{\underline{\beta}}^{(m+1)}$ are then input into the $(m+1)^{\text{th}}$ E-step, a new approximation $t^{(m+1)}\left(\underline{x}\right)$ of the sufficient statistic is obtained, and so on until the difference between $\hat{\underline{\beta}}^{(m)}$ and $\hat{\underline{\beta}}^{(m+1)}$ is sufficiently small, for some m, to indicate convergence of the parameter estimates.

Dempster, Laird and Rubin (1977) describe the algorithm for special cases where the distribution $f_{\underline{x}}\left(\underline{x}|\underline{\beta}\right)$ belongs to the regular exponential or curved exponential family. However, the most general level where the distribution of $\underline{x}$ and hence the likelihood function of $\underline{\beta}$ conditional on $\underline{x}$ is not specified is considered here. At the expectation step of the m^{th} iteration a function Q of $\underline{\beta}$ conditional on the current estimates $\underline{\hat{\beta}}^{(m)}$ is computed where

$$\underline{Q}\left(\underline{\beta}|\underline{\hat{\beta}}^{(m)}\right) = \underset{R}{E}\left[l_{\underline{x}}\left(\underline{\beta}|\underline{x}\right)|\underline{y},\underline{\hat{\beta}}^{(m)}\right] \tag{5.2}$$

That is, the expectation of the complete data log likelihood over the region $R = [\underline{x}:\underline{y} = \underline{y}(\underline{x})]$ is taken, using the observed data and the current parameter estimates. This leads to a "pseudo-complete data" problem which is solved during the M-step. Maximum likelihood estimates $\underline{\hat{\beta}}^{(i+1)}$ are chosen to maximise the expected log likelihood or equivalently to solve the equations

$$\frac{d}{d\underline{\beta}}\underline{Q}\left(\underline{\beta}|\underline{\hat{\beta}}^{(i)}\right) = \underline{0} \tag{5.3}$$

where the value of $\underline{\beta}$ which satisfies these equations is $\underline{\hat{\beta}}^{(m+1)}$.

5.2.2. Examples.

Dempster, Laird and Rubin give a simple example to illustrate the principles of the EM algorithm. The data is taken from Rao (1965).

In this example, 197 animals are split into five categories. A multinomial model is assumed in which the probability of being categorised in a given cell depends on the parameter π. The complete data $\underline{x} = \left(x_1, x_2, x_3, x_4, x_5\right)$ would consist of the 5 cell counts from which it would be possible to calculate the ML estimate of π, $\hat{\pi}$, by straightforward means. However, only four counts are observed. These are the total of cells 1 and 2 added

together, and the individual counts for cells 3, 4 and 5. So the observed data is

$\underline{y} = (y_1, y_2, y_3, y_4) = (x_1 + x_2, x_3, x_4, x_5)$.

Cell No.	1	2	3	4	5
Prob.	$\frac{1}{2}$	$\frac{\pi}{4}$	$\frac{1}{4}(1-\pi)$	$\frac{1}{4}(1-\pi)$	$\frac{\pi}{4}$
Count	125		18	20	34

The object is to obtain the ML estimate $\hat{\pi}$, using the four observed counts.

The E-step of the EM algorithm here consists of estimating expected values of x_1 and x_2 given the current estimate of π, $\hat{\pi}^{(m)}$. The probability of an animal being categorised in either cell 1 or cell 2 is $\frac{1}{2} + \frac{\pi}{4}$. The proportion of the total likely to be categorised in cell 1 is $\frac{\frac{1}{2}}{\frac{1}{2}+\frac{\pi}{4}} = \frac{2}{2+\pi}$ and in cell 2 is $\frac{\frac{\pi}{4}}{\frac{1}{2}+\frac{\pi}{4}} = \frac{\pi}{2+\pi}$. Therefore at iteration m,

$$x_1^{(i)} = 125 \times \frac{2}{2+\pi^{(m)}} \quad \text{and} \quad x_2^{(i)} = 125 \times \frac{\pi^{(m)}}{2+\pi^{(m)}} \tag{5.4}$$

Since $x_1^{(i)} = 125 - x_2^{(i)}$, a sufficient statistic $t^{(i)}(\underline{x})$ for the complete data is

$(x_2^{(i)}, x_3, x_4, x_5)$.

At the M-step an updated estimate of π is obtained by maximising the log likelihood function using the complete data values estimated during the E-step.

$$\begin{aligned} l(\pi|\underline{x}) &= \log f_{\underline{X}}(\underline{x}|\pi) \\ &= \log\left[\frac{(x_1+x_2+x_3+x_4+x_5)!}{x_1!x_2!x_3!x_4!x_5!}(\tfrac{1}{2})^{x_1}(\tfrac{\pi}{4})^{x_2}(\tfrac{1-\pi}{4})^{x_3}(\tfrac{1-\pi}{4})^{x_4}(\tfrac{\pi}{4})^{x_5}\right] \end{aligned}$$

Differentiating with respect to π and equating the result to 0 gives the ML estimate of π,

$$\hat{\pi} = \frac{x_2 + x_5}{x_2 + x_3 + x_4 + x_5}$$

Therefore, the M-step, at iteration m, consists of the calculation

$$\hat{\pi}^{(m+1)} = \frac{x_2^{(m)} + 34}{x_2^{(m)} + 34 + 18 + 20} \tag{5.5}$$

The algorithm computes the ML estimate of π using a starting value of 0.5 as follows:

Let $\hat{\pi}^{(0)} = 0.5$

Iteration (0): E-step.

$x_1^{(0)} = 100$ and $x_2^{(0)} = 25$, using (5.4).

M-step.

$$\hat{\pi}^{(1)} = \frac{25 + 34}{25 + 34 + 18 + 20} = 0.608247422 \text{ , using (5.5).}$$

The table below shows the values of $\hat{\pi}^{(m)}$ obtained after 8 iterations:

m	$\hat{\pi}^{(i)}$
0	0.500000000
1	0.608247423
2	0.624321051
3	0.626488879
4	0.626777323
5	0.626815632
6	0.626820719
7	0.626821395
8	0.626821484

The true maximum likelihood estimate of π, obtained analytically, is 0.626821498 (correct to 9 d.p.).

As this example shows EM can be used when missing data is from a multinomial distribution. In analysis of variance the models are assumed to be normal and linear. In this case, an unbalanced design can be made computationally straightforward by using EM to fill

in the design matrix so that it becomes balanced. Similarly, missing values in multivariate normal data can be replaced by calculating the expectations of their means, mean squares and mean products (with other variables) during the E-step of the algorithm. The complete data likelihood can then be maximised, using sufficient statistics calculated from these expectations, to obtain parameter estimates for the underlying multivariate distribution.

In the following example taken from McLachlan and Krishnan (1997) two observations are missing in a 3^2 designed experiment. The data is adapted from Cochran and Cox (1957) and is shown in the following table with a question mark indicating a missing response:

	No. of Lettuce Plants	**Nitrogen Level**	**Phosphorus Level**
i	(y_i)	(x_{1i})	(x_{2i})
1	?	-1	-1
2	409	-1	0
3	341	-1	1
4	413	0	-1
5	358	0	0
6	?	0	1
7	326	1	-1
8	291	1	0
9	312	1	1

In this experiment the response variable is y_i, the number of lettuce plants grown under a combination i, (i = 1,2,...,9) of two factors, nitrogen level, x_{1i}, and phosphorus level, x_{2i}. There are three levels of each of these factors denoted by -1,0 and 1. The following linear regression model is suggested for the data:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i$$

where $\underline{\beta} = (\beta_0, \beta_1, \beta_2)^T$ are parameters to be estimated and the error terms, e_i, are distributed normally with zero means and common variance σ^2.

For the complete data problem, i.e. the full 3^2 experiment, the least-squares estimate of $\underline{\beta}$ is given by the formula

$$\underline{\hat{\beta}} = (X^T X)^{-1} X^T Y$$

where X is the design matrix

$$\begin{bmatrix} 1 & -1 & -1 \\ 1 & -1 & 0 \\ 1 & -1 & 1 \\ 1 & 0 & -1 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & -1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

and Y is the vector of responses. This leads to the following simple results

$$\hat{\beta}_0 = \frac{1}{9}\sum_{i=1}^{9} y_i = \bar{y}$$

$$\hat{\beta}_1 = \frac{1}{6}\left[\sum_{i=7}^{9} y_i - \sum_{i=1}^{3} y_i\right]$$

$$\hat{\beta}_2 = \frac{1}{6}\left[y_3 + y_6 + y_9 - y_1 - y_4 - y_7\right]$$

The error variance σ^2 is estimated by the residual mean square.

By applying a version of the EM algorithm which exploits the simplicity of the complete data analysis, it is possible to calculate least squares estimates of the model parameters in the incomplete data case given above. The procedure which was suggested by Healy and Westmacott (1956) is as follows:

(1) Find starting values for the missing responses.

(2) Compute least-squares estimates of the model parameters using complete data methods.

(3) Calculate new values for the missing data given the parameter estimates computed in step 2.

(4) Update the missing response values with the new estimates computed in step 3.

(5) Return to step 2 and continue until the parameter estimates converge.

This procedure is now applied to the lettuce plant data. An initial estimate of $\underline{\beta}$ can be computed using observations y_2, y_4 and y_5. Using the model $\underline{Y} = X\underline{\beta}$ we obtain

$$\begin{bmatrix} 409 \\ 413 \\ 358 \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

which leads to $\underline{\beta}^T = (358, -51, -55)$. From this we predict

$$y_1 = 358 - 51(-1) - 55(-1) = 464$$

and

$$y_6 = 358 - 51(0) - 55(1) = 303$$

These are the starting values for the missing responses. In step 2 of the algorithm least squares estimates of $\underline{\beta}$ are computed using these starting values in place of the missing data. Thus we have, at iteration 1,

$$\hat{\beta}_0 = \bar{y} = 357.4445$$

$$\hat{\beta}_1 = \frac{1}{6}\left[\sum_{i=7}^{9} y_i - \sum_{i=1}^{3} y_i\right] = -47.5$$

$$\hat{\beta}_2 = \frac{1}{6}\left[y_3 + y_6 + y_9 - y_1 - y_4 - y_7\right] = -41.16667$$

The residual sum of squares is 1868.158. From these estimates new values of y_1 and y_6 are predicted, and so on until convergence. Convergence to four decimal places for the

parameter values and the residual sum of squares occurs after 19 iterations for this data set. The final estimates are

$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{y}_1$	$\hat{y}_6$	**RSS**
355.9690	-41.7830	-31.9458	429.6978	324.0233	885.6418

Dempster, Laird and Rubin discuss many other instances where EM is applied in incomplete data situations, including censored, truncated and grouped data, finite mixture models, hyper-parameter estimation, variance component estimation, factor analysis, discriminant analysis and time series analysis.

5.2.3. Convergence.

In Section 3 of their 1977 paper Dempster, Laird and Rubin discuss the convergence properties of the EM algorithm. A proof that the algorithm always converges to a maximum likelihood estimate of $\underline{\beta}$ given the incomplete data $\underline{y}$ is desirable. However the results given in the paper fall somewhat short of this.

Let $l(\underline{\beta}) = \log g_{\underline{Y}}(\underline{y}|\underline{\beta})$

Theorem 1 shows that, on each iteration of EM,

$$l\left(\underline{\hat{\beta}}^{(i+1)}\right) \geq l\left(\underline{\hat{\beta}}^{(i)}\right)$$

and if $Q\left(\underline{\hat{\beta}}^{(i+1)}|\underline{\hat{\beta}}^{(i)}\right) > Q\left(\underline{\hat{\beta}}^{(i)}|\underline{\hat{\beta}}^{(i)}\right)$

then $l\left(\underline{\hat{\beta}}^{(i+1)}\right) > l\left(\underline{\hat{\beta}}^{(i)}\right)$

In addition, if $\underline{\hat{\beta}}$ is a ML estimate of $\underline{\beta}$ then $l(\underline{\hat{\beta}})$ is a stationary point of the algorithm.

Conditions under which the sequence $l\left(\underline{\hat{\beta}}^{(1)}\right), l\left(\underline{\hat{\beta}}^{(2)}\right), \ldots, l\left(\underline{\hat{\beta}}^{(i)}\right) \ldots$ converges to $l(\underline{\hat{\beta}})$ are

discussed but a mistake in theorem 2 pointed out by Wu (1983) invalidates most of the subsequent proofs presented by Dempster, Laird and Rubin. Wu himself lists several convergence results, in particular stating that, if $l(\underline{\beta})$ is uni-modal with only one stationary point and $\frac{d}{d\underline{\beta}'} Q(\underline{\beta}'|\underline{\beta})$ is continuous in $\underline{\beta}$ and $\underline{\beta}'$, then $\underline{\hat{\beta}}^{(m)} \rightarrow \underline{\hat{\beta}}^{*}$ as $m \rightarrow \infty$ where $l(\underline{\hat{\beta}}^{*}) \geq l(\underline{\hat{\beta}}) \; \forall \; \underline{\hat{\beta}}$.

5.3. GLIM.

5.3.1. Iterative Re-Weighted Least Squares Estimation.

This section consists of an examination of the method of solution of maximum likelihood equations for generalized linear models used by the GLIM package (McCullagh and Nelder, 1989; Dobson, 1990; Payne, 1987). The algorithm described here is an essential component of the software devised to fit latent variable GLMs.

As before, it is assumed that the data vector $\underline{y}$ is a realisation of $\underline{Y}$, a vector of n independent random variables each with a probability distribution from the exponential family (see equation (3.1)), and with expected value $E(\underline{Y}) = \underline{\mu}$. The log likelihood, expressed as a function of the canonical parameter $\underline{\theta}$ conditional on the data $\underline{y}$, with ϕ known, is

$$l(\underline{\theta}|\phi, \underline{y}) = \sum_{i=1}^{n} l_i(\theta_i)$$

where

$$l_i(\theta_i) = \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} - c(y_i, \phi)$$

The i^{th} linear predictor is

$$\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j = \underline{x}_i^T \underline{\beta}$$

where $\underline{x}_i^T$ is the ith row of design matrix X. The linear predictor is related to the expected value μ_i by the link function $g(.)$ so that

$$\mu_i = g^{-1}(\eta_i) = g^{-1}\left(\underline{x}_i^T \underline{\beta}\right)$$

It is required to estimate the parameter vector $\underline{\beta}$. For maximum likelihood estimates, it is necessary to differentiate the log likelihood with respect to $\underline{\beta}$ to obtain the score vector $\underline{u}\left(\underline{\beta}\right)$. The j^{th} component of $\underline{u}$ is

$$u_j\left(\underline{\beta}\right) = \sum_i \frac{\partial l_i(\theta_i)}{\partial \beta_j} = \sum_i \frac{\partial l_i}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_j}$$

In Appendix B it is shown that

$$u_j\left(\underline{\beta}\right) = \sum_i \frac{(y_i - \mu_i)x_{ij}}{\mathrm{var}(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right) \tag{5.6}$$

For ML estimates the system of equations to be solved is

$$\underline{u}(\underline{\beta}) = \underline{0}$$

In all but the case of the normal model these equations are non-linear.

The Newton-Raphson method of solving a system of non-linear equations $\underline{F}(\underline{x}) = \underline{0}$, using suitable starting values $\underline{x}^{(0)}$, finds $\underline{x}^{(m)}$ at iteration $m \geq 1$ as follows:

$$\underline{x}^{(m)} = \underline{x}^{(m-1)} - \left[J\left(\underline{x}^{(m-1)}\right)\right]^{-1} \underline{F}\left(\underline{x}^{(m-1)}\right) \tag{5.7}$$

where $J(\underline{x})$ is the matrix of first derivatives with respect to $\underline{x}$ of the functions $\underline{F}(\underline{x})$ and both are evaluated at $\underline{x}^{(m-1)}$. The iterative system is run until a suitable convergence criterion is satisfied.

Using the Newton-Raphson method to solve the ML equations $\underline{u}(\hat{\underline{\beta}}) = \underline{0}$ where $\hat{\underline{\beta}}$ is the ML estimate of $\underline{\beta}$, the system (5.7) becomes

$$\underline{\hat{\beta}}^{(m)} = \underline{\hat{\beta}}^{(m-1)} - \left[J\left(\underline{\hat{\beta}}^{(m-1)}\right)\right]^{-1} \underline{u}\left(\underline{\hat{\beta}}^{(m-1)}\right)$$

given an initial approximation $\underline{\hat{\beta}}^{(0)}$. Here the elements of J are the second derivatives of the log likelihood function since the components of $\underline{u}$ are the first derivatives (both evaluated at $\underline{\beta} = \underline{\hat{\beta}}^{(m-1)}$). Hence J is the Hessian matrix H where

$$H(\underline{\hat{\beta}}) = \left[\frac{\partial^2 l}{\partial \underline{\beta}^2}\right] = [\underline{u}']$$

The information matrix is $\mathrm{I} = E\left[-\frac{\partial^2 l}{\partial \underline{\beta}^2}\right]_{\underline{\beta}=\underline{\hat{\beta}}^{(m-1)}}$.

In large samples H is approximately equal to the negative of I, so $-J$ can be replaced by I to give

$$\underline{\hat{\beta}}^{(m)} = \underline{\hat{\beta}}^{(m-1)} + \left[\mathrm{I}\left(\underline{\hat{\beta}}^{(m-1)}\right)\right]^{-1} \underline{u}\left(\underline{\hat{\beta}}^{(m-1)}\right)$$

This is known as Fisher's Method of Scoring. Pre-multiplying by I, we have

$$\left[\mathrm{I}\left(\underline{\hat{\beta}}^{(m-1)}\right)\right]\underline{\hat{\beta}}^{(m)} = \left[\mathrm{I}\left(\underline{\hat{\beta}}^{(m-1)}\right)\right]\underline{\hat{\beta}}^{(m-1)} + \underline{u}\left(\underline{\hat{\beta}}^{(m-1)}\right) \qquad (5.8)$$

From (5.6) we have that

$$\mathrm{I}_{jk} = -E\left[\frac{\partial}{\partial \beta_k}\sum_i \frac{(y_i - \mu_i)x_{ij}}{\mathrm{var}(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)\right]$$

which leads (see Appendix B) to

$$\mathrm{I}_{jk} = \sum_{i=1}^{n} x_{ij} w_i x_{ik} \text{ where } w_i = \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \frac{1}{\mathrm{Var}Y_i},$$

or, in matrix notation

$$\mathrm{I} = X^T W X$$

where X is the design matrix for the model and W is a diagonal matrix of weights with elements w_i.

Substituting for I in (5.8), we obtain

$$X^T W X \underline{\hat{\beta}}^{(m)} = X^T W X \underline{\hat{\beta}}^{(m-1)} + \underline{u}\left(\underline{\hat{\beta}}^{(m-1)}\right) \tag{5.9}$$

The j^{th} element of the p-vector on the right hand side of this equation can be written as

$$\sum_{i=1}^{n}\sum_{k=1}^{p} x_{ij} w_i x_{ik} \hat{\beta}_k^{(m-1)} + \sum_{i=1}^{n} w_i (y_i - \mu_i) x_{ij} \left(\frac{\partial \eta_i}{\partial \mu_i}\right)$$

By putting

$$\begin{aligned} z_i &= \sum_{k=1}^{p} x_{ik} \hat{\beta}_k^{(m-1)} + (y_i - \mu_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right) \\ &= \eta_i + (y_i - \mu_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right) \end{aligned}$$

it can be seen that the j^{th} elements are of the form

$$\sum_{i=1}^{n} x_{ij} w_i z_i$$

so (5.9) becomes

$$X^T W X \hat{\beta}^{(m)} = X^T W \underline{z} \tag{5.10}$$

which, with weights $W = V^{-1}\sigma_i^{-2} I_n$, are the normal equations for the general linear model (see Section 2.2 for notation) with dependent variable $\underline{z}$. When the responses are non-normal the weights are functions of the means and hence of $\underline{\beta}$.

The algorithm used by modelling software GLIM (Payne, 1987) to solve the maximum likelihood equations for generalized linear models comprises the following steps:

(i) Evaluate $\underline{z}$ and W at starting values $\underline{\hat{\beta}}^{(0)}$.

(ii) Evaluate $X^T W X$ and $X^T W \underline{z}$.

(iii) Solve (5.10) for $\underline{\hat{\beta}}^{(1)}$.

(iv) Repeat steps (i) to (iii) for $\underline{\hat{\beta}}^{(1)}$ to obtain $\underline{\hat{\beta}}^{(2)}$.

(v) Continue repeating steps (i) to (iii), obtaining $\hat{\underline{\beta}}^{(m)}$ from $\hat{\underline{\beta}}^{(m-1)}$,

$m = 3,4,5,\ldots$, until the convergence criterion is satisfied.

5.3.2. Using GLIM to Fit Generalized Linear Models.

GLIM allows the user to fit standard GLMs by specifying a particular error distribution from a range of exponential family error distributions (Section 3.2.1.1) and a link function (Section 3.2.1.3) to relate the linear predictors to the expected values, also selected from a range of standard functions. Once these two functions have been chosen GLIM can calculate the information it requires for the IRLS model fitting algorithm described above. The link function provides the formula for computing fitted values, i.e. $\underline{\mu}$, from the linear predictor. In addition when a standard link function is chosen $\frac{\partial \underline{\eta}}{\partial \underline{\mu}}$ can be automatically calculated and evaluated at the fitted values. This vector is required for the evaluation of $\underline{z}$ and W in equation (5.10) above. The form of the error distribution determines the formula for $\mathrm{Var}(\underline{Y}_i)$ which is also required for the evaluation of W. Finally GLIM computes a measure of goodness-of-fit known as the deviance which is equal to minus twice the log likelihood. For this it requires an equation to calculate the contribution of each data value to the total deviance. Again this is determined by the error distribution.

Alternatively a non-standard GLM may be fitted. In this case GLIM has the facility to use code supplied by the user to assign values to the four vectors described above; that is, the fitted values, the variances, $\frac{\partial \underline{\eta}}{\partial \underline{\mu}}$, and the deviances.

Finally, the details of the model to be fitted are specified. The components of the linear predictors; that is, the elements of the design matrix X and the parameters to be estimated, are declared. GLIM then fits the model by calculating values which maximise the

likelihood function of the model parameters conditional on the data set using the IRLS algorithm described in Section 5.3.1.

5.3.3. Using GLIM to Fit Latent Variable Generalized Linear Models.

The methodology for fitting latent variable GLMs using the EM algorithm and GLIM depends upon the use of the IRLS procedure during the maximisation phase of EM. To obtain ML estimates for the parameters of latent variable GLMs, the latent variables are regarded as missing data. The expectation step then computes expected complete data values, given current parameter estimates and the observed data. The maximisation step then maximises the complete data log likelihood using these expected values and produces updated parameter estimates which are used in the next E-step unless convergence has occurred.

It will be demonstrated in Chapter 6 that the model for the expected complete data is a GLM. This GLM can be fitted to the expected complete data values produced by the expectation step using standard software such as GLIM. The maximisation step of the EM algorithm can therefore be accomplished in the latent variable situation by performing the iterative IRLS procedure until convergence. The parameter estimates for the fitted complete data model are then taken as the new parameters for the next EM iteration. A new set of expected complete data values are calculated by the next E-step based on these new parameters. Then the GLM is fitted once again to the new expected complete data using IRLS. Iterations continue in this way until the EM procedure is deemed to have converged.

The EM algorithm can be programmed entirely within GLIM. A macro to run the control loop until convergence is required. In a single iteration this macro makes calls to other macros which implement the E-step, the M-step and a convergence check. Code for the E-step must be supplied but the M-step makes use of GLIM's built-in fitting algorithm. A facility for incorporating subroutines written in FORTRAN is also employed. The general

methodology is applicable to all latent variable models where the expected complete data log likelihood belongs to a GLM.

5.4. STANDARD ERRORS OF THE PARAMETER ESTIMATES.

5.4.1. The Problem.

The standard errors of the maximum likelihood parameter estimates $\hat{\underline{\beta}}$ are not easily obtainable when the EM algorithm is used to fit a latent variable GLM. The IRLS maximisation routine that GLIM implements uses the information matrix (the negative of the matrix of expected values of the second derivatives of the log likelihood function) to solve the likelihood equations. The inverse of the information matrix is an asymptotic approximation to the covariance matrix of $\hat{\underline{\beta}}$ and with this matrix readily available GLIM can easily produce approximate standard errors for each of its estimates. For a latent variable GLM, the estimates $\hat{\underline{\beta}}$ from the GLIM M-step are the result of a maximisation of the expected complete data log likelihood. The covariance matrix, as far as GLIM is concerned, is therefore the inverse of the information matrix based on the second derivatives of the expected complete data likelihood. However, the covariance matrix, as far as the fitting of latent variable GLMs is concerned, is the inverse of the information matrix based on the observed data likelihood. As a result the standard errors output by GLIM provide only lower bounds for the true standard errors, since the extra uncertainty introduced by the missing data increases the variability of the parameter estimates.

In order to calculate confidence intervals for the parameters of a latent variable GLM, it is necessary to obtain estimates of the standard errors of the ML estimates calculated with the fitting algorithm described in this chapter. With this objective in mind some asymptotic likelihood theory is reviewed in Section 5.4.2 in order to establish the theoretical sampling distribution of $\hat{\underline{\beta}}$ the ML estimate of $\underline{\beta}$. Then in Section 5.4.3 a review

of some alternative methods for finding standard errors of the parameter estimates is presented.

5.4.2. Asymptotic Likelihood Theory.

The following section contains a brief review of large sample likelihood theory (Kalbfleisch and Prentice, 1989, Dobson, 1990). As before, let the observed data log likelihood be

$$l(\underline{\beta}) = \log g_{\underline{Y}}(\underline{y}|\underline{\beta})$$

The 'score vector' $\underline{u} = (u_1, u_2, \ldots, u_p)^T$ is defined as the vector of first derivatives with respect to $\underline{\beta}$

$$\underline{u}(\underline{\beta}) = \frac{\partial}{\partial \underline{\beta}} l(\underline{\beta})$$

Asymptotically the Central Limit theorem applies to $\underline{u}$ which has a multivariate normal distribution in large samples with mean $\underline{0}$ and covariance matrix I, the information matrix. This is given by the following well-known results (see, for example, Dobson (1990), Appendix A):

(i) $\quad E(\underline{u}) = \underline{0}$

and

(ii) $\quad E(\underline{u}\underline{u}^T) = E[-\underline{u}']$

$$\Rightarrow \mathrm{I} = E[-\underline{u}']$$

where the covariance matrix of $\underline{u}$, $\mathrm{I} = \mathrm{I}(\underline{\beta})$, is the matrix of negatives of the second derivatives known as the 'information matrix'. Since it is a function of the unknown $\underline{\beta}$ it is usually approximated by evaluating at $\underline{\beta} = \hat{\underline{\beta}}$. In addition, in large samples the expectation over $\underline{Y}$ can be replaced by the actual data $\underline{y}$. That is, asymptotically,

$$\left[-\underline{u}'\right]_{\underline{Y}=\underline{y}} \to \mathrm{I}(\hat{\underline{\beta}}) \text{ as } n \to \infty$$

where n is the dimension of $\underline{y}$.

The Hessian matrix, $H(\hat{\underline{\beta}})$, of second derivatives evaluated at $\underline{y}$ and $\hat{\underline{\beta}}$ is

$$H(\hat{\underline{\beta}}) = \left[\frac{\partial^2 l}{\partial \underline{\beta}^2}\right] = \left[\underline{u}'\right]$$

so in large samples

$$\mathrm{I}(\hat{\underline{\beta}}) \approx -H(\hat{\underline{\beta}}) \tag{5.11}$$

The sampling distribution of $\hat{\underline{\beta}}$, the maximum likelihood estimate of $\underline{\beta}$ will now be considered. Using Taylor's approximation to the first order about the point $\underline{\beta} = \hat{\underline{\beta}}$,

$$\underline{u}(\underline{\beta}) \approx \underline{u}(\hat{\underline{\beta}}) + H(\hat{\underline{\beta}})(\underline{\beta} - \hat{\underline{\beta}})$$

Since $\hat{\underline{\beta}}$ is the ML estimate $\underline{u}(\hat{\underline{\beta}}) = \underline{0}$ and $H(\hat{\underline{\beta}})$ can be replaced by $-\mathrm{I}(\hat{\underline{\beta}})$ as in (5.11), then

$$(\hat{\underline{\beta}} - \underline{\beta}) \approx \mathrm{I}^{-1}\underline{u}(\underline{\beta}) \tag{5.12}$$

provided I is non-singular. Taking expectations of both sides gives

$$E(\hat{\underline{\beta}} - \underline{\beta}) = \underline{0}$$

That is, $\hat{\underline{\beta}}$ is an asymptotically unbiased estimate of $\underline{\beta}$.

The covariance matrix for $\hat{\underline{\beta}}$ can be derived from (5.12):

$$\begin{aligned} E\{(\hat{\underline{\beta}} - \underline{\beta})(\hat{\underline{\beta}} - \underline{\beta})^T\} &\approx E\{\mathrm{I}^{-1}\underline{u}(\underline{\beta})\underline{u}(\underline{\beta})^T(\mathrm{I}^{-1})^T\} \\ &= \mathrm{I}^{-1} \end{aligned}$$

since I is symmetric and $\mathrm{I} = E(\underline{u}\underline{u}^T)$.

Thus the ML estimate $\hat{\underline{\beta}}$ is asymptotically multivariate normal with mean $\underline{\beta}$ and covariance matrix I^{-1}. It follows that the standard error of the i^{th} component of $\hat{\underline{\beta}}$,

$i = 1,2,\ldots,p$, is given by the square root of the i^{th} element on the diagonal of I^{-1}, and a 95% confidence interval for β_i is

$$\hat{\beta}_i - 1.96\sqrt{\mathrm{I}^{-1}_{ii}} < \beta_i < \hat{\beta}_i + 1.96\sqrt{\mathrm{I}^{-1}_{ii}}$$

The variances and covariances of the components of $\underline{\hat{\beta}}$ are determined by the degree of curvature of the likelihood function $l_{\underline{Y}}(\underline{\beta}|\underline{y})$ around $\underline{\hat{\beta}}$. A flat likelihood function is associated with a large variance in the ML estimate since the difference between the ML estimate and the true parameter value may be large even for small differences in their likelihoods. Conversely a steeply curved likelihood function indicates small variations in the estimates. A large difference between the ML estimate and the true value becomes a much more unlikely event since the difference in the likelihoods is also large. Thus the standard errors of the components of $\underline{\hat{\beta}}$ are calculated from the second derivatives of the observed data likelihood function. For a one parameter model

$$\mathrm{var}(\hat{\beta}) = -\left(\frac{\partial^2 l}{\partial \beta^2}\right)^{-1}$$

Thus, the greater the curvature of the log likelihood function, the smaller the variance of the ML estimate.

The effect of the missing data is to increase the variability in the parameter estimates. The standard errors of estimates that maximise the complete data likelihood are smaller than those based on incomplete data; the greater the amount of missing data, the larger the standard errors of $\underline{\hat{\beta}}$.

One of the drawbacks of the EM algorithm is that it does not automatically or with any ease provide estimates of the standard errors of the ML parameter estimates $\underline{\hat{\beta}}$. This is because the method finds parameter estimates which maximise the observed (incomplete) data likelihood $l_{\underline{Y}}(\underline{\beta}|\underline{y})$ by actually maximising the expected complete data likelihood

function $l_{\underline{X}}\left(\underline{\beta}|\underline{x}\right)$. The information matrix obtained during the computation does not therefore relate to the relevant likelihood function.

The problem in making inferences about $\hat{\underline{\beta}}$ when using the EM algorithm lies in the calculation of the information matrix, I, for the incomplete data. The likelihood function for the incomplete data is frequently algebraically complex and the process of finding its first and second derivatives analytically can be extremely difficult.

5.4.3. Alternative Methods Of Calculating Standard Errors.

5.4.3.1. Analytic Methods.

The matrix of second derivatives of the observed data log likelihood is given by the expression

$$\frac{\partial^2}{\partial\underline{\beta}^2} l_{\underline{Y}}\left(\underline{\beta}|\underline{y}\right) = \sum_{i=1}^{N} \frac{\partial^2}{\partial\underline{\beta}^2} \ln \int_R p_{\underline{Y}_i|\Gamma}\left(\underline{y}_i|\gamma_i,\underline{\beta}\right) f_\Gamma\left(\gamma_i\right) d\gamma_i ; \qquad (5.13)$$

see Section 6.2.2.

The information matrix is

$$\mathrm{I} = \underset{\underline{Y}}{E}\left[-\frac{\partial^2}{\partial\underline{\beta}^2} l_{\underline{Y}}\left(\underline{\beta}|\underline{y}\right)\right] \qquad (5.14)$$

and the covariance matrix of $\hat{\underline{\beta}}$ is I^{-1}. Attempts at direct algebraic differentiation of the observed data log likelihood have not yielded any useful results.

The information matrix defined in (5.14) is the *expected* information matrix. The *observed* information matrix is the matrix of second derivatives evaluated at $\underline{Y} = \underline{y}$.

$$\mathrm{I}_O = \left[-\frac{\partial^2}{\partial\underline{\beta}^2} l_{\underline{Y}}\left(\underline{\beta}|\underline{y}\right)\right]_{\underline{Y}=\underline{y}} \qquad (5.15)$$

In large samples (5.14) and (5.15) are approximately equal. For GLMs with canonical link functions they are always equal.

Louis (1982) describes a method for calculating the observed information matrix when using the EM algorithm. He shows that the observed information matrix (5.15) can be calculated from the expected values of the first and second derivatives of the complete data

log likelihood. Expectations are taken with respect to the posterior distribution of $\underline{\gamma}$ conditional on the data and $\hat{\underline{\beta}}$ (see Section 6.2.4).

$$\mathrm{I}_o = \underset{\gamma|y}{E}\left[-\frac{\partial^2}{\partial \underline{\beta}^2} l_{\underline{Y},\underline{\gamma}}\left(\underline{\beta}|\underline{y},\underline{\gamma}\right)\right] - \underset{\gamma|y}{E}\left\{\left[\frac{\partial}{\partial \underline{\beta}} l_{\underline{Y},\underline{\gamma}}\left(\underline{\beta}|\underline{y},\underline{\gamma}\right)\right]\left[\frac{\partial}{\partial \underline{\beta}} l_{\underline{Y},\underline{\gamma}}\left(\underline{\beta}|\underline{y},\underline{\gamma}\right)\right]^T\right\}$$

when all terms are evaluated at $\underline{\beta} = \hat{\underline{\beta}}$. The first term represents the observed information matrix of the expected complete data and the second the expected square of the score functions of the complete data. Although the first term should be available as a by-product of the final M-step of the EM algorithm, it has so far proved difficult to express the second term in an algebraic form which can be easily evaluated.

A further method, suggested by Meilijson (1989), requires that the observed data be independent and identically distributed, and so is not applicable to conditionally independent responses.

5.4.3.2. The SEM Algorithm.

Meng and Rubin (1991) suggest a 'supplemented' EM algorithm (SEM) for calculating the variance/covariance matrix of $\hat{\underline{\beta}}$. The essential part of this method is the calculation of a p x p matrix DM, where p is the dimension of $\underline{\beta}$. The elements of DM are in some sense linear approximations, based on Taylor's theorem, to the individual convergence rates of the p parameters. DM is computed iteratively until convergence of each one of its elements is achieved. The rationale behind this computation is that the convergence rate of EM is dependent on the amount of missing data; the more missing data the slower the convergence of EM. It can be shown that the extra variability of the parameter estimates due to the missing data is a function of DM.

$$\mathrm{I}_o^{-1} = \mathrm{I}_{oc}^{-1}(1 - DM)^{-1}$$

where I_o is the observed information matrix of equation (5.15) and I_{oc} is the expected complete data observed information matrix obtainable from the final EM iteration.

Unfortunately the author's attempts to apply SEM have not met with great success. It was found that a large number of time-consuming extra iterations of EM were required to calculate the elements of DM. In addition, very small differences in some parameters over successive iterations led to instability in the numerical estimates of the convergence rates.

Further work is needed to assess whether SEM could be modified or adapted to suit the application being researched.

5.4.3.3. "Aitkin's Method".

Some empirical evidence of the sampling distributions of the maximum likelihood parameter estimates has been obtained from the simulation study described in Chapter 9. Estimates of standard errors obtained from this study have compared well with estimates obtained from a simple method suggested by Aitkin (1994) which makes use of the fact that, under certain conditions, the likelihood ratio test and the Wald test are equivalent. The significance of a single parameter β can be tested in two different ways. A z-statistic which has a standard normal distribution can be computed under the null hypothesis that $\beta = 0$ so that

$$z = \frac{\hat{\beta}}{s.e.(\hat{\beta})} \tag{5.16}$$

where $\hat{\beta}$ is the estimate of β. Alternatively a full model including the parameter β and a reduced model without β are both fitted. The change in deviance (minus twice the log likelihood) between the full and reduced models has a chi-squared distribution with 1 degree of freedom under an equivalent null hypothesis (i.e. that the two models fit equally well). The signed square root of the difference in deviance is therefore also a standard normal variable. Therefore (5.16) can be equated to the square root of the difference in deviance between the two models

$$\sqrt{\Delta D} = z = \frac{\hat{\beta}}{s.e.(\hat{\beta})}$$

which leads to

$$s.e.(\hat{\beta}) = \frac{\hat{\beta}}{\sqrt{\Delta D}}$$

This method is generally useful and easy to implement. However, there are obvious limitations when the standard error of a combination of parameters (e.g. $\hat{\beta}_1 - \hat{\beta}_2 + \hat{\beta}_3$) is required.

CHAPTER 6. A LATENT VARIABLE GLM FOR BINARY RESPONSES.

6.1. INTRODUCTION.

In Section 6.2 a general model for binary responses is considered. Three different likelihood functions associated with this model are examined. The maximisation of the likelihood of the observed response data which is required in order to fit the model has proved intractable by direct methods. The response data together with the values of the latent covariates are referred to as the 'complete data'. The log likelihood function for the complete data would normally be maximised by a standard fitting algorithm for GLMs if the covariates were known. Instead the expected complete data log likelihood function is considered. This likelihood function is shown to be also the likelihood function of a (different but related) GLM where the abilities are discrete and known. The expected complete data log likelihood function can therefore be maximised by the IRLS fitting algorithm.

6.2. THE BINARY RESPONSE MODEL.

The IRT models described in Section 4.2 are examples of GLMs for binary response data with single latent variables. In the following section this model is described in more general terms and a likelihood function is obtained for the observed data vector. Standard methods of maximum likelihood estimation cannot be applied directly to this function. Instead the expectation of the log likelihood of the complete data, i.e. the observed data vector $\underset{\sim}{y}$ *and* the latent covariates $\underset{\sim}{\gamma}$, is calculated and this likelihood is compared with the likelihood function of a GLM with fixed effects.

6.2.1. Defining the Latent Variable GLM.

The response y_{ij}, a realisation of random variable Y_{ij}, is the jth (j=1,2,...,J) observation (e.g. response to item j) on the i^{th} unit (e.g. subject i) (i=1,2,...,I). It is possible that not all J observations are recorded for every unit. In this case the total number of responses recorded for unit i is denoted $J(i)$ and the total number of units responding to item j is denoted $I(j)$. The expected value of the random variable is dependent on unknown parameter vector $\underline{\beta}_j$ and latent covariate γ_i (e.g. the ability of subject i), a realisation of latent variable Γ_i. It is assumed that the conditional distribution of Y_{ij} is binomial

$$Y_{ij} \sim \mathrm{Bi}\left(1, \pi_{ij}(\gamma_i)\right)$$

A non-canonical link function is chosen to model the relationship between the linear predictor and the conditional mean. Its inverse is

$$\pi_{ij} = c_j + \frac{1 - c_j}{1 + \exp(-\eta_{ij})}$$

where c_j is a known parameter representing the lower asymptote of the logistic function. When $c_j = 0$, the link function is canonical for the binomial distribution. Thus c_j represents the lowest value of the expected response which can arise from extremely low values of γ_i. In some models $c_j > 0$. For example, in an IRT model for a multiple-choice test response a positive value for this parameter reflects the underlying positive probability of a very low ability candidate correctly answering a question by guesswork alone.

The link function as defined above is directly comparable to the three-parameter IRT model discussed in Chapter 4. It must be emphasised however that in the treatment that follows the 'guessing' parameter c_j is not estimated (as it would be in most IRT applications). Instead it

is fixed at a pre-determined value. If this parameter were not fixed the model would not be a GLM and it would not be possible to apply GLM methodology to its fitting algorithm

The linear predictor associated with observation y_{ij}, given that $\Gamma_i = \gamma_i$, is

$$\eta_{ij} = \underline{x}_j^T \underline{\varphi}_j + \alpha_j \gamma_i$$

where $\underline{x}_j^T$ is the row of the fixed effects design matrix associated with responses to item j, $\underline{\varphi}_j$ is a vector of fixed effect parameters and α_j is the slope on γ_i. Comparing this to the IRT model given in equations (4.5) and (4.6) where

$$\eta_{ij} = a_j(\gamma_i - b_j)$$

it can be seen that α_j is equivalent to the discrimination parameter a_j and $\underline{x}_j^T \underline{\varphi}_j$ corresponds to $-a_j b_j$ where b_j is the difficulty parameter. The vector of unknown parameters in the model is denoted $\underline{\beta}$ where $\underline{\beta}_j^T = (\underline{\varphi}_j^T, \alpha_j)$.

If $\Gamma_i \sim N(0,1)$ and $\alpha_j = \alpha$ for all j, then $\alpha\Gamma_i \sim N(0, \alpha^2)$ so that α^2 is equivalent to the variance of the random effect.

If vector $\underline{\gamma}$ is assumed to contain observed values instead of unknown covariates, the model described is an ordinary fixed effect GLM. The likelihood function based on known $\underline{\gamma}$ and responses $\underline{y}$ can be thought of as the 'complete data' likelihood.

For $Y_{ij} \sim \text{Bi}(n_{ij}, \pi_{ij})$, the log likelihood expressed as a function of π_{ij} is

$$l_{\underline{Y}}(\pi_{ij}|y_{ij}) = y_{ij} \ln \pi_{ij} + (n_{ij} - y_{ij}) \ln(1 - \pi_{ij}) + C_1$$

where $C_1 = \ln\binom{n_{ij}}{y_{ij}}$ is a constant. So, assuming independent observations conditional on $\underline{\gamma}$

$$l_{\underline{Y},\underline{\Gamma}}\left(\underline{\pi}|\underline{y},\underline{\gamma}\right)=\sum_{i=1}^{I}\sum_{j=1}^{J(i)}\left(y_{ij}\ln\pi_{ij}+\left(n_{ij}-y_{ij}\right)\ln\left(1-\pi_{ij}\right)\right)+C_2 \qquad (6.1)$$

where $C_2=\sum_{i=1}^{I}\sum_{j=1}^{J(i)}\ln\binom{n_{ij}}{y_{ij}}$ is a constant.

With $n_{ij}=1$, for all i and j, and $\pi_{ij}=\pi_{ij}\left(\underline{\beta}_j,\gamma_i\right)$, the log likelihood of the complete data expressed as a function of the unknown parameters is

$$l_{\underline{Y},\underline{\Gamma}}\left(\underline{\beta}|\underline{y},\underline{\gamma}\right)=\sum_{i=1}^{I}\sum_{j=1}^{J(i)}\left(y_{ij}\ln\pi_{ij}\left(\underline{\beta}_j,\gamma_i\right)+\left(1-y_{ij}\right)\ln\left(1-\pi_{ij}\left(\underline{\beta}_j,\gamma_i\right)\right)\right)+C_2 \qquad (6.2)$$

with $C_2=0$.

This function is normally maximised using standard software such as GLIM to obtain maximum likelihood parameter estimates. However in a latent variable GLM $\underline{\gamma}$ is unknown. What is really required is the maximisation of the log likelihood of the observed data $\underline{y}$ which is not as straightforward, as the next section demonstrates.

6.2.2. The Observed Data Likelihood:

Since conditional independence of the responses is assumed, the probability function of the vector response variable $\underline{Y}$ conditional on item parameters $\underline{\beta}$ and latent vector $\underline{\gamma}$ is

$$p_{\underline{Y}|\underline{\gamma}}\left(\underline{Y}=\underline{y}|\underline{\gamma},\underline{\beta}\right)=\prod_{i=1}^{I}\prod_{j=1}^{J(i)}\pi_{ij}^{\,y_{ij}}\left(1-\pi_{ij}\right)^{1-y_{ij}} \qquad (6.3)$$

Since the γ_i are also independent the p.d.f. of $\underline{\Gamma}$ the latent variable vector is

$$f_{\underline{\Gamma}}\left(\underline{\gamma}\right)=\left(\frac{1}{\sqrt{2\pi}}\right)^{I}\prod_{i=1}^{I}\exp\left[-\frac{\gamma_i^{\,2}}{2}\right] \qquad (6.4)$$

The joint probability distribution of $\underline{Y}$ and $\underline{\Gamma}$ is

$$p_{\underline{Y},\underline{\Gamma}}\left(\underline{y},\underline{\gamma}|\underline{\beta}\right) = p_{\underline{Y}|\underline{\gamma}}\left(\underline{y}|\underline{\gamma},\underline{\beta}\right) f_{\underline{\Gamma}}\left(\underline{\gamma}\right) \tag{6.5}$$

and the marginal distribution of $\underline{Y}$ is obtained by integration with respect to $\underline{\gamma}$

$$p_{\underline{Y}}\left(\underline{y}|\underline{\beta}\right) = \int_{R} p_{\underline{Y}|\underline{\gamma}}\left(\underline{y}|\underline{\gamma},\underline{\beta}\right) f_{\underline{\Gamma}}\left(\underline{\gamma}\right) d\underline{\gamma} \text{ where } R = [\underline{\gamma}: \underline{y} = \underline{y}(\underline{\gamma})]$$

or, substituting (6.3) and (6.4),

$$p_{\underline{Y}}\left(\underline{y}|\underline{\beta}\right) = \left(\frac{1}{\sqrt{2\pi}}\right)^{I} \prod_{i=1}^{I} \int_{R} \left(\prod_{j=1}^{J(i)} \pi_{ij}^{\,y_{ij}} \left(1-\pi_{ij}\right)^{1-y_{ij}} \right) \exp\left[-\frac{\gamma_i^{\,2}}{2}\right] d\gamma_i \tag{6.6}$$

This integration cannot be performed analytically and so an approximation is introduced based upon a numerical quadrature rule of some form (see Chapter 7). Essentially, the integral is replaced by a weighted summation over K nodes γ_k, with weights w_k, (k=1,2,...,K) the values of which depend upon the chosen integration strategy. Equation (6.6) now becomes

$$p_{\underline{Y}}\left(\underline{y}|\underline{\beta}\right) \approx \left(\frac{1}{\sqrt{2\pi}}\right)^{I} \prod_{i=1}^{I} \sum_{k=1}^{K} \left(\prod_{j=1}^{J(i)} \pi_{jk}^{\,y_{ij}} \left(1-\pi_{jk}\right)^{1-y_{ij}} \right) \exp\left[-\frac{\gamma_k^{\,2}}{2}\right] w_k \tag{6.7}$$

where now $\pi_{jk} = c_j + \dfrac{1-c_j}{1+\exp(-\eta_{jk})}$ and $\eta_{jk} = \underline{x}_j^{\,T} \underline{\varphi}_j + \alpha_j \gamma_k$. The continuous distribution of the latent ability variable has been replaced by a discrete number, K, of ability points. Instead of associating each subject with an individual ability γ_i there is now a restricted range of K abilities γ_k.

The likelihood of the data $\underline{y}$ is given by equation (6.7) with $\underline{\beta}$ the variable. The log likelihood function for the observed data is therefore

$$l_{\underline{Y}}\left(\underline{\beta}|\underline{y}\right) = \ln p_{\underline{Y}}\left(\underline{y}|\underline{\beta}\right)$$

or

$$l_{\underline{Y}}\left(\underline{\beta}|\underline{y}\right) \approx I \ln \frac{1}{\sqrt{2\pi}} + \sum_{i=1}^{I} \ln \sum_{k=1}^{K} \left(\prod_{j=1}^{J(i)} \pi_{jk}^{\,y_{ij}} \left(1-\pi_{jk}\right)^{1-y_{ij}} \right) \exp\left[-\frac{\gamma_k^{\,2}}{2}\right] w_k \tag{6.8}$$

Theoretically this likelihood must be maximised in order to fit the latent variable GLM. The likelihood equations are a non-linear system of equations which require the matrix of second derivatives for a Newton-Raphson-type method of solution. In addition estimates of the standard errors of the parameter estimates depend on the second derivatives and also require the inversion of this matrix.. Expressions for the second derivatives are very complicated and alternative methods are sought for the solution of the likelihood equations. The EM algorithm provides an attractive alternative method as it does not require the differentiation of the complicated function shown in equation (6.8).In addition it is supported by proven convergence theorems (Wu, 1983). Results obtained by using EM are therefore more reliable.

The method of fitting the latent variable GLM using the EM algorithm incorporating GLIM was outlined in the previous chapter. The object is to maximise the observed data likelihood (6.8), that is the likelihood function of $\underline{y}$ (irrespective of $\underline{\gamma}$) , using the information in the complete data likelihood (6.2). To achieve this the EM algorithm maximises the expectation of the complete data log likelihood, instead of the actual complete data log likelihood. To derive an expression for the expected complete data likelihood, it is necessary to consider first the posterior distribution of the latent variable.

6.2.3. The Posterior Distribution of the Latent Variable:

The computation of the expected complete data likelihood requires the distribution of $\underline{\Gamma}$ conditional on data vector $\underline{y}$ and the computed parameter estimates $\hat{\underline{\beta}}$. This 'posterior' distribution is found using Bayes' Theorem

$$f_{\underline{\Gamma}|\underline{y}}\left(\underline{\gamma}|\underline{y},\hat{\underline{\beta}}\right)=\frac{p_{\underline{Y}|\underline{\gamma}}\left(\underline{y}|\underline{\gamma},\hat{\underline{\beta}}\right)f_{\underline{\Gamma}}\left(\underline{\gamma}\right)}{\int_{R} p_{\underline{Y}|\underline{\gamma}}\left(\underline{y}|\underline{\gamma},\hat{\underline{\beta}}\right)f_{\underline{\Gamma}}\left(\underline{\gamma}\right)d\underline{\gamma}} \qquad (6.9)$$

where the probability distribution of $\underline{Y}$ given $\underline{\Gamma} = \underline{\gamma}$ and of $\underline{\Gamma}$ are given by functions (6.3) and (6.4).

The integral in the denominator is a normalising constant which can be approximated by a weighted sum over K nodes as in equation (6.7). This leads to

$$f_{\underline{\Gamma}|\underline{y}}\left(\underline{\gamma}|\underline{y},\hat{\underline{\beta}}\right) \approx \frac{p_{\underline{Y}|\underline{\gamma}}\left(\underline{y}|\underline{\gamma},\hat{\underline{\beta}}\right) f_{\underline{\Gamma}}\left(\underline{\gamma}\right)}{\sum_{k=1}^{K} p_{\underline{Y}_i|\underline{\gamma}}\left(\underline{y}_i|\gamma_k,\hat{\underline{\beta}}\right) f_{\Gamma}\left(\gamma_k\right) w_k} \tag{6.10}$$

Substituting functions (6.3) and (6.4) in the above, the posterior probability distribution of the latent variable associated with subject i is

$$f_{\Gamma|\underline{y}_i}\left(\gamma_i|\underline{y}_i,\hat{\underline{\beta}}\right) = C_3^{-1} \prod_{i=1}^{I}\left[\prod_{j=1}^{J(i)} \pi_{ij}^{\,y_{ij}}\left(1-\pi_{ij}\right)^{1-y_{ij}}\right]\exp\left[-\frac{\gamma_i^{\,2}}{2}\right] \tag{6.11}$$

where the normalising constant C_3 is

$$C_3 \approx \sum_{k=1}^{K}\prod_{i=1}^{I}\left[\prod_{j=1}^{J(i)} \pi_{jk}^{\,y_{ij}}\left(1-\pi_{jk}\right)^{1-y_{ij}}\right]\exp\left[-\frac{\gamma_k^{\,2}}{2}\right] w_k .$$

6.2.4. The Expected Complete Data Likelihood.

The log likelihood function for the complete data $\underline{y}$ and $\underline{\gamma}$, as a function of $\underline{\beta}$, is obtained from the joint p.d.f. of the two variables $\underline{Y}$ and $\underline{\Gamma}$:

$$p_{\underline{Y},\underline{\Gamma}}\left(\underline{y},\underline{\gamma}|\underline{\beta}\right) = \left(\frac{1}{\sqrt{2\pi}}\right)^{I} \prod_{i=1}^{I}\left[\prod_{j=1}^{J(i)} \pi_{ij}^{\,y_{ij}}\left(1-\pi_{ij}\right)^{1-y_{ij}}\right]\exp\left[-\frac{\gamma_i^{\,2}}{2}\right]$$

Taking natural logarithms

$$l_{\underline{Y},\Gamma}\left(\underline{\beta}|\underline{y},\underline{\gamma}\right) = I\ln\frac{1}{\sqrt{2\pi}} + \sum_{i=1}^{I}\sum_{j=1}^{J(i)}\left[y_{ij}\ln\pi_{ij} + (1-y_{ij})\ln\left(1-\pi_{ij}\right)\right] - \sum_{i=1}^{I}\left[\frac{\gamma_i^{\,2}}{2}\right] \tag{6.12}$$

The function Q, the expected complete data log likelihood, is found by taking expectations of function (6.12) over the posterior distribution of $\underline{\Gamma}$ given the data and parameter estimates as in equation (6.9).

$$Q\left(\underline{\beta}|\underline{y},\underline{\hat{\beta}}\right)=\int_R\sum_{i=1}^{I}\sum_{j=1}^{J(i)}\left[y_{ij}\ln\pi_{ij}+(1-y_{ij})\ln\left(1-\pi_{ij}\right)\right]f_{\Gamma|\underline{y}_i}\left(\gamma_i|\underline{y}_i,\underline{\hat{\beta}}\right)d\gamma_i+C_4$$

where

$$C_4=\int_R\left[\ln\frac{1}{\sqrt{2\pi}}-\sum_{i=1}^{I}\frac{\gamma_i^{\,2}}{2}\right]f_{\Gamma|\underline{y}_i}\left(\gamma_i|\underline{y}_i,\underline{\hat{\beta}}\right)d\gamma_i$$

which is a constant. Evaluation of the integral required for the above expectation is impossible by analytic methods. Applying quadrature rules to approximate the integral using K nodes γ_k, with weights w_k gives

$$Q\left(\underline{\beta}|\underline{y},\underline{\hat{\beta}}\right)=\sum_{k=1}^{K}\sum_{i=1}^{I}\sum_{j=1}^{J(i)}\left[y_{ij}\ln\pi_{ij}+(1-y_{ij})\ln\left(1-\pi_{ij}\right)\right]f_{\Gamma|\underline{y}_i}\left(\gamma_k|\underline{y}_i,\underline{\hat{\beta}}\right)w_k+C_4 \qquad (6.13)$$

The conditional (given the data) posterior probability that the value of the latent variable associated with unit i is γ_k is denoted P_{ik} where $\sum_{k=1}^{K}P_{ik}=1$. P_{ik} can be expressed as

$$P_{ik}=\frac{p_{\underline{Y}_i|\gamma}\left(\underline{y}_i|\gamma_k,\underline{\hat{\beta}}\right)f_\Gamma(\gamma_k)w_k}{\sum_{k=1}^{K}p_{\underline{Y}_i|\gamma}\left(\underline{y}_i|\gamma_k,\underline{\hat{\beta}}\right)f_\Gamma(\gamma_k)w_k} \qquad (6.14)$$

where the conditional probability of response vector y_i is

$$p_{\underline{Y}_i|\gamma}\left(\underline{y}_i|\gamma_k,\underline{\beta}\right)=\prod_{j=1}^{J(i)}\pi_{ij}^{\,y_{ij}}\left(1-\pi_{ij}\right)^{1-y_{ij}}$$

and the probability of γ_k is $f_\Gamma(\gamma_k)w_k$ with

$$f_\Gamma(\gamma_k)=\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{\gamma_k^{\,2}}{2}\right]$$

Using the posterior distribution of the continuous variable $\underline{\Gamma}$ given in (6.10)

$$f_{\Gamma|\underline{y}_i}\left(\gamma_k|\underline{y}_i,\hat{\underline{\beta}}\right) \approx \frac{p_{\underline{Y}_i|\gamma}\left(\underline{y}_i|\gamma_k,\hat{\underline{\beta}}\right)f_{\Gamma}(\gamma_k)}{\sum_{k=1}^{K} p_{\underline{Y}_i|\gamma}\left(\underline{y}_i|\gamma_k,\hat{\underline{\beta}}\right)f_{\Gamma}(\gamma_k)w_k} \tag{6.15}$$

Substituting (6.15) in (6.14)

$$P_{ik} = f_{\Gamma|\underline{y}_i}\left(\gamma_k|\underline{y}_i,\hat{\underline{\beta}}\right)w_k \tag{6.16}$$

It is therefore possible to substitute (6.16) in (6.13) to give

$$\begin{aligned} Q\left(\underline{\beta}|\underline{y},\hat{\underline{\beta}}\right) &= \sum_{i=1}^{I}\sum_{j=1}^{J(i)}\sum_{k=1}^{K}\left[y_{ij}\ln\pi_{jk} + \left(1-y_{ij}\right)\ln\left(1-\pi_{jk}\right)\right]P_{ik} + C_4 \\ &= \sum_{k=1}^{K}\sum_{j=1}^{J}\left[\sum_{i=1}^{I(j)} y_{ij}P_{ik}\ln\pi_{jk} + \sum_{i=1}^{I(j)}\left(1-y_{ij}\right)P_{ik}\ln\left(1-\pi_{jk}\right)\right] + C_4 \end{aligned}$$

An insightful way to view this is to think of the continuous distribution of the latent variable being replaced by a discrete distribution with a finite number, K, of values $\gamma_1,\gamma_2,\ldots,\gamma_K$ and probability masses $P_{i1},P_{i2},\ldots,P_{iK}$ defined by the quadrature rule. The approximation to function Q, the posterior expectation, is effectively a discrete expectation with masses P_{ik} at nodes γ_k.

Summing over i gives

$$Q\left(\underline{\beta}|\underline{y},\hat{\underline{\beta}}\right) = \sum_{k=1}^{K}\sum_{j=1}^{J}\left[U_{jk}\ln\pi_{jk} + \left(N_{jk}-U_{jk}\right)\ln\left(1-\pi_{jk}\right)\right] + C_4 \tag{6.17}$$

where $N_{jk} = \sum_{i=1}^{I(j)} P_{ik}$ is interpreted as the expected number of responses to item j for subjects with latent attribute (e.g. ability) concentrated on the node γ_k and $U_{jk} = \sum_{i=1}^{I(j)} y_{ij}P_{ik}$ as the expected number of positive responses to item j dependent on latent attribute γ_k. When all the units have responded to all J items then the j index is redundant. Then N_k is simply the

expected number of units with latent attribute γ_k, given the data and current parameter estimates. However if the total number of responses to item j is dependent on j then the N_{jk} may vary over j. These expected sums of responses are the *expected complete data*.

By comparing the expected complete data log likelihood function shown in equation (6.17) with the complete data log likelihood function in equation (6.2), it is easily seen that (6.17) has the form of a log likelihood function of a GLM with responses $U_{jk} \sim \text{Bi}\left(N_{jk}, \pi_{jk}\right)$. The link function, which is

$$\pi_{jk} = c_j + \frac{1 - c_j}{1 + \exp(-\eta_{jk})}$$

and the linear predictor associated with observation U_{jk} which is

$$\eta_{jk} = \underline{x}_j^T \underline{\beta}_j + \alpha_j \gamma_k$$

are equivalent to those defined in Section 6.2.1 (the subject subscript i is replaced by k). This is the key to the methodology described in this thesis. Maximising the expected complete data log likelihood is equivalent to fitting the GLM whose log likelihood function is equation (6.17). Using this result we have successfully fitted latent variable GLMs using standard GLM fitting software. The software written for this purpose and its implementation is described in the next three chapters.

CHAPTER 7. CHOOSING A NUMERICAL INTEGRATION STRATEGY.

7.1. INTRODUCTION.

From Section 6.2.4 the posterior distribution of the latent variable $\underline{\Gamma}$ is

$$f_{\underline{\Gamma}|\underline{y}}\left(\underline{\gamma}|\underline{y},\hat{\underline{\beta}}\right)=\frac{p_{\underline{y}|\underline{\gamma}}\left(\underline{y}|\underline{\gamma},\hat{\underline{\beta}}\right)f_{\underline{\Gamma}}\left(\underline{\gamma}\right)}{\int_{R}p_{\underline{y}|\underline{\gamma}}\left(\underline{y}|\underline{\gamma},\hat{\underline{\beta}}\right)f_{\underline{\Gamma}}\left(\underline{\gamma}\right)d\underline{\gamma}} \tag{7.1}$$

It was also shown in this section that the expectation of the complete data likelihood with respect to this distribution, expressed as a function of $\underline{\beta}$, is

$$Q\left(\underline{\beta}|\underline{y},\hat{\underline{\beta}}\right)=\int_{R}\left[\ln p_{\underline{y},\underline{\Gamma}}\left(\underline{y},\underline{\gamma}|\underline{\beta}\right)\right]f_{\underline{\Gamma}|\underline{y}}\left(\underline{\gamma}|\underline{y},\hat{\underline{\beta}}\right)d\underline{\gamma} \tag{7.2}$$

Combining equations (7.1) and (7.2) the function Q can be written in the form

$$Q\left(\underline{\beta}|\underline{y},\hat{\underline{\beta}}\right)=\sum_{i=1}^{I}\int_{R}\frac{F_1\left(\gamma_i|\underline{\beta}\right)F_2\left(\gamma_i\right)\exp\left(-\gamma_i^2/2\right)}{\int_{R}F_2\left(\gamma_i\right)\exp\left(-\gamma_i^2/2\right)d\gamma_i}d\gamma_i \tag{7.3}$$

where

$$F_1\left(\gamma_i|\underline{\beta}\right)=\ln p_{\underline{y}_i,\Gamma_i}\left(\underline{y}_i,\gamma_i|\underline{\beta}\right)$$

and

$$F_2\left(\gamma_i\right)=p_{\underline{y}_i|\Gamma_i}\left(\underline{y}_i|\gamma_i,\hat{\underline{\beta}}\right)$$

The integrals above are analytically intractable. The algorithm for obtaining maximum likelihood parameter estimates involves numerical approximation to these integrals. The factors influencing the choice of integration strategy are outlined in this section. Various numerical

methods of integration are available. Simple Newton-Cotes rules are briefly considered in Section 7.2. Section 7.3 contains an introduction to Gaussian Quadrature with specific attention on Gauss-Legendre and Gauss-Hermite methods. The performances of these two quadrature rules are compared in Section 7.4.

7.2. NEWTON-COTES RULES.

The simplest numerical methods for approximating integrals are the various Newton-Cotes formulae such as the trapezoidal rule and Simpson's formula (see, for example, Froberg, 1969). These rules are derived by integrating interpolating polynomials through known values of the function at equidistant intervals. These values are weighted and summed, the weights being chosen to maximise accuracy. Newton-Cotes methods are appropriate in situations where the integrands are not known explicitly but discrete evaluated points are available, for example in tabulated form. The nodes (or abscissae) are therefore pre-determined by the data.

7.3. GAUSSIAN QUADRATURE.

7.3.1. Introduction.

When the function requiring numerical integration is known explicitly then some form of Gaussian quadrature may be appropriate. By making use of Lagrangian interpolating polynomials, these rules allow both the weights and the abscissae to be chosen to minimise the error. In general it is required to find nodes x_k and weights w_k, k=1,2,...,K, such that

$$\int_a^b a(x)f(x)dx = \sum_{k=1}^{K} w_k f(x_k) + R_K \qquad (7.4)$$

where

$$w_k = \int_a^b a(x)\frac{(x-x_1)\ldots(x-x_{k-1})(x-x_{k+1})\ldots(x-x_K)}{(x_k-x_1)\ldots(x_k-x_{k-1})(x_k-x_{k+1})\ldots(x_k-x_K)}dx$$

The error term is R_K

$$R_K = \int_a^b a(x)(x-x_1)\ldots(x-x_K)\{\alpha_0 + \alpha_1 x + \alpha_2 x^2 + \ldots\}dx$$

where $\alpha_0, \alpha_1, \alpha_2, \ldots$ are constants.

These methods rely on the existence of sets of K polynomials which are orthogonal with respect to the known weight function $a(x)$ over the interval $[a,b]$. In these sets the kth polynomial has k roots. The K abscissae are the roots of the K^{th} polynomial. Once they have been chosen, the weights are automatically determined. (For example: Davis and Rabinowitz, 1984; Froberg, 1973; Burdon and Faires, 1989). The roots of the polynomials used for different quadrature rules and the weights associated with them are extensively tabulated (e.g. Stroud and Sechrest, 1966; Abramowitz and Stegun, 1972).

7.3.2. Gauss-Legendre Rule.

The Gauss-Legendre integration rule makes use of the set of Legendre polynomials. This set is orthogonal on [-1,1] with respect to the weight function $a(x) = 1$. Therefore the approximation takes the form

$$\int_{-1}^{1} f(x)dx \approx \sum_{k=1}^{K} w_k f(x_k) \tag{7.5}$$

A transformation is used to translate $x \in [a,b]$ to $x' \in [-1,1]$. This is

$$x' = \left(\frac{1}{b-a}\right)(2x-a-b) \tag{7.6}$$

so that when the range of integration is changed to $[a,b]$ (7.5) becomes

$$\int_a^b f(x)dx \approx \left(\frac{b-a}{2}\right)\sum_{k=1}^{K} w_k f\left(\left(\frac{b-a}{2}\right)x_k' + \left(\frac{b+a}{2}\right)\right) = \left(\frac{b-a}{2}\right)\sum_{k=1}^{K} w_k f(x_k) \tag{7.7}$$

where $-1 \le x_k' \le 1$ and $a \le x_k \le b$, k=1,2,...,K.

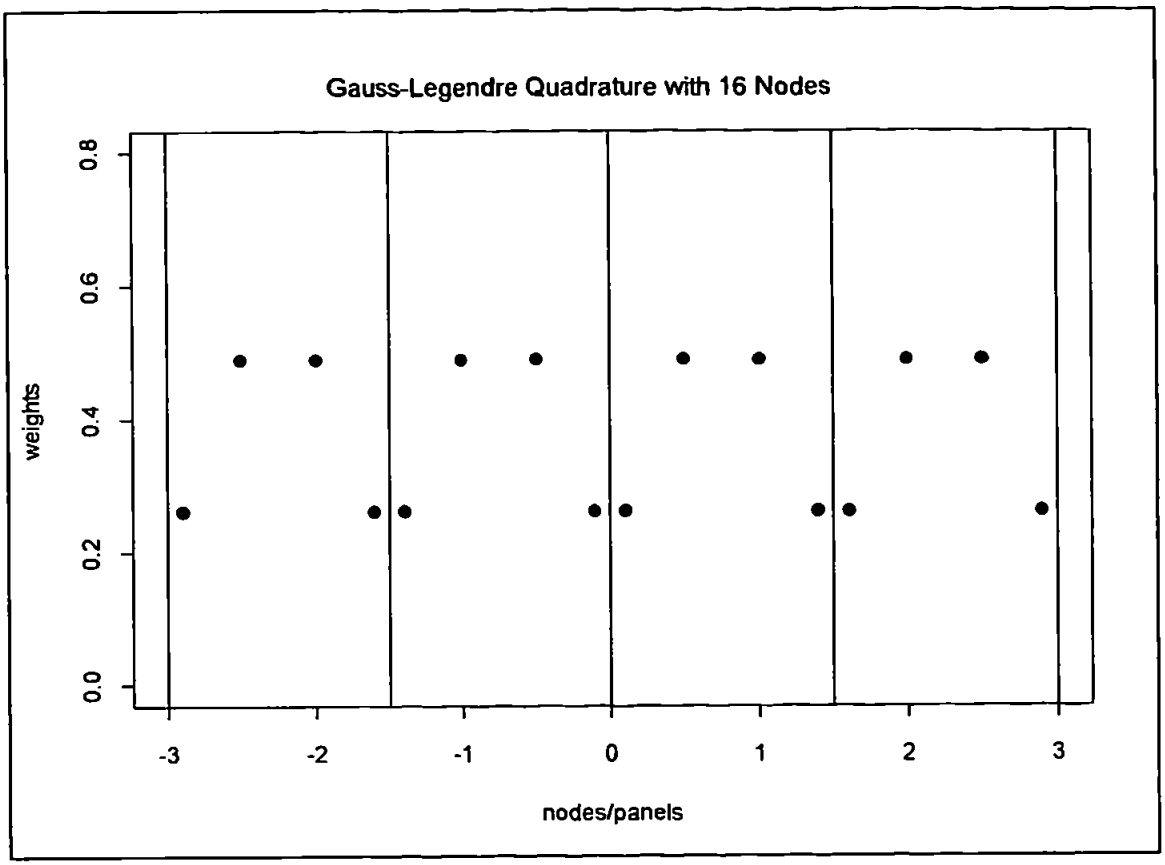


FIGURE 5. Gauss-Legendre Quadrature.

For a Gauss-Legendre approximation to the integral $\int_a^b f(x)dx$ using K nodes the following procedure is adopted. First the nodes and weights for a 4-point integration rule over the interval [-1,1] are obtained from tables. They are considered to constitute one panel. The number of panels ($K/4$) over the required range $[a,b]$ is obtained. The limits and widths of the ranges of the new panels are then calculated by dividing $[a,b]$ into ($K/4$) panels of equal width, and the original nodes are transformed to each of these new intervals using equation (7.6). The original weights are scaled to the width of the new panels, panels narrower than the original (<2) requiring smaller weights and vice versa. The procedure results in K nodes and weights over the range $[a,b]$. Figure 5 shows K=16, i.e. 4 panels, over [-3,3]. The widths of the panels

and the weights are 3/4 of their original size. The abscissae $x_k' \in [-1,1]$, k=1,2,3,4, have been transformed four times so that $x_k \in$ [-3,-1.5] for k=1,2,3,4, $x_k \in$[-1.5,0] for k=5,6,7,8, $x_k \in$[0,1.5] for k=9,10,11,12, and $x_k \in$[1.5,3] for k=13,14,15,16.

Using Gauss-Legendre quadrature in this manner the function Q in equation (7.3) can be approximated by

$$Q\left(\underline{\beta}|\underline{y},\hat{\underline{\beta}}\right) \approx \sum_{i=1}^{I}\sum_{k=1}^{K} F_1\left(\gamma_k|\underline{\beta}\right)\frac{F_2\left(\gamma_k\right)\exp\left(-\gamma_k^2/2\right)w_k}{\sum_{k=1}^{K} F_2\left(\gamma_k\right)\exp\left(-\gamma_k^2/2\right)w_k}$$

where the γ_k are the abscissae obtained by the procedure outlined above and the w_k are the associated scaled weights. In another form

$$Q\left(\underline{\beta}|\underline{y},\hat{\underline{\beta}}\right) \approx \sum_{i=1}^{N}\sum_{k=1}^{K} F_1\left(\gamma_k|\underline{\beta}\right)\frac{F_2\left(\gamma_k\right)W_k}{\sum_{k=1}^{K} F_2\left(\gamma_k\right)W_k} \tag{7.8}$$

where $W_k = \exp\left(-\gamma_k^2/2\right)w_k$.

7.3.3. Gauss-Hermite Rule.

In this case the K abscissae are the roots of the K^{th} Hermite polynomial H_K. These polynomials are orthogonal on the interval $(-\infty,\infty)$ with respect to weight function $a(x) = e^{-x^2}$. Therefore the nodes x_k and weights w_k are chosen so that

$$\int_{-\infty}^{\infty} e^{-x^2} f(x)dx \approx \sum_{k=1}^{K} w_k f\left(x_k\right) \tag{7.9}$$

The functions to be integrated in equations (7.1) and (7.2) are of the form

$$\int_R e^{-\frac{\gamma^2}{2}} F(\gamma)d\gamma$$

where $F(\gamma)$ is a function of γ. In order to employ the Gauss-Hermite rule a transformation of the variable, namely $\gamma = \sqrt{2}x$, is required. This gives

$$\int_{-\infty}^{\infty} e^{-x^2} F(\sqrt{2}x)dx \approx \sum_{k=1}^{K} w_k F(\sqrt{2}x_k) = \sum_{k=1}^{K} w_k F(\gamma_k)$$

where the x_k are the tabulated solutions of $H_K = 0$, $\gamma_k = \sqrt{2}x_k$ and the w_k are the associated weights, which are also tabulated. Using Gauss-Hermite quadrature the function Q in equation (7.3) can be approximated by

$$Q(\underline{\beta}|\underline{y},\underline{\hat{\beta}}) \approx \sum_{i=1}^{N}\sum_{k=1}^{K} F_1(\gamma_k|\underline{\beta}) \frac{F_2(\gamma_k)w_k}{\sum_{k=1}^{K} F_2(\gamma_k)w_k} \qquad (7.10)$$

Equations (7.10) and (7.8) are written in the same form so that the w_k and W_K are directly comparable.

7.3.4. Discussion.

It would appear at first sight that the Gauss-Hermite quadrature rule is the optimal choice for the required integration because of the presence in the integrand of the factor $e^{-\frac{\gamma^2}{2}}$, which can easily be transformed to the weight function e^{-x^2}. However there are various reasons why this is not in fact the case and why it is anticipated that a Gauss-Legendre rule gives a better performance.

The functions equivalent to $f(x)$ in equation (7.9) can be sharply peaked and asymmetrical (about zero) and generally dominate the integrand, which is therefore a 'lumpy' asymmetrical function. It is surmised that this is because the posterior distribution of the latent covariate γ is approximately $N(\mu,\sigma^2)$, the means and standard deviations varying over the

different units (subjects). Although the integrands may vary in their shapes the same set of nodes and weights is required to approximate these functions for all the units of observation.

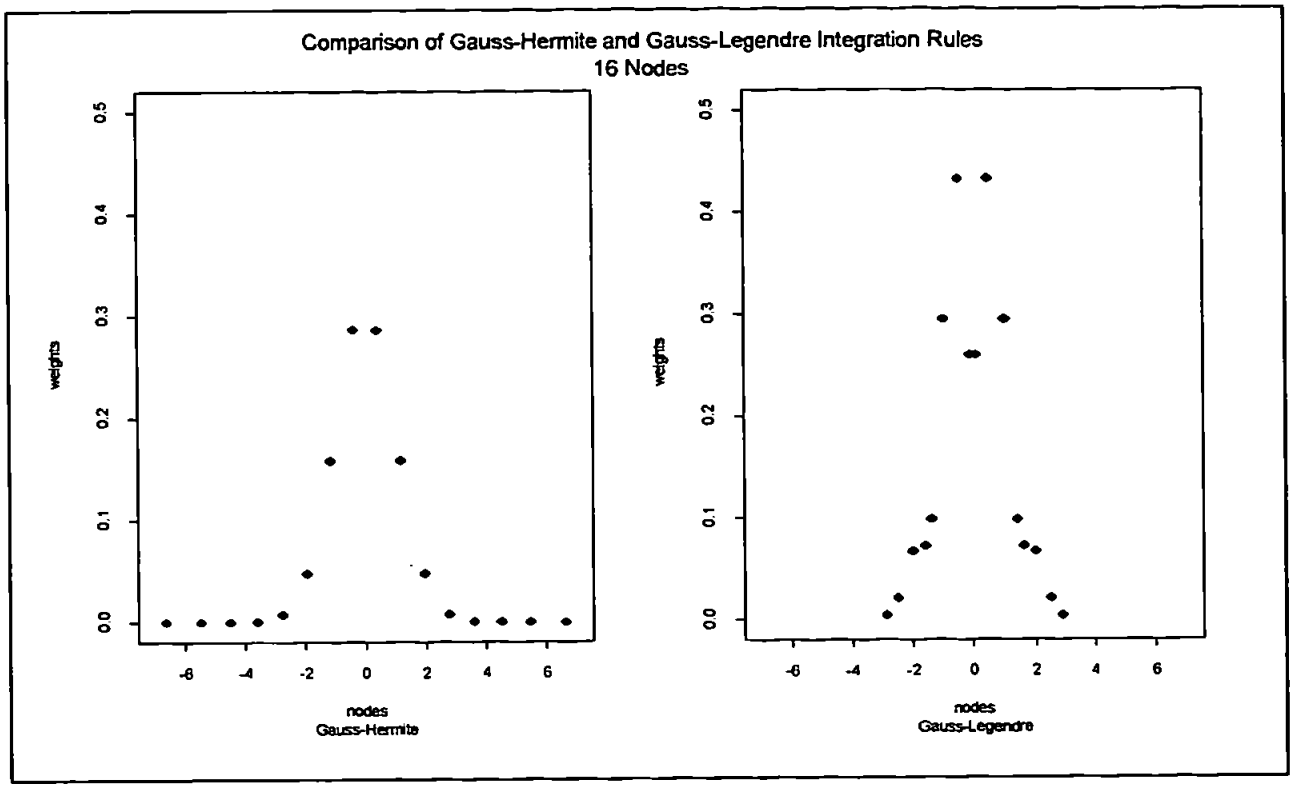


FIGURE 6. Comparison of Gauss-Legendre and Gauss-Hermite Rules showing nodes.

Figure 6 shows a comparison between a 16-node Gauss-Hermite rule with weights w_k (equation 7.10) where the interval of integration is $[-\infty,\infty]$ and a 16-node Gauss-Legendre rule with weights $W_k = \exp\left(-\gamma_k^2/2\right)w_k$ (see equation 7.8). Thus the W_k in Figure 6 correspond to the weights w_k shown in Figure 5 multiplied by the function $\exp\left(-\gamma_k^2/2\right)$.

For Gauss-Legendre approximation the interval of integration is, in this case, [-3,3]. This range is chosen to approximate the range of a standard normal distribution. The infinite

range of the Gauss-Hermite rule would therefore appear at first sight to be more appropriate. However, the graph on the left shows how the standard normal distribution dominates the Gauss-Hermite method with too many nodes of negligible weight 'wasted' in what equate to the tails of the distribution of γ. The Gauss-Legendre rule is based on panels of 4 nodes which are scaled and replicated over a finite interval and the right-hand graph shows a more useful distribution with a greater number of nodes spread over the region of interest. It is conjectured that this is more effective in approximating distributions where the data indicates a high probability that the variable γ is not near the centre of the standard normal distribution.

7.4. A COMPARATIVE STUDY.

7.4.1. Design.

In order to compare the relative performance of Gauss-Legendre and Gauss-Hermite integration rules a comparative study of two versions of the model-fitting software was undertaken. The first version was the software described elsewhere in this thesis, incorporating Gauss-Legendre quadrature, and the second used Gauss-Hermite quadrature but was in all other respects identical. Two data sets were simulated from the model

$$P(Y_{ij} = 1) = \pi_{ij} = \frac{1}{1 + \exp(-\eta_{ij})},\ i = 1,2,\ldots,I;\ j = 1,2,\ldots,J$$

where the linear predictors are

$$\eta_{ij} = 1 + \gamma_i$$

i.e. the true values of both the slope and intercept parameters were 1 for all j.

Data set 1 was the smaller, comprising 50 units (subjects) with 25 observations on each; data set 2 was larger with 400 units of 100 observations. The model fitting algorithm (described fully in Chapter 8) was run on each data set using both integration rules with 4, 8, 12, 16, 20,

24, 32 and 48 nodes. All the Gauss-Legendre rules used assumed the interval of integration to be [-3,3]. The algorithm was run on the smaller data set with a convergence criterion of 0.001 and on the larger set with 0.01.

It was hoped that both rules would converge towards similar estimates of both slope and intercept parameters as the number of nodes increased. Differences in the patterns of convergence were noted. Standard errors for the parameter estimates were taken from the results of the simulation study for similar sized data sets (see Table 8). These were compared with the errors in the estimates as the number of nodes increased. The results were plotted and the graphs can be seen in Figure 7.

7.4.2. Results.

In the smaller data set the integration rules produced estimates which converged towards the same parameter value, specifically 1.21 to 2 d.p for the intercept and 1.10 to 2 d.p for the slope. The Gauss-Legendre estimates were very close to this value when 12 nodes were used but the Gauss-Hermite values took longer to settle down as expected, requiring 20 nodes to attain a similar accuracy. In the large data set convergence to (0.95,1.00) was better with the Gauss-Legendre rule, again confirming expectations. A set of 20 nodes gave very good accuracy in this case but 32 and 48 points were needed to achieve a similar level of accuracy on the intercept and slope respectively when a Gauss-Hermite approximation was used. In terms of time the increased number of iterations meant that the Gauss-Hermite rules could take twice as long as the Gausss-Legendre rules to achieve the same accuracy.

I was noted that when 64 nodes were used the error in the Gauss-Legendre estimates increased again slightly. This problem was alleviated when the range of integration was

extended to [-4,4]. The estimates were well-behaved for $K < 64$ and 64 nodes was too large a number for practical purposes.

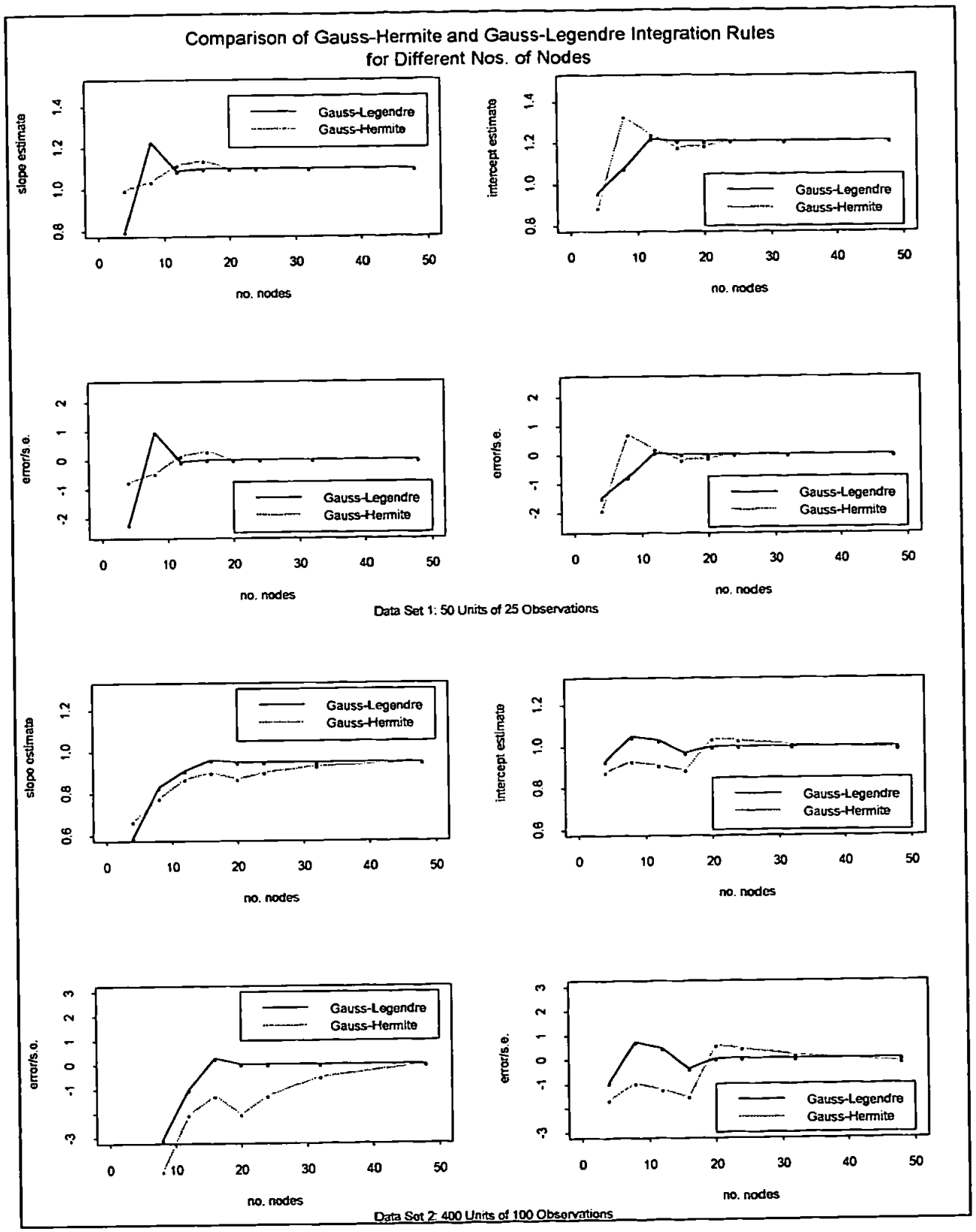


FIGURE 7. Comparison of Performances of Integration Rules.

7.4.3. Conclusion.

With Gauss-Legendre quadrature more accurate parameter estimates are obtained using a smaller number of nodes. This was particularly noticeable in the large data set that was tested where Gauss-Hermite approximation required an impracticably large number of nodes to achieve acceptable accuracy.

Gauss-Legendre quadrature may be an appropriate choice regardless of the properties of the function to be integrated. The Gauss-Hermite method works best in situations where the weight function e^{-x^2} dominates the integrand and the function $f(x)$ in equation (5.2.9) is relatively smooth. In this latent variable context however it seems likely that the $f(x)$ approximates to a normal distribution, usually with a non-zero mean, which dominates the integrand. The larger the sample size, particularly when there are a large number of observations per unit, the smaller the standard deviations of these posterior distributions and the greater the tendency for the individual means to deviate from zero. This is why the Gauss-Legendre quadrature rule with its more even spread of nodes and weights is able to out-perform Gauss-Hermite.

For example, in the IRT application a subject who correctly answers only a very few items in a large test has a very high probability of a low ability score. The posterior distribution of his ability is sharply peaked about a low mean. Other subjects may have equally sharply peaked distributions at the other end of the scale whilst others will be more central in relation to the range of integration. The Gauss-Legendre integration rule is better able to approximate this range of ability distributions than Gauss-Hermite.

CHAPTER 8. FITTING A LATENT VARIABLE GLM FOR BINARY RESPONSES.

8.1. INTRODUCTION.

In this chapter the methodology for fitting latent variable GLMs is applied to the model for binary response data developed in Chapter 6. It is shown how software written in GLIM to implement the EM algorithm can be used to fit a model of this type to a data set from an experiment in IRT.

In Section 8.2 the fitting methodology for latent variable GLMs using the EM algorithm and GLIM which was described in Chapter 5 is applied to the binary response model. In Section 8.3 there is a detailed description of the implementation of the model fitting software and in Section 8.4 it is specifically applied to two examples from Item Response Theory, a timed item test of mental arithmetic and a timed transitive inference test.

8.2. FITTING THE BINARY RESPONSE MODEL.

The general procedure for fitting a latent variable GLM using the EM algorithm and GLIM has been described in Chapter 5. When the response data is binary the algorithm proceeds as follows:

STEP 1: Choose a suitable number of quadrature points for the integral approximations and calculate the nodes γ_k and weights w_k, as defined by an appropriate integration rule (see Chapter 7).

STEP 2: Choose starting values for the parameter estimates $\hat{\underline{\beta}}^{(0)}$.

STEP 3: This step is the expectation step of the EM algorithm. Using response vector $\underline{y}$ and parameter estimate $\hat{\underline{\beta}}^{(m)}$ calculate N_{jk}, the expected number of units at node γ_k attempting item j, and U_{jk}, the expected number of units at node γ_k scoring 1 on observation j, for j=1,2,...,J and k=1,2,...,K, where

$$N_{jk} = \sum_{i=1}^{I(j)} P_{ik} \text{ and } U_{jk} = \sum_{i=1}^{I(j)} y_{ij} P_{ik} \tag{8.1}$$

From equation (6.14)

$$P_{ik} = \frac{p_{\underline{Y}_i|\gamma}\left(\underline{y}_i|\gamma_k, \hat{\underline{\beta}}^{(m)}\right) f_\Gamma(\gamma_k) w_k}{\sum_{k=1}^{K} p_{\underline{Y}_i|\gamma}\left(\underline{y}_i|\gamma_k, \hat{\underline{\beta}}^{(m)}\right) f_\Gamma(\gamma_k) w_k}$$

which will be written

$$P_{ik} = \frac{p_{ik}}{\sum_{k=1}^{K} p_{ik}}$$

where

$$p_{ik} = p_{\underline{Y}_i|\gamma}\left(\underline{y}_i|\gamma_k, \hat{\underline{\beta}}^{(m)}\right) f_\Gamma(\gamma_k) w_k$$

Expanding this expression gives

$$p_{ik} = (2\pi)^{-\frac{1}{2}} w_k \exp\left[-\frac{\gamma_k^2}{2} + \sum_{j=1}^{J}\left(y_{ij} \ln \pi_{jk} + \left(1 - y_{ij}\right)\ln\left(1 - \pi_{jk}\right)\right)\right]$$

where

$$\pi_{jk} = c_j + \frac{1 - c_j}{1 + \exp\left[-\eta_{jk}\left(\underline{\beta}_j^{(m)}\right)\right]} \tag{8.2}$$

and

$$\eta_{jk}\left(\underline{\beta}_j^{(m)}\right) = \underline{x}_j^T \underline{\varphi}_j^{(m)} + \alpha_j^{(m)} \gamma_k \tag{8.3}$$

STEP 4: Using the IRLS algorithm within GLIM, fit the model to the data

$\underline{U} = \left(U_{11}, U_{21}, \ldots, U_{J1}, U_{12}, U_{22}, \ldots, U_{J2}, \ldots, U_{1K}, U_{2K}, \ldots, U_{JK}\right)^T$ and

$\underline{N} = \left(N_{11}, N_{21}, \ldots, N_{J1}, N_{12}, N_{22}, \ldots, N_{J2}, \ldots, N_{21}, N_{2K}, \ldots, N_{JK}\right)^T$ assuming $U_{jk} \sim \text{Bi}\left(N_{jk}, \pi_{jk}\right)$ with link function (8.2) and linear predictor (8.3). This is the maximisation step of the EM algorithm.

STEP 5: Using the new parameters $\underline{\hat{\beta}}^{(m+1)}$ obtained from the M-step in the $(m+1)$th iteration of EM, repeat steps 3 and 4 until convergence.

A similar fitting algorithm for GLMs with random effects is described by Hinde (1988). However Hinde suggests using the P_{ik} as prior weights and expanding the original data vector of length n, ($n = I \times J$), to a vector of length ($K \times n$). Because of the summation over the I subjects that occurs in equation (6.17) the method outlined here has the advantage that it *reduces* the data vector from $n = I \times J$ to $n = J \times K$ (assuming $K < I$). This summation over i is possible when there are no model parameters indexed by i. In the IRT applications under consideration model parameters are normally functions of the item characteristics and so this reduction in what may be extremely large data sets is possible. As a result computer processing is speeded up and there is less strain on memory resources.

8.3. RUNNING THE MODEL FITTING SOFTWARE.

8.3.1. An Overview of the GLIM Program.

All the software routines required to fit a latent variable GLM for binary responses are implemented in a single GLIM program. A listing of this program is attached in Appendix C. A generalized linear model with the non-standard link function given in equation (8.2) is built into

the program code by declaring the *SOWN* directive. This allows the macro *FIT* to assign fitted values to the system vector *%fv* using the values of the linear predictors in the system vector *%lp*. Thus the model predicts that the number of subjects at ability node k answering item j correctly is

$$E(U_{jk}) = \mu_{jk} = N_{jk}\pi_{jk} = N_{jk}\left[c_j + \frac{1-c_j}{1+\exp(-\eta_{jk})}\right]$$

where π_{jk} is the probability of a correct response and N_{jk} is the number responding.

Similarly the macros *DIR*, *VAR* and *DEV* assign user-defined values to the derivatives $\frac{\delta\eta}{\delta\mu}$, the variance functions and the deviances respectively. The derivatives are given by the equation

$$\frac{\delta\eta_{jk}}{\delta\mu_{jk}} = \frac{1}{N_{jk}(\pi_{jk}-c_j)} + \frac{1}{N_{jk}(1-\pi_{jk})}$$

the variances

$$Var(U_{jk}) = N_{jk}\pi_{jk}(1-\pi_{jk})$$

and the deviances

$$D_{jk} = 2U_{jk}\log\left[\frac{U_{jk}}{N_{jk}\pi_{jk}}\right] + 2(N_{jk}-U_{jk})\log\left[\frac{N_{jk}-U_{jk}}{N_{jk}(1-\pi_{jk})}\right].$$

Macros *FIT* and *DEV* therefore define the logistic functions with guessing parameters (when the guessing parameter is zero it is the standard logistic model) which relate the linear predictors to the expected values and *VAR* and *DEV* define the binomial distribution from which the data are assumed to come.

The components of the binary response model defined above are therefore pre-set in the program. Although the structure of the linear predictor is decided by the user at run-time this is accomplished only within the limits set up in the GLIM code. The factors and covariates which

the user may wish to include in a model formula are also coded into the program. Factor levels are generated and assigned to vectors during the program's initialisation routines. The factors can be included in or excluded from any model formulae under consideration and their exact nature will depend on the individual application. Similarly values of any fixed covariates which may appear in the model are also assigned to vectors.

The program is designed to enable the user to fit a model and obtain the results and then repeat the fitting procedure with an alternative model formula as many times as required without having to re-run the program. There are also facilities to enter and change the number of quadrature nodes used in the numerical integration routines and to enter and change the tolerance levels which are the criteria of convergence. In this way the program can be used to obtain suitable starting values by iterating only a few times with a small number of quadrature points. This is normally a relatively quick process. The convergence criteria can then be made more stringent and the number of nodes increased to obtain more accurate estimates.

The program consists of a set of nested macros. By calling the highest level of macro the user can run the entire model fitting algorithm from start to finish with only the minimum of input. Alternatively more control can be gained by running a succession of lower level macros in the user's desired order (provided certain rules regarding sequence are obeyed). For example, instead of allowing the program to iterate back and forth between the expectation and maximisation steps of EM until convergence it is possible to run each step individually. This may be desirable if a close examination of the results of each step is required. It also allows the iterations to proceed until the user decides, with the results in view, that he or she wishes it to stop, rather than ending the algorithm by some predetermined criterion.

GLIM has the facility to call subroutines written in FORTRAN and data may be passed between these subroutines and the GLIM calling program. Several of the GLIM macros that

comprise the latent variable GLM model fitting software make use of this facility which is implemented with the *SPASS* command. This directive transfers control to a FORTRAN module 'PASS.F77' which in turn calls the required subroutine. The software has been designed so that when GLIM passes control to a FORTRAN subroutine the user is informed with a message on the screen indicating the nature of the processing that is taking place.

8.3.2. Running the GLIM Program.

8.3.2.1. Initialisation Routines.

The program is started from the package GLIM. The user types

$input FILENAME

where *FILENAME* is a file containing the GLIM program GLIRT1. When this is loaded the user is first asked for the name of a file to which the results of all the model fitting procedures can be written.

Enter name of file for output of parameter estimates

The program then sets some default values for the run. These are a default model which has a common intercept and slope on ability for every item, a default tolerance (0.001) and a default maximum accuracy (9 d.p.) for calculations. The first two of these may be altered by input from the keyboard during the run.

The user is then prompted:

Use macro NODES then INIT for initial estimates

Use macro LOOP to run EM algorithm

Alternatively typing *'$USE RUN'* will call all three macros automatically one after the other.

8.3.2.2. Calculating the Quadrature Nodes and Weights.

When the macro *NODES* is called control is immediately passed to a FORTRAN subroutine called LEGDAT and the following message is displayed:

subroutine LEGDAT - calculating standard normal nodes

Input no. of nodes (4,8,12,...,60)

Input min and max

The user must supply (i) the number of nodes required for the numerical integration procedure (only multiples of 4 between 4 and 60 are accepted) and (ii) a lower and upper limit for the range of integration. This will normally be (-3,3) or (-4,4) for the standard normal distribution. As in STEP 1 (Section 8.2) above the subroutine then calculates the required set of nodes and weights using Gauss-Legendre quadrature rules (see Chapter 7). A file 'Q.DAT' contains nodes and weights for a 4-point integration rule taken from tables (e.g. Stroud and Sechrest, 1966) and is available to the subroutine. Output is to a vector which is passed back to the GLIM calling program and to a file 'LEG4.DAT' which is only used for the purposes of further (related) research.

8.3.2.3. Entering the Model Formula.

When control is returned to GLIM, macro *NODES* prompts the user to specify the components of the linear predictor and a tolerance level:

You must enter a model specification and a tolerance level (%x)

Current model: "MMMM" Tol: "nn"

To specify a new model enter macro MODEL

reset %x if a new tolerance is required

The scalar *%x* contains the current tolerance value used to detect convergence of the EM algorithm. Its contents are displayed (*nn*) and can be changed by typing, for example,

Scal %x=0.0001 $

The current model formula is held in the named string *MODEL* and is displayed on the screen (*MMMM*). A new one is entered by typing, for example,

$MACRO MODEL A+B $

where *A* and *B* are vectors containing factor levels.

8.3.2.4. Starting the Algorithm.

When the macro *INIT* is run control is once again passed to a FORTRAN subroutine and the message

subroutine INIT -initialisation

appears on the screen. The vectors of nodes and weights calculated by the LEGDAT subroutine are passed as a single string from GLIM to INIT. This subroutine opens the file 'A2GLIRT.DAT' which contains the observed binary response data to which the model is to be fitted. The first two values on this file are the dimensions of the data, *I* and *J* (i.e. the number of subjects and number of items in the case of IRT data). Next to be read from this file are the *J* lower asymptotes or guessing parameters c_j which may be different for each item. Finally the *I* by *J* data matrix is read into an array.

The next step (STEP 2 of Section 8.2 above) is to assign initial values to the slope and intercept parameter vectors. At present suitable constants are assigned to both parameters within the FORTRAN code. The FORTRAN subroutine now has all the information it needs for STEP 3 the calculation of initial values for the expected complete data

$\underline{U} = \left(U_{11}, U_{21}, \ldots, U_{J1}, U_{12}, U_{22}, \ldots, U_{J2}, \ldots, U_{1K}, U_{2K}, \ldots, U_{JK}\right)^T$ and

$\underline{N} = \left(N_{11}, N_{21}, \ldots, N_{J1}, N_{12}, N_{22}, \ldots, N_{J2}, \ldots, N_{21}, N_{2K}, \ldots, N_{JK}\right)^T$ using equations (8.1) and (6.14). These values are passed back to GLIM where they are stored in vectors.

Before the first maximisation step is run to fit a model to the expected data values all the vectors required for the IRLS algorithm must be in place. A macro *PREP* is called which deletes any data vectors remaining from the fitting of a previous model and sets up new ones. The vector containing the data i.e. $\underline{U}$ must be declared as such (*SYVAR*). The data $\underline{N}$ are the 'binomial' denominators and are used in the specification of the binary response model through the user-defined macros *FIT, DIR, DEV and VAR.* GLIM requires that all the vectors involved in the fit must be of the same length as the data. This entails expanding the K-vector of ability nodes by repeating each node J times. Similarly the J guessing parameters must each be copied K times. Vectors containing factor levels and fixed covariates are then generated. The result is that the set of J items , complete with the factor levels and covariates associated with each item, is repeated K times, with one set of items at each ability node.

Two dummy sets of items are added to the end of the data vector. Thus $\underline{U}$ becomes $\underline{U} = \left(U_{11}, U_{21}, \ldots, U_{J1}, U_{12}, U_{22}, \ldots, U_{J2}, \ldots, U_{1K+1}, U_{2K+1}, \ldots, U_{JK+1}, U_{1K+2}, U_{2K+2}, \ldots, U_{JK+2}\right)^T$. The extra components represent dummy data points at ability levels of 0 and 1. For example $U_{2,K+1}$ represents the number of subjects at ability level 0 answering item 2 correctly and $U_{2,K+2}$ is the number of subjects at ability level 1 answering item 2 correctly. The last $2J$ values of the ability vector are set accordingly to J zeros followed by J ones. The corresponding components of $\underline{U}$ and $\underline{N}$ are set to 0s and 1s respectively. These data points are given zero weighting so that they do not contribute towards the parameter estimation. They are otherwise treated as bona fide data and have associated item factor levels and covariates. During the fitting algorithm values are assigned to the related components of the system vectors as for any other data value.

After fitting the model the system vector of linear predictors *%LP* will contain values of the intercept only for each item where the ability node is 0, and the sum of the intercept and the slope for each of the items where the ability node is 1. The slopes can therefore be found by simple subtraction. This allows the slopes and intercepts which may have complicated structures varying from item to item depending on the factors used in the model to be easily extracted and passed to the next E-step.

At this stage a message informing the user of the current model, the number of items, the number of nodes and the tolerance level is displayed. For example:

MODEL: BLOCK + DIFF.GAMMA

ITEMS: 60 NODES: 16 TOL:0.001

GLIM is now ready to fit the first model. This is controlled by macro *MAX*. The message

Maximisation Step: Iteration 1

is displayed and the GLIM fitting algorithm is implemented (STEP 4). At the end of its iterations GLIM displays the new parameter estimates along with standard errors and a scaled deviance (the latter two are irrelevant to the present purpose).

8.3.2.5. Continuing the EM Iterations.

At this point the program has completed the first EM iteration and produced some initial parameter estimates to use in the second expectation phase.. In order to run the EM algorithm to convergence the user may use macro *LOOP* (STEP 5). This macro will automatically test for convergence after every E-step. Alternatively he or she can use macros *ESTEP* and *MAX* alternately. Whichever way it is done the next stage is to call the FORTRAN subroutine ESTEP which implements the full expectation step of the EM algorithm. The calculations used to compute fresh values of $\underline{U}$ and $\underline{N}$ are those performed in subroutine INIT

(equation 8.1) but using the updated intercept and slope parameters from the previous M-step. As before a message appears on the screen whilst the routine is running:

subroutine ESTEP - expectation phase

When control returns to GLIRT1 the new expected complete data is read into GLIM vectors and the program is immediately ready for another fitting routine implemented by macro *MAX*.

8.3.2.6. Convergence.

A secondary task of the subroutine ESTEP is to calculate the 'fit statistic', $-2l$, where l is $l_{\underline{Y}}(\underline{\hat{\beta}}^{(m)}|\underline{y})$; that is the observed data log likelihood function evaluated at the current parameter estimates $\underline{\hat{\beta}}^{(m)}$ and the observed responses $\underline{y}$. This statistic is part of the log likelihood ratio statistic which can be used to compare the goodness-of-fit of different models. It is used here to detect convergence of the EM algorithm. It has been shown (Dempster, Laird and Rubin, 1977) that each iteration of EM increases the likelihood (Section 5.2.3). As the log likelihood increases, $-2l$ decreases. A sufficiently small decrease in $-2l$ indicates that convergence has occurred. Alternatively, a maximum has been reached if there is no significant increase in the log likelihood over the last iteration. If convergence is not indicated then the EM cycle continues.

The calculation and checking of the fit statistic is performed outside of the main GLIM program because of the higher level of accuracy it is possible to achieve using FORTRAN. The subroutine INIT opens a file called 'FIT.DAT' and writes to it one record consisting of a double precision variable. This variable is the fit statistic and is set to zero by the initialisation routine. Each time the expectation step is run 'FIT.DAT' is opened and the last value of the fit

statistic is read. Before the subroutine ends a new value is calculated and written to the file in place of the old one. It is also displayed on the screen for the user to see at each iteration:

FIT statistic: (nnnnnn.nnnnnn)

The difference between the new fit statistic and the previous one is compared with the tolerance level set by the user (this is passed to the subroutine from GLIRT1). If the difference in fits is less than the required tolerance then the current value of $-2l$ is passed back to GLIRT1 in place of the tolerance. On returning to GLIM, before starting the maximisation routine, macro *CHECK* sets a switch if it detects that the tolerance level passed to ESTEP has been changed. This switch then initiates macro *ENDUP* after the final model fit.

8.3.2.7. Restarting the Program.

Macro *ENDUP* automatically writes the results of the final IRLS fitting algorithm to the output file named at the start of the run. Details of the model formula, the number of quadrature points used, the convergence criterion and the final fit statistic are all written to the file together with the standard GLIM output which was also displayed on the screen. The message

use macro NODES for new nodes, or re-set tolerance (%x)

then use macro LOOP to re-run algorithm

is then displayed. The user can then run the algorithm again with an increased number of quadrature points and/or tighten up the convergence criterion. Alternatively he or she may exit from the program at this point. If further models are fitted the final results are appended to the file and can be printed when needed.

8.4. EXAMPLES FROM ITEM RESPONSE THEORY.

8.4.1. Example 1 - A Timed Item Test of Mental Arithmetic.

The latent variable GLM for binary response data, described in Chapter 6, was fitted, using the methodology described in this chapter, to response data arising from a timed mental arithmetic test (Wright *et al*, 1994). This application was described in the introduction to this thesis (Section 1.3).

8.4.1.1. The Data.

In this example 293 subjects answered a total of 60 test items. The subjects were presented with successive items on a computer screen. Each one consisted of an arithmetic equality (e.g. '$11-4+8=15$' or '$15-8+12=21$') which was either true or false, for a pre-set time period of either 4, 6 or 8 seconds. The item was then removed from the screen and the subject given 1.5 seconds to respond 'true' or 'false'. The items were divided into 2 replications of 30 item types. Within each replication 10 items, of which 5 were in fact true and 5 false, were presented at each of the three exposure times. In addition 5 different types of expression were devised for the arithmetic equalities, corresponding to 5 supposed levels of difficulty. Each group of 5 true or 5 false questions at each exposure time contained one item of each expression type, giving a total of 30 different types of item. The order of items was randomised within replication, and in effect a total of 4 different patterns of randomisation were used. Each 60-item test therefore consisted of two replications. The questions for this timed item arithmetic test are attached to this thesis in Appendix D.

8.4.1.2. Fitting the Model.

In Section 8.3.1 it was shown how the binomial distribution and link function which define the binary response model for this data are incorporated in a GLIM program. It was also stated that the components of the linear predictor are decided by the user at run-time and that these are dependent on structures set up in the GLIM code. Within the general framework of the latent variable GLM for binary response data, different models were fitted to the data using alternative specifications of the linear predictor. From equation (8.3) it can be seen that the linear predictor consists of two parts; the first is the intercept $\underline{x}_j^T \underline{\varphi}_j$ which corresponds to the difficulty parameter associated with item j; the second is the effect of the random variable. The slope on the random variable is the corresponding discrimination parameter. In the timed mental arithmetic test the structure of the intercept and slope parameters is dependent on a set of item characteristics. The intercept is a linear combination of fixed effect parameters from $\underline{\varphi}_j$. For the data described above a model for the difficulty parameter may include effects due to the expression type, the effect of the correct response being either true or false and the effect of the exposure time, all of which may contribute to the difficulty of an item. The first two of these are treated as factors with 5 and 2 levels respectively. The exposure time can be treated in two ways, as discrete or continuous. If it is treated as a discrete variable, it enters the model as a factor with 3 levels. If it is modelled as a continuous variable, its reciprocal is entered into the design matrix as a covariate. Smaller values of time are associated with more difficult items which also have larger (absolute value) difficulty parameters. In its reciprocal form the value of the time covariate gets bigger as time gets shorter. In this model, $\underline{\varphi}_j$ contains the slope on the reciprocal. The discrimination parameter can also appear in a model as a structured combination of these or other item

characteristics. However in this particular example interest lay chiefly in structuring the intercept.

To suit the requirements of the data set described in this example vectors containing the levels of five factors were set up in the GLIM software for use in the specification of the linear predictors of the models to be fitted. These were

(i) item difficulty (5 levels)

(ii) whether the equality presented is true or false (2 levels)

(iii) block (2 levels)

(iv) the individual item (J levels)

(v) time (3 levels).

The reciprocal of time could also be included as a covariate and a vector containing the values 1/4,1/6 and 1/8 was generated.

MODEL	D.F.	FIT STATISTIC
Null	2,399	16394.0
γ	2,398	15611.6
time + γ	2,396	14974.3
difficulty + γ	2,394	14590.3
time + difficulty + γ	2,391	13950.0
time. difficulty + γ	2,384	13831.8
item + γ	2,339	13433.5
item + item. γ	2,280	13319.8

TABLE 2. Results of Fitting Several Models to Timed Item Test Data.

8.4.1.3. Results.

Various models were fitted in which the difficulty parameter was structured to include different item characteristics. Each time a new model was fitted the final E-step produced a fit statistic *-2l* (equal to minus twice the log likelihood of the observed data) which was used to assess to what degree the inclusion or exclusion of parameters affected the fit of the model to the data. Some of the results are shown in Table 2. There appeared to be no significant effect attributable to truth/falsity.

A full model containing 120 parameters was fitted allowing both the difficulty and discrimination parameters to vary over the items. Theoretically the change in fit between the full and a restricted model with p parameters has asymptotically a χ^2_{120-p} distribution if the restricted model fits well. However comparison with χ^2_{120-p} resulted in rejection of the null hypothesis in all cases. Comparisons between the fit of the different models were therefore made using the ratio between the increase in the fit statistic and the decrease in parameters estimated. On this basis a 61 parameter model where the slope on ability was kept constant did not fit appreciably less well than the full model for the loss of 59 parameters. A constant slope on ability was therefore retained. It was concluded that, relatively, the best fit with the least number of parameters was achieved by a 9-parameter model which included the effects of the expression type and exposure time in the difficulty parameter (or intercept).

8.4.2. Example 2 - Transitive Inference Test.

8.4.2.1. The Data.

This test was given under similar conditions to the previous example. As part of a recruitment selection procedure 1273 British Army applicants were given an experimental

timed item inference test consisting of 44 items. The questions, randomised in 5 blocks of 8 and 1 block of 4, and a summary of the response data matrix are attached in Appendix E.

Again, an essential feature of the test was the controlled response time. Each item was presented to the subject for a controlled period of time , either 2,3,4 or 5 seconds during which period the subject was unable to respond. At the end of the set time response buttons appeared on the computer screen and the subject was asked to give his response immediately.

8.4.2.2. The Analysis.

In analysing the response data various models which allowed structured parameters for difficulty (i.e. intercept) and discrimination (i.e. slope on ability) were fitted. Eight different problem types were identified according to the form in which the item was composed. These eight forms are illustrated in Table 3.

It was hypothesised that some of these problem forms were easier to solve than others. Hence problem type was set up as a factor with 8 levels to be included as part of the difficulty

TYPE	PROBLEM FORM	DIFFICULTY
1	A is better than B. Who is better?	1
2	A is better than B. Who is worse?	2
3	A is worse than B. Who is better?	2
4	A is worse than B. Who is worse?	1
5	A is not as bad as B. Who is better?	3
6	A is not as bad as B. Who is worse?	2
7	A is not as good as B. Who is better?	2
8	A is not as good as B. Who is worse?	3

TABLE 3. Problem Forms used in Transitive Inference Test.

and/or discrimination parameters. Time was also set up as a factor with four levels. As in the first example (above) the reciprocal of time could also be included in a model as a covariate. In these cases a slope parameter could be estimated and the product "slope.1/time" included in the difficulty parameter for each item. For comparative purposes the software also included 'item' as a factor with 44 levels.

8.4.2.3. Results.

As in the previous example the relative differences in fit between the restricted models were assessed by comparing the changes in the fit statistics with the changes in number of parameters estimated. This was because very large values of the χ^2 goodness of fit statistics were observed for all of the models. which resulted in rejection of all the models. Furthermore we note that it has been suggested that the χ^2 distribution is not valid in the binary data situation.

First of all the difficulty and discrimination parameters were modelled using problem type ('probtype') and/or time as factors in combination and alone. It can be seen from the table of results (Table 4) that both these factors contributed relatively highly to the fit of the model when included in the intercept. Conversely neither one nor the other (nor both) improved the model fit very much when included in the structure of the discrimination parameter.

Using these two factors only model M14 'probtype + time + γ'with 12 parameters was considered relatively the best model in terms of fit and parsimony. However by entering the reciprocal of time as a covariate into the make-up of the intercept a reduction to 10 parameters could be achieved (M13). This resulted in only a small loss of fit.

	MODEL	D.F.	FIT STATISTIC
M1	γ	878	44552.1
M2	1/time + γ	878	43802.8
M3	probtype + γ	871	42588.2
M4	probtype +probtype.γ	864	42538.5
M5	probtype + time.γ	868	41874.8
M6	probtype + (time + probtype).γ	861	41716.4
M7	time + γ	875	41532.5
M8	time +time.γ	872	41311.8
M9	time + probtype.γ	868	41048.6
M10	time + (time + probtype).γ	865	40913.1
M11	**probdiff + 1/time + γ**	**875**	**40346.8**
M12	**probdiff + time + γ**	**873**	**40318.5**
M13	**probtype + 1/time + γ**	**870**	**40168.6**
M14	**probtype + time + γ**	**868**	**40131.0**
M15	probtype + time +time.γ	865	40121.6
M16	probtype + time + probtype.γ	861	40108.4
M17	probtype + time + (time + probtype).γ	858	40088.8
M18	item + γ	835	39665.7
M19	item + item.γ	792	39538.8

TABLE 4. Results of Fitting Several Models to Transitive Inference Test Data.

An attempt was also made to reduce the eight different problem types to three types based on an assessment of their difficulty. The value of parameters for the eight problem types obtained from a previous model was used to assist this process. Each problem consisted of a statement and a question. It was hypothesised that an item was easier if (a) the comparative words used in the statement and question were the same (forms 1,4,6,7) and (b) the statement did not contain a negative (forms 1,2,3,4). This gave three difficulty levels:

(i) Problems with consistent wording in statement and question and a positive statement

(ii) Problems with either consistent wording and a negative statement or inconsistent wording and a positive statement

(iii) Problems with both inconsistent wording and a negative statement.

Thus the eight problem types were reduced to three. The levels of this difficulty factor ('probdiff') are shown in the third column of Table 3. Model M12 shows the result of fitting a 7 parameter model with the intercept made up of a problem difficulty effect (3 levels) and a time effect (3 levels) and a constant slope on ability. If time appears as a covariate instead of a factor (Model M11) the number of parameters to be estimated is only 5.

In conclusion it was found that four models M11 (five parameters), M12 (seven parameters), M13 (ten parameters) and M14 (twelve parameters) fitted relatively well. Although the 12 parameter model is a better fit the parsimony of the 5 parameter model is surely appealing. The parameter estimates for these four models are also given in Appendix F. Figures 8 and 9 show item response curves for the fitted five and seven parameter models respectively. It is clear that at each of the three difficulty levels the items become easier and the probability of success greater as the time allowed to complete the item is increased. In addition the graphs show that the probability of success is highest for the (supposed) easiest items at difficulty level 1 for all levels of time allowed, and the probability of success is lowest for the (supposed) hardest items at difficulty level 3. The seven parameter model, which includes time as a factor, predicts that there is very little difference in probability of success between items with exposure times of 4 seconds and items with exposure times of 5 seconds, suggesting that the extra second does not confer much more advantage. This result, although apparent, is less evident in the five parameter model which includes a slope on the reciprocal of time. Both

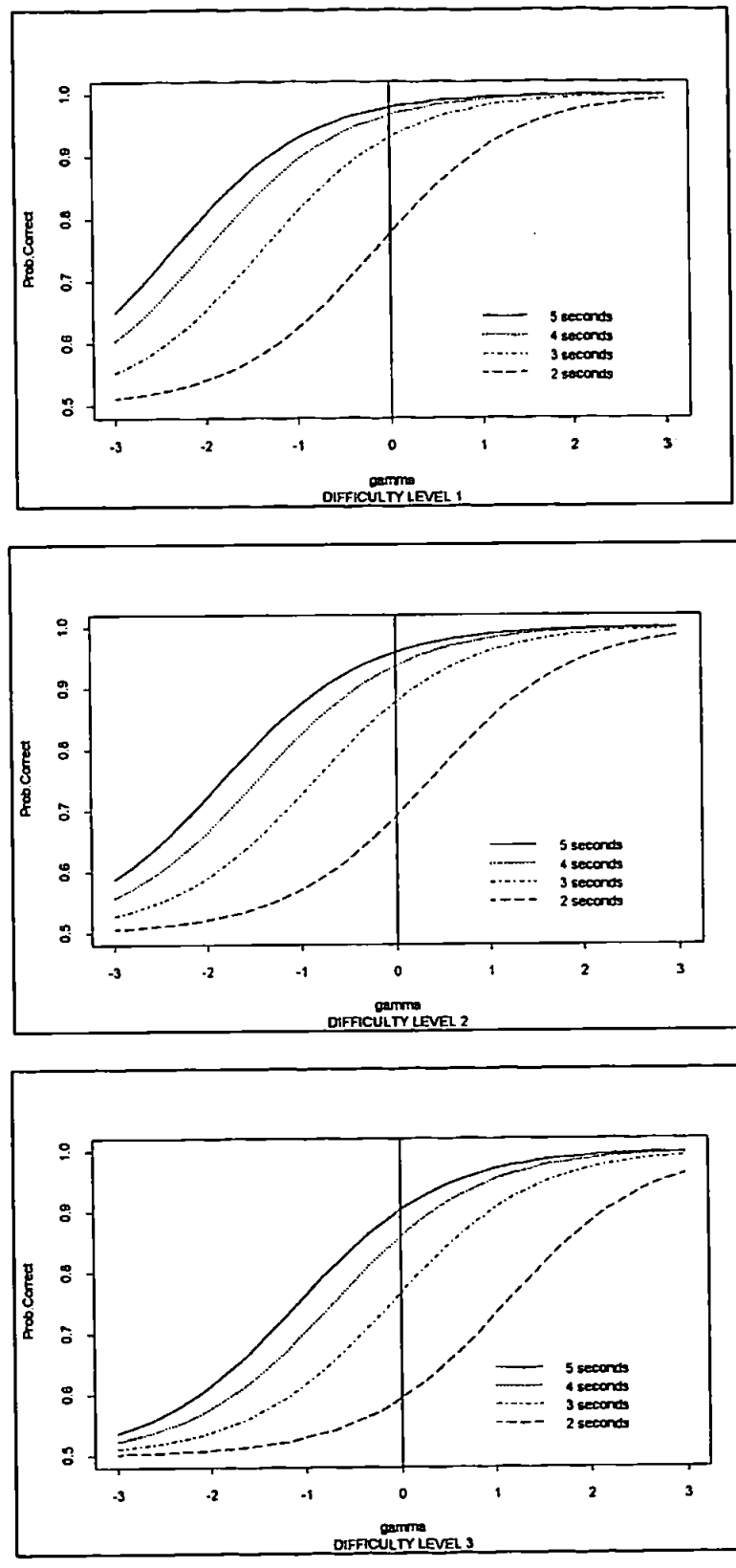


FIGURE 8. Item Response Curves for five parameter model.

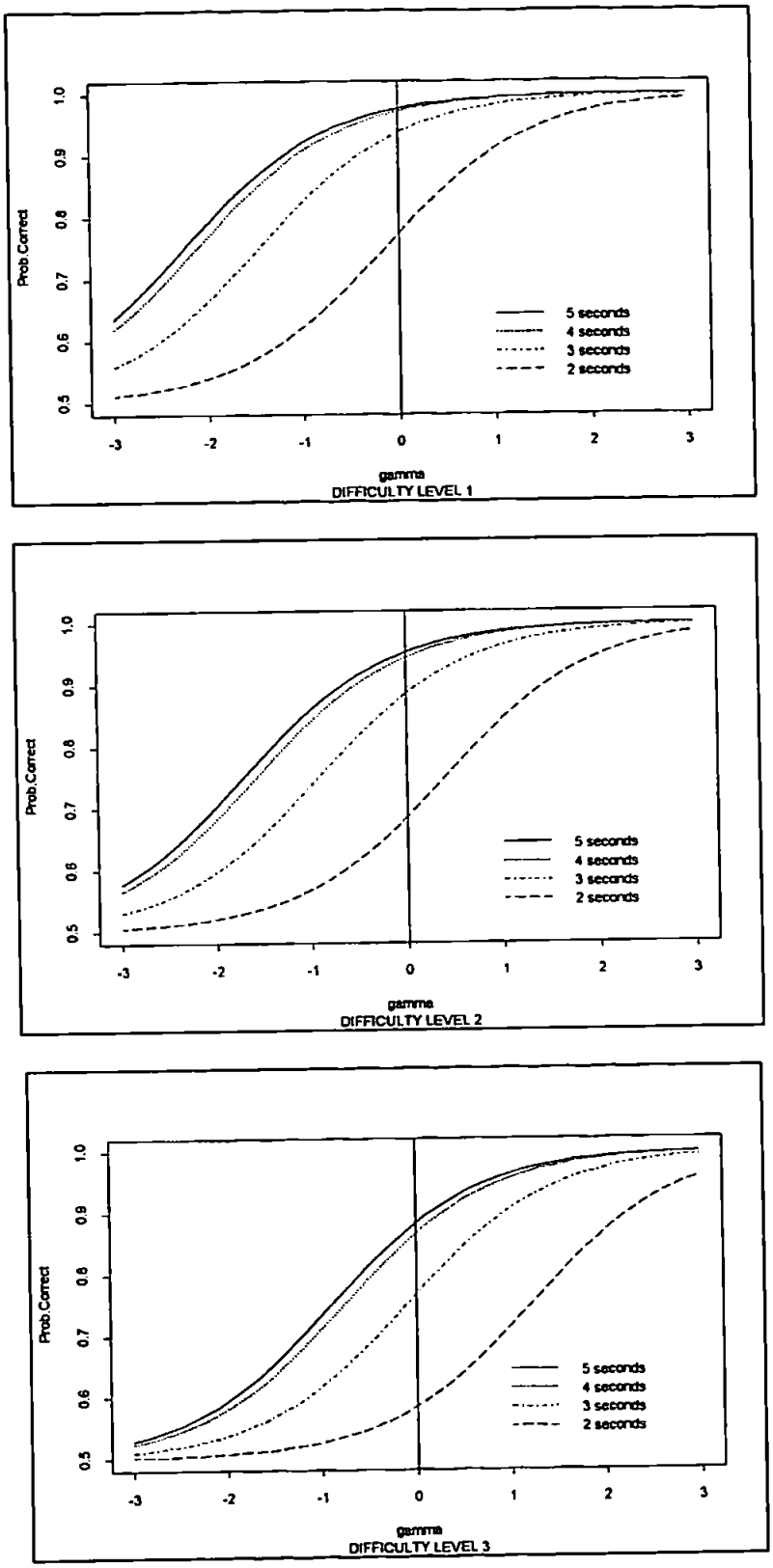


FIGURE 9. Item Response Curves for seven parameter model.

models predict that increasing the time allowed for an item from 2 to 3 seconds increases the probability of success by a far greater amount.

The item response curves in Figures 8 and 9 reflect the information in the following table which shows the mean number of correct responses given over all items at each difficulty level for each exposure time. For example, the mean number of correct answers to the most difficult items (level 3) when a time limit of 3 seconds is allowed is 929. This table does not of course take into account subject ability.

Difficulty	Time	Mean No. Correct
1	2	1001
1	3	1165
1	4	1216
1	5	1199
2	2	883
2	3	1106
2	4	1162
2	5	1177
3	2	777
3	3	929
3	4	1071
3	5	1177

CHAPTER 9. A SIMULATION STUDY.

9.1. OBJECTIVES.

The simulation study has been used frequently as a tool in the evaluation of statistical modelling techniques. For example, in Item Response Theory, as new methodology has been developed various simulation studies have been carried out to test and compare the properties of different models and fitting algorithms (see Chapter 4). One of the first such studies to investigate joint ML estimation under the three-parameter logistic model was undertaken by Frederick Lord in 1975. His data simulated a test of 90 items taken by 2,995 examinees. Later, Yen (1981) simulated samples of data from 1000 subjects and 36 test items in order to compare the three logistic models. Another study (Hulin, *et al*, 1982) was conducted to assess the accuracy of simultaneous item parameter and ability estimation in both the two and three-parameter models. Swaminathan and Gifford (1983) sought to investigate the properties of the three-parameter logistic model. They compared a maximum likelihood estimating procedure using the computer program LOGIST (Wood, Wingersky and Lord, 1978) with an alternative method of parameter estimation devised by Urry (1974) and implemented by the program ANCILLES. These are just a few examples: a review of the work done in this field including these and other simulation studies can be found in Baker (1987). Recently, Siegel (1996) used a simulated data set to compare the efficiency of multidimensional adaptive testing and one-dimensional adaptive testing in the measurement of abilities.

The simulation study described in this chapter was originally designed with a view towards investigating the sampling distributions of the estimators described in the model fitting procedure using the EM algorithm and GLIM. More specifically, it was hoped to gain insight into (a) the degree and nature of any bias in the parameter estimates and (b) the

precision of the estimates obtained. Of principle interest was the effect on both the bias and accuracy of sample size, that is the effect of both the number, I, of units (or clusters, groups or subjects) and the number, J, of observations (or items) within the units. A third important issue was the influence on the estimators of the known parameter c_j which represents the lower asymptote.

As the simulation study proceeded it became apparent that the results obtained were dependent not only on the sampling distributions of the estimators as required but also on the numerical artefacts of the computational procedures employed. The study was therefore extended to include an investigation of these extraneous factors and an assessment of their impact on this and future simulation studies. The work described in this chapter can therefore be regarded as a pilot study in which the ground is prepared for further investigations.

9.2. DESIGN.

Simulated data for the purposes of this study were generated from the model described in Section 6.2.1 where

$$P(Y_{ij} = 1) = \pi_{ij} = c_j + \frac{1 - c_j}{1 + \exp(-\eta_{ij})}, \; i = 1,2,\ldots,I; \; j = 1,2,\ldots,J$$

and the linear predictors, written in a simplified form, are

$$\eta_{ij} = \varphi_j + \alpha_j \gamma_i$$

The study proceeded in the following way. The slope and intercept parameters φ_j and α_j were set to some suitable fixed values. These values therefore constituted the true values of the parameters which the fitting procedure would attempt to recover, with varying sample sizes and values of the lower asymptote c_j. The values of I (no. units) used were 50, 100, 200 and 400. Larger sample sizes would have been a considerable drain on

computer time and probably not have added to the overall conclusions. The values of J (number of observations per unit) were set to 25, 50 and 100. These values directly reflected the quantity of information about the parameters contained in the data. A sample size of 25 might represent a large amount of information for a 2-parameter model but a small amount for a model with more parameters. Various attempts were made to run the simulations with J =200 but this resulted in considerable numerical problems in the software, mainly due to the small likelihoods generated. As 200 was also considered an unrealistically high value for J in most applications it was decided not to proceed with this value in the present study. The 12 different combinations of I and J were tested with the lower asymptote parameter set to 0 and to 0.5, for all j, giving a total of 24 different tests. The lower asymptote indicates the probability of obtaining a given response completely by chance. The maximum possible value is 0.5 for binary outcomes. The value of 0 corresponds to a situation where the expected response is not influenced by chance. (An example of this is a test situation where the subject must provide an answer rather than choose from given alternatives.)

For each individual simulation, I independent values of the random effect γ_i were generated from a standard normal distribution. The probabilities π_{ij}, were then calculated from the above model. The binomial error term was simulated by putting $y_{ij} = 1$ if $u_{ij} \leq \pi_{ij}$, otherwise $y_{ij} = 0$, where the u_{ij} are independently drawn from a uniform distribution on [0,1]. This implies that $P(Y_{ij} = 1) = \pi_{ij}$ and $P(Y_{ij} = 0) = 1 - \pi_{ij}$, as required. In each of the 24 different simulation situations, 100 independent sets of response data were generated and the model fitting procedure run on each set. The empirical distributions of the 100 estimates for each parameter in each situation were examined for normality. The means and standard errors were calculated and plotted against sample size.

For the pilot study the true values of both the slope and intercept parameters φ_j and α_j were set to 1 for all j. With the dimension of the unknown parameter vector restricted to 2, examination of the likelihood function could be accomplished by graphical methods. It was hoped that the simplicity of this model would reveal some of the major strengths and weaknesses of the procedure.

9.3. SUBSIDIARY ISSUES.

It was recognised that variables in the model fitting procedure other than sample size and parameter values might contribute towards the bias and accuracy of the final estimates. It was therefore necessary to investigate and as far as possible eliminate or at least stabilise their influence on the fitting procedure. Into this category came (1) the value of the parameter vector $(\hat{\alpha}^{(0)},\hat{\varphi}^{(0)})$ used to start the algorithm, (2) the tolerance used to detect convergence, and (3) the range and number of nodes used for the integration approximation.

Questions concerning the validity of results stemmed from the fact that there are two different vectors which estimate the true parameter (α,φ). These are (i) the true maximum likelihood estimate $(\hat{\alpha},\hat{\varphi})$ and (ii) the approximation $(\hat{\hat{\alpha}},\hat{\hat{\varphi}})$ which is actually obtained from the software. The three variables I, J and c_j, together with the K integration nodes and weights, determine the likelihood function and therefore its maximum $(\hat{\alpha},\hat{\varphi})$ but the tolerance level and starting parameters influence $(\hat{\hat{\alpha}},\hat{\hat{\varphi}})$, or, in other words, the accuracy with which $(\hat{\alpha},\hat{\varphi})$ is obtained. The empirical distributions of $\hat{\alpha}$ and $\ddot{\varphi}$ and their relationship to the true values were obviously of primary interest in the investigation. However the distributions obtained from the study were of $\hat{\hat{\alpha}}$ and $\hat{\hat{\varphi}}$ and it was not initially

possible to know with what precision $(\hat{\hat{\alpha}},\hat{\hat{\varphi}})$ estimated $(\hat{\alpha},\hat{\varphi})$, nor the extent of the effect of the error.

Initial choices were made for these three variables and then revised as further information came to light during the course of the study. The bases for the decisions that were made are outlined in this section.

9.3.1. Starting Values.

Preliminary observations of the effect of the starting values on the final estimates appeared to indicate that low starting values led to low estimates of the intercept parameter in particular. In order to investigate this an examination of the likelihood functions of various data sets was conducted. In each case, the likelihood of the data was calculated over a 2- dimensional grid of parameter values. The grid points were at intervals of 0.01 over the range 0.75-1.25 for both intercept and slope. The 'true' parameter, (1, 1), was therefore in the centre of the grid. Each likelihood value was then expressed as a proportion of the maximum on the grid and a contour map based on lines of equal proportions was plotted using the software S-PLUS (StatSci.,1991). Examination of these likelihood plots and of cross-sections taken horizontally and vertically through the maximum indicated that the 'hills' were typically of a shape elongated in the direction of the intercept axis. Thus a small increase in the likelihood function was associated with a much greater change in the intercept than in the slope. This finding was backed up by observations of the changing parameter estimates during convergence of the fitting algorithm, when much greater changes were generally observed in the intercept parameter between iterations than in the slope parameter. This meant that the final estimates tended to be nearer the maximum of the likelihood function for the slope than for the intercept; the small change in the likelihood needed to reach the maximum after the procedure converged would have produced more

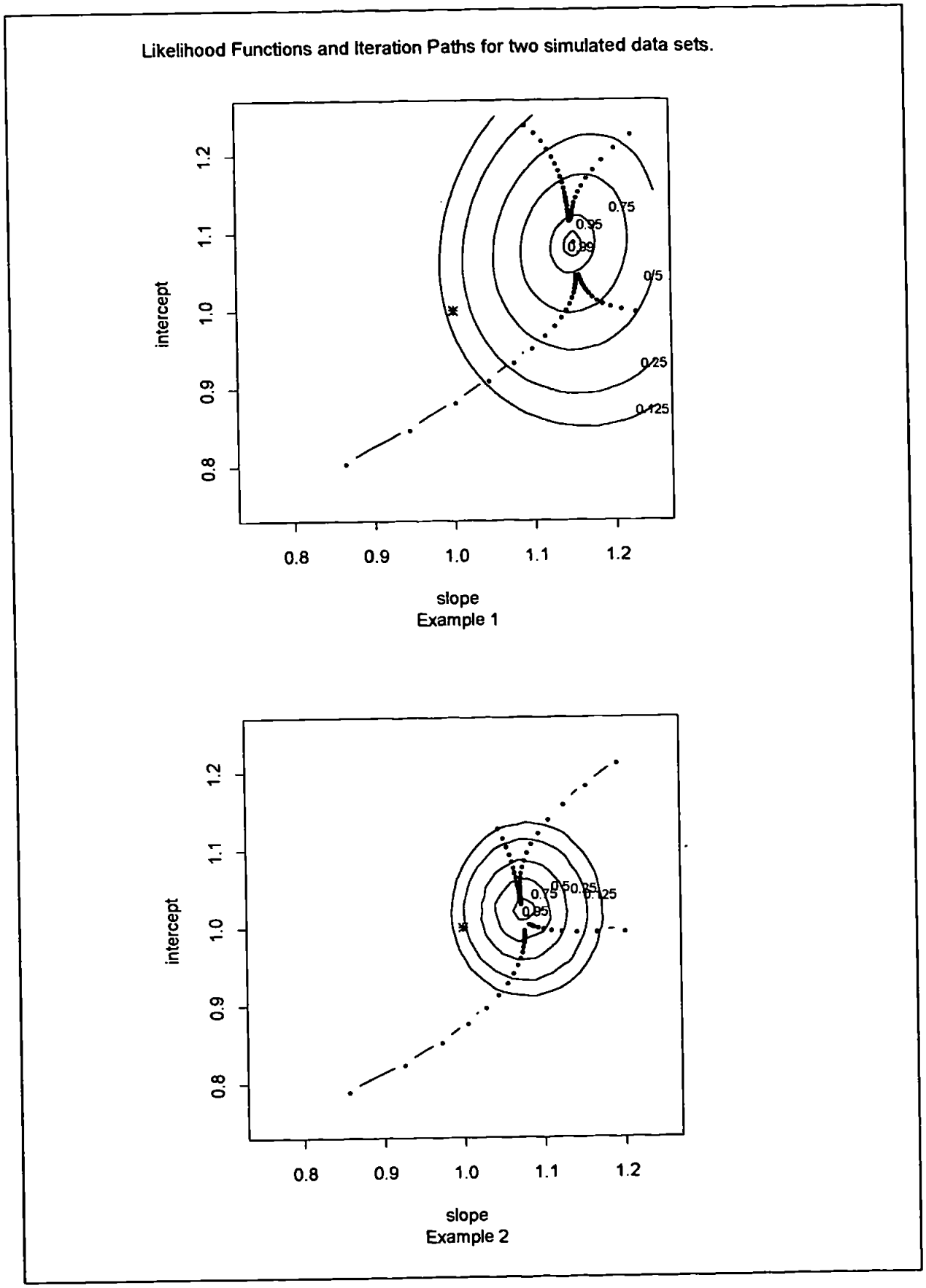


FIGURE 10. Likelihood contour maps showing iteration paths.

change in the estimate for the intercept than the slope. As a result of this property of the likelihood function the intercept estimates depended more heavily on the value of the starting vector since this determined the direction from which the maximum was approached.

In order to illustrate the dependency of the final estimates on their starting values, the fitting algorithm was run on the same data set using 4 different initial vectors. The estimates obtained at each iteration of the EM algorithm were plotted on the likelihood contour maps in order to show the 'paths' of convergence towards the maximum. Two examples are shown in Figure 10. Example 1 is a data set consisting of 50 observations on each of 100 subjects with a lower asymptote set to 0. In example 2 there are 100 observations on each of 400 subjects and a value of 0.5 for the lower asymptote. The values $(\hat{\alpha}^{(0)},\hat{\varphi}^{(0)})$ used were (1.25, 1.25), (0.75, 0.75), (0.75, 1.25) and (1.25, 1). In both examples all 4 fitting procedures converged to estimates of the maximum likelihood value of the slope parameter which were all within 0.01 of each other. The difference between the lowest and highest estimate of the intercept parameter was 0.07 in example 1 and 0.04 in example 2 (see Table 5). It was also noted that in example 1 the estimates all appear to be roughly equidistant from the maximum; in example 2 the 2 sets of estimates that start from an intercept equal to 1.25 appear to be nearer the true maximum.

The evidence indicated that constantly approaching the maximum of the likelihood function from a single fixed starting vector could result in biased estimates of the intercept in particular and, to a lesser extent, the slope. Because the tendency was either to consistently underestimate or consistently overestimate the ML estimate $(\hat{\alpha},\hat{\varphi})$ the expected value of the distribution of $(\hat{\hat{\alpha}},\hat{\hat{\varphi}})$ could not be assumed to be the same as that of $(\hat{\alpha},\hat{\varphi})$. Limited experimentation using the same 100 data sets with different fixed starting vectors did indeed confirm that bias could be induced in the intercept parameter by using

certain starting values. As the position of the maximum in relation to any given point on the grid could not be pre-determined, it was decided to overcome this problem by randomising the starting vector in the range

$$\left(0.75 \le \hat{\alpha}^{(0)} \le 1.25,\ 0.75 \le \hat{\varphi}^{(0)} \le 1.25\right)$$

This did have the effect of eliminating the bias in the example tested. The standard errors obtained when the starting values were randomised were not noticeably different from those obtained when fixed values were used. It should be noted however that only one combination of sample sizes was tested in order to produce this finding; it was assumed, without testing, that it would apply equally to all the simulation situations.

	EXAMPLE 1		EXAMPLE 2	
Start	Slope Est.	Intercept Est.	Slope Est.	Intercept Est.
(1.25, 1.25)	1.15	1.12	1.07	1.04
(0.75, 0.75)	1.15	1.05	1.08	1.00
(0.75, 1.25)	1.15	1.12	1.07	1.03
(1.25, 1)	1.16	1.05	1.08	1.01

TABLE 5. Parameter estimates obtained for different starting values.

9.3.2. Convergence Criteria.

In order to detect convergence and stop the model fitting algorithm, the fit statistic, $-2l$, where l is the observed data log likelihood at the current parameter estimates, was calculated. When the difference between two consecutive values of this statistic was less than a given tolerance value then the algorithm was assumed to have reached a maximum value of the likelihood function.

In a real data situation the model fitting algorithm can be run continuously with decreasing tolerances until either a satisfactory accuracy in the parameter values is obtained or no further improvement is possible. However, as far as the simulation study was concerned, a tolerance that would lead to convergence in a reasonable number of iterations whilst producing estimates that were close to those that maximised the likelihood was required. Since a set of simulations using large sample sizes might run for over 24 hours, it was thought worthwhile to sacrifice some accuracy for the sake of reducing the overall run times. Some experimentation was therefore undertaken in order to assess how to fix the tolerance to achieve a reasonable level of accuracy in all conditions without incurring the cost of excessive run times.

The magnitude of the fit statistic depended on the sample sizes and could be calculated as, very roughly, $2I = I \times J$, (bearing in mind that the true log likelihoods differed from their computed values by an additive constant). This product was reduced by approximately 0.8 when the lower asymptote was set to 0.5. The fit statistic was found to be accurate only to 6 or 7 significant figures. Setting the convergence criterion to detect small differences in the 7th significant figure therefore sometimes resulted in a failure to converge. Since the accuracy of the estimates therefore depended on the sample sizes through the fit statistic, fixing the tolerance at a constant value for all the simulation situations would have resulted in a greater accuracy in the larger sample estimates than in the smaller data sets. The tolerance therefore needed to be set to different levels according to the magnitude of the fit statistic.

It was eventually decided to calculate a tolerance equal to $I \times J \times 5 \times 10^{-6}$. This level was chosen to ensure that convergence would occur in relatively few iterations for all starting values and sample sizes. These tolerances appeared to give similar levels of accuracy in the two examples discussed in section 9.3.1. However, after noting the rather

strange behaviour of the standard errors in the results table (see Figure 11) some of which increased rather than decreased with sample size, it was decided to look more closely at the accuracy of the estimates of $(\hat{\alpha},\hat{\varphi})$. It was felt that some proportion of the error in the parameter estimates might be attributable to the numerical constraints of estimation rather than the true variation in the ML estimates. Firstly the calculation of the likelihood ratio statistic on which convergence was based was reviewed. More accuracy in this statistic was achieved by handling the calculation and checking for convergence within the FORTRAN subroutines of the model-fitting software. Secondly, having established a more accurate fit statistic, tests were conducted to see whether it was possible to improve on the estimates previously obtained. Table 6 shows the results of using the new calculations on the two example data sets described in section 9.3.1. Two different starting values were used for each example. The largest tolerance shown in each case is that which gave the best estimate using the old fit statistic (restricting the choice to powers of 10). The figures in brackets are the previous best estimates.

Although the tighter convergence criteria made little difference to the estimation of the slope parameter the new procedure allowed more accurate estimation of the intercept. At the cost of vastly increased numbers of iterations particularly for the more 'distant' starting value (0.75,1.25), accuracy to 3 decimal places on both parameters was obtained with tolerances of 10^{-6} for the smaller data set and 10^{-5} for the larger. In view of these results, the 2 sets of simulations corresponding to the sample sizes used in the two examples were run again using the same simulated data and same random starting values as before. However the tolerance levels were reduced to be reasonably confident of accuracy to 2 decimal places on the parameter estimates. The object was to discover the effect of the greater accuracy on the standard errors of $(\hat{\alpha},\hat{\varphi})$.

Example 1	True ML estimate. (1.1494, 1.0842)					
	Start: (1.25,1)			Start: (0.75,1.25)		
Tolerance	Slope Est.	Intcpt Est.	# Iterations	Slope Est.	Intcpt Est.	# Iterations
.0001	1.150 (1.150)	1.082 (1.078)	44 (33)	1.149 (1.149)	1.086 (1.090)	81 (46)
.00001	1.150	1.083	57	1.149	1.085	93
.000001	1.149	1.084	70	1.149	1.084	106
Example 2	True ML estimate. (1.0744, 1.0211)					
	Start: (1.25,1)			Start: (0.75,1.25)		
Tolerance	Slope Est.	Intcpt Est.	# Iterations	Slope Est.	Intcpt Est.	# Iterations
.001	1.075 (1.075)	1.019 (1.017)	24 (20)	1.074 (1.074)	1.024 (1.026)	58 (32)
.0001	1.075	1.020	32	1.074	1.022	67
.00001	1.074	1.021	41	1.074	1.021	76

TABLE 6. Parameter estimates obtained from different convergence criteria.

The results of these tests compared with the results of the original simulations (see Tables 8 and 9) showed that the improved accuracy made almost no difference to either the means or the standard errors rounded to 2 decimal places. In fact, with the narrower criteria the standard errors showed small increases in example 1 and small decreases in example 2. The run times for the 100 simulations were vastly longer and became prohibitive in the large samples. The conclusion was that it was reasonable to accept the validity of the results obtained with the original convergence criteria. The more accurate calculation of the fit statistic was however incorporated into the software for all future implementations.

9.3.3. Range and number of nodes.

The greater the number of nodes, K, used to approximate the integral

$$\int p_{\underline{Y}_i|\gamma}\left(\underline{y}_i|\gamma_i,\underline{\beta}\right)f_\theta(\gamma_i)d\gamma_i$$

by

$$\sum_{k=1}^{K} p_{\underline{Y}_i|\gamma}\left(\underline{y}_i|\gamma_k,\underline{\beta}\right)f_\gamma(\gamma_k)w_k$$

the longer the calculations during both the expectation and the maximisation steps. In large samples a high number of nodes could cause excessively long run times. Since the likelihood function includes the above integral it depends in part on the choice of K which in turn, together with the specified range of integration, determines the value of both the nodes, γ_k, and the weights, w_k. It was thought necessary to examine whether increasing K resulted in better approximations of the true parameter value (1,1). Several different data sets were examined. For each set a starting vector and a tolerance were fixed and the model fitting

	EXAMPLE 1		EXAMPLE 2	
No. Nodes	Slope Est.	Intercept Est.	Slope Est.	Intercept Est.
8	1.05	1.24	1.06	1.00
12	1.16	1.05	1.08	1.01
16	1.15	1.07	1.08	1.02
20	1.16	1.07	1.08	1.02
40	1.17	1.07	1.08	1.02

TABLE 7. Parameter estimates obtained from different numbers of nodes.

algorithm was run using 8,12,16,20 and 40 nodes within the range (-3,3). Table 7 shows the resulting parameter estimates for the two example data sets used previously. It is chiefly noticeable how little difference there is between the estimates except in the case of 8 nodes in example 1. In addition increasing the number of nodes does not appear to move the estimates towards the true values in either example. On the basis of this and similar results it was decided to perform the integration with 12 nodes (or 3 panels) between the values of -3 and 3, suitable limits for a standard normal distribution.

9.4. RESULTS:

9.4.1. Bias.

The results of the first part of the simulation study are tabulated in Table 8 with all figures rounded to 2 decimal places. The entries for the slopes and intercepts are the means and standard deviations of the 100 estimates of each parameter obtained in each simulation. The results showed no indication of bias in the estimates of either the slope or the intercept parameter even in the smallest samples. The largest deviation of a mean from the true parameter value was 0.04 and the smallest standard error 0.04 so the departures from the true values were not significant in any instance. Results obtained when the accuracy of the fit statistic was improved are shown in Table 9 and verify the first table of figures shown (see Section 9.3.2.).

		Lower Asymptote = 0				Lower Asymptote = 0.5			
Obs(J).	Units(I)	slope	s.e.	intcpt	s.e.	slope	s.e.	intcpt	s.e.
25	50	0.99	0.13	0.98	0.17	0.99	0.19	1.00	0.16
25	100	1.00	0.09	0.99	0.10	0.98	0.13	0.98	0.13
25	200	1.00	0.07	1.00	0.08	1.00	0.10	0.99	0.09
25	400	0.99	0.05	1.00	0.05	1.00	0.07	1.01	0.06
50	50	0.99	0.12	0.96	0.15	1.01	0.15	1.03	0.15
50	100	0.99	0.08	0.99	0.10	1.00	0.10	1.01	0.11
50	200	1.01	0.06	1.01	0.08	1.00	0.08	0.99	0.07
50	400	1.01	0.04	1.00	0.06	1.01	0.05	1.00	0.05
100	50	0.99	0.11	1.01	0.15	0.99	0.13	0.98	0.13
100	100	0.99	0.08	1.01	0.12	1.00	0.09	0.99	0.10
100	200	0.98	0.05	1.01	0.08	1.01	0.06	0.99	0.08
100	400	0.98	0.04	1.00	0.07	1.01	0.05	1.00	0.06
TRUE VALUES		1		1		1		1	

TABLE 8. Simulation results showing mean parameter estimates and their standard errors for different sample sizes and lower asymptotes.

	Tol.	Slope	s.e	Intercept	s.e.
I=100, J=50, c_j=0	.0125	0.993	0.081	0.993	0.097
	.0001	0.993	0.083	0.996	0.106
	.000001	0.993	0.084	0.996	0.107
I=400, J=100, c_j=0.5	0.2	1.005	0.050	0.998	0.061
	.001	1.005	0.049	1.000	0.058
	.00001	1.005	0.049	1.000	0.058

TABLE 9. Selected simulation results obtained for different convergence criteria using improved fit statistic.

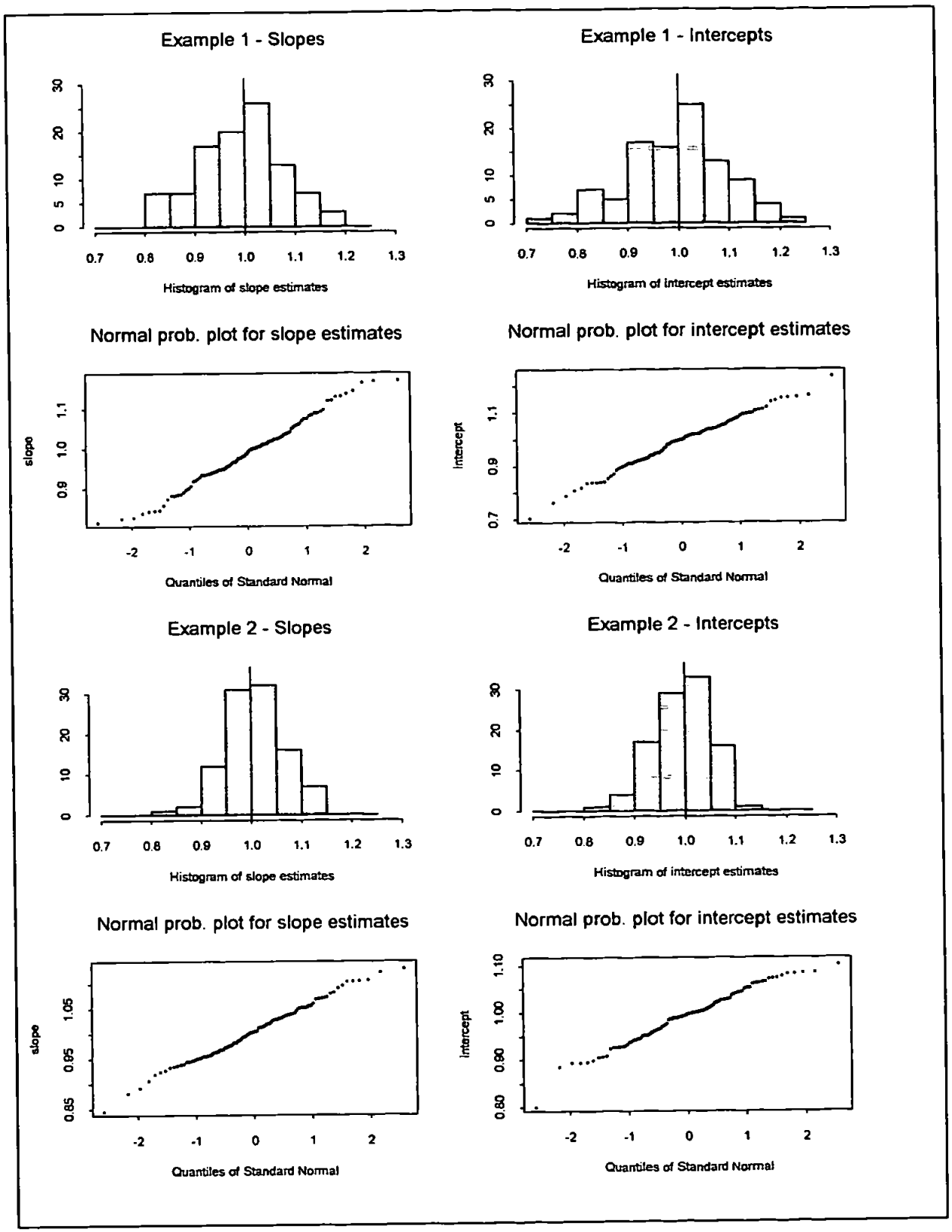


FIGURE 11. Histograms and Normal Probability Plots for parameter estimates from two simulated data sets.

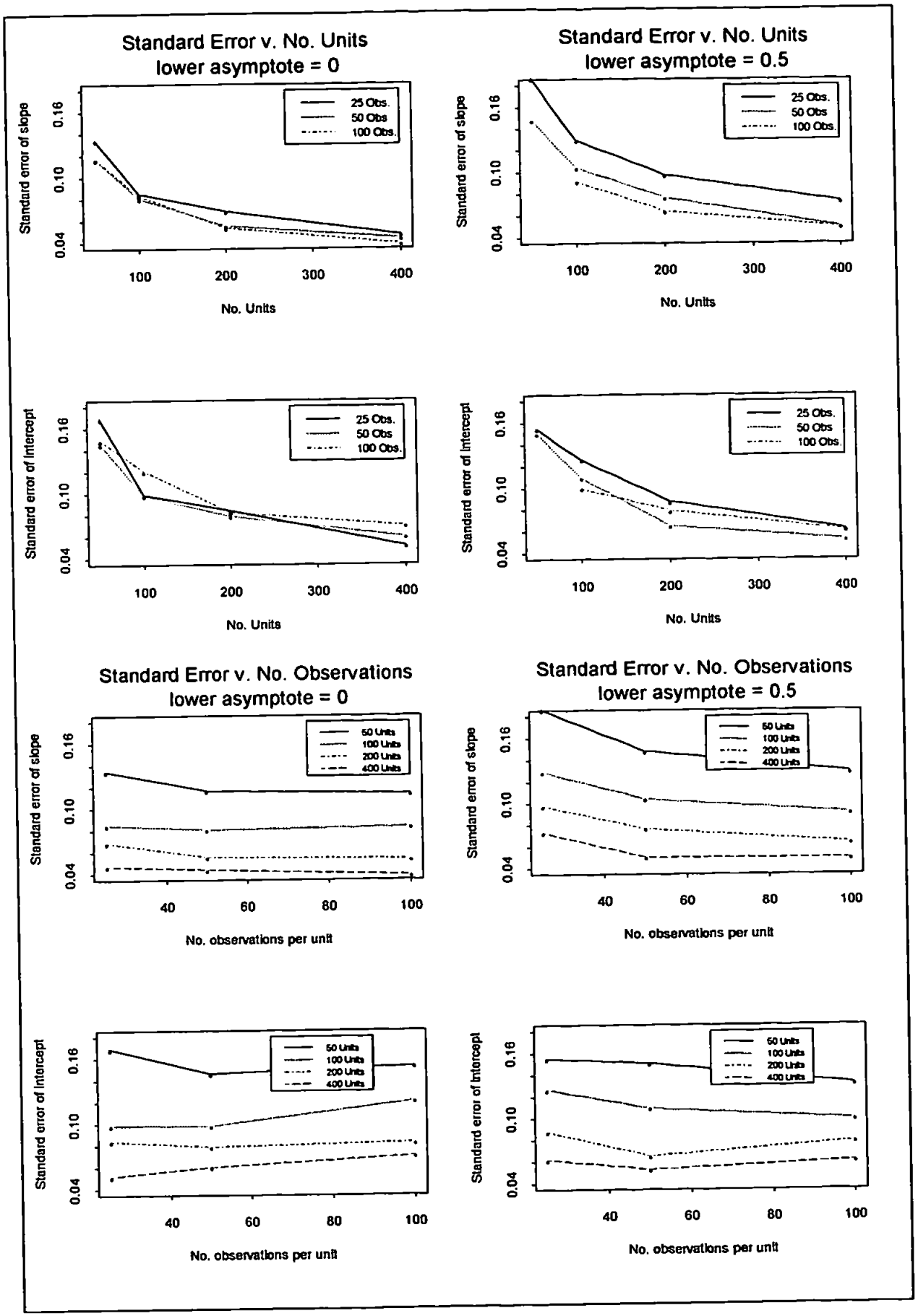


FIGURE 12. Standard errors of parameter estimates.

9.4.2. Normality.

Histograms of the parameter estimates and normal probability plots were obtained in each separate simulation. Figure 11 shows these graphs for both the intercept and slope parameters in two example simulation situations, corresponding to the individual examples previously examined in this section. In example 1, the simulated data consisted of 100 replications of samples of 50 observations on each of 100 units with a lower asymptote equal to 0. In example 2 the simulation consisted of larger data sets with 400 units, 100 observations per unit and a lower asymptote set to 0.5. The probability plots confirm the normality of the distributions of both in both examples. The histograms reveal that all four means are close to 1 and that smaller variances are observed in the larger sample.

9.4.3. Standard Errors.

Figure 12 illustrates in more detail how the standard errors varied with sample size. The top four graphs show how the standard errors of both parameters decrease as the number of units in the sample increases. This distinct trend is observed with both values of the lower asymptote, and appears however many measurements are made on each unit. A possible conclusion to be drawn from this is that the accuracy of the fitting procedure is sensitive to the normality of the distribution of the random effect in the sample. As the approximation to a standard normal distribution improves with the larger values of I so do the parameter estimates. The slope parameter is estimated with greater precision (i.e. smaller standard errors) when the lower asymptote is set to 0. When $c_j = 0.5$ the amount of information in the observations about the random effect is effectively reduced by a half, since this is equivalent to a half of the observations occurring by chance rather than as a result of the influence of the random effect. Therefore less accurate estimation of the slope

parameter (also the standard deviation of the random effect) could be expected. The lower asymptote does not have a similar effect on the intercept.

The four lower plots in Figure 12 show the behaviour of the standard errors as the number of observations within a unit is increased. This time there is no overall tendency for larger numbers of observations to be associated with smaller standard errors. Increasing the number of observations beyond 50 per unit has some effect on the precision of the slope parameter when $I = 50$, 100 or 200 and when the lower asymptote is 0.5. For the intercept, the standard errors increase when $J > 50$ except for $I = 50$ or 100 and $c_j = 0.5$. Thus increasing the number of observations appears to improve the precision of the estimate only over the samples where there is less information in the data. These results lead to the possibility that saturation levels can be reached. Increasing the number of observations beyond a certain point which depends on I cannot improve the precision of the estimates.

In the 2-parameter model all the observations are contributing to the estimation of (α, φ). (For example, in the IRT application this corresponds to a situation where all the test items are of identical difficulty and discrimination.) If there were more parameters in the model presumably there would be larger standard errors for small values of J but the precision could be expected to improve as the number of observations increased.

CHAPTER 10. GENERALIZING LATENT VARIABLE GLMS FOR EXPONENTIAL RESPONSES.

10.1. INTRODUCTION.

In Chapter 6 the methodology for latent variable GLMs was applied to binary response data. In Section 10.2 of this chapter the same procedures are adapted hypothetically to Poisson data thought to be dependent upon latent variables. An expression is derived for the log likelihood of the expected complete data and this is compared to the log likelihood of the standard GLM for Poisson data where covariates are all fixed and known. As in the binomial case the form of the expected complete data likelihood allows parameters to be estimated with standard maximisation routines designed for the GLM.

Binomial and Poisson variables are both discrete and both have fixed relationships between their means and variances. Normally distributed data have neither of these properties. Although other methods for estimating random effects in normal models are obviously well developed (e.g. Searle, 1971; Harville, 1975; Draper and Smith, 1981; Hocking, 1985), the methodology for latent variable GLMs is applied to the normal case in Section 10.3 for theoretical interest. Normal data which can be assumed to have constant variance is considered first, followed by the more general case where the variance is allowed to differ between observations.

In the final part of this chapter, Section 10.4, a latent variable GLM is considered in its most general form. Without specifying a distribution for the response data other than that it is from the exponential family, an expression for the expected complete data log likelihood is derived. This is compared as before with the log likelihood of the standard GLM, this time in its

general form. As a result of this process general rules which govern the relationship between the two likelihood functions are deduced.

10.2. A MODEL FOR POISSON RESPONSES.

A latent variable GLM is appropriate in situations where the response data are realisations of Poisson variables dependent upon some unknown random covariate. For example the data might be road traffic accident counts (Wright and Barnett, 1991) in which an important contributory variable was not measured at the time that the data was collected. Overdispersion in Poisson models has been examined by Hinde (1982) and Aitkin and Francis (1966). Brillinger and Preisler (1983) looked at counts of red blood cells which depended on a latent covariate. Machine failures where some kind of propensity to failure on the part of the machine underlies the data could also be modelled in this way; a similar latent variable might contribute to counts of flaws in different fabric samples.

10.2.1. The Poisson Response Latent Variable GLM.

As in Section 6.2.1 it will be assumed that the data has a nested structure with J observations available on each of I units. In the IRT example in section 8.4 there is a link between all the j^{th} responses (i.e. $y_{1j}, y_{2j}, \ldots, y_{Ij}$ for any j) in that they are all responses to the same item. It is then attributes at item level which are represented by the parameter $\underline{\beta}_j$.

Similarly in the Poisson case the j^{th} count on one unit is in some sense the same type of response as the j^{th} count on the others. For example, in a transport application, there might be J counts of accidents recorded on different days of the week at each of I junctions so that y_{ij} is

the count at junction i on day j and $\underline{\beta}_j$ is the vector of parameter values (e.g. visibility due to weather) associated with day j.

As before the response y_{ij} is a realisation of random variable Y_{ij} and represents the j^{th} (j=1,2,...,J) observation on the ith unit (i=1,2,...,I). It is possible that not all J observations are recorded for every unit. As before the total number of responses recorded for unit i is denoted $J(i)$ and the total number of units responding to item j is denoted $I(j)$. The expected value of Y_{ij} is dependent on unknown parameter vector $\underline{\beta}_j$ and latent covariate γ_i, a realisation of latent variable Γ_i which contributes to the model at the unit (e.g. junction) level. This time it is assumed that the conditional distribution of Y_{ij} is Poisson.

$$Y_{ij} \sim Po\left(\lambda_{ij}(\gamma_i)\right)$$

The canonical link function for the Poisson distribution is

$$\eta_{ij}(\gamma_i) = \ln \lambda_{ij}(\gamma_i)$$

so that parameter λ_{ij} conditional on γ_i is given by the inverse link function

$$\lambda_{ij}|\gamma_i = \exp(\eta_{ij})$$

The linear predictor associated with observation y_{ij} can be assumed to be the same as in the binary response model

$$\eta_{ij} = \underline{x}_j^T \underline{\varphi}_j + \alpha_j \gamma_i$$

where $\underline{x}_j^T$ is the row of the fixed effects design matrix associated with response y_{ij}, $\underline{\varphi}_j$ is a vector of fixed effect parameters and α_j is the discrimination parameter, or slope on γ_i. As in Section 6.2.1 the distribution of Γ_i is assumed standard normal.

For $Y_{ij} \sim Po\left(\lambda_{ij}\right)$, the log likelihood expressed as a function of λ_{ij} is

$$l_y\left(\lambda_{ij}|y_{ij}\right) = y_{ij}\ln\lambda_{ij} - \lambda_{ij} + C_{P1}$$

where $C_{P1} = -\ln(y_{ij}!)$ is a constant. So, assuming independent observations conditional on $\underline{\gamma}$

$$l_{\underline{y},\underline{\gamma}}\left(\underline{\lambda}|\underline{y},\underline{\gamma}\right) = \sum_{i=1}^{I}\sum_{j=1}^{J(i)}\left(y_{ij}\ln\lambda_{ij} - \lambda_{ij}\right) + C_{P2} \qquad (10.1)$$

where C_{P2} is a constant.

Function (10.1) is therefore the equivalent Poisson likelihood to equation (6.2), the likelihood for binary responses. With known $\underline{\gamma}$ this is the conventional likelihood function for a Poisson response GLM and the values of $\underline{\beta}$ which maximise it can be computed with software such as GLIM.

10.2.2. The Observed Data Likelihood.

The observed data likelihood for the Poisson model is obtained, as in Section 6.2.2, by taking the joint distribution of $\underline{Y}$ conditional on $\underline{\beta}$ and $\underline{\gamma}$, integrating with respect to $\underline{\gamma}$ and then approximating the integral using a Gaussian quadrature rule as before. Taking logs of the result gives the following likelihood function for the observed data, equivalent to (6.8):

$$l_{\underline{y}}\left(\underline{\beta}|\underline{y}\right) \approx I\ln\frac{1}{\sqrt{2\pi}} + \sum_{i=1}^{I}\ln\sum_{k=1}^{K}\left(\prod_{j=1}^{J(i)}\frac{e^{-\lambda_{jk}}\lambda_{jk}^{\,y_{ij}}}{y_{ij}!}\right)\exp\left[-\frac{\gamma_k^{\,2}}{2}\right]w_k \qquad (10.2)$$

It is assumed that this function is as difficult to maximise as the likelihood for binary responses. It is therefore necessary to examine the expected complete data log likelihood for the Poisson model following the reasoning in Section 6.2.4.

10.2.3. The Expected Complete Data Likelihood.

The log likelihood function for the complete data is again found by first taking natural

logarithms of the joint p.d.f. of the two variables $\underline{Y}$ and $\underline{\Gamma}$:

$$l_{\underline{y},\underline{\gamma}}(\underline{\beta}|\underline{y},\underline{\gamma}) = I \ln\frac{1}{\sqrt{2\pi}} + \sum_{i=1}^{I}\sum_{j=1}^{J(i)}\left[y_{ij}\ln\lambda_{ij} - \lambda_{ij}\right] + C_{P2} - \sum_{i=1}^{I}\left[\frac{\gamma_i^2}{2}\right] \qquad (10.3)$$

Taking expectations of function (10.3) over the posterior distribution of $\underline{\Gamma}$ given the data and parameter estimates as in equation (6.9) we obtain

$$Q\left(\underline{\beta}|\underline{y},\hat{\underline{\beta}}\right) = \int_R \sum_{i=1}^{I}\sum_{j=1}^{J(i)}\left[y_{ij}\ln\lambda_{ij} - \lambda_{ij}\right] f_{\Gamma|\underline{y}_i}\left(\gamma_i|\underline{y}_i,\hat{\underline{\beta}}\right)d\gamma_i + C_{P4}$$

where R is simply the region over which γ_i is defined and

$$f_{\Gamma|\underline{y}_i}\left(\gamma_i|\underline{y}_i,\hat{\underline{\beta}}\right) = {C_{P3}}^{-1}\prod_{i=1}^{I}\prod_{j=1}^{J(i)}\frac{e^{-\lambda_{ij}}\lambda_{ij}^{\,y_{ij}}}{y_{ij}!}\exp\left[-\frac{\gamma_i^2}{2}\right]$$

with normalising constant

$$C_{P3} = \int_R \prod_{i=1}^{I}\prod_{j=1}^{J(i)}\frac{e^{-\lambda_{ij}}\lambda_{ij}^{\,y_{ij}}}{y_{ij}!}\exp\left[-\frac{\gamma_i^2}{2}\right]d\gamma_i$$

and C_{P4} also a constant.

The integral is again approximated as the sum of K weighted function evaluations at nodes γ_k, with weights w_k:

$$Q\left(\underline{\beta}|\underline{y},\hat{\underline{\beta}}\right) \approx \sum_{k=1}^{K}\sum_{i=1}^{I}\sum_{j=1}^{J(i)}\left[y_{ij}\ln\lambda_{jk} - \lambda_{jk}\right] f_{\Gamma|\underline{y}_i}\left(\gamma_k|\underline{y}_i,\hat{\underline{\beta}}\right)w_k + C_{P4} \qquad (10.4)$$

As in Section 6.2.4 the conditional posterior probability of discrete variable γ_k is denoted P_{ik} where

$$P_{ik} = f_{\Gamma|\underline{y}_i}\left(\gamma_k|\underline{y}_i,\hat{\underline{\beta}}\right)w_k$$

As in the binary case the expected complete data log likelihood in (10.4) is in effect a discrete approximation to the posterior expectation with masses P_{ik} at nodes γ_k. It can be written

$$Q\left(\underline{\beta}|\underline{y},\hat{\underline{\beta}}\right)=\sum_{i=1}^{I}\sum_{j=1}^{J(i)}\sum_{k=1}^{K}\left[y_{ij}\ln\lambda_{jk}-\lambda_{jk}\right]P_{ik}+C_{P4}$$

$$=\sum_{k=1}^{K}\sum_{j=1}^{J}\left[\sum_{i=1}^{I(j)}y_{ij}P_{ik}\ln\lambda_{jk}-P_{ik}\lambda_{jk}\right]+C_{P4}$$

Summing over i gives

$$Q\left(\underline{\beta}|\underline{y},\hat{\underline{\beta}}\right)=\sum_{k=1}^{K}\sum_{j=1}^{J}\left[U_{jk}\ln\lambda_{jk}-N_{jk}\lambda_{jk}\right]+C_{P4} \qquad (10.5)$$

where $N_{jk}=\sum_{i=1}^{I(j)}P_{ik}$ is interpreted as the expected number of observations of type j dependent upon latent effect γ_k and $U_{jk}=\sum_{i=1}^{I(j)}y_{ij}P_{ik}$ as the expected total count over all observations of type j dependent on latent effect γ_k. For example, N_{jk} might be the expected number of counts made on day j conditional on γ_k and U_{jk} the expected value of the total of all the counts made on day j conditional on γ_k.

If the U_{jk} were Poisson variables with parameters $N_{jk}\lambda_{jk}$, $\underline{U}$ would have, by comparison with equation (10.1), a complete data log likelihood function of the form

$$l_{\underline{U},\underline{\gamma}}\left(\underline{\lambda}|\underline{U},\underline{\gamma}\right)=\sum_{k=1}^{K}\sum_{j=1}^{J}\left(U_{jk}\ln\left(N_{jk}\lambda_{jk}\right)-N_{jk}\lambda_{jk}\right)+C_{P2'}$$

This can be written

$$l_{\underline{U},\underline{\gamma}}\left(\underline{\lambda}|\underline{U},\underline{\gamma}\right)=\sum_{k=1}^{K}\sum_{j=1}^{J}\left(U_{jk}\ln\lambda_{jk}-N_{jk}\lambda_{jk}\right)+\sum_{k=1}^{K}\sum_{j=1}^{J}U_{jk}\ln N_{jk}+C_{P2'} \qquad (10.6)$$

The expression $\sum_{k=1}^{K}\sum_{j=1}^{J}U_{jk}\ln N_{jk}$ which is not dependent on $\underline{\beta}$ can be absorbed into the constant and does not effect the maximum likelihood parameter estimates. By comparing (10.5) and (10.6) it can be seen that the expected complete data likelihood function (10.5) has the form of a log likelihood function of a GLM with responses $U_{jk}\sim Po\left(N_{jk}\lambda_{jk}\right)$. The link

function and linear predictor are found by replacing γ_i by γ_k in the GLM for the complete data and changing the i subscript to k. Thus the link function is

$$N_{jk}\lambda_{jk} = \exp(\eta_{jk}(\gamma_k))$$

and the linear predictor associated with observation U_{jk} is

$$\eta_{jk} = \underline{x}_j^T \underline{\varphi}_j + \alpha_j \gamma_k$$

As in the binary response example maximising the expected complete data log likelihood is equivalent to fitting the GLM whose log likelihood function is equation (10.6) to the expected complete data $\underline{U}$ and $\underline{N}$.

10.3. A MODEL FOR NORMAL RESPONSES.

In this section a similar analysis is applied to continuous normal response data where a latent variable is thought to enter into the model. An expression for the expected complete data log likelihood is derived for theoretical interest. The question of whether this is a reasonable model to use in approaching the problem of ML estimation in such a situation is not discussed.

10.3.1. A Latent Variable GLM for Normal Responses.

All the assumptions made in Sections 6.2.1 and 10.2.1 continue to hold in this section. The data again has a nested structure with J observations available on each of I units. The nested structure of the data, responses y_{ij} $(j=1,2,...,J)$, $(i=1,2,...,I)$, $I(j)$ and $J(i)$, vector $\underline{\beta}_j$ and latent covariate γ_i, a realisation of Γ_i are all as previously defined. However this time the Y_{ij}, conditional on the γ_i, are continuous normal variables with, it is assumed at first, constant variance:

$$Y_{ij}|\gamma_i \sim N\left(\mu_{ij}(\gamma_i), \sigma^2\right)$$

In a GLM the canonical link function for the normal distribution is the identity function

$$\eta_{ij}(\gamma_i) = \mu_{ij}(\gamma_i)$$

with the linear predictor as before

$$\eta_{ij} = \underline{x}_j^T \underline{\beta}_j + \alpha_j \gamma_i$$

where $\underline{x}_j^T$, $\underline{\beta}_j$ and α_j are all as in both the previous examples. It is also assumed that

$\Gamma_i \sim N(0,1)$.

For $Y_{ij}|\gamma_i \sim N\left(\mu_{ij}, \sigma^2\right)$, the log likelihood is, assuming independent observations conditional on $\underline{\gamma}$,

$$l_{\underline{y},\underline{\gamma}}\left(\underline{\mu}|\underline{y},\underline{\gamma}\right) = \frac{1}{\sigma^2}\sum_{i=1}^{I}\sum_{j=1}^{J(i)}\left(y_{ij}\mu_{ij} - \frac{\mu_{ij}^2}{2}\right) + C_{N2} \qquad (10.7)$$

where C_{N2} is a constant.

Function (10.7) is therefore the conventional likelihood function for a normal response GLM which, if $\underline{\gamma}$ were known, could be maximised for $\underline{\beta}$ with the standard software. If this is GLIM and the model has known constant variance σ^2 there is a facility to assign this value to the scale parameter (Payne, 1987). If the variance is unknown it can be estimated from the deviance (Payne, 1987).

10.3.2. The Expected Complete Data Likelihood.

For the complete data likelihood logarithms of the joint p.d.f. of $\underline{Y}$ and $\underline{\Gamma}$ are taken:

$$l_{\underline{y},\underline{\gamma}}(\underline{\beta}|\underline{y},\underline{\gamma}) = I \ln\frac{1}{\sqrt{2\pi}} + \sum_{i=1}^{I}\sum_{j=1}^{J(i)}\frac{1}{\sigma^2}\left(y_{ij}\mu_{ij} - \frac{\mu_{ij}^2}{2}\right) + C_{N2} - \sum_{i=1}^{I}\left[\frac{\gamma_i^2}{2}\right] \qquad (10.8)$$

Expectations of function (10.8) are taken over the posterior distribution of $\underline{\Gamma}$ given the data

and parameter estimates:

$$Q\left(\underline{\beta}|\underline{y},\hat{\underline{\beta}}\right)=\int_{R}\sum_{i=1}^{I}\sum_{j=1}^{J(i)}\frac{1}{\sigma^2}\left(y_{ij}\mu_{ij}-\frac{\mu_{ij}^{\ 2}}{2}\right)f_{\Gamma|\underline{y}_i}\left(\gamma_i|\underline{y}_i,\hat{\underline{\beta}}\right)d\gamma_i+C_{N4} \qquad (10.9)$$

where

$$f_{\Gamma|\underline{y}_i}\left(\gamma_i|\underline{y}_i,\hat{\underline{\beta}}\right)=C_{N3}^{\ -1}\prod_{i=1}^{I}\prod_{j=1}^{J(i)}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(y_{ij}-\mu_{ij})^2}{2\sigma^2}\right]\exp\left[-\frac{\gamma_i^{\ 2}}{2}\right]$$

with normalising constant

$$C_{N3}\approx\sum_{k=1}^{K}\prod_{i=1}^{I}\prod_{j=1}^{J(i)}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(y_{ij}-\mu_{jk})^2}{2\sigma^2}\right]\exp\left[-\frac{\gamma_k^{\ 2}}{2}\right]w_k\,.$$

Once again approximating the integral in (10.9) as the sum of K function evaluations at nodes γ_k, weighted by w_k:

$$Q\left(\underline{\beta}|\underline{y},\hat{\underline{\beta}}\right)\approx\frac{1}{\sigma^2}\sum_{k=1}^{K}\sum_{i=1}^{I}\sum_{j=1}^{J(i)}\left(y_{ij}\mu_{jk}-\frac{\mu_{jk}^{\ 2}}{2}\right)f_{\Gamma|\underline{y}_i}\left(\gamma_k|\underline{y}_i,\hat{\underline{\beta}}\right)w_k+C_{N4} \qquad (10.10)$$

It is now possible to replace the continuous conditional posterior distribution of the latent variable by the discrete posterior probabilities P_{ik} where

$$P_{ik}=f_{\Gamma|\underline{y}_i}\left(\gamma_k|\underline{y}_i,\hat{\underline{\beta}}\right)w_k$$

so that

$$\begin{aligned}Q\left(\underline{\beta}|\underline{y},\hat{\underline{\beta}}\right)&=\frac{1}{\sigma^2}\sum_{i=1}^{I}\sum_{j=1}^{J(i)}\sum_{k=1}^{K}\left(y_{ij}\mu_{jk}-\frac{\mu_{jk}^{\ 2}}{2}\right)P_{ik}+C_{N4}\\&=\frac{1}{\sigma^2}\sum_{k=1}^{K}\sum_{j=1}^{J}\left[\sum_{i=1}^{I(j)}\left(y_{ij}P_{ik}\mu_{jk}-\frac{P_{ik}\mu_{jk}^{\ 2}}{2}\right)\right]+C_{N4}\end{aligned}$$

Summing over i gives

$$Q\left(\underline{\beta}|\underline{y},\hat{\underline{\beta}}\right)=\frac{1}{\sigma^2}\sum_{k=1}^{K}\sum_{j=1}^{J}\left[U_{jk}\mu_{jk}-N_{jk}\frac{\mu_{jk}^{\ 2}}{2}\right]+C_{N4} \qquad (10.11)$$

The interpretations of U_{jk} and N_{jk} are similar to the binary and Poisson cases. $N_{jk} = \sum_{i=1}^{I(j)} P_{ik}$ is the expected number of observations of type j dependent upon latent effect γ_k and $U_{jk} = \sum_{i=1}^{I(j)} y_{ij} P_{ik}$ is the expected sum of all observations j dependent on latent effect γ_k. For example, N_{jk} might be the expected number of measurements made under experimental conditions j conditional on γ_k and U_{jk} the expected value of the sum of all the measurements made under experimental conditions j conditional on γ_k.

It is now proposed that U_{jk} is a hypothetical random variable with $U_{jk} \sim N\left(N_{jk}\mu_{jk}, N_{jk}\sigma^2\right)$. The vector $\underline{U}$ has by comparison with equation (10.7), a complete data log likelihood function of the form

$$l_{\underline{U},\underline{\gamma}}\left(\underline{\mu}|\underline{U},\underline{\gamma}\right) = \frac{1}{\sigma^2}\sum_{k=1}^{K}\sum_{j=1}^{J} N_{jk}^{-1}\left[U_{jk}N_{jk}\mu_{jk} - \frac{\left(N_{jk}\mu_{jk}\right)^2}{2}\right] + C_{N2'}$$

which is the same as

$$l_{\underline{U},\underline{\gamma}}\left(\underline{\mu}|\underline{U},\underline{\gamma}\right) = \frac{1}{\sigma^2}\sum_{i=1}^{I}\sum_{j=1}^{J}\left(U_{jk}\mu_{jk} - N_{jk}\frac{\mu_{jk}^2}{2}\right) + C_{N2'} \qquad (10.12)$$

Apart from the difference in the constant terms (10.11) and (10.12) are the same. That is the expected complete data likelihood function (10.11) has the form of a log likelihood function of a GLM with responses $U_{jk} \sim N\left(N_{jk}\mu_{jk}, N_{jk}\sigma^2\right)$ where the link function is

$$N_{jk}\mu_{jk} = \eta_{jk}(\gamma_k)$$

and the linear predictor associated with observation U_{jk} is

$$\eta_{jk} = \underline{x}_j^T \underline{\beta}_j + \alpha_j\gamma_k$$

The variances are specified by declaring σ^2 as the value of the scale parameter if σ^2 is known, or estimating it from the model deviance if it is unknown. To create variances of the form $\frac{\sigma^2}{w_{jk}}$ where $w_{jk} = N_{jk}^{-1}$ a vector of reciprocals of N_{jk} must also be declared as prior weights (Payne, 1987). As in the binary and Poisson response examples, maximising the expected complete data log likelihood is equivalent to fitting the GLM whose log likelihood function is equation (10.12) to the expected complete data $\underline{U}$ and $\underline{N}$.

10.3.3. Normal Models with Non-Constant Variance.

More generally the distribution of the random variable from Section 10.3.1. can be written

$$Y_{ij}|\gamma_i \sim N\left(\mu_{ij}(\gamma_i), \frac{\sigma^2}{w_{ij}}\right)$$

where w_{ij} are known prior weights. The special case when all the w_{ij} are equal to 1 is discussed above in Sections 10.3.1 and 10.3.2. The likelihood function equivalent to (10.7) is of the form

$$l_{\underline{y},\underline{\gamma}}\left(\underline{\mu}|\underline{y},\underline{\gamma}\right) = \frac{1}{\sigma^2}\sum_{i=1}^{I}\sum_{j=1}^{J(i)} w_{ij}\left(y_{ij}\mu_{ij} - \frac{\mu_{ij}^2}{2}\right) + C_{\tilde{N}2} \qquad (10.13)$$

where $C_{\tilde{N}2}$ is a constant.

After approximating the integral and replacing the continuous posterior distribution of the latent variable by its discrete form, the expected complete data log likelihood becomes

$$Q\left(\underline{\beta}|\underline{y},\hat{\underline{\beta}}\right) \approx \frac{1}{\sigma^2}\sum_{i=1}^{I}\sum_{j=1}^{J(i)}\sum_{k=1}^{K} w_{ij}\left(y_{ij}\mu_{jk} - \frac{\mu_{jk}^{\,2}}{2}\right)P_{ik} + C_{\tilde{N}4}$$

$$= \frac{1}{\sigma^2}\sum_{k=1}^{K}\sum_{j=1}^{J}\left[\sum_{i=1}^{I(j)}\left(w_{ij}y_{ij}P_{ik}\mu_{jk} - \frac{w_{ij}P_{ik}\mu_{jk}^{\,2}}{2}\right)\right] + C_{\tilde{N}4}$$

Summing over i gives a variable term

$$\frac{1}{\sigma^2}\sum_{k=1}^{K}\sum_{j=1}^{J}\left[U_{jk}\mu_{jk} - N_{jk}\frac{\mu_{jk}^{\,2}}{2}\right] \qquad (10.14)$$

and the expected complete data is $N_{jk} = \sum_{i=1}^{I(j)} w_{ij}P_{ik}$ and $U_{jk} = \sum_{i=1}^{I(j)} y_{ij}w_{ij}P_{ik}$. Thus the sums of discrete posterior probabilities which constituted the expected complete data in the previous examples are in this case weighted by the w_{ij}. It is easily seen that (10.14) is the variable part of (10.12) which is the log likelihood function of $U_{jk} \sim N\left(N_{jk}\mu_{jk}, N_{jk}\sigma^2\right)$. Thus the model for the expected complete data is fitted with prior weights $N_{jk}^{-1} = \left[\sum_{i=1}^{I(j)} w_{ij}P_{ik}\right]^{-1}$.

10.4. THE GENERAL EXPONENTIAL MODEL.

In this section it is shown that under certain conditions the expected complete data log likelihood function can be derived for a latent variable GLM with response data from any exponential family distribution and that this function always has the form of the log likelihood function of a different but related GLM which can be fitted with the standard IRLS algorithm.

10.4.1. The Latent Variable GLM for Exponential Responses.

Once again y_{ij} represents the j^{th}, (j=1,2,...,J), observation on the i^{th} unit/subject, (i=1,2,...,I). The random variable Y_{ij} is assumed to have a distribution from the exponential family which can be written in the form of equation (3.1)

$$f_Y\left(y_{ij};\theta_{ij},\phi\right)=\exp\left\{\left(y_{ij}\theta_{ij}-b(\theta_{ij})\right)/a_{ij}(\phi)+c(y_{ij},\phi)\right\} \qquad (10.15)$$

with canonical parameter θ_{ij} and $E(Y_{ij})=b'(\theta_{ij})$. The J-vector of observations on unit i, $\underline{y}_i$, is associated with latent covariate γ_i and its expected value is dependent on γ_i and unknown parameter vector $\underline{\beta}_j$ only. In this model no parameters are indexed by i, and the units are differentiated only by the value of the latent effect. The link function and linear predictor are expressed in their general forms in Chapter 3, Section 3.3, equations (3.3) and (3.5).

In the equivalent GLM with known $\underline{\gamma}$, the log likelihood for $\underline{\theta}$, (with ϕ and y_{ij} known) is, assuming independent observations conditional on $\underline{\gamma}$,

$$l_{\underline{y},\underline{\gamma}}\left(\underline{\theta}|\underline{y},\underline{\gamma},\phi\right)=\sum_{i=1}^{I}\sum_{j=1}^{J(i)}\left(y_{ij}\theta_{ij}-b(\theta_{ij})\right)/a_{ij}(\phi)+C_{E1} \qquad (10.16)$$

with C_{E1} a constant.

10.4.2. The Generalized Expected Complete Data Log Likelihood.

To derive the expected complete data log likelihood without distributional assumptions let the probability function of the vector response variable $\underline{Y}$ conditional on item parameters $\underline{\beta}$ and latent vector $\underline{\gamma}$ be $f_{\underline{Y}|\underline{\gamma}}\left(\underline{y}|\underline{\gamma},\underline{\beta}\right)$ and the p.d.f. of $\underline{\Gamma}$ the latent variable vector be $f_{\underline{\Gamma}}\left(\underline{\gamma}\right)$. Assuming independent responses conditional on the latent variables, independent latent

variables, and taking natural logarithms of the joint distribution gives a complete data likelihood of the form

$$l_{\underline{y},\underline{\gamma}}(\underline{\beta}|\underline{y},\underline{\gamma}) = \sum_{i=1}^{I}\sum_{j=1}^{J(i)}\left[\ln f_{\underline{Y}|\gamma}\left(y_{ij}|\gamma_i,\underline{\beta}_j\right)\right]+\sum_{i=1}^{I}\ln f_{\Gamma}(\gamma_i) \qquad (10.17)$$

The function Q, the expected complete data log likelihood, is found by taking expectations of (10.17) over the posterior distribution of $\underline{\Gamma}$ given the data and parameter estimates.

$$Q\left(\underline{\beta}|\underline{y},\hat{\underline{\beta}}\right) = \int_R \sum_{i=1}^{I}\sum_{j=1}^{J(i)}\left[\ln f_{\underline{Y}|\gamma}\left(y_{ij}|\gamma_i,\underline{\beta}_j\right)\right] f_{\Gamma|\underline{y}_i}\left(\gamma_i|\underline{y}_i,\hat{\underline{\beta}}\right)d\gamma_i + C_{E4}$$

where C_{E4} is constant and the conditional posterior distribution of $\underline{\Gamma}$ is

$$f_{\Gamma|\underline{y}_i}\left(\gamma_i|\underline{y}_i,\hat{\underline{\beta}}\right) = \frac{f_{\underline{Y}_i|\gamma}\left(\underline{y}_i|\gamma_i,\hat{\underline{\beta}}\right)f_{\underline{\Gamma}}(\gamma_i)}{\int_R f_{\underline{Y}_i|\gamma}\left(\underline{y}|\gamma_i,\hat{\underline{\beta}}\right)f_{\underline{\Gamma}}(\gamma_i)d\gamma_i}$$

The integration is then approximated by a weighted sum of function evaluations at nodes γ_k. As a result the value range of the continuous latent variable Γ_i is replaced by a set of discrete nodes indexed by k, (k=1,2,...,K), so function Q becomes

$$Q\left(\underline{\beta}|\underline{y},\hat{\underline{\beta}}\right) = \sum_{i=1}^{I}\sum_{j=1}^{J(i)}\sum_{k=1}^{K}\ln f_{\underline{Y}|\gamma}\left(y_{ij}|\gamma_k,\underline{\beta}_j\right) f_{\Gamma|\underline{y}_i}\left(\gamma_k|\underline{y}_i,\hat{\underline{\beta}}\right)w_k + C_{E4} \qquad (10.18)$$

where

$$f_{\Gamma|\underline{y}_i}\left(\gamma_k|\underline{y}_i,\hat{\underline{\beta}}\right) \approx \frac{f_{\underline{Y}_i|\gamma}\left(\underline{y}_i|\gamma_k,\hat{\underline{\beta}}\right)f_{\Gamma}(\gamma_k)}{\sum_{k=1}^{K} f_{\underline{Y}_i|\gamma}\left(\underline{y}_i|\gamma_k,\hat{\underline{\beta}}\right)f_{\Gamma}(\gamma_k)w_k} \qquad (10.19)$$

Without making any assumptions about the distributions of either the response variable or the latent variable, a discrete posterior conditional probability distribution for γ_k is defined by mass points P_{ik} where

$$P_{ik} = \frac{f_{\underline{Y}_i|\gamma}\left(\underline{y}_i|\gamma_k, \underline{\hat{\beta}}\right) f_{\Gamma}(\gamma_k) w_k}{\sum_{k=1}^{K} f_{\underline{Y}_i|\gamma}\left(\underline{y}_i|\gamma_k, \underline{\hat{\beta}}\right) f_{\Gamma}(\gamma_k) w_k}$$

Substituting in (10.19) gives

$$P_{ik} = f_{\Gamma|\underline{y}_i}\left(\gamma_k|\underline{y}_i, \underline{\hat{\beta}}\right) w_k \tag{10.20}$$

with the term P_{ik} denoting the posterior probability that response vector $\underline{y}_i$ depends on γ_k given parameters $\underline{\hat{\beta}}$. Substituting (10.20) in (10.18) we obtain

$$Q\left(\underline{\beta}|\underline{y}, \underline{\hat{\beta}}\right) = \sum_{i=1}^{I} \sum_{j=1}^{J(i)} \sum_{k=1}^{K} \ln f_{Y|\gamma}\left(y_{ij}|\gamma_i, \underline{\beta}_j\right) P_{ik} + C_{E4}$$

If the assumption that the distribution of Y_{ij} is from the exponential family is now made, then the expected complete data log likelihood becomes

$$Q\left(\underline{\beta}|\underline{y}, \underline{\hat{\beta}}\right) = \sum_{i=1}^{I} \sum_{j=1}^{J(i)} \sum_{k=1}^{K} \left[\frac{y_{ij}\theta_{jk} - b(\theta_{jk})}{a_{ij}(\phi)} + c(y_{ij}, \phi) \right] P_{ik} + C_{E4} \tag{10.21}$$

Rewriting and summing over i, we obtain

$$Q\left(\underline{\beta}|\underline{y}, \underline{\hat{\beta}}\right) = \sum_{j=1}^{J} \sum_{k=1}^{K} \left[\theta_{jk} \sum_{i=1}^{I(j)} \left[a_{ij}(\phi)\right]^{-1} y_{ij} P_{ik} - b(\theta_{jk}) \sum_{i=1}^{I(j)} \left[a_{ij}(\phi)\right]^{-1} P_{ik} \right] + C_{E5}$$

or

$$Q\left(\underline{\beta}|\underline{y}, \underline{\hat{\beta}}\right) = \sum_{j=1}^{J} \sum_{k=1}^{K} \left[\theta_{jk} U_{jk} - b(\theta_{jk}) N_{jk}\right] + C_{E5} \tag{10.22}$$

where

$$U_{jk} = \sum_{i=1}^{I(j)} \left[a_{ij}(\phi)\right]^{-1} y_{ij} P_{ik}$$

and

$$N_{jk} = \sum_{i=1}^{I(j)} \left[a_{ij}(\phi)\right]^{-1} P_{ik}$$

As previously seen the U_{jk} can be interpreted as the expected total response for all observations of type j conditional on γ_k. For example for binomial data it is a total number of successful outcomes and for Poisson data it is a total count. The N_{jk} are the expected number of responses of type j conditional on latent effect γ_k.

A vector $\underline{V} = \left(V_{11}, V_{21}, \ldots, V_{J1}, V_{12}, V_{22}, \ldots, V_{J2}, \ldots, V_{1K}, V_{2K}, \ldots, V_{JK}\right)^T$ in which $V_{jk} = U_{jk} N_{jk}^{-1}$ can be formed from the expected complete data and interpreted as the vector of mean expected responses for observations of type j conditional on γ_k. These data can be treated as if they were realisations of random variables from exponential family distributions with canonical parameters θ_{jk}. This is essentially the same parameter as in (10.15). It is now dependent on fixed γ_k but defines the same relationship between the mean and the linear predictor as in the distribution of the response data. In addition the values N_{jk}^{-1} are assigned to the functions $a_{jk}(\phi)$; that is the scale parameter is set to 1 and the N_{jk} are set up as prior weights (Payne, 1987). Using a standard generalized linear model for $\underline{V}$ under these conditions, we obtain the log likelihood function based on equation (10.16) as

$$l_{\underline{V},\underline{\gamma}}\left(\underline{\theta}|\underline{V},\underline{\gamma}\right) = \sum_{j=1}^{J}\sum_{k=1}^{K}\left[\frac{U_{jk}N_{jk}^{-1}\theta_{jk} - b(\theta_{jk})}{N_{jk}^{-1}}\right] + C_{E1},$$

This can be rewritten as

$$l_{\underline{V},\underline{\gamma}}\left(\underline{\theta}|\underline{V},\underline{\gamma}\right) = \sum_{j=1}^{J}\sum_{k=1}^{K}\left[U_{jk}\theta_{jk} - N_{jk}b(\theta_{jk})\right] + C_{E1},$$

which is the likelihood function in (10.22) up to the constant term.

10.4.3. Summary.

During the E-Step of the EM algorithm a vector $\underline{V}$ of expected complete data is computed. Component V_{jk} of the vector is the mean expected response for observations of type j dependent on covariate γ_k. These data values are calculated using a discrete posterior probability distribution of the latent variable conditional on the data and current parameter estimates. This continuous distribution is approximated by a set of masses P_{ik} at nodes γ_k. An updated estimate of parameter vector $\underline{\beta}$ is calculated during the M-step by maximising the log likelihood of a fixed effects GLM for expected data $\underline{V}$. This GLM has the same error distribution, canonical parameter and linear predictor as the original latent variable GLM. Where unknown values of the latent variable (indexed by i) appeared in the latent variable model the fixed effects model has discrete known covariates γ_k. In addition the fixed effects model requires prior weights equivalent to the expected number of responses of type j which are conditional on γ_k i.e. on ${N_{jk}}^{-1}$ where $N_{jk} = \sum_{i=1}^{I(j)} \left[a_{ij}(\phi) \right]^{-1} P_{ik}$.

APPENDIX A

The GLM and the Exponential Family of Distributions

Under the GLM the distribution of the random response variable Y_i is restricted to the exponential family and can be expressed in the following form:

$$f_Y(y_i;\theta_i,\phi) = \exp\{(y_i\theta_i - b(\theta_i))/a_i(\phi) + c(y_i,\phi)\} \qquad (A.1)$$

where $a_i(.)$, $b(.)$ and $c(.)$ are specific functions, θ_i is known as the canonical parameter, and ϕ is a known scale parameter constant over observation $\underline{y}$.

Now the log likelihood for θ_i, with ϕ and y_i known is

$$l_Y(\theta_i;y_i,\phi) = (y_i\theta_i - b(\theta_i))/a_i(\phi) + c(y_i,\phi) \qquad (A.2)$$

$$\text{So} \quad \frac{\partial l}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a_i(\phi)} \qquad (A.3)$$

$$\text{Now} \quad E\left(\frac{\partial l}{\partial \theta_i}\right) = 0 *$$

$$\Rightarrow \quad E\left(\frac{Y_i - b'(\theta_i)}{a_i(\phi)}\right) = 0$$

$$\Rightarrow \quad \frac{\mu_i - b'(\theta_i)}{a_i(\phi)} = 0$$

$$\Rightarrow \quad \mu_i = b'(\theta_i)$$

From (A.3)

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

$$\text{Now} \quad -E\left(\frac{\partial^2 l}{\partial \theta_i^2}\right) = E\left[\left(\frac{\partial l}{\partial \theta_i}\right)^2\right] *$$

$$\Rightarrow \qquad \frac{b''(\theta)}{a_i(\phi)} = E\left[\left(\frac{Y_i - b'(\theta_i)}{a_i(\phi)}\right)^2\right]$$

$$= \frac{E\left[\left(Y_i - \mu_i\right)^2\right]}{\left(a_i(\phi)\right)^2}$$

$$\Rightarrow \qquad Var(Y_i) = b''(\theta_i)a_i(\phi)$$

* The proof of these well-known results can be found, for example, in Dobson (1990), Appendix A.

Example A1. Normal distribution:

Let $Y_i \sim N(\mu_i, \sigma^2)$. Then

$$f_Y(y_i; \mu_i, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_i - \mu_i)^2}{2\sigma^2}\right]$$

This can be written as

$$f_Y(y_i; \mu_i, \sigma^2) = \exp\left[\frac{y_i\mu_i - \mu_i^2/2}{\sigma^2} - \frac{1}{2}\left(\frac{y_i^2}{\sigma^2} + \ln 2\pi\sigma^2\right)\right]$$

Comparing this with (A.1):

$$\theta_i = \mu_i$$

$$\phi = \sigma^2$$

$$b(\theta_i) = \frac{\mu_i^2}{2} = \frac{\theta_i^2}{2}$$

$$c(y_i, \phi) = -\frac{1}{2}\left[\frac{y_i^2}{\phi} + \ln 2\pi\phi\right]$$

Therefore,

$$E(Y_i) = b'(\theta_i) = \theta_i = \mu_i$$

$$Var(Y_i) = b''(\theta_i)a_i(\phi) = \sigma^2$$

Example A2. Binomial Distribution:

1. In this treatment the random variable Y_i is the proportion and Y_i^* the number of successes out of n_i trials. Let $Y_i = \frac{Y_i^*}{n_i}$ where $Y_i^* \sim Bi(n_i, \pi_i)$. Then

$$p_Y(y_i; n_i, \pi_i) = \binom{n_i}{n_i y_i} \pi_i^{n_i y_i} (1-\pi_i)^{n_i - n_i y_i}$$

This can be written as

$$p_Y(y_i; n_i, \pi_i) = \exp\left[n_i y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right) + n_i \ln(1-\pi_i) + \ln\binom{n_i}{n_i y_i} \right]$$

Comparing this with (A.1):

$$\theta_i = \ln \frac{\pi_i}{1-\pi_i}$$

$$\phi = 1, \; a_i(\phi) = \frac{1}{n_i}$$

$$b(\theta_i) = -\ln(1-\pi_i) = \ln(1+e^{\theta_i})$$

$$c(y_i, \phi) = \ln\binom{n_i}{n_i y_i}$$

Therefore

$$E(Y_i) = b'(\theta_i) = \frac{1}{1+e^{-\theta_i}} = \pi_i$$

$$Var(Y_i) = b''(\theta_i) a_i(\phi) = n_i \pi_i (1-\pi_i)$$

Example A2. Poisson Distribution:

Let $Y_i \sim Po(\lambda_i)$ Then

$$p_Y(y_i; \lambda_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}$$

This can be written as

$$p_Y(y_i;\lambda_i) = \exp[y_i \ln \lambda_i - \lambda_i - \ln(y_i!)]$$

Comparing this with (A.1):

$$\theta_i = \ln \lambda_i$$

$$\phi = 1$$

$$b(\theta_i) = e^{\theta_i}$$

$$c(y_i, \phi) = -\ln(y_i!)$$

Therefore

$$E(Y_i) = b'(\theta_i) = e^{\theta_i} = \lambda_i$$

$$Var(Y_i) = b''(\theta_i)a_i(\phi) = e^{\theta_i} = \lambda_i$$

Other members of the exponential family are the negative binomial, the gamma and the inverse Gaussian distributions.

APPENDIX B

MODELS IN ITEM RESPONSE THEORY

LINK FUNCTION OF THREE-PARAMETER MODEL.

In the three-parameter logistic model the probability of subject *i* responding correctly to item *j* is

$$\pi_{ij} = c_j + \frac{1-c_j}{1+e^{-\eta_{ij}}} \qquad \text{(B.1)}$$

where $\eta_{ij} = a_j\left(\gamma_i - b_j\right)$

Re-arranging (B.1) $\Rightarrow$ $(\pi_{ij} - c_j)(1+e^{-\eta_{ij}}) = 1-c_j$

$\Rightarrow$ $\pi_{ij} - c_j + \pi_{ij}e^{-\eta_{ij}} - c_j e^{-\eta_{ij}} = 1-c_j$

$\Rightarrow$ $\pi_{ij} + e^{-\eta_{ij}}(\pi_{ij} - c_j) = 1$

$\Rightarrow$ $e^{-\eta_{ij}} = (1-\pi_{ij})(\pi_{ij} - c_j)^{-1}$

$\Rightarrow$ $-\eta_j = \ln(1-\pi_{ij})(\pi_{ij} - c_j)^{-1}$

$\Rightarrow$ $\underline{\eta_{ij} = \ln\frac{\pi_{ij} - c_j}{1-\pi_{ij}}}$

APPENDIX C

Exponential Family

Score Vector and Information Matrix

The log likelihood of canonical parameter θ_i given response y_i a single realisation of the random response variable Y_i with a probability distribution from the exponential family is

$$l_i(\theta_i) = \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} - c(y_i, \phi) \qquad \text{(B.1)}$$

Also

$$\mu_i = b'(\theta_i) \qquad \text{(B.2)}$$

$$Var(Y_i) = b''(\theta_i)a(\phi) \qquad \text{(B.3)}$$

The jth element of the score vector $\underline{u}$ is

$$u_j(\underline{\beta}) = \sum_i \frac{\partial l_i(\theta_i)}{\partial \beta_j} = \sum_i \frac{\partial l_i}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_j}$$

The object of the following is to find

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_j} \qquad \text{(B.4)}$$

(i) From (B.1)

$$\frac{\partial l_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)} = \frac{y_i - \mu_i}{a(\phi)} \qquad \text{(B.5)}$$

(ii) From (B.2)

$$\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i)$$

$$\Rightarrow \quad \frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{b''(\theta_i)} \tag{B.6}$$

(iii) $$\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$$

$$\Rightarrow \quad \frac{\partial \eta_i}{\partial \beta_j} = x_{ij} \tag{B.7}$$

Substituting (B.5), (B.6) and (B.7) in (B.4)

$$\frac{\partial l_i}{\partial \beta_j} = \frac{(y_i - \mu_i)}{a(\phi)} \cdot \frac{1}{b''(\theta_i)} \frac{\partial \mu_i}{\partial \eta_i} \cdot x_{ij}$$

Using (B.3) this becomes

$$\frac{\partial l_i}{\partial \beta_j} = \frac{(y_i - \mu_i)x_{ij}}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \tag{B.8}$$

Therefore

$$\underline{u_j(\underline{\beta}) = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)}$$

To find the elements of the information matrix I let weights w_i be

$$w_i = \frac{1}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2$$

Substituting in (B.8),

$$\frac{\partial l_i}{\partial \beta_j} = (y_i - \mu_i)x_{ij} w_i \left(\frac{\partial \eta_i}{\partial \mu_i} \right)$$

The information matrix is

$$\mathrm{I} = E\left[-\frac{\partial}{d\beta_j \partial\beta_k}\right]$$

The jk^{th} element of the information matrix is therefore

$$\mathrm{I}_{jk} = E\left[-\frac{\partial}{\partial\beta_k}\left\{\sum_{i=1}^{n}(y_i-\mu_i)x_{ij}w_i\left(\frac{\partial\eta_i}{\partial\mu_i}\right)\right\}\right]$$

$$\Rightarrow \qquad \mathrm{I}_{jk} = -E\left[\sum_{i=1}^{n}\frac{\partial}{\partial\beta_k}\left\{(y_i-\mu_i)x_{ij}w_i\left(\frac{\partial\eta_i}{\partial\mu_i}\right)\right\}\right]$$

Using the product rule

$$\mathrm{I}_{jk} = -E\left[\sum_{i=1}^{n}(y_i-\mu_i)\frac{\partial}{\partial\beta_k}\left(x_{ij}w_i\left(\frac{\partial\eta_i}{\partial\mu_i}\right)\right)+\sum_{i=1}^{n}x_{ij}w_i\left(\frac{\partial\eta_i}{\partial\mu_i}\right)\frac{\partial}{\partial\beta_k}(y_i-\mu_i)\right]$$

Since $E(y_i-\mu_i)=0$ the first term in this expression is 0, and since $\frac{\partial y_i}{\partial\beta_k}=0$ the second term reduces to

$$-E\left[\sum_{i=1}^{n}x_{ij}w_i\left(\frac{\partial\eta_i}{\partial\mu_i}\right)\frac{\partial}{\partial\beta_k}(-\mu_i)\right]$$

$$\Rightarrow \mathrm{I}_{jk} = \sum_{i=1}^{n}x_{ij}w_i\left(\frac{\partial\eta_i}{\partial\mu_i}\right)\frac{\partial\mu_i}{\partial\beta_k}$$

Using (B.7) we have that

$$\frac{\partial\mu_i}{\partial\beta_k} = \frac{\partial\mu_i}{\partial\eta_i}\frac{\partial\eta_i}{\partial\beta_k} = \frac{\partial\mu_i}{\partial\eta_i}x_{ik}$$

$$\Rightarrow I_{jk} = \sum_{i=1}^{n} x_{ij} w_i x_{ik}$$

$$\Rightarrow \underline{I = X^T W X}$$

APPENDIX D

DI. GLIM PROGRAM - GLIRT1.

```
$SUBFILE GLIRT1
$c
$print;'Enter name of file for output of parameter estimates' : $
$output 13 $
$output 2 $
$MACRO MODEL gamma $ENDMAC
$cal %x=0.001 $
$warn
$acc 9 $
$print; 'Use macro NODES then INIT for initial estimates';
'Use macro LOOP to run EM algorithm'; $
$c
$MACRO RUN
$use NODES $
$use INIT $
$use LOOP $
$ENDMAC
$c
$MACRO NEWNODES
$use NODES $
$use LOOP $
$ENDMAC
$c
$MACRO NODES
$c
$c calls subroutine LEGDAT to calculate nodes and weights for
$c integral approximation and stores them in vector V1
$c
$c
$del v1  $
$var 63 p1  $
$print; 'subroutine LEGDAT - calculating standard normal nodes'; $
$pass 1 p1  $
$cal %k=p1(1):      ! no. of nodes
    %n=%k+3 $     ! length of V1
$var %n v1:
    %n sa $
$c store weights and nodes in vector V1
$cal sa = %gl(%n,1):
    v1=p1(sa) $
$del p1 $
$cal %y=1  $        ! ensures PREP is run before MAX
$print; 'You must enter a model specification and a tolerance level (%x)';
 'Current model: ' MODEL '  Tol: ' %x ;
$print 'To specify a new model enter macro MODEL';
 'reset %x if a new tolerance is required'; $
$ENDMAC
$c
```

```
$MACRO INIT
$c
$c calls INIT subroutine to calculate initial estimates for nos.
$c done and nos. correct
$c
$c prepare vector p2
$c
$cal %a=%k*401+205:
    %b=%a-%n $
$var %a p2  $
$var %b p0  $
$cal p0=0  $
$ass p2=v1,p0  $
$del p0  $
$print; 'subroutine INIT - initialisation'; $
$pass 2 p2 $
$cal %s=5+%k:
    %j=p2(%s) $     ! store no. items
$c
$c store guessing parameters in vector gps
$c
$var %j sub:
    %j gps $
$cal sub=%gl(%j,1):
    sub=sub+%k+5:
    gps=p2(sub) $
$c
$use PREP  $
$c
$c store initial estimates in vectors n and y, then move to P3
$c
$cal %t=1:
    %s=%k+%j+6:
    %v=%s+%k*%j:
    %e=%k*%j $
$arg MEXP p2 $
$while %e MEXP  $
$del p2  $
$use MAX  $
$ENDMAC
$c
$MACRO PREP
$c
$c sets up the data required for fitting the model in standard
$c length vectors gamma, g, item, block, diff, tf and wt;
$c vectors n and y are given initial values of 1's and 0's resp.
$c
$del gamma g item block diff time tf wt n y c1 c2 sa y1  $
$c
$print; 'MODEL: ' model  $
$print  'ITEMS:' *i %j 'NODES: ' *i %k ' TOL: ' %x ;  $
$c set no. units
```

```
$c
$cal %u = (%k+2)*%j  $
$units %u $
$c
$c  move nodes to vector gamma
$c
$cal sub = %gl(%j,1):
      %s = 4:
      %e = %k:
     gamma=0 $
$while %e MGAMMA  $
$cal gamma(sub) = 0:
      sub = sub+%j:
      gamma(sub)=1 $
$c
$c move guessing params to data vector g
$c
$cal sub = %gl(%j,1):
      %e = %k+2 $
$cal g=0  $
$while %e MGP  $
$c
$c generate item nos. in vector item
$c
$cal item = %gl(%j,1)  $
$c
$c generate factor levels for block, difficulty and true/false
$c
$cal block = %gl(2,10):
       diff = %gl(5,1) :
        tf = %gl(2,15) $
$c
$c generate values of co-variate 1/time
$c
$cal time = %gl(3,5):
      time = 1/(2*(time+1)) $
$c
$c assign weights to vector wt (= 1 for elements 1 to k*j,
$c                          = 0 for elements k*j+1 to (k+2)*j)
$c
$cal %c = %k*%j:
      %d = %j*2 $
$var %c c1:
      %d c2 $
$cal c1 = 1:
      c2 = 0 $
$ass wt=c1,c2  $
$c
$c initialise expected nos. done and correct, vector n and vector y
$c
$cal n = 1:
      y = 0 $
```

```
$c
$c identify y-variable y, factors item, block,, difficulty and
$c true/false, and weights
$c
$yvar y  $
$factor item %j block 2 diff 5 tf 2  $
$weight wt  $
$cal %z=1:
     %i=0:
     %y=0 $
$c
$ENDMAC
$c
$MACRO MGAMMA
$c
$c moves k nodes to vector gamma in blocks of length j
$c
$cal gamma(sub) = V1(%s):
      sub = sub + %j:
      %s = %s+1:
      %e = %e-1 $
 $ENDMAC
 $c
 $MACRO MGP
 $c
 $c moves j guessing params for 1 to k+2
 $c
 $cal g(sub)=gps:
      sub=sub+%j $
 $cal %e = %e-1  $
 $ENDMAC
 $c
 $MACRO LOOP
 $while %y PREP  $
 $cal %z=1:
      %p=0 $
 $while %z EMALG  $
 $use ENDUP  $
 $ENDMAC
 $c
 $MACRO EMALG
 $use ESTEP  $
 $use CHECK  $
 $use MAX  $
 $ENDMAC
 $c
 $MACRO ESTEP
 $c
 $c passes vector P3 to subroutine ESTEP
 $ass P3=%x,V1,n,y,V2  $
 $print; 'subroutine ESTEP - expectation phase'; $
 $pass 3 P3  $
```

```
$c
$c move expected nos.done and correct from vector P3 to n and y
$c
$cal %s = %k+5:
     %v = %s+(%k+2)*%j:
     %t=1:
     %e=%k*%j $
$arg MEXP p3  $
$while %e MEXP
$c
$ENDMAC
$c
$MACRO CHECK
$c
$c checks for convergence
$c
$cal  %z=%eq(p3(1),%x):
      %p=P3(1) $
$ENDMAC
$c
$MACRO MAX
$c
$c fit current model
$c
$cal %i=%i+1  $
$print ; 'Maximization Step:  Iteration: ' *i %i ;  $
$own fit dir var dev
$scale 1  $
$cal %lp = %if(y>0,%log(y/(n-y+0.5)),-15) $
$cycle 50$
$fit #model  $
$dis e $
$c
$c move intercepts and slope+intercepts to vector V2
$c
$cal %m=%j*2  $
$var %m v2:
     %m sb $
$cal sb=%gl(%m,1):
     sb=sb+%k*%j:
     v2=%lp(sb) $
$ENDMAC
$c
$c macros for model fitting follow:
$c
$MACRO FIT $cal % fv=n*(g+(1-g)/(1+%exp(-%lp)))
$ENDMAC
$c
$MACRO DIR $cal %dr = 1/(%fv-n*g) +1/(n-%fv)
$ENDMAC
$c
$MACRO VAR $cal %va = %fv*(1-%fv/n)$
```

```
$ENDMAC
$c
$MACRO DEV $cal %di = -2*(y*%log(%fv) + (n-y)*%log(n-%fv))
$cal y1 = %ge(y,0.00000001)*y+(1-%ge(y,0.00000001)) $
$cal %di = %di+2*y1*%log(y1) $
$cal y1 = n-y $
$cal y1 = %ge(y1,0.00000001)*y1+(1-%ge(y1,0.00000001)) $
$cal %di = %di+2*y1*%log(y1) $
$ENDMAC
$c
$MACRO ENDUP
$output 13  $
$print 'MODEL: ' model ; 'NODES: ' *i %k ; 'TOL: ' %x ; 'FINAL FIT: ' %p  ;
'NO. ITERATIONS: ' %i $
$dis d e  $
$output 2  $
$print
  'use macro NODES for new nodes, or re-set tolerance(%x)';
 'then use macro LOOP to re-run algorithm'
  $
$ENDMAC
$c
$MACRO MEXP
$c
$c moves data to vectors n and y from vector %1=P2 (after IVNIT) or
$c %1=p3 (after ESTEP)
$c
$cal n(%t) = %1(%s):
     y(%t)=%1(%v):
     %t=%t+1:
     %s= %s+1:
     %v=%v+1:
     %e=%e-1 $
 $ENDMAC
 $c
 $RETURN
```

D2. FORTRAN SUBROUTINE PASS.

```
C*****************************************************************
C--- GLIM 3.77  (copyright)1984 Royal Statistical Society, London ---
C---------------------------------------------------------------
C
      SUBROUTINE PASS (OPT,RARRAY,RLEN,CARRAY,CLEN,RMV,IFT,IFTA)
      INTEGER OPT,CARRAY(*),CLEN,RLEN,IFT,IFTA(2)
      REAL RARRAY(*),RMV

      EXTERNAL LEGDAT,INIT,ESTEP,WRWARN,fop
      IF (OPT.EQ.1) CALL LEGDAT(RARRAY,RLEN)
      IF (OPT.EQ.2) CALL INIT(RARRAY,RLEN)
      IF (OPT.EQ.3) CALL ESTEP(RARRAY,RLEN)
      IF (OPT.LE.0)
     - CALL WRWARN('the PASS subroutine is not implemented',38)
      RETURN
      END
```

D3. FORTRAN SUBROUTINE LEGDAT.

```
SUBROUTINE LEGDAT(rarray,rlen)
    INTEGER p,q,rlen,sub
    DOUBLE PRECISION pnodes(2), pwei(2),min,max,x1,x2,y(4)
    REAL rarray(rlen)
    OPEN(9,FILE = 'LEG4.DAT')
    READ(9,7001) pnodes,pwei
 7001 FORMAT(F17.15)
c
c     Read no. of nodes and range
c
    WRITE(*,9999)
 9999 FORMAT('    Input no of nodes(4,8,12,...,60)')
   10 READ(*,*) p
     IF(p.lt.4) GO TO 10
     IF (p.gt.60) GO TO 10
     q=MOD(p,4)
     IF (q.ne.0) GO TO 10
     rarray(1)=p
     p=p/4
     WRITE(*,9998)
 9998 FORMAT('    Input min and max')
   20 READ(*,*) min,max
     IF(min.ge.max) GO TO 20
     DO 9 i = 1,2
       pwei(i) = pwei(i)*(max-min)/dble(2*p)
       rarray(i+1)=pwei(i)
    9 CONTINUE
     sub=3
     DO 1 i = 1,p
       x1 = min+(max-min)*dble(i-1)/dble(p)
       x2 = x1 + (max-min)/dble(p)
       DO 2 j=1,4
         k=ABS(j-2.5)+0.5
         y(j)=(x2-x1)*pnodes(k)
         IF(j.lt.3) y(j)=-y(j)
         y(j)=((x1+x2)+y(j))/2
         sub=sub+1
         rarray(sub)=y(j)
    2  CONTINUE
    1 CONTINUE
     CLOSE(9)
     RETURN
     END
```

D4. FORTRAN SUBROUTINE JINIT.F77.

```
SUBROUTINE INIT(rarray,reclen)
c
c
c       INITIALIZATION PROGRAM
c
c
      INTEGER subj,items,nodes,u(1000,200),reclen,sub
      DOUBLE PRECISION temp,asum(1000),
     +       n(60,200),p(1000,60),weight(60),
     +        ccoeff(200),tcoeff(200),avg,gamma(60),
     +        y(60,200),diff,lp,lastfit
      REAL   g(200),rarray(reclen)
      COMMON /com/ n,p,y,u,asum
c
      OPEN(9,FILE='a2glirt.dat')
      lastfit=9d6
      OPEN(17,FILE='fit.dat')
      WRITE(17,1000) lastfit
 1000 FORMAT(F20.10)
      CLOSE(17)
c
c
c         read item data
c
c      read no. subjects + no. items
c
      READ(9,*) subj,items
c
c      read guessing parameters
c
      READ(9,*) (g(j),j=1,items)
c
c       read responses
c
      DO 1 i=1,subj
       READ(9,5002)(u(i,j),j=1,items)
 5002   FORMAT (i1,199i1)
    1 CONTINUE

c
c         quadrature formula data
c
      nodes=rarray(1)
      DO 22 k=1,nodes
       sub=MOD(k,4)
       IF(sub.lt.2) THEN
        sub=3
              ELSE
        sub=2
       ENDIF
```

```
      gamma(k)=rarray(k+3)
      weight(k)=rarray(sub)
   22 CONTINUE
c
c       initialise slope and intercept parameters
c
     DO 2 j = 1,items
      ccoeff(j) =1
      tcoeff(j) =1
    2 CONTINUE
c
c
c       calculate expected no. done and no. correct
c
     DO 6 i = 1,subj
      asum(i)=0.0d00
      avg = 0.0d00
      DO 4 k = 1,nodes
       p(i,k) = (-(gamma(k)**2)/2)
       DO 3 j=1,items
        lp=ccoeff(j)+tcoeff(j)*gamma(k)
        IF(u(i,j).eq.1) p(i,k) = p(i,k)+
     *     log(g(j)+(1-g(j))/(1+exp(-lp)))
        IF(u(i,j).eq.0) p(i,k) = p(i,k)+
     *       log(1-g(j))-log(1+exp(lp))
     3  CONTINUE
        avg = avg + p(i,k)/nodes
     4  CONTINUE
       DO 5 k=1,nodes
        diff=p(i,k)-avg
        IF(diff.gt.88) diff = 88
        temp =  exp(diff)
        p(i,k) = weight(k)*temp
        asum(i) = asum(i)+ p(i,k)
     5  CONTINUE
     6 CONTINUE
      WT = 1
      DO 9 k = 1,nodes
       DO 8  j = 1,items
        n(k,j) = 0.0d00
        y(k,j) = 0.0d00
        DO 7 i = 1,subj
         IF(u(i,j).eq.0) n(k,j) = n(k,j)+p(i,k)/asum(i)
         IF(u(i,j).eq.1) then
          n(k,j) = n(k,j)+p(i,k)/asum(i)
          y(k,j) = y(k,j)+p(i,k)/asum(i)
         END IF
      7   CONTINUE
      8  CONTINUE
      9 CONTINUE
c
c     add following to GLIM vector: subjects,items, guessing params,
```

```
c     expected nos.done, expected nos. correct.
c
      sub=nodes+4
      rarray(sub)=subj
      rarray(sub+1)=items
      sub=sub+2
      DO 10 j = 1,items
        rarray(sub)=g(j)
        sub=sub+1
   10 CONTINUE
      DO 11 k=1,nodes
        DO 11 j=1,items
          rarray(sub)=n(k,j)
          sub=sub+1
   11 CONTINUE
      DO 12 k=1,nodes
        DO 12 j=1,items
          rarray(sub)=y(k,j)
          sub=sub+1
   12 CONTINUE
      close(9)
      RETURN
      END
```

D5. FORTRAN SUBROUTINE ESTEP.

```
SUBROUTINE ESTEP(rarray,reclen)
c
c
c       E-STEP PROGRAM
c
c
      INTEGER subj,items,nodes,u(1000,200),reclen,asub,bsub
      REAL g(200),rarray(reclen),calc,diff,tol
      DOUBLE PRECISION  N(60,200),weight(60),lastfit,
     +          ccoeff(200),tcoeff(200),gamma(60),
     +          y(60,200),lp,fit,sum(1000), p(1000,60)
      COMMON /com/ N,P,Y,U,SUM
c
      OPEN(9,FILE='a2glirt.dat')
      OPEN(17,FILE='fit.dat')
      READ(17,*) lastfit
      CLOSE(17)
      OPEN(17,FILE='fit.dat')
c
c         read item data
c
      READ(9,*) subj,items
c
c       read guessing parameters
c
      READ(9,*) (g(j),j = 1,items)
c
c
c       read item responses
c
      DO 1 i = 1,subj
        sum(i) = 0.0d00
        READ(9,5002)(u(i,j),j=1,items)
 5002   FORMAT(i1,199i1)
    1 CONTINUE
c
c      extract quadrature formula data from GLIM array
c
       nodes=rarray(2)
       DO 22 k=1,nodes
        asub=MOD(k,4)
         IF (asub.lt.2) THEN
          asub=4
        ELSE
          asub=3
        ENDIF
        gamma(k)=rarray(k+4)
        weight(k)=rarray(asub)
  22 CONTINUE
c
```

```
c
c       extract item parameter estimates from GLIM array
c
      asub=5+nodes+2*(2+nodes)*items
      bsub=asub+items
      DO 2 j = 1,items
       ccoeff(j)=rarray(asub)
       tcoeff(j)=rarray(bsub)-ccoeff(j)
       asub=asub+1
       bsub=bsub+1
    2 CONTINUE
c
c       calculate expected no. done and no. correct
c
      DO 6 i = 1,subj
       DO 4 k = 1,nodes
        p(i,k) = (-(gamma(k)**2)/2)
        DO 3 j = 1,items
         lp = ccoeff(j)+tcoeff(j)*gamma(k)
         IF(u(i,j).eq.1) p(i,k) = p(i,k)+
     *                log( g(j) + (1-g(j))/(1 + exp(-lp)))
         IF(u(i,j).eq.0) p(i,k) = p(i,k)+
     *                log(1-g(j)) - log(1+exp(lp))
      3   CONTINUE
      4  CONTINUE
         DO 5 k=1,nodes
          p(i,k) = weight(k)*exp(p(i,k))
          sum(i) = sum(i)+ p(i,k)
      5  CONTINUE
      6 CONTINUE
c
c         check convergence
c
      tol=rarray(1)
      fit=0.0d00
      DO 98 i=1,subj
       fit = fit-2*log(sum(i))
     98 CONTINUE
      diff=lastfit-fit
      IF (diff.lt.tol) rarray(1)=fit
      WRITE(*,1000) fit
   1000 FORMAT(' FIT statistic: ',F12.6)
      WRITE(17,1001) fit
   1001 FORMAT(F18.10)
      DO 9 k = 1,nodes
       DO 8  j = 1,items
        n(k,j) = 0.0d00
        y(k,j) = 0.0d00
        DO 7 i = 1,subj
         IF(u(i,j).eq.0) n(k,j) = n(k,j)+p(i,k)/sum(i)
         IF(u(i,j).eq.1) then
         n(k,j) = n(k,j)+p(i,k)/sum(i)
```

```
      y(k,j) = y(k,j)+p(i,k)/sum(i)
      END IF
7     CONTINUE
8   CONTINUE
9 CONTINUE
  asub=5+nodes
  bsub=asub+(nodes+2)*items
  DO 10 k = 1,nodes
    DO 10 j=1,items
      rarray(asub)=n(k,j)
      rarray(bsub)=y(k,j)
      asub=asub+1
      bsub=bsub+1
10 CONTINUE
  CLOSE(9)
  CLOSE(17)
  RETURN
  END
```

APPENDIX E

QUESTIONS FOR TIMED ITEM TEST OF MENTAL ARITHMETIC.

Item No.	Problem	Difficulty/Type	Time	True/False
1	8+3+9=20	1	4	T
2	12-3+8=17	2	4	T
3	11+15+7=33	3	4	T
4	15-8+12=19	4	4	T
5	16+19-27=8	5	4	T
6	9+3+6=18	1	6	T
7	11-4+7=14	2	6	T
8	12+16+6=34	3	6	T
9	16-8+12=20	4	6	T
10	16+18-25=9	5	6	T
11	7+4+8=19	1	8	T
12	13-4+7=16	2	8	T
13	12+16+7=35	3	8	T
14	16-7+13=22	4	8	T
15	16+15-23=8	5	8	T
16	8+4+9=22	1	4	F
17	11-3+8=14	2	4	F
18	11+16+7=32	3	4	F
19	14-8+12=16	4	4	F
20	15+19-27=5	5	4	F
21	8+3+6=15	1	6	F
22	12-4+7=17	2	6	F
23	11+16+6=31	3	6	F
24	15-8+12=21	4	6	F
25	15+18-25=6	5	6	F
26	6+4+8=16	1	8	F
27	12-4+7=17	2	8	F
28	11+16+7=32	3	8	F
29	15-7+13=23	4	8	F
30	15+16-23=65+	5	8	F
31	5+6+9=20	1	4	T
32	11-4+8=15	2	4	T
33	11+16+7=34	3	4	T
34	14-7+12=19	4	4	T
35	15+18-27=6	5	4	T
36	8+5+6=19	1	6	T
37	11-3+8=16	2	6	T
38	13+16+7=36	3	6	T
39	14-6+12=20	4	6	T
40	14+19-25=8	5	6	T
41	6+5+8=19	1	8	T
42	12-3+7=16	2	8	T
43	11+17+8=36	3	8	T
44	14-8+13=19	4	8	T
45	17+14-23=5	5	8	T
46	5+6+7=16	1	4	F
47	11-4+6=15	2	4	F

Item No.	Problem	Difficulty/Type	Time	True/False
48	11+16+5=34	3	4	F
49	14-7+9=14	4	4	F
50	15+18-25=6	5	4	F
51	8+5+4=19	1	6	F
52	11-3+6=16	2	6	F
53	13+16+5=32	3	6	F
54	14-6+13=23	4	6	F
55	15+19-24=18	5	6	F
56	6+3+9=20	1	8	F
57	12-3+5=16	2	8	F
58	11+17+4=30	3	8	F
59	14-8+12=16	4	8	F
60	17+14-23=10	5	8	F

APPENDIX F

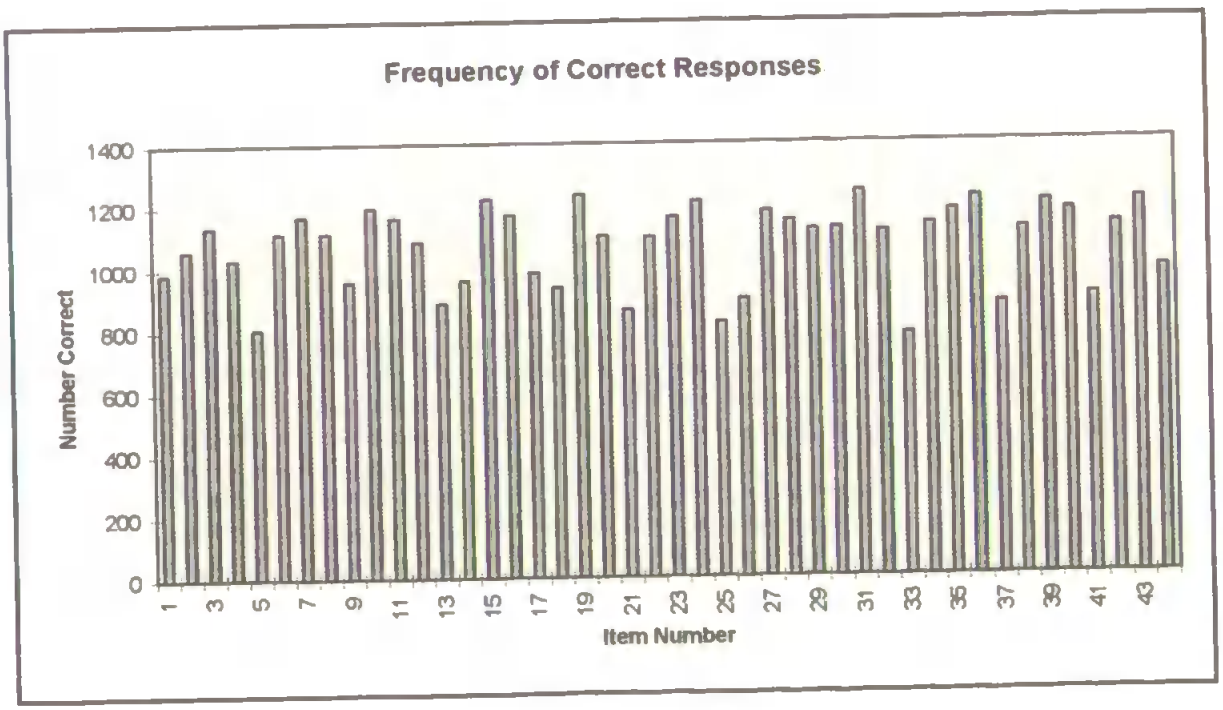
F1. QUESTIONS FOR TRANSITIVE INFERENCE TEST.

Item	Time (secs)	Question	Type	Difficulty
1	2	Paul is slower than John. Who is slower ? Paul John	4	1
2	3	Bill is slower than Joe. Who is faster ? Bill Joe	3	2
3	5	Chris is happier than Mike. Who is happier ? Mike Chris	1	1
4	4	Pete is not as bad as Ian. Who is better ? Pete Ian	5	3
5	2	Dave is not as short as Sid. Who is shorter ? Dave Sid	6	2
6	3	Chris is not as strong as Tom. Who is stronger ? Chris Tom	7	2
7	5	Tom is taller than Chris. Who is shorter ? Chris Tom	2	2
8	4	Phil is not as tall as Mike. Who is shorter ? Phil Mike	8	3
9	2	Chris is better than Dave. Who is worse ? Dave Chris	2	2
10	3	Steve is heavier than John. Who is heavier ? Steve John	1	1
11	5	Sid is dimmer than Ian. Who is brighter ? Ian Sid	3	2
12	4	Tom is not as bad as Phil. Who is better ? Tom Phil	5	3
13	2	Bill is not as bright as John. Who is brighter ? Bill John	7	2
14	3	Bob is not as old as John. Who is younger ? John Bob	8	3
15	5	Sid is shorter than Paul. Who is shorter ? Sid Paul	4	1
16	4	John is not as sad as Steve. Who is sadder ? Steve John	6	2
17	2	Fred is shorter than Bill. Who is taller ? Fred Bill	3	2
18	3	Chris is not as bad as John. Who is better ? John Chris	5	3
19	5	Paul is happier than George. Who is happier ? Paul George	1	1
20	4	Phil is not as tall as Dave. Who is shorter ? Dave Phil	8	3
21	2	George is not as dim as Dave. Who is dimmer ? Dave George	6	2
22	3	John is not as heavy as Bob. Who is heavier ? John Bob	7	2
23	5	Sid is stronger than Steve. Who is weaker ? Sid Steve	2	1
24	4	Steve is shorter than John. Who is shorter ? Steve John	4	1
25	2	Fred is not as young as Paul. Who is younger ? Paul Fred	6	2
26	3	Bill is not as weak as Joe. Who is stronger ? Joe Bill	5	3
27	5	Chris is not as heavy as Sid. Who is heavier ? Sid Chris	7	2
28	4	Fred is weaker than Pete. Who is stronger ? Pete Fred	3	2
29	2	George is sadder than Bill. Who is sadder ? George Bill	4	1
30	3	Chris is heavier than Sid. Who is lighter ? Sid Chris	2	2
31	5	Fred is brighter than John. Who is brighter ? Fred John	1	1
32	4	Fred is not as tall as Bob. Who is shorter ? Bob Fred	8	3

Item	Time (secs)	Question	Type
33	2	Joe is not as slow as Ian. Who is faster ? Joe Ian	5
34	3	Sid is lighter than Pete. Who is lighter ? Pete Sid	4
35	5	Dave is not as happy as Chris. Who is sadder ? Dave Chris	8
36	4	Bill is older than Paul. Who is older ? Bill Paul	1
37	2	Steve is not as sad as Bill. Who is sadder ? Steve Bill	6
38	3	Chris is not as bright as Pete. Who is brighter ? Pete Chris	7
39	5	Tom is brighter than Ian. Who is dimmer ? Tom Ian	2
40	4	George is lighter than Dave. Who is heavier ? George Dave	3
41	2	John is faster than Joe. Who is faster ? Joe John	1
42	3	Joe is weaker than Sid. Who is stronger ? Joe Sid	3
43	5	Paul is worse than Mike. Who is worse ? Paul Mike	4
44	4	Steve is not as slow as Paul. Who is faster ? Steve Paul	5

F2. SUMMARY OF RESPONSE DATA MATRIX FOR TRANSITIVE INFERENCE TEST.

ITEM NO.	NO. CORRECT	ITEM NO.	NO. CORRECT
1	985	23	1160
2	1060	24	1211
3	1135	25	820
4	1031	26	894
5	804	27	1176
6	1114	28	1146
7	1167	29	1116
8	1114	30	1123
9	956	31	1238
10	1195	32	1110
11	1161	33	777
12	1084	34	1135
13	886	35	1177
14	958	36	1220
15	1220	37	875
16	1170	38	1115
17	983	39	1202
18	935	40	1171
19	1234	41	901
20	1101	42	1128
21	860	43	1208
22	1098	44	986



F3. PARAMETER ESTIMATES.

MODEL (M11): probdiff + 1/time + gamma

DEGREES OF FREEDOM: 875

FIT STATISTIC: 40346.81

PARAMETER ESTIMATES:

Problem Difficulty (3 Levels):	1. (Easiest) 5.037
	2. 4.341
	3. (Most Difficult) 3.351
Slope on reciprocal of time:	-9.646
Slope on ability:	1.322

MODEL (M12): probdiff + time + gamma

DEGREES OF FREEDOM: 873

FIT STATISTIC: 40318.52

PARAMETER ESTIMATES:

Problem Difficulty (3 Levels):	1. (Easiest) 2.982
	2. 2.259
	3. (Most Difficult) 1.173
Time (4 levels):	1. Two seconds: -2.787
	2. Three seconds:-1.043
	3. Four seconds: -0.1784
	(4. Five seconds: 0.0)
Slope on ability:	1.324

MODEL (M13): probtype + 1/time + gamma

DEGREES OF FREEDOM: 870

FIT STATISTIC: 40168.61

PARAMETER ESTIMATES:

Problem Type (8 Levels):	1. (Easiest) 4.664		
	2. 4.251	3. 4.233	4. 5.010
	5. 2.846	6. 3.771	7. 4.286
	8. (Most Difficult) 3.594		
Slope on reciprocal of time:	-9.046		
Slope on ability:	1.336		

MODEL (M14): probtype + time + gamma

DEGREES OF FREEDOM: 868

FIT STATISTIC: 40131.03

PARAMETER ESTIMATES:

Problem Type (8 Levels):	1. (Easiest) 2.718		
	2. 2.355	3. 2.188	4. 3.037
	5. 0.7164	6. 1.760	7. 2.300
	8. (Most difficult) 1.495		
Time (4 levels):	1. Two seconds: -2.560		
	2. Three seconds: -0.9741		
	3. Four seconds: -0.04996		
	(4. Five seconds: 0.0)		
Slope on ability	1.337		

REFERENCES

Abramowitz, M. and Stegun, I.A. (1972) Editors, *Handbook of Mathematical Functions.* New York: Dover Publications.

Aitkin, M. (1994) An EM algorithm for overdispersion in generalized linear models. *Proceedings of the 9th International Workshop on Statistical Modeling.*

Aitkin, M., Anderson, D.A. and Hinde, J.P. (1981) Statistical modeling of data on teaching styles. *Journal of the Royal Statistical Society*, Series A, **144**, 419-461.

Aitkin, M. and Francis, B.J. (1996) Fitting overdispersed generalized linear models by nonparametric maximum likelihood. *GLIM Newsletter.*

Albert, J.H. (1992) Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics*, **17**, 251-269.

Alvey, N. (1977) *GENSTAT.* Oxford: Numerical Algorithms Group.

Andersen, E.B. (1984) *The Statistical Analysis of Categorical Data.* Berlin: Springer.

Anderson, D.A. (1988) Some models for overdispersed binomial data. *Australian Journal of Statistics*, **30**, 125-148.

Anderson, D.A. and Aitken, M.A. (1985) Variance component models with binary response: interviewer variability. *Journal of the Royal Statistical Society*, Series B, **47**, 203-210.

Anderson, D.A. and Hinde, J.P. (1988) Random effects in generalized linear models and the EM algorithm. *Communications in Statistics - Theory and Methods*, **17**, 3847-3856.

Anderson, T.W. (1984, 2nd ed.) *An Introduction to Multivariate Statistical Analysis*. New York: Wiley.

Baker, F.B. (1987) Methodological review: item parameter estimation under the one-, two-, and three-parameter logistic models. *Applied Psychological Measurement*, **11**, 111-141.

Barnett, V.D. and Wright, D.E. (1992) Biomedical applications for a generalized linear functional Poisson model. *Journal of Applied Statistics*, **19**, 41-47

Bartholomew, D.J. (1987) *Latent Variable Models and Factor Analysis*. London: Griffin.

Birnbaum, A. (1962) On the foundations of statistical inference (with discussion). *Journal of the American Statistical Association*, **57**, 269-326

Birnbaum, A. (1969) Statistical theory for logistic mental test models with a prior distribution of ability. *Journal of Mathematical Psychology*, **6**, 258-276.

Bock, R.D. and Aitkin, M. (1981) Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika*, **46**, 443-459.

Bock, R.D. and Lieberman, M. (1970) Fitting a response model for *n* dichotomously scored items. *Psychometrika*, **35**, 179-197.

Breslow, N.E. and Clayton, D.G. (1993) Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9-25.

Brillinger, D.R. and Preisler, H.K. (1983) Maximum likelihood estimation in a latent variable problem. *Studies in Econometrics, Time Series, and Multivariate Statistics* (S. Karlin, T. Amemiya and L.A. Goodman. Eds.) New York: Academic Press.

Burden, R.L and Faires, J.D (1989, 4th Ed.) *Numerical Analysis*. Boston, Mass: PWS-Kent.

Carroll, R.J. and Ruppert, D. (1982) Robust estimation in heteroscedastic linear models. *The Annals of Statistics*, **10**, 429-441.

Clayton, D.G. (1994) Generalized linear mixed models. *Proceedings of the 9th International Workshop on Statistical Modeling*.

Cochran,W.G., and Cox, G. (1957) *Experimental Designs*. New York: Wiley.

Conaway, M. (1990) A random effects model for binary data. *Biometrics*, **46**, 317-328.

Czado, C. (1994) Modeling overdispersion in binomial regression. *Proceedings of the 9th International Workshop on Statistical Modeling.*

Davis, P.J and Rabinowitz, P (1984, 2nd Ed.) *Methods of Numerical Integration*. New York: Academic Press.

Dempster, A.P, Laird, N.M, and Rubin, D.B (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, Series B, **39**, 1-38.

Diggle, P.J., Liang, K-Y, and Zeger, S.L. (1994) *Analysis of Longitudinal Data.* Oxford: Oxford University Press.

Dobson, A.J. (1990) *An Introduction to Generalized Linear Models.* London: Chapman and Hall.

Draper, N. R. and Smith, H. (1981, 2nd Ed.) *Applied Regression Analysis*. New York: Wiley.

Drum, M.L. and McCullagh, P. (1993) REML estimation with exact covariance in the logistic mixed model. *Biometrics*, **49**, 677-689.

Engel, B. and Keen, A. (1992) A simple approach for the analysis of generalized linear mixed models. *Statistica Neerlandica*, **48**, 1-22.

Fellner, W.H. (1986) Robust estimation of variance components. *Technometrics*, **28**, 51-60.

Fellner, W.H. (1987) Sparse matrices and the estimation of variance components by likelihood methods. *Comm. Statist.* B., **16**, 439-63.

Froberg, C-E. (1973, 2nd Ed.) *Introduction to Numerical Analysis*. Reading, Massachusetts: Addison Wesley.

Gelfand, A.E. and Smith, A.F.M. (1990) Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398-409.

Gilmour, A.R., Anderson, R.D. and Rae, A.L. (1985) The analysis of binary data by a generalized linear mixed model. *Biometrika*, **72**, 593-599.

Goldstein, H. (1986) Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, **73**, 43-56.

Goldstein, H. (1991) Nonlinear multilevel models, with an application to discrete response data. *Biometrika*, **78**, 45-51.

Goldstein, H. (1995) *Multilevel Statistical Models* (2nd Ed.). London: Arnold.

Goldstein, H. and Rasbash, J. (1996) Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society*, Series A, **159**, 505-513.

Green, P.J. (1987) Penalized likelihood for general semi-parametric regression models. *International Statistical Review*, **55**, 245-259.

Hagenaars, J.A. (1993) Loglinear models with latent variables. *(Sage University Papers , series on Quantitative Applications in the Social Sciences, series no. 07-094)* Newbury Park, CA: Sage.

Hambleton, R.K. and Swaminathan, H. (1985) *Item Response Theory - Principles and Applications*. Boston: Kluwer-Nijhoff.

Hambleton, R.K., Swaminathan, H. and Rogers, H.J. (1991) *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage.

Harville, D.A. (1977) Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, **72**, 320-340.

Healy, M.J.R., and Westmacott, M. (1956) Missing values in experiments analyzed on automatic computers. *Applied Statistics*, **5**, 203-206.

Henderson, C.R. (1975) Best linear unbiased estimation and prediction under a selection model. *Biometrics*, **31**, 13-22.

Hinde, J. (1982) Compound Poisson regression models. *GLIM 82*. (R. Gilchrist. Ed.), New York: Springer Verlag.

Hocking, R.R. (1985) *The Analysis of Linear Models*. Monterey: Brooks/Cole.

Hulin, C.L, Lissak, R.I. and Drasgow, F. (1982) Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo Study. *Applied Psychological Measurement*, **6**, 249-260.

Im, S. and Gianola, D. (1988) Mixed models for binomial data with an application to lamb mortality. *Applied Statistics*, **37**, 196-204.

Jenrich, R.I. and Sampson, P.F. (1976) Newton-Raphson and related algorithms for maximum likelihood variance component estimation. *Technometrics*, **18**, 11-17.

Kalbfleisch, J.D. and Prentice, R.L. (1980) *The Statistical Analysis of Failure Time Data*. New York: John Wiley.

Karim, M.R. and Zeger, S.L. (1992) Generalized linear models with random effects; Salamander mating revisited. *Biometrics*, **48**, 631-644.

Kendall, M.G. and Stuart, A. (1979, 4th Ed.) *The Advanced Theory of Statistics*. Vol. 2. London: Griffin.

Kimura, D.K. (1992) Functional comparative calibration using an EM algorithm. *Biometrics*, 48, 1263-1271.

Kuk, A.Y.C. (1995) Asymptotically unbiased estimation in generalized linear models with random effects. *Journal of the Royal Statistical Society*, Series B, **57**, 395-407.

Lee, Y. and Nelder, J.A. (1996) Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society*, Series B, **58**, 619-678.

Liang, K-Y and Zeger, S.L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13-22.

Lindsay, J.K. (1993) *Models for Repeated Measurements*. Oxford: Clarendon Press.

Longford, N.T. (1988) *VARCL: Software For Variance Components Analysis Of Data And Hierarchically Nested Random Effects (Maximum Likelihood)*. Princeton, New Jersey: Educational Testing Service.

Lord, F.M. (1975) *Evaluation With Artificial Data Of A Procedure For Estimating Ability And Item Characteristic Curve Parameters*. (Research Bulletin 75-33). Princeton, New Jersey: Educational Testing Service.

Lord, F.M. and Novick, M.R. (1968) *Statistical Theories Of Mental Test Scores*. Reading, MA: Addison-Wesley

Louis, T.A. (1982) Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society*, Series B, **44**, 226-233.

Meng, X-L. and Schilling, S. (1996) Fitting full-information item factor models and an empirical investigation of bridge sampling. *Journal of the American Statistical Association*, **91**, 1254-1267.

Mislevy, R.J. (1984) Estimating latent distributions. *Psychometrika*, **49**, 359-381.

Mislevy, R.J. (1985) Estimation of latent group effects. *Journal of the American Statistical Association*, **80**, 993-997.

Mislevy, R.J. (1986) Bayes modal estimation in item response models. *Psychometrika*, 51, pp 177-195.

Mislevy, R.J. (1989) *PC-BILOG: Item Analysis and Test Scoring with Binary Logistic Models*. Mooresville, IN: Scientific Software.

Mislevy, R.J. and Bock, R.D. (1984) *BILOG: Item Analysis and Test Scoring with Binary Logistic Models*. Mooresville, IN: Scientific Software.

Nelder, J.A. and Mead, R. (1965) A Simplex method for function minimization. *Comput. J.*, 7, 308-313.

Nelder, J.A. and Wedderburn, R.W.M. (1972) Generalized linear models. *Journal of the Royal Statistical Society*, Series A, **135**, 370-384.

O'Hagan, A. (1976) On posterior, joint and marginal modes. *Biometrika*, **63**, 329-333.

Owen, R. J. (1975) A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, **70**, 351-356.

Patefield, W.M. (1981) Multivariate linear relationships: maximum likelihood estimation and regression bounds. *Journal of the Royal Statistical Society*, Series B, **43**, 342-352.

Patterson, H.D. and Thompson, R. (1971) Recovery of interblock information when block sizes are unequal. *Biometrika*, **58**, 545-553.

Payne, C.D. (1987) Editor, *The GLIM System Release 3.77 Manual - Edition 2.* Oxford: The Royal Statistical Society and NAG Ltd.

Pickles, A., Pickering, K. and Taylor, C. (1996) Reconciling recalled dates of developmental milestones, events and transitions: a mixed generalized linear model with random mean and variance functions. *Journal of the Royal Statistical Society*, Series A, **159**, 225-234.

Prosser, R., Rasbash, J. and Goldstein, H. (1993) *ML3 Software for Three-Level Analysis: User's Guide for Version 2*. London: Institute of Education.

Rao, C.R. (1973) *Linear Statistical Inference and its Applications*. New York: John Wiley.

Rasbash, J, Yang, M., Woodhouse, G. and Goldstein,H. (1995) *MLn: Command Reference Guide*. London: Institute of Education.

Rigdon, S.E. and Tsutakawa, R.K. (1983) Parameter estimation in latent trait models. *Psychometrika*, **48**, 567-574.

Rodriguez, G. and Goldman, N. (1995) An assessment of estimation procedures for multilevel models with binary data. *Journal of the Royal Statistical Society*, Series A, **158**, 73-89.

Schall, R. (1991) Estimation in GLMs with random effects. *Biometrika*, **78**, 719-727.

Searle, S.R. (1971) *Linear Models*. New York: John Wiley.

Segall, D.O. (1996) Multidimensional adaptive testing. *Psychometrika*, 61, pp 331-354.

Spearman, C. (1904) General intelligence, objectively determined and measured. *American Journal of Psychology*, **15**, 201-93.

Stiratelli, R., Laird, N. and Ware, J.H. (1984) Random effects models for serial observations with binary response. *Biometrics*, **40**, 961-971.

Stroud, A.H. and Sechrest, D. (1966) *Gaussian Quadrature Formulas*. Englewood Cliffs, New Jersey: Prentice Hall.

Sun, L., Hsu, J.S.J., Guttman, I. And Leonard, T. (1996) Bayesian Methods for Variance Component Models. *Journal of the American Statistical Association*, **91**, 743-751.

Swaminathan, H. and Gifford, J.A. (1982) Bayesian estimation in the Rasch model. *Journal of Educational Statistics*, 7, 175-191.

Swaminathan, H. and Gifford, J.A. (1983) Estimation of parameters in the three-parameter latent trait model. *New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing* (Weiss, D.J. Ed.), New York: Academic Press.

Swaminathan, H. and Gifford, J.A. (1985) Bayesian estimation in the two-parameter logistic model. *Psychometrika*, **50**, 349-364.

Swaminathan, H. and Gifford, J.A. (1986) Bayesian estimation in the three-parameter logistic model. *Psychometrika*, **51**, 589-601.

Thissen, D. (1982) Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika*, **47**, 175-186

Tierney, L. and Kadane, J.B. (1986) Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, **86**, 82-86.

Tsutakawa, R.K. (1984) Estimation of two-parameter logistic item response curves. *Journal of Educational Statistics*, **9**, 263-276.

Tsutakawa, R.K. (1988) A mixed model for analyzing geographic variability in mortality rates. *Journal of the American Statistical Association*, **83**, 37-42.

Urry, U.V. (1974) Approximations to item parameters of mental test models and their uses. *Educational and Psychological Measurement*, **34**, 253-269.

Williams, D.A. (1982) Extra-binomial variation in logistic linear models. *Applied Statistics*, **31**, 144-148.

Williams, E.J. (1969) Regression methods in calibration problems. *Bulletin of the International Statistical Institute*, **43**, 17-28.

Wolfinger, R. (1993) Laplace's approximation for nonlinear mixed models. *Biometrika*, **80**, 791-795.

Wood, R.L., Wingersky, M.S. and Lord, F.M. (1978) *LOGIST: A Computer Program for Estimating Examinee Ability and Item Characteristic Curve Parameters.* (Research Memorandum 76-6) (revised). Princeton, New Jersey: Educational Testing Service.

Wright, D.E. and Barnett, V.D. (1991) *Fitting Predictive Accident Models in GLIM with Uncertainty in the Flow Estimates.* Transport and Road Research Laboratory, Dept. of Transport. Contractor Report 286.

Wright, D.E., Creagh-Osborne, J.E., Tapsfield, P. and Kyllonen, P. (1994) A timed item test of mental arithmetic. The Technical Cooperation Program. London: M.O.D.

Wu, C.F.J. (1983) On the convergence properties of the EM algorithm. *The Annals of Statistics*, **11**, 95-103.

Yen, W.M. (1981) Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, **5**, 245-262.

Zeger, S.L. and Karim, M.R. (1991) Generalized linear models with random effects; a Gibbs sampling approach. *Journal of the American Statistical Association*, **86**, 79-86.

Zeger, S.L., Liang, K-Y. and Albert, P. (1988) Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, **44**, 1049-1060.