# Some Problems in the Theory & Application of Graphical Models

Andrew Wilfred Roddam

Linacre College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy at the University of Oxford

Trinity Term 1999

**Some Problems in the Theory & Application of Graphical Models**

**Thesis submitted for the degree of Doctor of Philosophy**

**Andrew Wilfred Roddam, Linacre College**

**Trinity Term 1999**

## Abstract

A graphical model is simply a representation of the results of an analysis of relationships between sets of variables. It can include the study of the dependence of one variable, or a set of variables on another variable or sets of variables, and can be extended to include variables which could be considered as intermediate to the others. This leads to the concept of representing these chains of relationships by means of a graph; where variables are represented by vertices, and relationships between the variables are represented by edges. These edges can be either directed or undirected, depending upon the type of relationship being represented.

The thesis investigates a number of outstanding problems in the area of statistical modelling, with particular emphasis on representing the results in terms of a graph. The thesis will study models for multivariate discrete data and in the case of binary responses, some theoretical results are given on the relationship between two common models. In the more general setting of multivariate discrete responses, a general class of models is studied and an approximation to the maximum likelihood estimates in these models is proposed.

This thesis also addresses the problem of measurement errors. An investigation into the effect that measurement error has on sample size calculations is given with respect to a general measurement error specification in both linear and binary regression models.

Finally, the thesis presents, in terms of a graphical model, a re-analysis of a set of childhood growth data, collected in South Wales during the 1970s. Within this analysis, a new technique is proposed that allows the calculation of derived variables under the assumption that the joint relationships between the variables are constant at each of the time points.

# Acknowledgements

It is my great pleasure to acknowledge the support and direction of my supervisor, Sir David Cox, who provided me with a great deal of inspiration and direction during the past three and a half years. Without his continual support and encouragement this thesis would not have been possible. I would also like to thank Professor Nanny Wermuth for her never ending enthusiasm and some very useful discussions.

I would like to thank my parents for their belief and continual support and encouragement; I know that you thought I would never finish education, but this is it. I would also like to express my deepest gratitude to my grandmother who never doubted me and was always there when I needed someone to talk to: I dedicate this thesis to you.

In the Department of Statistics, I would like to thank Rene, Yih-Choung and Francisco for sharing an office with me, Vivien and Rita for help with administrative matters, Susan and Dave for computing support, Professor Brian Ripley for advice on making S-Plus do what I wanted it to do, and Mario for numerous chats and advice. I would like to thank all my other friends who are too numerous to mention by name, but you know who you all are.

I would like to thank the Engineering and Physical Sciences Research Council for financial support in the form of an earmarked studentship.

Finally, I would like to thank my partner Vicky; you made the first year exciting and the past two years worthwhile. Without you the whole process would have been so much harder.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

The central theme underlying this thesis is the concept of a graphical model. A graphical model is, in many ways, simply a representation of the results of an analysis of relationships between sets of variables. It can include the study of the dependence of one variable, or a set of variables on another variable or sets of variables. It can also be extended to include variables that could be considered as intermediate to the others, in the sense that they can be considered as responses to one set of variables, and as explanatory to another set. This leads to the idea of representing these chains of relationships by means of a graph. This particular form of representation, and its underlying theory, has created the field of 'graphical modelling'.

Graphical modelling is the diagrammatic representation of a set of relationships between variables in a data set. The variables in a data set are represented by vertices, and relationships between the variables are represented by edges. These edges can be either directed or undirected depending upon the type of relationship being represented. There are a number of specific types of graphical model (Cox & Wermuth, 1996, Chapter 2) and this thesis will concentrate on the joint-response chain graph model. In this type of representation, variables are arranged into boxes from left to right and within a particular box variables are to considered on an equal footing. The key factor is that variables in any box are considered as responses to variables in boxes to its right. Within a particular box, relationships

between the variables are represented either by full undirected lines or by dashed undirected lines. The absence of an undirected dashed line between two variables indicates that they are marginally independent, and in the case of the variables being normally distributed this would imply that their estimated covariance was zero. Marginal independence says that for two random quantities, $X$ and $Y$, the distribution of $X$ is unchanged when we are given information about $Y$. The absence of an undirected full line between two variables would indicate that these two variables were conditionally independent of all other variables within a box (and in boxes to the right). Conditional independence says for three random quantities, $X$, $Y$ and $Z$, that $X$ and $Y$ become independent once we know the value of $Z$. In all of these definitions, the random quantities could be either a single variable or a set of variables. Between boxes there are directed arrows and these can be either full or dashed line arrows, although all arrows pointing to any box are of the same type. Dashed arrows towards node $Y_i$ indicate that regressions of $Y_i$ on variables in the boxes to the right of $Y_i$ are being considered, while full-line arrows mean that the regression is both on variables within the same box as $Y_i$ and in boxes to the right. A particularly good illustration of the distinction between these two types of edges is given in Figure 2.9 of Cox & Wermuth (1996).

In recent years there has been a large volume of work on graphical models, the result of which has been a large number of papers and four expository texts – Whittaker (1990), Edwards (1995), Cox & Wermuth (1996) and Lauritzen (1996). However, the field is by no means a new one: the ideas date back to those of the geneticist Sewall Wright in the 1920s, and were further developed until the late 1970s within the fields of path analysis, structural equation models and covariance selection models. It was only then that researchers started to investigate the links between these areas, and the new field of graphical models was born.

Within graphical modelling, there are a number of very different research areas. Of the four books cited previously, all have a slightly different approach to the material. A substantial proportion of the research has been directed to the problem that, given a particular data set, how do we analyse it? In particular, how does the analysis proceed if the interest is in assessing the relationships between the

variables under study? This is a classical problem in applied statistics, and the parallels between the research in these two fields is extensive. Most of the work presented in Whittaker (1990), Edwards (1995) and Cox & Wermuth (1996), has lain in interpreting the results of relatively standard statistical analyses in terms of a graphical model. However, a large amount of research has been carried out in the probabilistic properties of these graphical structures (Lauritzen, 1996). This thesis will concentrate more on the former than on the latter and will investigate the usefulness, in terms of graphical modelling, of methods to analyse data sets. In particular, this thesis will concentrate on the graphical chain model (Cox & Wermuth, 1996, Chapter 2) and methods for analysing data with the aim of representing the results of such analyses in a chain graph. This is simply an alternative representation of the results of standard statistical analyses and therefore this thesis is really split into two main sections. The first section concentrates on the more theoretical aspects of modelling that have direct implications for the types of analysis that may be carried out, when the aim of the analysis is to investigate relationships between variables. The second section concentrates on a practical application of standard analyses to a real data set, and representing the results of these analyses in terms of a chain graph model.

Although there has been a large upsurge of research into graphical models, there are still a number of theoretical problems in applied statistics to be addressed, and the aim of this thesis is to investigate a number of these. One of the main issues that still receives a great deal of attention in the literature is that of multivariate responses in regression problems. A number of authors have tackled this issue in the case of continuous data, but in the case of discrete responses there is still little agreement about the most fruitful approach. There are two common distributions for multivariate binary data, one of which (the quadratic exponential distribution) assumes that all interactions greater than order 2 are 0, whereas the alternative (the multivariate normal dichotomy) allows the interactions to vary. Chapter 2 of this thesis investigates the relationship between these two distributions, and in the case of a three-dimensional binary random variable, provides results on the required sample size for estimating a three-factor interaction in the multivariate

normal dichotomy. This allows a general proposal to be made for those situations where the two distributions will give similar results, and in those cases where the results have the potential to be quite different. This has important consequences for the modelling of multivariate binary data, as the result shows that we have only to be concerned about the choice of distribution in those cases where it is likely that the answers will be different.

Another area requiring further investigation, again in the context of multivariate discrete data, is the extension of models for binary responses to general discrete responses. In chapter 3 a brief review of models for multivariate discrete data is given, with the emphasis on those models that can be considered as either marginal or conditional models. There has long been an interest in fitting marginal models to data sets but this had originally been avoided due to the computational problems associated with them. Chapter 3 therefore considers a marginal model proposed by Glonek & McCullagh (1995). A new approximation to the maximum likelihood estimates of the parameters of this model is presented; the approximation provides a considerable computational advantage when fitting these models. This approximation is extended to a class of models proposed by Glonek (1996) which have both marginal and conditional components, a property which is very desirable for interpreting the results of a modelling procedure in terms of a graphical model.

A subject that has been of interest to statisticians for many years is that of errors-in-variables. This is typically a regression model where, in addition to the response variable being measured with error, one or more of the explanatory variables are also measured with error. It is well known that, in the case of linear regression, this results in estimated parameters being attenuated towards zero, and, in the case of other models, estimated parameters are systematically biased. This has a number of consequential problems when one is interested in assessing relationships between variables (as in graphical modelling); in addition to the need to adjust for potential bias, measurement error is an additional source of variation and therefore sample sizes required to detect significant associations have to be considerably larger than in the case where variables are measured without error. In

addition, classical estimation procedures are no longer applicable and alternative procedures, which take account of measurement error, have to be used. In chapter 4 we investigate the issue of sample size calculations when one of the explanatory variables is going to be measured with error, with respect to a measurement error model proposed by Reeves et al. (1998). This is a very general measurement error model specification and estimation procedure, which is applicable to both continuous responses (linear regression), and binary responses (logistic regression), and where the explanatory variables measured with error can be either discrete or continuous.

Finally, in this thesis we apply the technique of graphical modelling to the reanalysis of a data set on the growth of children between birth and 5 years old. The data were collected in South Wales in the 1970s, and recorded a number of measures of growth, although in this chapter only the height and weight measurements will be analysed. In addition to growth measurements on the children, numerous maternal/paternal variables were recorded, and these will be investigated in order to establish which background factors affect the growth of children. Initially, the height and weight measurements will be modelled separately, allowing a comparison between the two marginal models and the factors affecting them. There is, however, a strong interest in how the relationship between height and weight develop when considered as a joint response. There are a number of ways in which this could be investigated: one could consider fitting a multivariate linear model, but this assumes that the correlation between height and weight is constant, which is not necessarily a reasonable assumption. A more insightful method would be to consider a derived variables analysis, similar to that proposed by Cox & Wermuth (1992), which would allow an understanding of the joint relationship between the two measures of growth. However, this method was only considered for a single observation of the response vector; in this particular application, we have repeat measurements of the children's height and weight. An extension to the derived variables methodology is proposed which takes account of the repeated measurements, and estimates the joint relationships between height and weight, assuming that these relationships are constant across the observations. Both the-

oretical and practical issues will be explored, and applied to the childhood growth data set.

This thesis is, in many ways, a collection of topics structured around the underlying theme of ways to analyse data with a particular emphasis on interpreting the results in terms of a graphical model. However, all the techniques discussed within this thesis can be applied to more general situations than graphical models, and it is hoped that this will be apparent within each chapter. The chapters are self-contained and each one contains an introduction to the subject area, as well as a discussion of the results.

# Chapter 2

# Multivariate Binary Distributions

## 2.1 Introduction

When considering multivariate binary data, there are two main classes of distributions commonly used; the quadratic exponential binary distribution (Cox & Wermuth, 1994), and the dichotomised Gaussian distribution (Pearson, 1909). In the more recent literature, the dichotomised Gaussian distribution has been used in the structural equation modelling/latent variable literature, and the quadratic exponential distribution in data modelling situations. However, it is of interest to consider situations in which the two classes of distributions would give similar results.

It is important to investigate the relationship between these two classes of distributions since, given a particular data set, it is sometimes more convenient to consider one of the two possibilities. But, in those situations where these two classes of distribution give similar results, the choice of which distribution to use when fitting a model to the data is less important. This has important consequences in the graphical modelling field, as it is essential to understand whether one would draw the same inference, in terms of association relationships, from a particular data set. Understanding the relationship between these two distributions will give insight into those sets of data where the same results will be

obtained irrespective of which distribution is used.

Let us consider a multivariate binary random variable $\boldsymbol{I} = (I_1, \ldots, I_q)$ taking values $(i_1, \ldots, i_q)$, which without loss of generality can be assumed to be equal to $-1, 1$. The quadratic exponential binary distribution specifies that the joint probability density for $\boldsymbol{I}$ is defined as

$$\log P(\boldsymbol{I} = \boldsymbol{i}) = \mu + \Sigma_r \alpha_r i_r + \Sigma_{r>s} \alpha_{rs} i_r i_s, \tag{2.1}$$

where $\mu$ is a normalising constant, and the $\alpha$'s are unknown parameters. This distribution can be seen as a binary analogue to the multivariate normal, since with $n$ independent, identically distributed observations, the likelihood has a full exponential family form. Also, the conditional distribution of two components has a similar qualitative interpretation to concentrations in the continuous case. Unfortunately it does not retain its exact form under marginalisation, for example consider marginalising equation (2.1) with respect to $I_p$, that is compute the distribution of $(I_1, \ldots, I_{p-1})$, we have that

$$\begin{aligned} \log P\{I_r = i_r (r = 1, \ldots, q-1)\} &= \mu + \log 2 + \Sigma \alpha_r i_r + \Sigma_{r>s} \beta_{rs} i_r i_s \\ &\quad + \log \cosh(\alpha_q + \Sigma \beta_{rq} i_r). \end{aligned}$$

In addition it also assumes that all interactions of order 3 and higher are zero.

The dichotomised Gaussian distribution, in which $\boldsymbol{U} \sim N_q(0, \Sigma)$ where $\Sigma$ is a correlation matrix, defines

$$I_r = \begin{cases} 1 & (U_r > a_r), \\ -1 & (U_r \leq a_r), \end{cases}$$

where the $a_r$'s are called the cut-points. It is only by convention that the underlying continuous variables are chosen such that they have zero mean and are marginally normal. This distribution is often considered a 'natural' representation of reality, if one believes that there is a continuous variable underlying discrete observations. Due to the standard properties of normality it retains exact form under marginalisation and conditioning on the $U_r$'s, and can be useful if one is interested in partitioning the covariance matrix (as in structural equation modelling). However, using the discrete observations causes a loss in efficiency in

estimating the parameters of this distribution, as compared with the continuous observations. The procedure for estimating the cut-points $a_r$ and $\Sigma$ proceeds by estimating the cut-points marginally for each variable by comparing the observed proportion falling within each category with the standard normal density function. The correlation matrix can then be estimated using polychoric correlations (Bollen, 1989, pp 433-447). Although this is an appealing procedure because of its simplicity, it is not clear that it is the most efficient form of estimation.

One of the main differences between the multivariate normal dichotomy and the quadratic exponential binary distribution is that the multivariate normal dichotomy allows the possibility of a three-factor interaction when $q \geq 3$, although both distributions have the same number of parameters. Questions of interest, which would allow insight into how different these two classes of distribution are, include:

- Is the three-factor interaction zero?
- If it is not zero, what size of sample would you need to be able to detect it in given situations?
- How does the three-factor interaction depend on the correlation matrix?

The aim of this chapter is to investigate these issues; we start by proposing an approximation formula for approximate median dichotomy of the multivariate normal distribution in two and three dimensions. We investigate numerically the accuracy of this approximation. We then outline the method for estimating a three-factor interaction and use the approximation formulæ obtained to gain some insight into the above questions in the three-dimensional case. Some numerical results will be given.

## 2.2 Approximate Median Dichotomy

Using the definition of the multivariate normal dichotomy given in the previous section, let $\pi^{\boldsymbol{a}}_{-,\ldots,-}$ be $P(A_j = -1 \ \forall \ j)$ using cut-points $\boldsymbol{a} = (a_1, \ldots, a_n)$. If we

restrict our interest to median dichotomy, so that $\boldsymbol{a} = (0, \ldots, 0)$, then $\pi^{\boldsymbol{a}}_{-,\ldots,-}$ is referred to as the orthant probability, and there are a number of known results in this case. In particular,

$$\begin{aligned} P_q &= \pi^{\boldsymbol{a}}_{-,\ldots,-} \\ &= \Phi_q(\boldsymbol{a}; \Sigma) \\ &= \frac{1}{2^q}(1 + s_q(\Sigma)), \end{aligned}$$

where $s_q(\Sigma)$ is referred to as the orthant function. There are a number of theoretical results for this function (McFadden, 1955, 1956), although only for $q < 4$ are there closed-form solutions (in the case $q = 4$ there is an approximate closed form solution but it is only valid in the equicorrelated case). In higher dimensions, a number of approximations have been proposed, but they nearly always require infinite series summations (Steck, 1962).

One possibility for investigating potential differences between the quadratic exponential and the multivariate normal dichotomy is to attempt to approximate the probabilities formed by dichotomising a multivariate normal distribution. One way to form the approximation is to consider a Taylor Series expansion about the median, thus using the known results for median dichotomy. An alternative method would be to use the technique proposed by Cox & Wermuth (1991), but their approximation formula is relatively complex algebraically and would not allow very detailed insight into the questions posed above.

### 2.2.1 Two-Dimensional Case

If we median dichotomise a bivariate normal distribution, then Sheppard (1898) showed that

$$P_2 = \frac{1}{4}\left(1 + \frac{2}{\pi}\sin^{-1}\rho\right).$$

We need, therefore, to consider the approximate expansion of the following integral, around the point (0,0):

$$\Phi_2(a_1, a_2; \Sigma) = \int_{-\infty}^{a_1} \boldsymbol{dx_1} \int_{-\infty}^{a_2} \boldsymbol{dx_2}\ \phi_2(x_1, x_2; \Sigma).$$

If we differentiate this with respect to $a_1$,

$$\begin{aligned}\frac{\partial \Phi_2}{\partial a_1} &= \int_{-\infty}^{a_2} \boldsymbol{dx_2}\ \phi_2(a_1, x_2; \Sigma) \\ &= \int_{-\infty}^{a_2} \boldsymbol{dx_2}\ \phi(a_1) f_{X_2|X_1}(x_2|a_1) \\ &= \phi(a_1)\Phi\left\{\frac{a_2 - \rho a_1}{\sqrt{(1-\rho^2)}}\right\},\end{aligned}$$

since $X_2|X_1 = a_1 \sim N(\rho a_1, 1-\rho^2)$, where $\rho = \mathrm{Corr}(X_1, X_2)$. The first derivative with respect to $a_2$ is obtained in the same way, and by symmetry is

$$\frac{\partial \Phi_2}{\partial a_2} = \phi(a_2)\Phi\left\{\frac{a_1 - \rho a_2}{\sqrt{(1-\rho^2)}}\right\}.$$

Calculation of the second derivatives with respect to $a_1, a_2$ gives

$$\begin{aligned}\frac{\partial^2 \Phi_2}{\partial a_1^2} &= \phi'(a_1)\Phi\left\{\frac{a_2 - \rho a_1}{\sqrt{(1-\rho^2)}}\right\} + \phi(a_1)\phi\left\{\frac{a_2 - \rho a_1}{\sqrt{(1-\rho^2)}}\right\}\left\{-\frac{\rho}{\sqrt{(1-\rho^2)}}\right\}, \\ \frac{\partial^2 \Phi_2}{\partial a_1 a_2} &= \phi(a_1)\phi\left\{\frac{a_2 - \rho a_1}{\sqrt{(1-\rho^2)}}\right\}\frac{1}{\sqrt{(1-\rho^2)}}\end{aligned}$$

and evaluating all of these at the point $\boldsymbol{a} = (0,0)$ gives

$$\begin{aligned}\frac{\partial \Phi_2}{\partial a_1} = \frac{\partial \Phi_2}{\partial a_2} &= \frac{1}{2\sqrt{(2\pi)}}, \\ \frac{\partial^2 \Phi_2}{\partial a_1^2} = \frac{\partial^2 \Phi_2}{\partial a_2^2} &= -\frac{\rho}{2\pi\sqrt{(1-\rho^2)}}, \\ \frac{\partial^2 \Phi_2}{\partial a_1 a_2} &= \frac{1}{2\pi\sqrt{(1-\rho^2)}}.\end{aligned}$$

We therefore approximate $\pi^{\boldsymbol{a}}_{--}$, using a Taylor Series expansion around the point $\boldsymbol{a} = (0,0)$,

$$\pi^{\boldsymbol{a}}_{--} \approx \frac{1}{4}\{1 + s_2(\rho)\} + \frac{1}{2\sqrt{(2\pi)}}(a_1 + a_2) + \frac{1}{4\pi\sqrt{(1-\rho^2)}}\{-\rho(a_1^2 + a_2^2) + 2a_1 a_2\}. \tag{2.2}$$

In order to assess the adequacy of this approximation, it needs to be compared to the actual bivariate normal probabilities. This can be done using the algorithm of Schervish (1984) which allows the calculation of any multivariate normal probability to within a specified error bound.

If one uses equation (2.2) to calculate $\pi^{\boldsymbol{a}}_{--}$ for a variety of correlations and cut-point values $\boldsymbol{a}$, then it is fairly clear that this equation produces a number of

probabilities that are negative. It is therefore necessary to consider ways of stabilising this approximation such that the resultant probabilities are all positive. One way to solve this problem is to consider the function,

$$f(a_1, a_2) = k \exp\{c_1(a_1 + a_2) + c_2(a_1^2 + a_2^2) + c_3 a_1 a_2\}, \tag{2.3}$$

and then, using a Taylor Series expansion around the point $\boldsymbol{a} = (0, 0)$, equate the coefficients $k, c_1, c_2$ and $c_3$ so that this new expansion gives the same approximation as that given by equation (2.2) with error $O(a^3)$.

Using the standard expansion of an exponential function, the resulting approximation is

$$f(a_1, a_2) \approx k \left\{ 1 + c_1(a_1 + a_2) + \frac{1}{2}(2c_2 + c_1^2)(a_1^2 + a_2^2) + (c_3 + c_1^2) a_1 a_2 \right\}.$$

Therefore, setting

$$\begin{aligned} k &= \frac{1}{4}\{1 + s_2(\rho)\}, \\ c_1 &= \frac{1}{2k\sqrt{(2\pi)}}, \\ c_2 &= \frac{1}{2}\left\{ -\frac{\rho}{2\pi k\sqrt{(1-\rho^2)}} - c_1^2 \right\}, \\ c_3 &= \frac{1}{2\pi k\sqrt{(1-\rho^2)}} - c_1^2, \end{aligned}$$

recovers the approximation obtained in equation (2.2); using these values in the function defined in equation (2.3) will therefore produce an approximation for $\pi^{\boldsymbol{a}}_{--}$ which is now guaranteed to be positive.

We can compare this new approximation with $\rho = 0(0.2)0.8, 0.9$, and cut-points $a_j = -0.5(0.1)0.5$ for $j = 1, 2$, to the numerical computation method to verify its numerical stability. The easiest way to look at this is to consider a plot, for each value of $\rho$, of the ratio of the numerical value to the approximation; this is shown in Figure 2.1.

We can see from this graph that the approximation derived is close for all but extreme values of correlations, generally being within $\pm 2\%$. Even at the most extreme value of the correlation considered, $\rho = 0.9$, the approximation was still

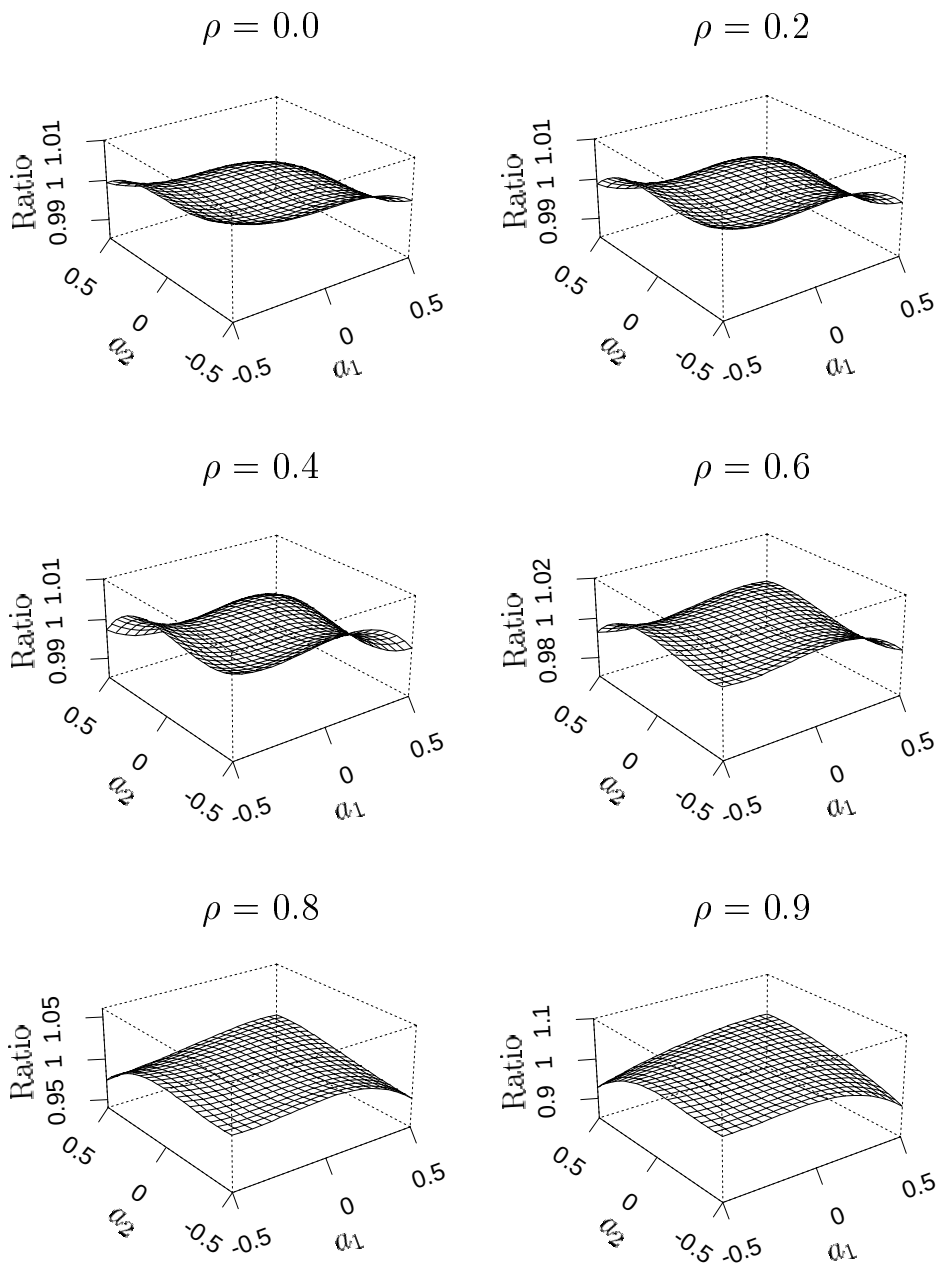



Figure 2.1: Series of plots showing the ratio of the numerical value to the approximation for $\pi^{\underline{a}}_{--}$ using various cut-points and correlation coefficients.

within 8%, which is an adequate performance given that we are approximating a high degree of curvature with a quadratic expression.

The other quadrant probabilities $\pi^{\boldsymbol{a}}_{-+}, \pi^{\boldsymbol{a}}_{+-}, \pi^{\boldsymbol{a}}_{++}$ can be obtained using the following relationships,

$$\pi^{\boldsymbol{a}}_{-+} + \pi^{\boldsymbol{a}}_{--} = \pi^{a_1}_{-},$$
$$\pi^{\boldsymbol{a}}_{--} + \pi^{\boldsymbol{a}}_{+-} = \pi^{a_2}_{-},$$

where $\pi^{a_i}_{-}$ is $P(Z \leq a_i)$ where $Z \sim N(0,1)$. Therefore,

$$\begin{aligned} \pi^{\boldsymbol{a}}_{-+} &\approx \Phi(a_1) - \pi^{\boldsymbol{a}}_{--}, \\ \pi^{\boldsymbol{a}}_{+-} &\approx \Phi(a_2) - \pi^{\boldsymbol{a}}_{--}, \\ \pi^{\boldsymbol{a}}_{++} &\approx 1 - \pi^{\boldsymbol{a}}_{--} - \pi^{\boldsymbol{a}}_{-+} - \pi^{\boldsymbol{a}}_{+-}. \end{aligned} \tag{2.4}$$

It is now possible to compare these approximations to the actual probabilities, calculated numerically from a bivariate normal distribution. The results for $\pi^{\boldsymbol{a}}_{++}$ are identical to those obtained for $\pi^{\boldsymbol{a}}_{--}$ due to symmetry in the approximation. However, the results for $\pi^{\boldsymbol{a}}_{-+}$ and $\pi^{\boldsymbol{a}}_{+-}$ (these are also the same due to symmetry) are much worse (Figure 2.2). This is due to the fact that, in estimating these two probabilities, one is estimating something very small, particularly for large values of the correlation co-efficient, since the majority of the density is then concentrated along the major axis. It should, therefore, not be surprising that a second order approximation does not perform too well in such circumstances. However, one must note that for correlations less than 0.5 the approximation is normally within 5%, and for correlations greater than 0.5 it slowly deteriorates until, at very high correlations, it is totally unstable. In using this approximation for comparison this should not concern us too much, since it is very rare in practice to observe very high values of the correlation co-efficient and given the stability of our approximation at low to moderate values, it should work adequately.

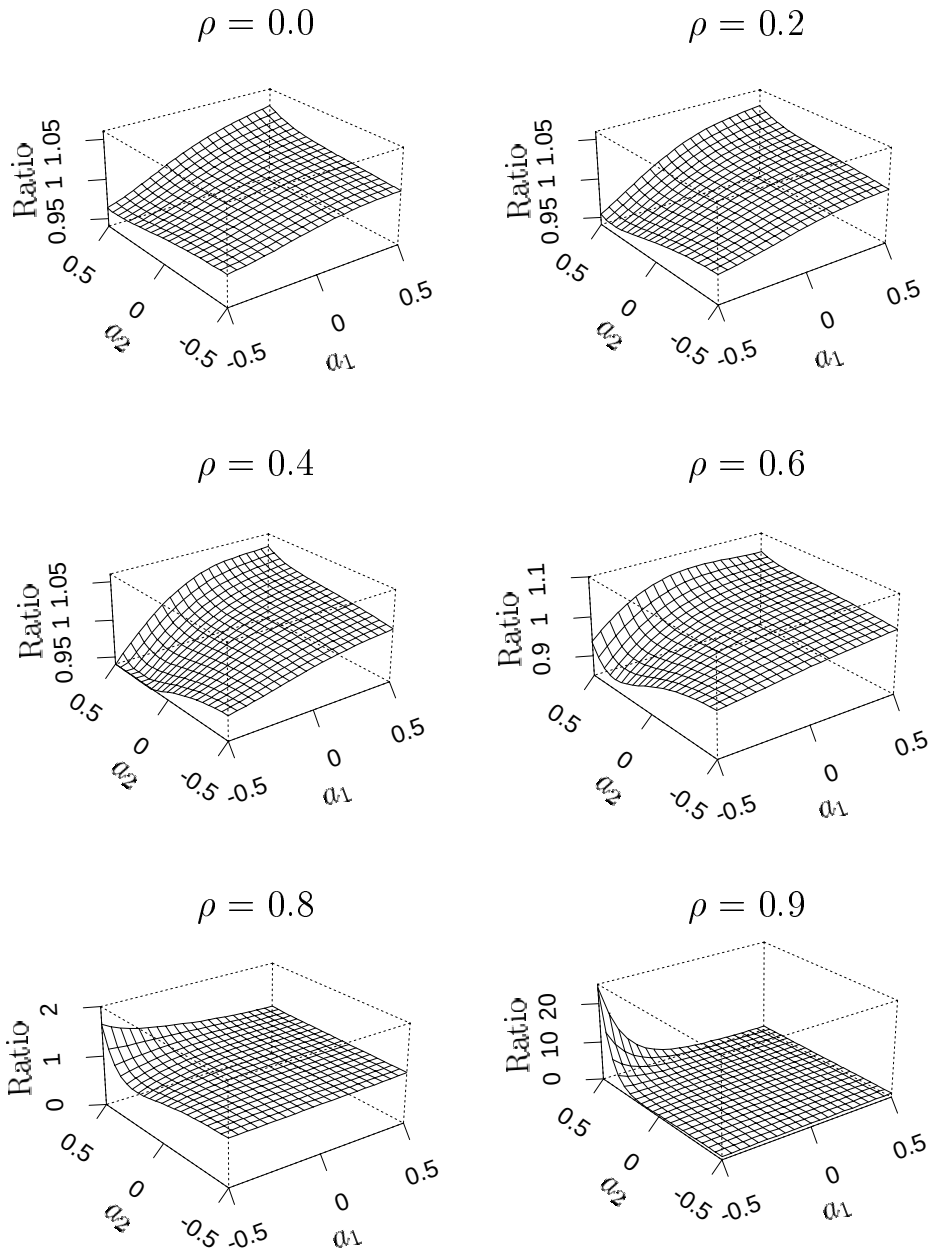



Figure 2.2: Series of plots showing the ratio of the numerical value to the approximation for $\pi^{\boldsymbol{a}}_{-+}$ using various cut-points and correlation coefficients.

### 2.2.2 Three-Dimensional Case

In three dimensions it would appear that Kendall (1941) is the most usual reference for obtaining the orthant probability by median dichotomising a three-dimensional normal distribution. However, in this paper he notes that the results were not new and refers the reader to page 175 of Aitken & Turnbull (1931) and Aitken & Gonin (1935). The result however shows that median dichotomy of a three-dimensional normal distribution gives an orthant probability of

$$P_3 = \frac{1}{8}\left\{1 + \frac{2}{\pi}(\sin^{-1}\rho_{12} + \sin^{-1}\rho_{13} + \sin^{-1}\rho_{23})\right\}.$$

Therefore, we can consider a direct extension to the ideas developed above. If we Taylor Series expand a three-dimensional normal random variate around its cut-points $\boldsymbol{a}$, then we need to evaluate both the first and second partial derivatives with respect to each cut-point.

Differentiating the cumulative three-dimensional normal density with respect to $a_1$ gives

$$\begin{aligned}
\frac{\partial \Phi_3}{\partial a_1} &= \int_{-\infty}^{a_2} \boldsymbol{dx_2} \int_{-\infty}^{a_3} \boldsymbol{dx_3}\ \phi_3(a_1, x_2, x_3; \Sigma) \\
&= \int_{-\infty}^{a_2} \boldsymbol{dx_2} \int_{-\infty}^{a_3} \boldsymbol{dx_3}\ \phi(a_1)\phi_2\left(x_2 - \rho_{12}a_1, x_3 - \rho_{13}a_1; \Sigma_{23.1}\right) \\
&= \phi(a_1)\Phi_2(a_2 - \rho_{12}a_1, a_3 - \rho_{13}a_1; \Sigma_{23.1}), \qquad (2.5)
\end{aligned}$$

where $\Sigma_{23.1}$ is the conditional covariance matrix of $(X_2, X_3)$ given $X_1$. Using a similar notation, differentiating with respect to $a_2, a_3$ gives

$$\begin{aligned}
\frac{\partial \Phi_3}{\partial a_2} &= \phi(a_2)\Phi_2(a_1 - \rho_{12}a_2, a_3 - \rho_{23}a_2; \Sigma_{13.2}), \\
\frac{\partial \Phi_3}{\partial a_3} &= \phi(a_3)\Phi_2(a_1 - \rho_{13}a_3, a_2 - \rho_{23}a_3; \Sigma_{12.3}).
\end{aligned}$$

Evaluating these expressions at the cut-point $\boldsymbol{a} = (0, 0, 0)$ gives

$$\begin{aligned}
\frac{\partial \Phi_3}{\partial a_i} &= \phi(0)\Phi_2(0, 0; \Sigma_{jk.i}) \text{ where } i \neq j \neq k \\
&= \phi(0)P_2(\Sigma_{jk.i}),
\end{aligned}$$

where $P_2$ was defined previously.

Calculation of the second derivatives gives

$$\begin{aligned}\frac{\partial^2\Phi_3}{\partial a_1\partial a_2} &= \int_{-\infty}^{a_3} \boldsymbol{dx_3}\ \phi_3(a_1,a_2,x_3;\Sigma)\\ &= \phi_2(a_1,a_2;\rho_{12})\Phi\left(\frac{a_3-\mu_{3.21}}{\sigma_{3.21}}\right),\end{aligned}$$

where $\mu_{3.21}, \sigma_{3.21}$ are the conditional mean and variance of $X_3$ given $X_1 = a_1$ and $X_2 = a_2$, and can be shown to be equal to

$$\begin{aligned}\mu_{3.21} &= \frac{\rho_{13}-\rho_{12}\rho_{23}}{1-\rho_{12}^2}a_1 + \frac{\rho_{23}-\rho_{12}\rho_{13}}{1-\rho_{12}^2}a_2,\\ \sigma_{3.21} &= \frac{1-\rho_{12}^2-\rho_{13}^2-\rho_{23}^2+2\rho_{12}\rho_{13}\rho_{23}}{1-\rho_{12}^2}.\end{aligned}$$

Similarly, it can be shown that

$$\begin{aligned}\frac{\partial^2\Phi_3}{\partial a_i\partial a_j} &= \int_{-\infty}^{a_k} \boldsymbol{dx_k}\ \phi_3(a_i,a_j,x_k;\Sigma)\ \ (k\neq i,j)\\ &= \phi_2(a_i,a_j;\rho_{ij})\Phi\left(\frac{a_k-\mu_{k.ij}}{\sigma_{k.ij}}\right).\end{aligned}$$

Evaluating this expression at the cut-point $\boldsymbol{a} = (0,0,0)$ we obtain

$$\frac{\partial^2\Phi_3}{\partial a_i\partial a_j} = \phi_2(0,0;\rho_{ij})\Phi(0),$$

where $\phi_2(0,0;\rho_{ij}) = \left(2\pi\sqrt{(1-\rho_{ij}^2)}\right)^{-1}$.

To calculate the partial derivative with respect to $a_1^2$, using equation (2.5) we obtain the following:

$$\frac{\partial^2\Phi_3}{\partial a_1^2} = \phi'(a_1)\Phi_2(a_2-\rho_{12}a_1, a_3-\rho_{13}a_1;\Sigma_{23.1}) + \phi(a_1)\frac{\partial\Phi_2}{\partial a_1}.$$

Noting that we can re-write $\Phi_2(a_2-\rho_{12}a_1, a_3-\rho_{13}a_1;\Sigma_{23.1})$ as

$$\begin{aligned}\Phi_2(a_2-\rho_{12}a_1, a_3-\rho_{13}a_1;\Sigma_{23.1}) &= \Phi_2(z_1,z_2;\Sigma_{23.1})\\ &= \int_{-\infty}^{z_1}\boldsymbol{dx_2}\int_{-\infty}^{z_2}\boldsymbol{dx_3}\ \phi_2(x_1,x_2;\Sigma_{23.1}),\end{aligned}$$

where $z_1 = a_2-\rho_{12}a_1$ and $z_2 = a_3-\rho_{13}a_1$, we obtain, using the chain rule,

$$\begin{aligned}\frac{\partial\Phi_2}{\partial a_1} &= \frac{\partial\Phi_2}{\partial z_1}\frac{\partial z_1}{\partial a_1} + \frac{\partial\Phi_2}{\partial z_2}\frac{\partial z_2}{\partial a_1}\\ &= -\left(\rho_{12}\frac{\partial\Phi_2}{\partial z_1} + \rho_{13}\frac{\partial\Phi_2}{\partial z_2}\right),\end{aligned} \tag{2.6}$$

and

$$\begin{aligned}\frac{\partial \Phi_2}{\partial z_1} &= \int_{-\infty}^{z_2} \boldsymbol{dx_3}\phi_2(z_1, x_3, \Sigma_{23.1}) \\ &= \phi(z_1)\Phi\left\{\frac{z_2 - \rho' z_1}{\sqrt{(1-\rho'^2)}}\right\},\end{aligned}$$

where $\rho'$ is the correlation between $X_2$ and $X_3$ given $X_1 = a_1$. Observing that we are only interested in approximate median dichotomy, it is not necessary to calculate $\rho'$ explicitly, since substituting $a_1 = a_2 = 0$ into the above expression will yield

$$\frac{\partial \Phi_2}{\partial z_1} = \phi(0)\Phi(0),$$

and using a similar calculation it can be shown that

$$\frac{\partial \Phi_2}{\partial z_2} = \phi(0)\Phi(0).$$

Therefore, re-substituting these results into equation (2.6) we find that

$$\frac{\partial \Phi_2}{\partial a_1} = -\phi(0)\Phi(0)(\rho_{12} + \rho_{13}).$$

Using these calculations, it is easily deduced that

$$\frac{\partial \Phi_2}{\partial a_i} = -\phi(0)\Phi(0)(\rho_{ij} + \rho_{ik}) \text{ where } i \neq j \neq k.$$

Therefore, we can evaluate the general second order derivatives, at the cut-points $\boldsymbol{a} = (0, 0, 0)$ obtaining

$$\begin{aligned}\frac{\partial^2 \Phi_3}{\partial a_i^2} &= \phi'(0)\Phi_2(0, 0; \Sigma_{jk.i}) - \phi^2(0)\Phi(0)(\rho_{ij} + \rho_{ik}) \text{ where } i \neq j \neq k \\ &= \phi'(0)P_2(\Sigma_{jk.i}) - \phi^2(0)\Phi(0)(\rho_{ij} + \rho_{ik}).\end{aligned}$$

We can form a Taylor Series approximation for $\pi^{\boldsymbol{a}}_{---}$ using the above expressions obtaining

$$\begin{aligned}\pi^{\boldsymbol{a}}_{---} = P_3 &+ \frac{1}{\sqrt{(2\pi)}}\left\{P_2\left(\frac{\rho_{23} - \rho_{12}\rho_{13}}{\sqrt{(1-\rho_{12}^2)(1-\rho_{13}^2)}}\right)a_1\right. \\ &\left. + P_2\left(\frac{\rho_{13} - \rho_{12}\rho_{23}}{\sqrt{(1-\rho_{12}^2)(1-\rho_{23}^2)}}\right)a_2 + P_2\left(\frac{\rho_{12} - \rho_{13}\rho_{23}}{\sqrt{(1-\rho_{13}^2)(1-\rho_{23}^2)}}\right)a_3\right\} \\ &- \frac{1}{8\pi}\left\{(\rho_{12} + \rho_{13})a_1^2 + (\rho_{12} + \rho_{23})a_2^2 + (\rho_{13} + \rho_{23})a_3^2\right\} \\ &+ \frac{1}{4\pi}\left\{\frac{a_1 a_2}{\sqrt{(1-\rho_{12}^2)}} + \frac{a_1 a_3}{\sqrt{(1-\rho_{13}^2)}} + \frac{a_2 a_3}{\sqrt{(1-\rho_{23}^2)}}\right\}. \qquad (2.7)\end{aligned}$$

Using the following set of identities, we can obtain all of the other relevant cell probabilities

$$\begin{aligned}
\pi^{\boldsymbol{a}}_{---} + \pi^{\boldsymbol{a}}_{--+} &= \pi^{a_1,a_2}_{--} \\
\pi^{\boldsymbol{a}}_{---} + \pi^{\boldsymbol{a}}_{-+-} &= \pi^{a_1,a_3}_{--} \\
\pi^{\boldsymbol{a}}_{---} + \pi^{\boldsymbol{a}}_{+--} &= \pi^{a_2,a_3}_{--} \\
\pi^{\boldsymbol{a}}_{---} + \pi^{\boldsymbol{a}}_{--+} + \pi^{\boldsymbol{a}}_{-+-} + \pi^{\boldsymbol{a}}_{-++} &= \Phi(a_1) \\
\pi^{\boldsymbol{a}}_{---} + \pi^{\boldsymbol{a}}_{--+} + \pi^{\boldsymbol{a}}_{+--} + \pi^{\boldsymbol{a}}_{+-+} &= \Phi(a_2) \\
\pi^{\boldsymbol{a}}_{---} + \pi^{\boldsymbol{a}}_{-+-} + \pi^{\boldsymbol{a}}_{+--} + \pi^{\boldsymbol{a}}_{++-} &= \Phi(a_3) \\
\Sigma\pi_{ijk} &= 1. \qquad (2.8)
\end{aligned}$$

As in the previous section, we can use the algorithm of Schervish (1984) to check the validity of the above approximations. However, it is almost impossible to give a complete enumeration here of the possible situations for this approximation. In Table 2.1, the ratio of the approximation to the numerical value is given in the symmetrical case (all correlations and cut-points equal); while in Table 2.2 the same ratio is given in the non-symmetrical case for a number of different correlation and cut-point structures. In general, it appears that for values of $|a_i| \leq 0.5$ and $|\rho| < 0.7$, this approximation is within 10% of the true values, and in a lot of situations much better. For $\rho > 0.7$ the approximation deteriorates quite badly, due mainly to the quadratic expression failing to detect the full degree of curvature in the density. As with the case of the two-dimensional approximation, there is poor performance at extremes of the distribution function which should not concern us too much, since in real data situations high values of the correlation co-efficient, and/or small cell probabilities, are quite rare.

Table 2.1: Ratio of the numerical value to the approximation of $\pi^{a}_{\_\_\_}$ in the symmetrical case for a number of different values of the cut-points.

| Correlation | Cut-points | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | -0.5 | -0.4 | -0.3 | -0.2 | -0.1 | 0.0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 |
| 0.0 | 1.194 | 1.064 | 1.017 | 1.003 | 1.000 | 1.000 | 1.000 | 1.000 | 1.001 | 1.005 | 1.011 |
| 0.2 | 1.021 | 1.005 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.001 | 1.004 | 1.009 | 1.017 |
| 0.4 | 1.000 | 1.000 | 1.001 | 1.001 | 1.000 | 1.000 | 1.001 | 1.002 | 1.006 | 1.013 | 1.022 |
| 0.6 | 1.037 | 1.022 | 1.012 | 1.005 | 1.001 | 1.000 | 1.001 | 1.005 | 1.012 | 1.022 | 1.035 |
| 0.7 | 1.077 | 1.044 | 1.023 | 1.010 | 1.002 | 1.000 | 1.002 | 1.008 | 1.018 | 1.031 | 1.049 |
| 0.8 | 1.144 | 1.082 | 1.042 | 1.017 | 1.004 | 1.000 | 1.003 | 1.013 | 1.028 | 1.048 | 1.073 |
| 0.9 | 1.290 | 1.165 | 1.083 | 1.034 | 1.008 | 1.000 | 1.007 | 1.025 | 1.052 | 1.088 | 1.131 |

Table 2.2: Ratio of the numerical value to the approximation of $\pi^{\boldsymbol{a}}_{\_\_\_}$ in the non-symmetrical case with 4 different correlation structures and 3 different cut-point structures.

| | $\boldsymbol{a} = (a_1, a_1, a_1)$ | | | | $\boldsymbol{a} = (a_1, a_1, -a_1)$ | | | | $\boldsymbol{a} = (a_1, a_1 + 0.3, -a_1 + 0.3)$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Correlation Structure | | | | Correlation Structure | | | | Correlation Structure | | | |
| Cut-point $a_1$ | $1^a$ | $2^b$ | $3^c$ | $4^d$ | $1^a$ | $2^b$ | $3^c$ | $4^d$ | $1^a$ | $2^b$ | $3^c$ | $4^d$ |
| -0.5 | 1.027 | 1.026 | 1.077 | 1.087 | 0.945 | 1.005 | 1.010 | 1.051 | 1.002 | 1.128 | 1.040 | 1.153 |
| -0.4 | 1.016 | 1.014 | 1.042 | 1.045 | 0.981 | 1.011 | 1.014 | 1.035 | 1.023 | 1.107 | 1.038 | 1.119 |
| -0.3 | 1.008 | 1.008 | 1.021 | 1.022 | 0.997 | 1.009 | 1.011 | 1.020 | 1.027 | 1.081 | 1.031 | 1.088 |
| -0.2 | 1.004 | 1.004 | 1.008 | 1.009 | 1.001 | 1.004 | 1.006 | 1.009 | 1.023 | 1.057 | 1.023 | 1.064 |
| -0.1 | 1.001 | 1.001 | 1.002 | 1.002 | 1.001 | 1.001 | 1.002 | 1.002 | 1.017 | 1.037 | 1.017 | 1.046 |
| 0.0 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.011 | 1.024 | 1.015 | 1.035 |
| 0.1 | 1.001 | 1.001 | 1.002 | 1.002 | 1.001 | 1.001 | 1.002 | 1.002 | 1.010 | 1.018 | 1.017 | 1.033 |
| 0.2 | 1.004 | 1.005 | 1.007 | 1.008 | 1.005 | 1.003 | 1.008 | 1.007 | 1.015 | 1.020 | 1.026 | 1.038 |
| 0.3 | 1.010 | 1.013 | 1.016 | 1.019 | 1.013 | 1.005 | 1.019 | 1.015 | 1.025 | 1.030 | 1.041 | 1.053 |
| 0.4 | 1.019 | 1.024 | 1.028 | 1.033 | 1.023 | 1.004 | 1.033 | 1.024 | 1.043 | 1.047 | 1.063 | 1.076 |
| 0.5 | 1.031 | 1.039 | 1.043 | 1.053 | 1.034 | 0.998 | 1.050 | 1.032 | 1.067 | 1.071 | 1.092 | 1.108 |

[a] $\rho_{12} = \rho_{13} = 0.6$, $\rho_{23} = 0.2$
[b] $\rho_{12} = \rho_{13} = 0.6$, $\rho_{23} = -0.2$
[c] $\rho_{12} = 0.8$, $\rho_{13} = 0.4$, $\rho_{23} = 0.2$
[d] $\rho_{12} = 0.8$, $\rho_{13} = 0.4$, $\rho_{23} = -0.2$

## 2.3 The Three-Factor Interaction

As discussed in the introduction, one of the main differences between the quadratic exponential distribution and the multivariate normal dichotomy is that the latter allows the possibility of a three-factor interaction when $q \geq 3$. The three-factor interaction between three binary variables is defined as

$$\log\left(\frac{\text{Odds Ratio}(I_1, I_2|I_3 = 1)}{\text{Odds Ratio}(I_1, I_2|I_3 = -1)}\right).$$

We can rewrite each of these conditional odds ratios in terms of the cell probabilities as follows

$$\begin{aligned} \text{OR}(I_1, I_2|I_3 = 1) &= \frac{\pi^{\boldsymbol{a}}_{+++}\pi^{\boldsymbol{a}}_{--+}}{\pi^{\boldsymbol{a}}_{+-+}\pi^{\boldsymbol{a}}_{-++}}, \\ \text{OR}(I_1, I_2|I_3 = -1) &= \frac{\pi^{\boldsymbol{a}}_{---}\pi^{\boldsymbol{a}}_{++-}}{\pi^{\boldsymbol{a}}_{+--}\pi^{\boldsymbol{a}}_{-+-}}. \end{aligned} \tag{2.9}$$

### 2.3.1 Size of the Three-Factor Interaction

For the three-factor interaction to be zero, we require that these two conditional odds ratios are equal. This is easier to investigate if we switch to the symmetrical case where $\Sigma = \rho J + (1-\rho)I$ and $a_r = a$. Using the sets of relationships given in equation (2.4) and equation (2.8), we can rewrite these two conditional odds ratios in terms of $\Phi(a), \pi^{\boldsymbol{a}}_{--}$ and $\pi^{\boldsymbol{a}}_{---}$ as follows

$$\begin{aligned} \frac{\pi^{\boldsymbol{a}}_{+++}\pi^{\boldsymbol{a}}_{--+}}{\pi^{\boldsymbol{a}}_{+-+}\pi^{\boldsymbol{a}}_{-++}} &= \frac{(\Phi(a) + \pi^{\boldsymbol{a}}_{---} - 2\pi^{\boldsymbol{a}}_{--})\pi^{\boldsymbol{a}}_{---}}{(\pi^{\boldsymbol{a}}_{--} - \pi^{\boldsymbol{a}}_{---})^2}, \\ \frac{\pi^{\boldsymbol{a}}_{---}\pi^{\boldsymbol{a}}_{++-}}{\pi^{\boldsymbol{a}}_{+--}\pi^{\boldsymbol{a}}_{-+-}} &= \frac{(1 - \pi^{\boldsymbol{a}}_{---} - 3\Phi(a) + 3\pi^{\boldsymbol{a}}_{--})(\pi^{\boldsymbol{a}}_{--} - \pi^{\boldsymbol{a}}_{---})}{(\Phi(a) + \pi^{\boldsymbol{a}}_{---} - 2\pi^{\boldsymbol{a}}_{--})^2}. \end{aligned}$$

We therefore require

$$\frac{(\Phi(a) + \pi^{\boldsymbol{a}}_{---} - 2\pi^{\boldsymbol{a}}_{--})\pi^{\boldsymbol{a}}_{---}}{(\pi^{\boldsymbol{a}}_{--} - \pi^{\boldsymbol{a}}_{---})^2} = \frac{(1 - \pi^{\boldsymbol{a}}_{---} - 3\Phi(a) + 3\pi^{\boldsymbol{a}}_{--})(\pi^{\boldsymbol{a}}_{--} - \pi^{\boldsymbol{a}}_{---})}{(\Phi(a) + \pi^{\boldsymbol{a}}_{---} - 2\pi^{\boldsymbol{a}}_{--})^2},$$

or equivalently

$$(\Phi(a) + \pi^{\boldsymbol{a}}_{---} - 2\pi^{\boldsymbol{a}}_{--})^3\pi^{\boldsymbol{a}}_{---} = (1 - \pi^{\boldsymbol{a}}_{---} - 3\Phi(a) + 3\pi^{\boldsymbol{a}}_{--})(\pi^{\boldsymbol{a}}_{--} - \pi^{\boldsymbol{a}}_{---})^3. \tag{2.10}$$

In the symmetrical case, the approximation formula derived in equation (2.2) and equation (2.7) become

$$\begin{aligned} \pi^{\boldsymbol{a}}_{--} &\approx P_2 + \frac{a}{\sqrt{(2\pi)}} + \frac{a^2}{2\pi}\left(\frac{1-\rho}{1+\rho}\right)^{1/2}, \\ \pi^{\boldsymbol{a}}_{---} &\approx P_3 + P_2\frac{3a}{\sqrt{(2\pi)}}\frac{\rho}{1+\rho} + \frac{3a^2}{4\pi}\left(\frac{1}{\sqrt{(1-\rho^2)}} - \rho\right). \end{aligned} \tag{2.11}$$

Except for the trivial case of $a = 0$ (i.e. median dichotomy) and $\rho = 0$, equation (2.10) will not generally be satisfied. We can see this by noting that equation (2.10) will be approximately equal if

$$\pi^{\boldsymbol{a}}_{---} \approx 1 - \pi^{\boldsymbol{a}}_{---} - 3\Phi(a) + 3\pi^{\boldsymbol{a}}_{--},$$

since the other two terms in equation (2.10) are of $O(a^3)$ and will thus be sufficiently small. Using the approximations from equation (2.11) in this equation requires

$$2P_3 - 3P_2 + \frac{3a}{\sqrt{(2\pi)}}\left(2P_2\frac{\rho}{1+\rho} - 1\right) + \frac{3a^2}{2\pi}\left\{\frac{\rho\left[1-\sqrt{(1-\rho^2)}\right]}{\sqrt{(1-\rho^2)}}\right\} \approx 1 - 3\Phi(a), \tag{2.12}$$

and it is clear that this approximation is not valid. Only the right-hand side of the approximation is a function of $\rho$ and thus as $|\rho| \to 1$, the approximation will be less and less accurate. Thus for small correlations, it is true that the three-factor interaction is approximately zero, but as the correlation increases, then the three-factor interaction no longer remains even approximately constant. In fact, as $\rho \to 1$ and $a$ increases, the three factor interaction tends to negative infinity. This can be seen using the numerical algorithm of Schervish (1984) to produce exact values of the three-factor interaction in the symmetrical case for various values of the correlation $\rho$ and cut-point $a$; these are given in Table 2.3 and it is quite clear from this table that the three-factor interaction is not constant.

Extending the above discussion to the non-symmetrical case is quite direct and the same result applies. Equation (2.12) remains in the same general form, although if the correlations and cut-points are not equal, the algebraic form becomes more complex. However, the left-hand side of the approximation will remain a function of the correlations and the cut-points, and the right-hand side will only be a

Table 2.3: Calculated values of the three-factor interaction in the symmetrical case.

| | Cut-point ($a$) | | | | |
|---|---|---|---|---|---|
| Correlation ($\rho$) | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 |
| 0.2 | -0.010 | -0.020 | -0.030 | -0.041 | -0.051 |
| 0.4 | -0.030 | -0.051 | -0.083 | -0.117 | -0.139 |
| 0.6 | -0.051 | -0.105 | -0.151 | -0.211 | -0.261 |
| 0.8 | -0.083 | -0.163 | -0.248 | -0.329 | -0.400 |
| 0.9 | -0.105 | -0.198 | -0.315 | -0.416 | -0.528 |

function of the cut-points. As with the symmetrical case, as the cut-points and correlations increase, then the three-factor interaction no longer remains even approximately constant.

### 2.3.2 Estimating the Three-Factor Interaction

In the previous discussion it was shown that in general the three-factor interaction is not constant. It is of interest, therefore, to look at the estimation of the three-factor interaction and to determine the size of sample required to statistically detect this interaction. The usual estimator of the three-factor interaction (Bishop et al., 1975, Chapter 2) is

$$\hat{\Delta}_3 = \log\{\mathrm{OR}(I_1, I_2|I_3 = 1)\} - \log\{\mathrm{OR}(I_1, I_2|I_3 = -1)\},$$

which can be re-written as

$$\hat{\Delta}_3 = \Sigma_o \log(\hat{\pi}_{ijk}) - \Sigma_e \log(\hat{\pi}_{ijk}),$$

where $\Sigma_o, \Sigma_e$ are over odd and even combinations respectively. The asymptotic variance of this estimator is

$$\begin{aligned} \mathrm{Var}(\hat{\Delta}_3) &= \frac{1}{n} \sum \frac{1}{\pi_{ijk}} \\ &= \frac{c_3(\boldsymbol{\pi})}{n}. \end{aligned}$$

Table 2.4: Sample size required to detect a three-factor interaction in the symmetrical case calculated using equation (2.13).

| | Cut-point ($a$) | | | | |
|---|---|---|---|---|---|
| Correlation ($\rho$) | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 |
| 0.2 | 809800 | 207870 | 96548 | 57774 | 40048 |
| 0.4 | 96399 | 24574 | 11283 | 6643 | 4509 |
| 0.6 | 33232 | 8439 | 3850 | 2247 | 1507 |
| 0.8 | 18051 | 4573 | 2078 | 1206 | 804 |
| 0.9 | 15839 | 4010 | 1820 | 1054 | 701 |

For a three-factor interaction to be detectable, we require that the sample size $n$ satisfies the following condition

$$\tilde{n} \approx \frac{c_3(\boldsymbol{\pi})}{\hat{\Delta}_3^2}, \tag{2.13}$$

and if we assume that the estimator of the three-factor interaction is asymptotically Normally distributed, we require

$$\tilde{n} \approx \frac{4c_3(\boldsymbol{\pi})}{\hat{\Delta}_3^2},$$

to allow us to detect statistical significance with appropriate power.

Although it is possible to use the approximation formulæ in equation (2.13), little insight will be gained into the sample size required to detect a three-factor interaction and in particular, the algebra becomes very cumbersome. It is more informative to consider using the numerical algorithm of Schervish (1984) to calculate this sample size in a number of situations. Firstly, consider the symmetrical case where $a_r = a$ and $\Sigma = \rho J + (1-\rho)I$. The sample sizes required to allow us to detect a three-factor interaction for various values of the correlation co-efficient $\rho$ and cut-point $a$ are given in Table 2.4. The values in this table are calculated using equation (2.13), and thus, to be able to statistically detect the three-factor interaction in these cases, the values need to be increased by a factor of 4.

In the non-symmetrical case, we can perform similar sample size calculations in a number of illustrative situations. It is impossible to enumerate completely all

possible situations but the results given in Table 2.5 provide a set of illustrative examples. As with the symmetrical case, the sample sizes are calculated using equation (2.13), and thus require to be increased by a factor of 4 in order for us to be able to statistically detect the interaction.

Table 2.5: Sample size required to detect a three-factor interaction in the non-symmetrical case with 4 correlation structures and 3 cut-point structures calculated using equation (2.13). NA indicates a sample size in excess of 1000000.

| | $\boldsymbol{a} = (a_1, a_1, a_1)$ | | | | $\boldsymbol{a} = (a_1, a_1, -a_1)$ | | | | $\boldsymbol{a} = (a_1, a_1 + 0.3, -a_1 + 0.3)$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Correlation Structure | | | | Correlation Structure | | | | Correlation Structure | | | |
| Cut-point $a_1$ | $1^a$ | $2^b$ | $3^c$ | $4^d$ | $1^a$ | $2^b$ | $3^c$ | $4^d$ | $1^a$ | $2^b$ | $3^c$ | $4^d$ |
| -0.5 | 14710 | 511 | 15179 | 961 | 1368 | 595 | 2719 | 1477 | 1437 | 12153 | 1794 | NA |
| -0.4 | 21849 | 544 | 22497 | 732 | 1639 | 423 | 3650 | 862 | 1339 | 4827 | 1826 | 462283 |
| -0.3 | 37351 | 703 | 38392 | 696 | 2363 | 381 | 5765 | 647 | 1380 | 2177 | 1983 | 115181 |
| -0.2 | 81734 | 1240 | 83904 | 942 | 4569 | 496 | 11919 | 711 | 1610 | 1121 | 2324 | 33437 |
| -0.1 | 321537 | 4253 | 329825 | 2665 | 16673 | 1373 | 45311 | 1715 | 2212 | 666 | 2991 | 11330 |
| 0 | NA | NA | NA | NA | NA | NA | NA | NA | 3891 | 464 | 4358 | 4497 |
| 0.1 | 321538 | 4253 | 329825 | 2665 | 16673 | 1373 | 45311 | 1715 | 10980 | 393 | 7625 | 2104 |
| 0.2 | 81734 | 1240 | 83905 | 942 | 4569 | 496 | 11919 | 711 | 189918 | 438 | 18458 | 1175 |
| 0.3 | 37351 | 703 | 38392 | 696 | 2363 | 381 | 5765 | 647 | 38745 | 772 | 106149 | 804 |
| 0.4 | 21849 | 544 | 22497 | 732 | 1639 | 423 | 3650 | 862 | 7043 | 4378 | 685606 | 713 |
| 0.5 | 14710 | 511 | 15179 | 961 | 1368 | 595 | 2719 | 1477 | 3219 | 10871 | 36187 | 950 |

[a] $\rho_{12} = \rho_{13} = 0.6$, $\rho_{23} = 0.2$
[b] $\rho_{12} = \rho_{13} = 0.6$, $\rho_{23} = -0.2$
[c] $\rho_{12} = 0.8$, $\rho_{13} = 0.4$, $\rho_{23} = 0.2$
[d] $\rho_{12} = 0.8$, $\rho_{13} = 0.4$, $\rho_{23} = -0.2$

Except in a limited number of circumstances, most of which are empirically unlikely, we are not going to be able to detect a three-factor interaction, since the required sample size is far greater than those usually seen in real data problems. The sample sizes where we are most likely to be able to detect a three-factor interaction occur with the 'anomalous' case with one negative $\rho$. This therefore suggests that, in the three-dimensional case, both the quadratic exponential binary distribution and the multivariate normal dichotomy will produce very similar results.

## 2.4 Discussion

This chapter has provided a simple approximation to the orthant probabilities in the case of approximate median dichotomy of a two- and three-dimensional multivariate normal distribution. The approximation was shown to perform well in a number of situations; those situations where the approximation performed poorly are unlikely to be seen in real data problems. Although it is possible to extend these approximations to higher dimensions, the accuracy will decrease because the approximation formula for the four-dimensional orthant probabilities is most accurate in the equicorrelated case (Steck, 1962). In the case of an unequal correlation matrix the results are less accurate, and attempting to use these in a Taylor series expansion would thus give poor results.

The main question that this chapter set out to address was whether there was any difference between the quadratic exponential binary distribution and the multivariate normal dichotomy. An investigation, using the derived approximations, showed that there is a difference in the properties of these two distributions. The quadratic exponential binary distribution requires that the three-factor (and higher order) interactions are zero whereas the multivariate normal dichotomy places no such constraints on the interaction terms, but they are determined by $\Sigma$ and hence are not free parameters. The approximation formula showed that in some situations the three-factor interaction in the multivariate normal dichotomy is approximately zero, but that in general it is not. However, it was also shown

that to detect interactions of this magnitude would require very large sample sizes. In most cases, the sample sizes observed in practice would be too small to be able to detect statistically a three-factor interaction.

This result suggests that in most situations with $q = 3$ the two distributions will give broadly similar results. This result has only been shown in the three-dimensional case, although it would be most unlikely that if one moved to a higher dimensional situation the sample size required to detect order three (or higher order) interactions would decrease. However, it is certainly possible that in higher dimensions there may be much more different but this would require further investigation. In terms of a graphical model, this result suggests that one would make the same inference, in terms of marginal and conditional independence relationships, irrespective of which distribution was used in the modelling procedure.

There is, however, one question remaining. In the case of a three-dimensional distribution there are seven independent probabilities and both the quadratic exponential binary and dichotomised Gaussian distributions have only six parameters. We know that the quadratic exponential binary distribution requires that the three-factor interaction is zero, and this chapter has shown that this is not the case for the dichotomised Gaussian distribution. Therefore, what is the corresponding constraint for the dichotomised Gaussian distribution? However, it is still not known what the corresponding constraint is; it may be that it has no direct interpretation. There appears to be no obvious condition placed on the cell probabilities, and it is therefore likely that the constraint imposed by the dichotomised Gaussian distribution is simply a functional relationship in terms of the cell probabilities, although this requires further investigation.

# Chapter 3

# Models For Multivariate Discrete Data

## 3.1 Introduction

A topic of great interest, particularly in the area of association models, is that of modelling multivariate responses with respect to potential explanatory variables. For continuous data there is a considerable literature on the subject, particularly with regard to multivariate Gaussian distributions and multivariate linear models. Where there are departures from normality, there exists a good deal of research for the continuous case; although if we turn our attention to discrete responses, the literature is less comprehensive and there remains no agreement between authors as to the relative merits of the proposed techniques. The two main techniques for the analysis of multivariate discrete data are:

- **Conditional Modelling:** specifies the conditional distribution of each component of the response variable, given the covariates and the remaining components of the response.
- **Marginal Modelling:** specifies marginal distributions for components of the response, given the covariates, and is more useful if one is interested in studying the effect of the covariates given the response.

One of the major drawbacks of conditional models is that they are dependent upon the dimension of the response, $y$. Thus if we exclude components of $y$, we will have different kinds of marginal distributions for the remaining components – the models cannot be considered 'upwardly compatible'. However they have achieved quite extensive acceptance in the literature, due to their nice theoretical properties, and ease of specification and estimation.

Marginal models were initially disliked due to the computational problems associated with them; however these have been addressed in a number of papers, particularly those by Liang & Zeger (1986) and Zeger & Liang (1986) who introduced marginal modelling techniques for longitudinal data. Zeger et al. (1988) also introduced a marginal model utilising the theory of generalised estimating equations for the analysis of longitudinal data. A more general technique was proposed by Liang et al. (1992) for multivariate categorical data, although this also used the theory of estimating functions. A detailed insight into marginal modelling was given by Glonek & McCullagh (1995), who proposed a model for multivariate categorical data which is then fitted using a standard maximum likelihood procedure.

Grizzle et al. (1969) set out a specification for analysing discrete data collected on several populations, using the following transformation

$$F(\pi) = X\beta,$$

where $F$ and $X$ are arbitrarily chosen to suit the purposes of the analysis and therefore include both conditional and marginal models, and $\pi$ is the vectorised form of the table of classification probabilities arranged in lexicographical order. They noted that if logarithmic relationships were considered, then one could refine $F$ and use the transformation

$$F(\pi) = K \log A\pi = X\beta. \tag{3.1}$$

This particular transformation has been reviewed by a number of authors and it will be the subject of the remainder of this chapter. Firstly we consider the class of multivariate logistic models proposed by Glonek & McCullagh (1995). We

then consider a non-iterative alternative to the maximum likelihood estimates of the parameters, before finally extending this approximate procedure to a more general class of models.

## 3.2 Multivariate Logistic Regression

Glonek & McCullagh (1995) introduce the following class of multivariate logistic regression models for an arbitrary number of discrete ordinal or nominal response variables. They define the multivariate logistic transformation to be of the form

$$\eta = C^T \log(L\pi) = X\beta, \tag{3.2}$$

where $X$ is the design matrix, $\beta$ are the parameters to be estimated and $L$ and $C$ are matrices defined recursively as follows. Initially set $L_0 = C_0 = (1)$ and suppose we have the matrices $L_d$ and $C_d$ defined for the $d$-dimensional case, where $d$ is the dimension of the response vector. We wish to add an extra response variable, $A_{d+1}$ which has dimension $r$. The new $L$ and $C$ matrices are defined by the following rules, where $1_p$ represents a $p$ dimensional vector of 1's

$$L_{d+1} = \begin{pmatrix} L_d \otimes 1^T_{r(d+1)} \\ L_d \otimes \bar{L} \end{pmatrix} \quad \text{and} \quad C_{d+1} = \begin{pmatrix} C_d & 0 \\ 0 & C_d \otimes \bar{C} \end{pmatrix},$$

where if $A_{d+1}$ is nominal we define

$$\bar{C} = \begin{pmatrix} I_{r-1} \\ -1^T_{r-1} \end{pmatrix} \quad \text{and} \quad \bar{L} = I_r, \tag{3.3}$$

else if $A_{d+1}$ is ordinal we define

$$\bar{C} = I_{r-1} \otimes \begin{pmatrix} 1 \\ -1 \end{pmatrix} \quad \text{and} \quad \bar{L} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 1 & \dots & 1 & 1 \\ 1 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 1 & 1 \\ \vdots & & & & & \vdots \\ 1 & 1 & 1 & \dots & 1 & 0 \\ 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}. \tag{3.4}$$

As an illustration, consider $d = 2$ with $r_1 = r_2 = 2$. Applying the above transformation to the vector of cell probabilities $\pi = (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})$ results in the following transformed cell probabilities

$$\left(\log \pi_{..}, \text{logit}\pi_{1.}, \text{logit}\pi_{.1}, \log \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}\right),$$

where for example, $\pi_{1.}$ indicates summation over the second subscript. If we now consider the case where $d = 2$ and $r_1 = 2, r_2 = 3$ and it is assumed that the latter of the variables is ordinal, then applying the transformation defined above to the vector of cell probabilities results in

$$\left(\log \pi_{..}, \text{logit}\pi_{1.}, \text{logit}\pi_{.1}, \text{logit}(\pi_{.1} + \pi_{.2}), \log \frac{\pi_{11}(\pi_{22} + \pi_{23})}{\pi_{21}(\pi_{12} + \pi_{13})}, \log \frac{\pi_{23}(\pi_{11} + \pi_{13})}{\pi_{13}(\pi_{21} + \pi_{22})}\right).$$

The estimation of $\beta$ is by classical maximum likelihood using the Fisher-scoring algorithm to form $\hat{\beta}$. One of the problems with this method is that it is very computer intensive since it requires two iterative algorithms. The first is used to estimate $\hat{\beta}$ and the second to invert the multivariate logistic transformation as it is not analytically invertible. It is, however, possible to use the technique proposed by Cox & Wermuth (1990) to approximate the maximum likelihood estimator in a non-iterative manner.

### 3.2.1 Approximate Maximum Likelihood Estimation

Cox & Wermuth (1990) discuss how to obtain approximations to maximum likelihood estimators when we can specify an extended and reduced model. The extended model is defined by the parameter $\phi = (\theta, \gamma)$, where $\theta$ and $\gamma$ can be either scalar or vector parameters. It is assumed that under the extended model, it is possible to calculate the maximum likelihood estimates directly and to obtain their asymptotic covariance matrix. We obtain the reduced model by fixing $\gamma$ at some given point, $\gamma_0$. If we let the maximum likelihood estimates under the extended model be $\hat{\phi} = (\hat{\theta}_e, \hat{\gamma}_e)$ with asymptotic covariance matrix

$$\text{Cov}(\hat{\phi}) = \begin{bmatrix} \sum_{\theta\theta} & \sum_{\theta\gamma} \\ \sum_{\gamma\theta} & \sum_{\gamma\gamma} \end{bmatrix},$$

then we observe that the approximate maximum likelihood estimates in the reduced model are $\tilde{\theta_r} = \hat{\theta}_e - \sum_{\theta\gamma}\sum_{\gamma\gamma}^{-1}(\hat{\gamma}_e - \gamma_0)$. The approximate covariance matrix for the parameter estimates in the reduced model can be calculated using $\mathrm{Cov}(\tilde{\theta_r}) = \sum_{\theta\theta} - \sum_{\theta\gamma}\sum_{\gamma\gamma}^{-1}\sum_{\gamma\theta}$. It is this technique that we use to fit the multivariate logistic model.

### 3.2.2 Fitting a Multivariate Logistic Model

If we consider response variables $A_1, \ldots, A_n$, such that $A_i$ is either a nominal or ordinal variable with dimension $r_i$, then our overall response vector $\pi$, has dimension $r_1 r_2 \ldots r_n$. The multivariate logistic transformation is defined in equation (3.2), and for each combination of explanatory variables $X_i$ we observe a table of multinomial frequencies, $y_i \sim M(n_i, \pi_i)$. One possible extended model to consider here is the saturated model which allows the cell probabilities to vary freely. Therefore under our extended model, we fit the observed cell proportions to form $\hat{\pi_e}$. If we consider

$$\hat{\eta}_e = C^T \log(L\hat{\pi_e}) = Z\hat{\phi},$$

where Z is a partitioned matrix such that $Z = (X|\,\mathrm{I})$, and $\hat{\phi} = (\hat{\beta}_e, \hat{\gamma}_e)^T$ is the maximum likelihood estimate under the extended model, the multivariate logistic model is achieved when we set $\gamma_e = 0$. To avoid problems with cell counts being zero, it appeared useful to add 0.5 to each cell, thus increasing the overall sample size by 50% of the total number of cells.

To utilise the theory of Section 3.2.1, we need the asymptotic covariance matrix of $\hat{\pi_e}$. This can be calculated by assuming that each cell of the $i$th contingency table has an independent Poisson distribution with rate $\kappa_i\pi_{ij}$, the estimate $\widehat{\kappa_i\pi_{ij}}$ is the actual count in each cell (Cox & Snell, 1989, Chapter 4). Here $\hat{\pi}_{ij}$ is the estimated proportion in the cell, and $\hat{\kappa}_i$ is the total number of observations in the $i$th table. Hence

$$\mathrm{Var}(\hat{\pi}_{ij}) = \frac{\hat{\pi}_{ij}}{\hat{\kappa}_i} \Rightarrow \mathrm{Cov}(\hat{\pi}_i) = \mathrm{Diag}\left(\frac{\hat{\pi}_{ij}}{\hat{\kappa}_i}\right)$$

and therefore

$$\mathrm{Cov}(\hat{\pi}_e) = \mathrm{B}(\mathrm{Cov}(\hat{\pi}_i)),$$

where B is a block diagonal matrix.

We consider applying the multivariate logistic transformation to the data. If we let the Jacobian of this transformation be $J$, then the asymptotic covariance matrix of $\hat{\eta}_e$ is $\mathrm{Cov}(\hat{\eta}_e) = J^T \mathrm{Cov}(\hat{\pi}_e) J = K$, say. If we let $L$ be the inverse of $Z$ obtained by suitable row-reduction (and re-sizing) of the matrix $Z$, then the asymptotic covariance matrix of $\hat{\phi}$ is $\mathrm{Cov}(\hat{\phi}) = LKL^T$. Thus we calculate the estimator $\tilde{\beta}$ and its approximate covariance matrix, using the method described in Section 3.2.1.

Once we have the approximate maximum likelihood estimates, we need to invert the multivariate logistic transformation so that we can obtain the estimated cell probabilities/fitted values. An algorithm was given in Glonek & McCullagh (1995) to perform this inversion; however there is a typographical error in this algorithm and therefore a corrected and slightly modified version is given here. To invert $\eta = C^T \log(L \exp \nu)$, we can use a modified Newton-Raphson type procedure by applying the following algorithm:

1. Begin with an initial approximation $\nu_0$ such that $1^T \exp(\nu_0) = 1$

2. Take

$$\nu_n = \nu_{n-1} - (C^T D_{n-1}^{-1} L \mathrm{Diag}(\exp \nu_{n-1}))^{-1} (C^T \log(L \exp \nu_{n-1}) - \eta)$$

   where $D_{n-1} = \mathrm{Diag}(L \exp \nu_{n-1})$

3. Normalise the result $\nu_n$ such that $1^T \exp(\nu_n) = 1$ and iterate until convergence.

### 3.2.3 Examples

We now consider applying the technique proposed above and the original version proposed in Glonek & McCullagh (1995) to some data sets. The *Mathematica*

code used to perform the calculations described above is given in the Appendix and the numerical procedure was implemented using Dr G Glonek's C code. In all of these examples, no goodness of fit statistics are presented since the models fitted in the following examples are equivalent to previous models reported in the literature. In addition the approximate maximum likelihood procedure produces goodness of fit statistics which are virtually equivalent to the maximum likelihood procedure.

### Coal Miners Data

The first example used to illustrate the above technique is the coal-miners data (described in, McCullagh & Nelder, 1989, pp 230-232), which relates breathlessness and wheeze and their interaction to age in a group of coal-miners. We take breathlessness ($a$) and wheeze ($b$) as the joint response, and age as the explanatory variable. We use the mid-points of each age group and transform, so that $x = (\text{age} - 42)/5$, to come up with categories $x = -4, -3, -2, -1, 0, 1, 2, 3, 4$. We fit the following model to the data set, shown in Table 3.1.

$$\begin{aligned} \eta_a &= \beta_0 + \beta_1 x, \\ \eta_b &= \beta_2 + \beta_3 x, \\ \eta_{ab} &= \beta_4 + \beta_5 x. \end{aligned}$$

The parameter estimates obtained by both methods are given in Table 3.2.

As we can see from Table 3.2 the approximation procedure proposed above produces estimates generally within one-tenth of the standard error of those calculated using the iterative approach to maximum likelihood estimation. Furthermore the standard errors obtained by the non-iterative procedure are almost identical to those obtained using the iterative procedure.

This model implies that for the population in this study the estimated odds of contacting breathlessness increase by a factor of 1.67 per unit increase in $x$ or equivalently an annual odds increase of 1.11 (or 11%). The corresponding annual increase for contracting wheeze is 6.7%. The observed decline in odds-ratio with

Table 3.1: Coal-miners who are smokers without radiological pneumoconiosis, classified by age, breathlessness and wheeze.

| Age Group | Breathlessness | | No breathlessness | |
|---|---|---|---|---|
| | Wheeze | No Wheeze | Wheeze | No Wheeze |
| 20-24 | 9 | 7 | 95 | 1841 |
| 25-39 | 23 | 9 | 105 | 1654 |
| 30-34 | 54 | 19 | 177 | 1863 |
| 35-39 | 121 | 48 | 257 | 2357 |
| 40-44 | 169 | 54 | 273 | 1778 |
| 45-49 | 269 | 88 | 324 | 1712 |
| 50-54 | 404 | 117 | 245 | 1324 |
| 55-59 | 406 | 152 | 225 | 967 |
| 60-64 | 372 | 106 | 132 | 526 |

Table 3.2: Parameter estimates & S.E.'s for coal-miners data.

| | Iterative Procedure | | Non-Iterative Procedure | |
|---|---|---|---|---|
| Parameter | Estimate | SE | Estimate | SE |
| $\beta_0$ | -2.2625 | 0.02989 | -2.2558 | 0.02978 |
| $\beta_1$ | 0.5145 | 0.01207 | 0.5123 | 0.01204 |
| $\beta_2$ | -1.4878 | 0.02056 | -1.4871 | 0.02055 |
| $\beta_3$ | 0.3254 | 0.00887 | 0.3253 | 0.00887 |
| $\beta_4$ | 3.0219 | 0.06973 | 3.0201 | 0.06944 |
| $\beta_5$ | -0.1314 | 0.02844 | -0.1308 | 0.02834 |

Table 3.3: Joint distribution of visual impairment by age and race combination taken from the Baltimore eye survey study.

| Eye | | Caucasian | | | | African-American | | | |
|---|---|---|---|---|---|---|---|---|---|
| Left | Right | 40-50 | 51-60 | 61-70 | 70+ | 40-50 | 51-60 | 61-70 | 70+ |
| - | - | 602 | 541 | 752 | 606 | 729 | 551 | 452 | 307 |
| - | + | 15 | 16 | 37 | 67 | 21 | 23 | 21 | 37 |
| + | - | 11 | 15 | 31 | 60 | 19 | 24 | 22 | 29 |
| + | + | 4 | 9 | 11 | 79 | 10 | 14 | 28 | 56 |

respect to age is a curious feature of these data that may be attributable to censoring or study design.

**Baltimore Eye Data**

We consider the example analysed by Liang et al. (1992) on the results of the Baltimore Eye Survey, originally reported by Tiesch et al. (1990). We have a bivariate binary response which indicates whether there is visual impairment in the left ($a$) and/or the right ($b$) eye – the data are presented in Table 3.3. The explanatory variables considered are race and age – there are four categories for age and two for race. We encode age by $A = (\text{age} - 60)$, to come up with categories $A = -15, -5, 5, 15$. We encode race as 0 or 1 representing Caucasian and African Americans respectively, and consider fitting the following model:

$$\begin{aligned}
\eta_a &= \beta_0 + \beta_1 A + \beta_2 A^2 + \beta_3 R + \beta_4 (A \times R) + \beta_5 (A^2 \times R), \\
\eta_b &= \beta_6 + \beta_7 A + \beta_8 A^2 + \beta_9 R + \beta_{10} (A \times R) + \beta_{11} (A^2 \times R), \\
\eta_{ab} &= \beta_{12} + \beta_{13} R.
\end{aligned}$$

This model was suggested partially by Liang et al. (1992), and partially by some preliminary data analysis. From fitting this model using both procedures we obtain the estimates presented in Table 3.4.

We again note that both parameter and standard error estimates from the pro-

Table 3.4: Parameter estimates & S.E.'s for visual impairment data.

| | Iterative Procedure | | Non-Iterative Procedure | |
|---|---|---|---|---|
| Parameter | Estimate | SE | Estimate | SE |
| $\beta_0$ | -3.1810 | 0.14242 | -3.1710 | 0.13708 |
| $\beta_1$ | 0.0664 | 0.00837 | 0.0659 | 0.00789 |
| $\beta_2$ | 0.0026 | 0.00085 | 0.0026 | 0.00082 |
| $\beta_3$ | 0.6708 | 0.18994 | 0.6659 | 0.18479 |
| $\beta_4$ | -0.0060 | 0.01095 | -0.0055 | 0.01047 |
| $\beta_5$ | -0.0018 | 0.00115 | -0.0018 | 0.00112 |
| $\beta_6$ | -3.0642 | 0.13565 | -3.0730 | 0.13212 |
| $\beta_7$ | 0.0621 | 0.00774 | 0.0619 | 0.00741 |
| $\beta_8$ | 0.0027 | 0.00081 | 0.0027 | 0.00078 |
| $\beta_9$ | 0.5128 | 0.18578 | 0.5267 | 0.18193 |
| $\beta_{10}$ | -0.0002 | 0.01034 | -0.0001 | 0.00999 |
| $\beta_{11}$ | -0.0012 | 0.00111 | -0.0013 | 0.00109 |
| $\beta_{12}$ | 2.4581 | 0.16812 | 2.4598 | 0.16361 |
| $\beta_{13}$ | 0.4460 | 0.24568 | 0.4394 | 0.24047 |

cedure proposed above agree very closely with the estimates from the iterative procedure, most differences being much less than one-tenth of the standard error.

This model suggests there is a quadratic increase in visual impairment with age in both the left and right eyes marginally, with the substantial increase in risk after age 60, although there is not a significant race/age interaction. There is however a significant difference between races in both eyes with African-Americans having significantly more visual impairment than Caucasians. The increase in the odds ratio is dependent only upon race and suggests that the odds of developing impairment in both eyes is more likely to occur in African-Americans compared to Caucasians.

### Patients' Pain Level Data

The final example to which we consider applying this technique is on the relationship between patients' pain level and their medication requirements in a randomised clinical trial on surgery for duodenal ulcer. This trial was reported in Price et al. (1970) and subsequently analysed by Williams & Grizzle (1972) and Dale (1986). The data for this example consist of the type of operation each patient received, along with their pain level and medication requirements. We study the joint distribution of pain and medication using operation type as an explanatory variable. For this example we take both of the responses to be of ordinal type, allowing the correct contrasts to be generated using the appropriate $C$ and $L$ matrices. The data are presented in Table 3.5.

Following Dale (1986) and some preliminary data analysis, we fit the following model

$$\begin{aligned} \eta_i &= \beta_i \text{ for } i = 1, \ldots, 5, \\ \eta_j &= \beta_j + \beta_9 x \text{ for } j = 6, 7, 8, \\ \eta_k &= \beta_{k-3} + \beta_9 x \text{ for } k = 9, 10, 11, \end{aligned}$$

where the explanatory variable $x$ is an encoded variable contrasting operation type, $x = 0$ for operations VP, VA and RA and $x = 1$ for operation VH. The pa-

Table 3.5: Number of patients' classified by type of operation, level of pain and medication requirements.

| Operation | Pain level | Medication Requirements | | | |
|---|---|---|---|---|---|
| | | Never | Seldom | Occasionally | Regularly |
| VP | None | 170 | 7 | 8 | 0 |
| | Slight | 18 | 5 | 8 | 3 |
| | Moderate | 7 | 0 | 4 | 14 |
| VA | None | 170 | 7 | 5 | 2 |
| | Slight | 22 | 7 | 8 | 1 |
| | Moderate | 8 | 1 | 8 | 9 |
| VH | None | 176 | 8 | 5 | 2 |
| | Slight | 26 | 6 | 5 | 5 |
| | Moderate | 14 | 3 | 2 | 9 |
| RA | None | 181 | 6 | 6 | 2 |
| | Slight | 17 | 12 | 7 | 3 |
| | Moderate | 10 | 2 | 3 | 11 |

Table 3.6: Parameter estimates & S.E.'s for pain data.

| | Iterative Procedure | | Non-Iterative Procedure | |
|---|---|---|---|---|
| Parameter | Estimate | SE | Estimate | SE |
| $\beta_1$ | 1.0821 | 0.07120 | 1.0880 | 0.07134 |
| $\beta_2$ | 2.1565 | 0.10144 | 2.1215 | 0.10008 |
| $\beta_3$ | 1.4250 | 0.07913 | 1.4339 | 0.07940 |
| $\beta_4$ | 1.9004 | 0.09299 | 1.8612 | 0.09174 |
| $\beta_5$ | 2.7152 | 0.12917 | 2.6204 | 0.12398 |
| $\beta_6$ | 2.7450 | 0.19432 | 2.6844 | 0.19378 |
| $\beta_7$ | 2.9979 | 0.20897 | 2.9922 | 0.20594 |
| $\beta_8$ | 3.7371 | 0.29392 | 3.7565 | 0.28404 |
| $\beta_9$ | -0.7413 | 0.34153 | -0.6352 | 0.33988 |

rameterisation of this model is such that $\eta_1 \ldots \eta_5$ are the marginal logits of the various levels of the response variables, and $\eta_6 \ldots \eta_{11}$ are log-odds ratios marginalised over appropriate levels of the response variables (or GCR's defined by Dale, 1986). The marginal parameters in this model compare each level of the variable with its previous level, for example $\eta_2$ models the marginal logit of slight pain versus no pain. The GCR's can be interpreted as an odds ratio of conditional events, for example the parameter $\eta_6$ represents the odds ratio of medication being never and seldom given pain is none and slight versus the odds ratio of medication being never and seldom given pain is moderate; the other parameters are defined similarly.

The parameter estimates obtained from the procedure suggested above and the iterative procedure are given in Table 3.6. Again we note from Table 3.6 that the non-iterative procedure has estimates that are very close, with only one difference greater than one-half of the standard error of the estimated parameters. The differences here are slightly larger than what was observed in the previous two examples, and this is probably related to having to adjust for zero counts in the observations. These estimates are also in agreement with the results suggested in

Dale (1986) who fits an almost identical model to that considered here.

For this model, the marginal parameters are of less interest than the GCR parameters. For example, in the non-VH patients, the odds on requiring medication at least occasionally given that the pain is 'higher' is about 18 times the odds on requiring medication at least occasionally given that the pain level is 'lower' (here, higher and lower refer to the none, slight and moderate scale). Similarly, for patients never requiring medication using the dichotomy at occasionally/regularly then the fitted odds ratio is approximately 46. This provides evidence that lower pain and infrequent medication requirements occurred together in this study. Further, these data suggest that the association between pain and medication frequency for this model is about half as large for operation VH as for the other operations.

## 3.3 More General Multivariate Models

One of the main drawbacks to the multivariate logistic model, and marginal models in general compared to conditional (log-linear) models, is that it is not possible to determine conditional independence relationships between the observed variables. In the setting of graphical models, we are interested in determining both marginal and conditional independence relationships. In particular, we are not only interested in determining relationships between the response and a set of explanatory variables, but also between components of the response. In this case, models which have both marginal and conditional components will be more appropriate.

### 3.3.1 Models which have Marginal and Conditional Components

A number of recent papers (for example, Zhao & Prentice, 1990; Glonek, 1996), discuss models that lie between the extremes of marginal and conditional models.

The parameterisation considered in the paper by Glonek (1996) is a combination of multivariate logistic contrasts, and the complementary subset of log-linear contrasts. The multivariate logistic contrasts are chosen with respect to the lower order interactions (e.g. all main effects and two-way interactions), and hence the log-linear contrasts consider the higher order interactions (e.g all three-way interactions and higher). Fitzmaurice & Laird (1993) also considered a model that is a mixture of marginal and log-linear parameters, although their models only considered the main effects as marginal components and set all the two-way and higher order interactions to be log-linear. However, as noted by Glonek (1996), this approach is not always easily interpretable, as the models in which researchers are often interested tend to have a combination of both marginal and log-linear parameters, which at least marginally model all two-way interactions; this lack of interpretability was demonstrated using the data from the Six Cities study (Ware et al., 1984). It would therefore seem that in principle, the class of regression models suggested by Glonek (1996) combines the desirable properties of both conditional and marginal models.

The specification of these joint marginal and conditional models proposed by Glonek (1996) is as follows

$$\eta = C_1^T \log(L_1 \pi), \qquad \lambda = C_2^T \log \pi,$$

and in the regression setting, we generally consider the following models, as it would be difficult to contemplate the need for more general models

$$\eta = X_\eta \beta_\eta, \qquad \lambda = X_\lambda \beta_\lambda.$$

In both of these equations, $\eta$ represents a multivariate logistic transformation (the marginal component), while $\lambda$ represents a log-linear transformation (the conditional component). It is clear that these models can be represented in the form of equation (3.2), and can therefore be fitted using the approximation procedure described above with the following alternative definitions of the $C$ and $L$ matrices.

If we have $d$ response variables, $A_1, \ldots, A_d$, having $r_1, \ldots, r_d$ levels respectively, we associate a pair of matrices $\overline{C}_i$ and $\overline{L}_i$ with variable $A_i$. If $A_i$ is nominal the

matrices are given by equation (3.3), whereas if $A_i$ is ordinal the matrices are given by equation (3.4). If we let $G = \{B_1, \ldots, B_g\}$ be any collection of subsets of $D = \{1, \ldots, d\}$ that is closed with respect to set inclusion, then for each such $G$ we define matrices $C_1, C_2$ and $L_1$ as follows.

For $\alpha = 1, 2, \ldots, g$, define the matrices

$$L_{(\alpha)} = \bigotimes_{i=1}^{d} M_{i\alpha} \text{ and } C_{(\alpha)} = \bigotimes_{i \in B_\alpha} \overline{C}_i$$

where

$$M_{i\alpha} = \begin{cases} \overline{L}_i & \text{if } i \in B_\alpha, \\ 1_{r_i}^T & \text{otherwise.} \end{cases}$$

Then set

$$L_1 = \begin{pmatrix} L_{(1)} \\ L_{(2)} \\ \vdots \\ L_{(g)} \end{pmatrix} \text{ and } C_1 = \begin{pmatrix} C_{(1)} & 0 & \ldots & 0 \\ 0 & C_{(2)} & \ldots & 0 \\ \vdots & & & \vdots \\ 0 & \ldots & 0 & C_{(g)} \end{pmatrix}.$$

To specify $C_2$, let

$$K(B) = \bigotimes_{i=1}^{d} J_i(B)$$

where

$$J_i(B) = \begin{cases} \begin{pmatrix} I_{r_i - 1} \\ -1_{r_i - 1}^T \end{pmatrix} & \text{if } i \in B, \\ r_i^{-1} 1_{r_i} & \text{otherwise} \end{cases}$$

for each $B \in 2^D$. Let $\overline{G}^c = \{E_1, E_2, \ldots, E_h\}$ be the complement of $\overline{G}$ in $2^D$ and set

$$C_2 = (K(E_1)\ K(E_2) \ldots K(E_h)).$$

As an illustration of this type of model specification, consider the case of three binary responses, and suppose that we wish to specify that the first order components are marginal components and that all higher order components are log-linear components. The resulting transformation maps

$$\pi \to (\eta^{A_1}, \eta^{A_2}, \eta^{A_3}, \lambda^{A_1 A_2}, \lambda^{A_1 A_3}, \lambda^{A_2 A_3}, \lambda^{A_1 A_2 A_3}),$$

where

$$\begin{aligned}
\eta^{A_1} &= \text{logit}(\pi_{1..}), \\
\lambda^{A_1A_2} &= \log \frac{\pi_{11*}\pi_{22*}}{\pi_{21*}\pi_{12*}}, \\
\lambda^{A_1A_2A_3} &= \log \frac{\pi_{111}\pi_{221}\pi_{212}\pi_{122}}{\pi_{211}\pi_{121}\pi_{112}\pi_{222}},
\end{aligned}$$

and other quantities are defined analogously. In all of these equations . indicates summation over the subscript and $*$ denotes the geometric mean taken over the subscript.

To formulate this model specification in a form similar to equation (3.2), which will allow us to use the same approximate maximum likelihood estimation procedure, we simply set

$$C = \text{Diag}(C_1, C_2), L = \text{Diag}(L_1, \text{I}) \text{ and } X = \text{Diag}(X_\eta, X_\lambda).$$

The results using this model specification and the previous approximate estimation procedure appear to be as good as with the previous model specification which only considered marginal components (see example below).

### 3.3.2 Example

As an example of the application of the approximate maximum likelihood procedure to this more general class of models, we consider the Six Cities Data which was originally reported in Ware et al. (1984), and has since been analysed by Zeger et al. (1988), Fitzmaurice & Laird (1993) and Glonek (1996). The data consist of a binary response indicating the presence or absence of wheeze at ages seven, eight, nine and ten for each of 537 children from Ohio, USA. The explanatory variable in this data set is whether or not the child's mother smoked during the first year of the study. The data are given in Table 3.7.

In the original treatment, both Zeger et al. (1988) and Fitzmaurice & Laird (1993) considered modelling only the first and second order logistic components, and set all higher order interactions to be zero. However, in this example, we fit the same

Table 3.7: The Six Cities data: child's wheeze status.

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

model as in Glonek (1996); the parameterisation of which is

$$\begin{aligned}\eta &= (\eta^{A_1}, \eta^{A_2}, \eta^{A_3}, \eta^{A_4}, \eta^{A_1,A_2}, \eta^{A_1,A_3}, \eta^{A_1,A_4}, \eta^{A_2,A_3}, \eta^{A_2,A_4}, \eta^{A_3,A_4}) \\ \lambda &= (\lambda^{A_1,A_2,A_3}, \lambda^{A_1,A_2,A_4}, \lambda^{A_1,A_3,A_4}, \lambda^{A_2,A_3,A_4}, \lambda^{A_1,A_2,A_3,A_4}),\end{aligned}$$

where

$$\begin{aligned}\eta^{A_i} &= \begin{cases} \mu_i & \text{for non-smoking mothers,} \\ \mu_i + \tau & \text{for smoking mothers,} \end{cases} \\ \eta^{A_i,A_j} &= \alpha \\ \lambda &= 0.\end{aligned}$$

The parameter estimates obtained from the original analysis by Glonek (1996) and those obtained by the approximation procedure described above are given in Table 3.8. The estimates produced by the approximation procedure again compare favourably with those produced by the iterative procedure.

This model suggests that the odds of developing wheeze in children of non-smoking mothers is approximately constant at all ages, while the odds for children in smoking mothers is approximately 30% higher (although this is not statistically

Table 3.8: Parameter estimates & S.E.'s for the Six Cities data.

| | Iterative Procedure | | Non-Iterative Procedure | |
|---|---|---|---|---|
| Parameter | Estimate | SE | Estimate | SE |
| $\mu_1$ | -1.77 | 0.136 | -1.75 | 0.138 |
| $\mu_2$ | -1.68 | 0.133 | -1.67 | 0.129 |
| $\mu_3$ | -1.76 | 0.135 | -1.78 | 0.132 |
| $\mu_4$ | -2.12 | 0.149 | -2.10 | 0.153 |
| $\tau$ | 0.274 | 0.178 | 0.275 | 0.180 |
| $\alpha$ | 2.05 | 0.175 | 2.04 | 0.176 |

significant). The joint parameter $\alpha$ suggests that the marginally, the association between two given measurements is constant, i.e. the evolution over time is stationary. Since all the log-linear parameters in this model were set to zero, this implies a number of conditional independence relationships between the observations at each of the time points. However, these provide little insight into the main objectives of the study.

### 3.3.3 Interpretation & Limitations

A number of problems using this specification were noted by Glonek (1996), the main one being that these models are only weakly upwardly compatible. In addition this class of models is not variation independent. However, they still provide a very useful class of models for analysing multivariate discrete data; allowing more than just a first order marginal component is an important feature for modelling situations. In the case of a graphical model, it is not only of interest to investigate marginal relationships between the response and explanatory variables, but it is also of interest to study the set of relationships between the response variables – this is possible by investigating the higher order conditional components. In the longitudinal data setting, the second order marginal components correspond to the structural evolution of the process over time.

One problem remains. If the aim of the analysis is to investigate conditional relationships, then the obvious way to proceed is to fit a log-linear model. However, if there is also an interest in studying the marginal relationships, or the evolution of the measurements over time, then it is clear that a marginal model should be fitted. Thus in any given situation, one could postulate that the most sensible analysis is to consider fitting a combination of marginal and conditional components, and therefore the class of models discussed in this section is an appropriate one to consider.

## 3.4 Discussion

This chapter has proposed an approximate fitting procedure for a class of models for multivariate discrete responses. This technique was first developed for the class of multivariate logistic models proposed by Glonek & McCullagh (1995) and extended to the class of models proposed by Glonek (1996). The approximation procedure was shown to perform very well compared to the more classical iterative procedure and a number of examples were given. However, the approximation is not limited to these two classes of models; any model that can be formulated as in equation (3.1) can utilise the approximation procedure described in Section 3.2.2. This is therefore a very general approximation procedure, useful in a wide range of situations. In the problems considered in this chapter, the computing time is almost equivalent. However, the approximation provides a considerable computational advantage when fitting models to high-dimensional responses since the computation methods for the solution of sparse matrices are far more efficient than numerical minimisation methods for high dimensional surfaces. In very high dimensional problems, the iterative method becomes prohibitively cumbersome in terms of computing time whereas the approximation presented here is more efficient. In addition in simple problems it is possible to observe the constraints that a model places on the fitted cell probabilities compared to the saturated model.

All the models described in this chapter are related to global association parameters; these are models that look for relationships between variables. An alternative

class of models are the local association models which are interested in the relationships between the levels of a variable when analysed with other variables. A number of authors have considered the case of local association models – for example Goodman (1981, 1991) and Wermuth & Cox (1998) – and also the relationship between local and global models (Dale, 1984). In general local models are useful only for ordered responses. In the case of an arbitrary categorical variable, it is unclear how a local model could be useful, as the results would depend on how the levels of the variables are arranged. These models are particularly useful for ordinal data, as they allow one to identify levels of a variable which can be merged together, thus reformulating the problem in fewer dimensions. Alternatively, this can be interpreted as finding independence relationships between levels of a variable, i.e. finding local independence relationships. This is in contrast to the models previously discussed, where it is only possible to investigate relationships between variables.

Although the specification of local and global models is such that they are designed to test two different sets of hypotheses, the model of Wermuth & Cox (1998) can be formulated in the setting of equation (3.1). Thus the approximation procedure proposed in Section 3.2.2 can also be applied to this local independence model, again offering a major saving in the computational burden required to fit these models.

For any data set, there are thus a number of possibilities for the modelling procedure. For a set of ordered categorical variables, it may initially be profitable to consider a set of local association models in an attempt to reduce the dimensionality of the problem. The analysis can then be extended to consider a global association model to assess the relationships between variables. However, the actual approach adopted for any given problem will be fixed by the questions that require answering by the study.

# Chapter 4

# Measurement Error Analysis

## 4.1 Introduction

An important aspect when fitting a statistical model is to ensure that estimated regression coefficients, which are often used to infer association relationships, are unbiased. If one or more of the explanatory variables in the data is measured with error – a problem often referred to as an errors-in-variables problem – this generally biases estimated regression coefficients. The consequences of measurement error are twofold:

- Firstly, estimated regression coefficients are typically biased. Thus incorrect conclusions about association relationships can be made if measurement error is not taken into account.
- Secondly, if one attempts to correct for measurement error in the estimation procedure then the standard error of the estimated regression coefficients are often larger than they would be if measurement error had not been accounted for.

The extensive literature on this topic is summarised by Fuller (1987), for measurement error in linear regression models, and quite recently, Carroll et al. (1995), review measurement error in a variety of non-linear problems. However, the com-

putations required in this latter works is quite intensive, making it difficult to obtain theoretical insights into the techniques. Although these computer intensive techniques may give more accurate corrections to the problem of measurement error, for the work considered here it is preferable to consider simpler models which allow more direct algebraic and numeric calculations.

In nearly all of the preceding work on this subject, the assumption has been that there are two main types of measurement error model. If we assume that $X_t$ is the true value of the explanatory variable under study and $X_s$ is the observed or surrogate value then we have:

- **Classical Measurement Error** We attempt to measure $X_t$ directly, but are unable to due to errors in the measurement procedure. This equivalently says that $X_s = X_t + \epsilon_{s.t}$, where $\epsilon_{s.t}$ is independent of $X_t$.

- **Berkson Measurement Error** This model is applicable to experiments where we have a pre-determined explanatory variable, e.g. dose of drug, but due to errors in the administration of the drug the patient receives a different amount. Its relevance to observational studies was pointed out by Cochran (1968) and discussed in detail by Reeves et al. (1998). This would be represented by the following model $X_t = X_s + \epsilon_{t.s}$, where $\epsilon_{t.s}$ is independent of $X_s$.

However, nearly all the proposed correction techniques have assumed that we have only one of these two possibilities; in many situations it is necessary to consider a model which combines both types of error model. The most recent paper to consider both these types of error models is Reeves et al. (1998), which proposes a measurement error model of which Berkson and classical measurement error models are special cases. This paper considered both linear and binary regression models and developed a simple procedure for correcting regression coefficients for measurement error. Hence it is possible to use the formulation presented in this paper to investigate some properties of measurement error models.

As already noted, one of the main problems measurement error creates is to decrease the precision of estimated regression coefficients. It is important to understand the relationship between the amount of measurement error and the relative decrease in precision, as this will be an important consideration for determining sample sizes prior to carrying out a study where it is known that an explanatory variable will be measured with error.

There are two ways to estimate a sample size prior to carrying out a study. The first, more classical method involves significance tests and often includes terminology such as 'power', 'significance level' and 'minimal detectable difference'; for a good general introduction see Desu & Raghavarao (1990). Such a classical methodology is now being replaced by sample size calculations based on confidence intervals, discussed in (Cox, 1958, Chapter 8), and by Armitage & Berry (1987) and Greenland (1988). Most work on sample size calculations does not consider the possibility of measurement error in the explanatory variables; see, however Freedman et al. (1990), McKeown-Eyssen & Tibshirani (1994), and Devine & Smith (1998). All these papers have considered the classical approach to sample size calculation; this chapter will investigate the effects that measurement error has on sample size calculations when the sample size calculation is based on confidence intervals.

The chapter is organised as follows: in Section 4.2 the measurement error model and results of Reeves et al. (1998) are reviewed, in Section 4.3 the implications for sample size calculations in the linear model are given. In Section 4.4 sample size implications are given for logistic regression models in both cohort and case-control studies. The chapter concludes with a brief discussion.

## 4.2 Measurement Error Model Specification

The structure of the measurement error model described in Reeves et al. (1998) is

$$\begin{aligned} \boldsymbol{X}_s &= \tilde{\boldsymbol{X}} + \boldsymbol{\epsilon}_s, \\ \boldsymbol{X}_t &= \tilde{\boldsymbol{X}} + \boldsymbol{\epsilon}_t, \end{aligned} \tag{4.1}$$

where $(\tilde{\boldsymbol{X}}, \boldsymbol{\epsilon}_s, \boldsymbol{\epsilon}_t)$ are independent $q \times 1$ random variables with mean vectors $(\boldsymbol{\mu}, \mathbf{0}, \mathbf{0})$ and covariance matrices $(\tilde{\Sigma}, \Sigma_s, \Sigma_t)$. Classical measurement error models are a special case of the above representation and correspond to setting all the elements of $\Sigma_t$ to zero; whereas Berkson measurement error models correspond to setting all the elements of $\Sigma_s$ to zero. Reeves et al. (1998) considered two ways in which to specify the lack of dependence between $(\tilde{\boldsymbol{X}}, \boldsymbol{\epsilon}_s, \boldsymbol{\epsilon}_t)$; they were assumed to be either independently normally distributed random variables, or to be mutually uncorrelated. In what follows in this chapter we will assume the stronger of the two assumptions, i.e. that they are independently normally distributed random variables.

If we are in the linear model setting where we have a response variable $Y$, then our interest is in the relationship between this variable and a set of explanatory variables $\boldsymbol{X}_t$, some of which have may have been measured with error. However, we have observed only the error contaminated variables $\boldsymbol{X}_s$, but we still wish to make inferences with respect to the true variables. Thus we are interested in the regression equation

$$\mathrm{E}(Y|X_t) = \boldsymbol{X}_t \beta_t,$$

where $\beta_t$ is the $q \times 1$ vector of true regression coefficients. Reeves et al. (1998) showed that if we estimate $\beta_s$ from the following regression equation

$$\mathrm{E}(Y|X_s) = \boldsymbol{X}_s \beta_s,$$

then

$$\hat{\beta}_s^T = \hat{\beta}_t^T \Gamma_{t.s} = \hat{\beta}_t^T \tilde{\Sigma} (\tilde{\Sigma} + \Sigma_s)^{-1}. \tag{4.2}$$

If the response variable $Y$ is binary rather than continuous and we are still interested in modelling the relationship between $Y$ and the explanatory variables $\boldsymbol{X}_t$, but we observe only the surrogate values $\boldsymbol{X}_s$, then we are interested in the regression model

$$\Pr(Y = 1|\boldsymbol{X}_t) = \Lambda(\boldsymbol{X}_t\beta_t),$$

where as before, $\beta_t$ is the $q \times 1$ vector of true regression coefficients, and $\Lambda$ denotes the logistic function, $\Lambda(t) = \mathrm{e}^t/(1 + \mathrm{e}^t)$. It was also shown in Reeves et al. (1998) that if, instead of using the true values of the explanatory variables in the above equation, we instead used the surrogate values and estimate $\beta_s$ then

$$\hat{\beta}_s^T = \hat{\beta}_t^T \Gamma_{t.s}(1 + k^2 \hat{\beta}_t^T \Sigma_{t.s} \Sigma_{t.s}^T \hat{\beta}_t)^{-1/2}, \tag{4.3}$$

where $k$ is some constant, $\Gamma_{t.s} = \tilde{\Sigma}(\tilde{\Sigma}+\Sigma_s)^{-1}$ and $\Sigma_{t.s}$, is the conditional covariance matrix of $\boldsymbol{X}_t$ given $\boldsymbol{X}_s$.

Clearly, the algebra is simpler in the case of a linear regression model than for a binary regression model, and thus it is only in the case of a linear regression model that we can easily derive exact results regarding the influence of measurement error on the precision of regression coefficient estimates.

## 4.3 The Linear Measurement Error Model

In Section 4.2 we saw the effect of this particular class of measurement error models on a linear regression relationship. It is fairly clear from equation (4.2) that if we fail to correct for measurement error, we have biased estimates of the regression coefficients. However, if we use an appropriate correction procedure although we obtain unbiased parameter estimates, there will be greater uncertainty associated with them. Therefore, the issue is how much greater is this uncertainty, as this will play an important role when designing studies where it is perceived that measurement error may be a problem.

One possibility for attempting to understand the effects of measurement error is to study the asymptotic relative efficiency (ARE) or Pitman efficiency (Cox &

Hinkley, 1974, Chapter 9) of two hypothesis tests: one in the case where we have observed the explanatory variables without error and thus can directly estimate the regression coefficients; the other in the case where we have had to correct our estimated regression coefficients for the effects of measurement error. Let us call these tests $T_t$ and $T_s$, where $T_t$ is the hypothesis test in the case of no measurement error, and $T_s$ is the hypothesis test in the case where we have corrected for measurement error. In the context of graphical chain models where we are interested in assessing both marginal and conditional independence relationships, we are often interested in testing whether there is no association between two variables: an equivalent test that a particular regression coefficient is zero. Thus we can consider the ARE for these two hypothesis tests for a particular regression coefficient in our model being zero.

Let us assume that our hypothesis tests are asymptotically $N(\mu(\theta), \sigma^2(\theta))$. Then we use the definition of the ARE as

$$\text{ARE}(T_t : T_s) = \left\{ \frac{\mu'_{T_t}(\theta_0)}{\mu'_{T_s}(\theta_0)} \right\}^2 \left\{ \frac{\sigma^2_{T_s}(\theta_0)}{\sigma^2_{T_t}(\theta_0)} \right\},$$

where $\theta_0$ is the null value and $\mu'$ is the derivative of $\mu(\theta)$ with respect to $\theta$.

### 4.3.1 ARE in a One Variable Model

Let us consider the simplest case of the linear model, that is one which has only one explanatory variable which is measured with error. If we could observe the explanatory variable without error our regression equation would be

$$Y = \alpha_t + \beta_t X_t + \zeta$$

and the error term, $\zeta$, would have variance $\sigma^2_\zeta$. Therefore our estimate of the regression coefficient $\beta_t$ would be the usual least squares estimate with variance

$$\frac{\sigma^2_\zeta}{\sum (X_{ti} - \overline{X}_t)^2}.$$

If, however, we only observed an error contaminated version of $X_t$, then when we fit the following model

$$Y = \alpha_s + \beta_s X_s + \zeta_s,$$

we can form an estimate of $\beta_t$ using the results given in equation (4.2). The variance of this estimate is

$$\mathrm{Var}(\hat{\beta}_{tc}) = \frac{\sigma^2_{\zeta_s}}{\sum(X_{si} - \overline{X}_s)^2 \gamma^2_{t.s}},$$

where $\hat{\beta}_{tc}$ is the estimate of $\beta_t$ corrected for measurement error, $\sigma^2_{\zeta_s}$ is the variance of the residual error term $\zeta_s$ and can be shown to equal $\sigma^2_\zeta + \beta^2_t \sigma^2_{t.s}$ (see Reeves et al., 1998, for details), and $\gamma_{t.s}$ is as defined in equation (4.2).

In the case of observing the explanatory variable without error, we can form the test statistic

$$T_t = \frac{\hat{\beta}_t}{\mathrm{Var}(\hat{\beta}_t)^{1/2}},$$

to test the hypothesis that $\beta_t = 0$ (this is the generalised likelihood ratio test that $\beta_t$ is zero). Similarly, in the case of observing error contaminated data, we can form the test statistic

$$T_s = \frac{\hat{\beta}_{tc}}{\mathrm{Var}(\hat{\beta}_{tc})^{1/2}},$$

to again test the hypothesis that $\beta_t = 0$. Hence we calculate

$$\begin{aligned}
\mu_{T_t}(\beta_t) &= \beta_t \left( \frac{\sum(X_{ti} - \overline{X}_t)^2}{\sigma^2_\zeta} \right)^{1/2} \\
\mu_{T_s}(\beta_t) &= \beta_t \left( \frac{\sum(X_{si} - \overline{X}_s)^2 \gamma^2_{t.s}}{\sigma^2_\zeta + \beta^2_t \sigma^2_{t.s}} \right)^{1/2} \\
\sigma^2_{T_t} &= 1 \\
\sigma^2_{T_s} &= 1,
\end{aligned}$$

thus

$$\mu'_{T_t}(\beta_t) = \left( \frac{\sum(X_{ti} - \overline{X}_t)^2}{\sigma^2_\zeta} \right)^{1/2} \tag{4.4}$$

$$\mu'_{T_s}(\beta_t) = \left( \frac{\sum(X_{si} - \overline{X}_s)^2 \gamma^2_{t.s}}{\sigma^2_\zeta + \beta^2_t \sigma^2_{t.s}} \right)^{1/2} - \beta^2_t \gamma_{t.s} \sigma^2_{t.s} \left( \frac{\sum(X_{si} - \overline{X}_s)^2}{\sigma^2_\zeta + \beta^2_t \sigma^2_{t.s}} \right)^{1/2}. \tag{4.5}$$

Using equation (4.4) and equation (4.5) evaluated at the point $\beta_t = 0$, and that asymptotically $\sum(X_{ti} - \overline{X}_t)^2 \to \mathrm{Var}(X_t)/n$ and $\sum(X_{si} - \overline{X}_s)^2 \to \mathrm{Var}(X_s)/n$, we

calculate

$$\begin{aligned} \text{ARE}(T_t : T_s) &= \frac{1}{\gamma_{t.s}^2}\left(\frac{\text{Var}(X_t)}{\text{Var}(X_s)}\right) \\ &= \frac{\tilde{\sigma}^4}{\tilde{\sigma}^4 + \tilde{\sigma}^2(\sigma_s^2 + \sigma_t^2) + \sigma_s^2\sigma_t^2}. \end{aligned} \tag{4.6}$$

In order to investigate the ARE for the one variable model, the ARE surface was plotted (Figure 4.1) by fixing $\tilde{\sigma}^2 = 1$ and allowing $\sigma_t^2/\tilde{\sigma}^2$ and $\sigma_s^2/\tilde{\sigma}^2$ to vary between zero and five. It should be noted from this plot that, as either the amount of classical measurement error or Berkson measurement error increases, the ARE decreases. This implies that to estimate $\beta_t$ with the same precision, the sample size needs to increase accordingly. Another feature of this plot is that, as the amount of background error $\tilde{\sigma}^2$ increases, the ARE also increases for fixed levels of measurement error. This is as one would expect since if the noise associated with a particular explanatory variable is much greater in magnitude than the measurement error, then measurement error has less effect. Although this treatment has only considered testing the hypothesis that $\beta_t = 0$ (since this is equivalent to $\beta_s = 0$), the discussion can easily be extended to other cases. However, although the algebra becomes more complicated, the same general conclusions are reached and in particular, numerical work has shown that as $\beta_t$ increases, the ARE decreases for fixed amounts of measurement error.

### 4.3.2 ARE in a Two Variable Model

To extend the previous discussion, we consider a linear regression model with two explanatory variables. We also consider the special case in which only one of the explanatory variables is measured with error. The specification of the measurement error model is, therefore, as follows:

$$\tilde{\Sigma} = \begin{pmatrix} \tilde{\sigma}_1^2 & \rho\tilde{\sigma}_1\tilde{\sigma}_2 \\ \rho\tilde{\sigma}_1\tilde{\sigma}_2 & \tilde{\sigma}_2^2 \end{pmatrix} \quad \Sigma_s = \begin{pmatrix} \sigma_s^2 & 0 \\ 0 & 0 \end{pmatrix} \quad \Sigma_t = \begin{pmatrix} \sigma_t^2 & 0 \\ 0 & 0 \end{pmatrix}.$$

In this model, we are interested in testing the hypothesis that the estimated regression coefficient for the measurement error contaminated variable is zero.

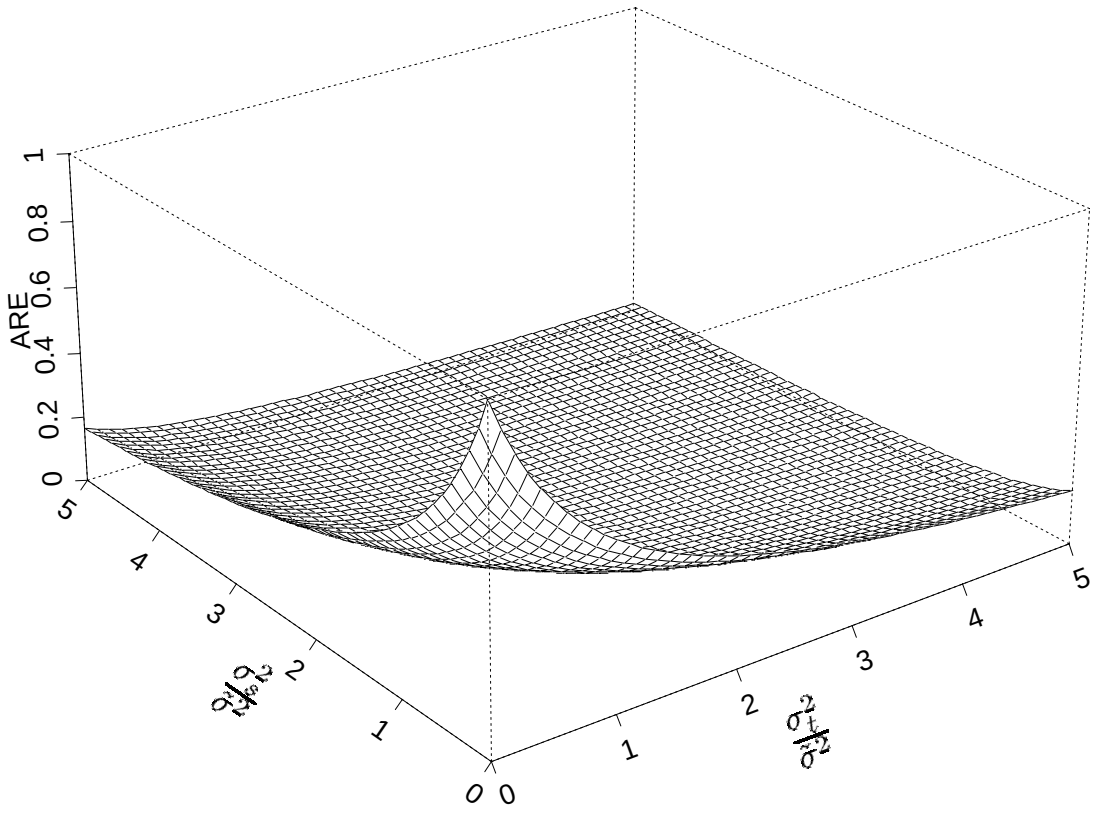



Figure 4.1: Asymptotic relative efficiency surface comparing the hypothesis test that $\hat{\beta}_t = 0$ in error-free and error-contaminated data, in a linear model with only one explanatory variable, fixing $\tilde{\sigma}^2 = 1$.

We think of the measurement-error-free variable as representing a variable that is easy to measure accurately, but in which we are not directly interested; we simply need to adjust for it in the analysis.

We calculate the ARE in a similar way to that for the one-variable model; first noting that in the model with no measurement error

$$\mathrm{Var}(\hat{\beta}_t) = \sigma_\zeta^2 (X_t^T X_t)^{-1},$$

and in the model with measurement error

$$\mathrm{Var}(\hat{\beta}_{tc}) = (\Gamma_{t.s}^{-1})^T (\beta_t^T \{\tilde{\Sigma}[\mathrm{I} - (\mathrm{I} + \tilde{\Sigma}^{-1}\Sigma_s)^{-1}] + \Sigma_t\}\beta_t + \sigma_\zeta)(X_s^T X_s)^{-1}\Gamma_{t.s}^{-1}.$$

We also note that the hypothesis tests in which we are interested are identical to the one variable case and thus the calculations are identical, except that we replace appropriate terms by their vector/matrix equivalents, and differentiate with respect to $\beta_{t1}$. Thus the ARE for this model specification is

$$\mathrm{ARE}(T_t : T_s) = \frac{\tilde{\sigma}_1^4 (1-\rho^2)^2}{(\tilde{\sigma}_1^2 - \tilde{\sigma}_1^2\rho^2 + \sigma_s^2)(\tilde{\sigma}_1^2 - \tilde{\sigma}_1^2\rho^2 + \sigma_t^2)}. \tag{4.7}$$

It is easiest to assess the properties of this ARE by plotting an ARE surface by fixing $\tilde{\sigma}^2 = 1$ and allowing $\sigma_t^2/\tilde{\sigma}^2$ and $\sigma_s^2/\tilde{\sigma}^2$ to vary between zero and five for various values of the correlation $\rho$. Figure 4.2 shows this ARE surface for $\rho = 0.2$ (upper plot) and $\rho = 0.8$ (lower plot).

The results from the two-variable model are similar to those obtained from the one-variable model. From Figure 4.2 we can see that, as either the amount of classical measurement error or Berkson measurement error increases, then the ARE decreases. In addition, as the magnitude of the background error in the explanatory variable increases relative to the amount of measurement error so does the ARE. As one would expect, as the correlation between the error free and error contaminated variable increases, then the ARE decreases more rapidly; due to the correlation between these two variables creating an additional problem in estimating the effect of either variable on the response variable.

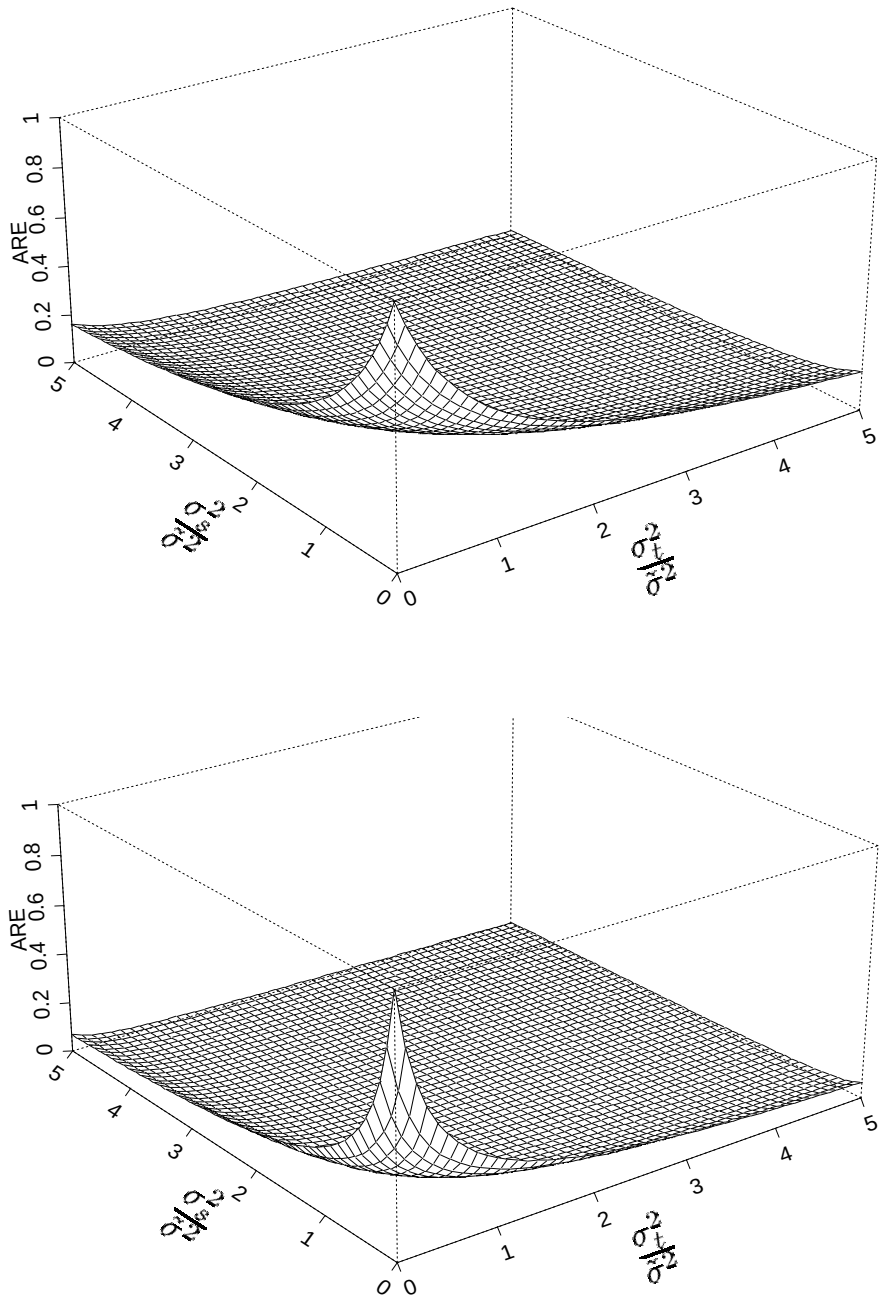



Figure 4.2: Asymptotic relative efficiency surface comparing the hypothesis test that $\hat{\beta}_{t1} = 0$ in error-free and error-contaminated data, in a linear model with two explanatory variables only one of which is measured with error, fixing $\tilde{\sigma}^2 = 1$. The upper plot sets $\rho = 0.2$ and the lower plot sets $\rho = 0.8$.

## 4.4 The Binary Regression Model

As with the linear model case, we see from Section 4.2 that failing to correct for measurement error in the binary case produces biased regression coefficients. Correcting for measurement error produces an unbiased estimate of the regression coefficients although this estimate is less precise than that which would have been estimated using error-free data. From the form of equation (4.3), it is clear that it will be difficult to carry out ARE calculations without making a number of simplifying assumptions. However, it is possible to carry out a number of simulations that will allow us to obtain an approximate numerical ARE in a variety of illustrative situations.

To obtain a numerical approximation to the ARE we simulate error-free data with a particular sample size 1000 times, and calculate the average of the estimated standard errors for the true regression coefficient in each simulation set. We also simulate error-contaminated data with the same sample size 1000 times, and again calculate the average of the estimated standard errors for the corrected regression coefficient in each simulation set. The ratio of the squared standard errors of the regression coefficients from the error-free and the error-contaminated data will be an approximate ARE. This method is illustrated in the following sections on both prospective and retrospective data with a range of regression coefficients and measurement error variances.

### 4.4.1 Cohort Simulation Study

To simulate prospective data, we assumed that there were 10 groups with equal numbers of individuals in each group. We assumed that the true covariate values $X_t$ for the $i$th group were generated from a normal distribution, $N(\mu_i, 1)$, where $\mu_i$ came from a normal distribution, $N(0, 1.5)$. For each individual the surrogate variable $X_s$, was generated by taking the true value and adding to it a realisation from a normal distribution, $N(0, \sigma_m^2)$. To create Berkson error within the sample, 10% of the individuals in each group had their surrogate values replaced with

Table 4.1: Approximate ARE comparing the hypothesis test that $\beta_t$ is equal to its simulated value, in error-free and error-contaminated data from a prospective study with only one explanatory variable.

| True Regression | Measurement Error Variance | | | | | |
|---|---|---|---|---|---|---|
| Coefficient $\beta_t$ | 0.25 | 0.50 | 0.75 | 1.00 | 1.25 | 1.50 |
| 0.25 | 0.887 | 0.804 | 0.732 | 0.676 | 0.616 | 0.580 |
| 0.50 | 0.874 | 0.762 | 0.690 | 0.621 | 0.561 | 0.518 |
| 0.75 | 0.835 | 0.708 | 0.613 | 0.543 | 0.488 | 0.435 |
| 1.00 | 0.788 | 0.648 | 0.536 | 0.463 | 0.390 | 0.329 |
| 1.25 | 0.731 | 0.565 | 0.447 | 0.364 | 0.261 | 0.227 |
| 1.50 | 0.669 | 0.483 | 0.355 | 0.268 | 0.200 | 0.171 |
| 1.75 | 0.598 | 0.385 | 0.274 | 0.176 | 0.148 | 0.122 |
| 2.00 | 0.514 | 0.264 | 0.123 | 0.126 | 0.115 | 0.108 |

mean of the remaining individuals within that group. For each individual the binary response variable, $Y$, was formed such that $\Lambda(\alpha_t) = 0.1$, where $\Lambda$ denotes the logistic function.

For the numerical ARE calculation described above, a sample size of 500 was used, and parameter values $\beta_t = 0.25(0.25)2.00$, $\sigma_m^2 = 0.25(0.25)1.50$ were investigated.

The approximate ARE for the combination of parameters from this model is given in Table 4.1. From this table, it is clear that measurement error in a prospective model acts in a similar way as in the linear model. As the amount of classical measurement error increases, the ARE rapidly decreases; further as the regression coefficient increases the ARE also decreases for fixed amounts of measurement error. Although Table 4.1 does not consider differing amounts of Berkson error, due to the symmetry of the model specification in equation (4.1), it is clear that as the amount of Berkson measurement error increases, the ARE will correspondingly decrease. It is, therefore, very important for investigators to take account of measurement error when considering the design of a prospective study since the required sample size will be larger when measurement error is

present.

### 4.4.2 Case-Control Simulation Study

To simulate matched pair case-control data, the matching was with respect to a variable $V$, a categorical variable with five levels, assumed to be known without error and independent of $X_t$. The marginal distribution of $X_t$ is the same as for the prospective data simulation procedure. We also assume that each of the categories of the matching variable $V$ has probability 1/5 in the population. It was also assumed that

$$\mathrm{P}(Y = y|X_t, V) = \Lambda\{(\alpha_t + \beta_t X_t + \delta)y\},$$

where $\delta$ represents the effect of the variable $V$ taking values -0.50(0.25)0.50 and where $\alpha_t$ is chosen such that the probability of a positive response is 0.1 when $X_t = 0$ and $V = 3$. To generate artificial matched case-control data, the following procedure was adopted. For the case, generate $X_t$ and $V$ from their marginal distributions and then generate the response variable $Y$. If $Y = 1$, it was assigned as a case; if not the procedure was repeated using new values of $X_t$ and $V$. To generate a matching control, the same procedure was adopted until a value of $Y = 0$ was achieved, for which the value of $V$ was the same as that for the corresponding case. The error-prone variables $X_s$ were generated in the same way as for prospective data.

For the numerical ARE calculation described previously, we consider samples of size 500, and parameters $\beta_t = 0.25(0.25)2.00$ and $\sigma_m^2 = 0.25(0.25)1.50$.

The approximate ARE for the combination of parameters described above is given in Table 4.2. The results for the retrospective simulation are very similar to those obtained from the prospective simulation. The main difference is that in this simulation the ARE is always higher than the corresponding prospective study. This is as expected, since including an extra variable (the matching variable) causes some of the measurement error associated with the exposure variable to be transferred to the variable that is not measured with error. From Table 4.2 we

Table 4.2: Approximate ARE comparing the hypothesis test that $\beta_t$ is equal to its simulated value, in error-free and error-contaminated data from a retrospective study with two explanatory variables only one of which is measured with error.

| True Regression | Measurement Error Variance | | | | | |
|---|---|---|---|---|---|---|
| Coefficient $\beta_t$ | 0.25 | 0.50 | 0.75 | 1.00 | 1.25 | 1.50 |
| 0.25 | 0.905 | 0.828 | 0.763 | 0.707 | 0.659 | 0.617 |
| 0.50 | 0.895 | 0.812 | 0.744 | 0.688 | 0.639 | 0.597 |
| 0.75 | 0.869 | 0.770 | 0.694 | 0.632 | 0.580 | 0.537 |
| 1.00 | 0.821 | 0.697 | 0.606 | 0.536 | 0.479 | 0.430 |
| 1.25 | 0.757 | 0.600 | 0.489 | 0.405 | 0.335 | 0.269 |
| 1.50 | 0.668 | 0.446 | 0.344 | 0.220 | 0.182 | 0.167 |
| 1.75 | 0.571 | 0.307 | 0.195 | 0.170 | 0.154 | 0.140 |
| 2.00 | 0.334 | 0.252 | 0.175 | 0.161 | 0.142 | 0.131 |

clearly see that, as both the amount of classical measurement error and the true regression coefficient increase, the ARE decreases.

## 4.5 Discussion

The results obtained in this chapter give an insight into the impact of measurement error on the precision of estimated regression coefficients. In the linear model case, this was demonstrated by deriving the ARE of a hypothesis test that the estimated coefficient was zero in error-contaminated data versus error-free data. In the case of a binary error model, some illustrative simulation studies allowed approximate numerical determinations of the ARE for the hypothesis test that the estimated regression coefficient was equal to its true value.

This chapter has highlighted the importance of thinking carefully about measurement error during the planning stages of a study. It was demonstrated that if any of the variables are likely to be measured with error, then classical sample size

calculations no longer apply and the required sample size is always going to be larger than initially thought. This is important if the aim of the study is to determine association relationships between the response and explanatory variables. If measurement error is not taken into account, and the effect of the exposure is small, it is likely that the sample will fail to detect it. In terms of the graphical model, this will imply that certain edges between variables are missing when they should be present and incorrect inferences regarding the chain of associations can be drawn.

All these calculations and simulations were carried out on the assumption that there was some knowledge about the relative magnitude of the measurement error process. This sort of knowledge would have to come from some preliminary (validation) study to allow it to be successfully incorporated into the framework presented here.

Although the results in this chapter are based on some very simple situations, i.e. on one- or two-variable models, they can easily be extended to more complex situations. For any given combinations of parameters we can calculate an approximate ARE by simulating error-free and error-prone data, and finding the ratio of the squared standard errors for the appropriate regression coefficient. This ARE can then be used to assess the effects that measurement error will have on the precision of estimated parameters, and also allow an appropriate adjustment to be made to the sample size calculation.

## 4.6 Acknowledgements

I am grateful to Dr Gillian Reeves for providing access to the Fortran programs for implementing the correction technique discussed in Reeves et al. (1998).

# Chapter 5

# Analysis of Childhood Growth

This chapter considers the re-analysis of a set of data collected during the 1970s to assess the impact of milk supplementation on childhood growth in the first five years of life. The initial results from this study showed that there was no difference between the control and treatment groups when analysing the various measures of growth considered, allowing us therefore to treat the data as that from a longitudinal study, and enabling us to investigate a range of factors affecting the growth and development of children.

## 5.1 Background to the Study

During 1971, the UK Government withdrew from all but the very poorest families the entitlement that every pregnant woman and every child up to age five had access to a half-price pint of milk every day. This study, which began in 1972 and was initiated by the then Department of Health and Social Security, was designed to investigate whether this governmental action had any effect on the growth of children during the first five years of life.

The study was based in two small towns in South Wales: Barry, a seaside town, and Caerphilly, a largely industrial town. The social-class distribution in Barry was almost identical to that of England and Wales, whereas Caerphilly had a

slightly lower proportion of people in the highest social classes. It would therefore seem reasonable to assume that this is a representative sample with respect to the social-class distribution of the population of England and Wales.

In each town, contact was made through general practitioners with just over 700 newly pregnant women. Only 37 (2.5%) refused to co-operate and it can therefore be concluded that this study is based upon a representative sample of the general population. During the study there was a small amount of dropout due to people not being pregnant, moving away, being lost to follow-up, or having miscarriages, twins or malformed infants (all of which were excluded). By the age of five, there were 951 children remaining in the study. An analysis of this drop-out suggested that it was completely at random and not a limiting factor of the study. The women were visited at the time of first reporting the pregnancy and a questionnaire was completed at the 20th and 36th weeks into the pregnancy. The infants were visited at ten days, six weeks, and at three, six, nine and 12 months, and thereafter at six monthly intervals until their fifth birthdays. The growth measurements recorded were:

- **Weight** Weight at delivery was recorded from hospital records and weight during the study was recorded using a beam balance.
- **Length** Crown to heel length was recorded at ten days to allow time for caputs to resolve and was measured using a stadiometer at every visit until the age of two years after which standing height was measured.
- **Head Circumference** The maximum occipito-frontal measurement was taken with a disposable paper tape on the tenth day and then again at every visit.
- **Skinfold Thickness** Mid triceps skinfold was measured with Harpenden calipers at ten days and again at each visit.

Various other measures were made on each child, including:

- Data on the occurrence and severity since the previous visit of various illnesses.
- Data on aspects of child behaviour using the questionnaire of Richman & Graham (1971).

In addition to the measurements on the children there were a number of parental measurements. The maternal height and weight were recorded using a stadiometer and a beam balance, and approximately 70% of the fathers were seen at home and had their height and weight measured with a tape and spirit level, and a spring balance. The mother's age and child's parity were also recorded. In addition to these physical measurements a number of others were taken, including the mother's smoking status during pregnancy, her blood pressure, any vitamin and mineral supplements she was taking, whether she had oedema and/or albumin, and her haemaglobin levels, with most of these variables being recorded both at 20 and 36 weeks. A measure of maternal behaviour using the Eysenck Personality Inventory (Eysenck, 1972) was made when the child was five years old. At various times during the study a number of socio-demographic variables were measured, including the household composition, the social class of the father, the earnings of the father and mother, the amount the family spent on food, whether the child was in a single parent family, and the marital status of the mother at the end of five years. There was an attempt by the original investigators to try and discover the relationships between maternal nutrition and its impact on childhood growth; however they considered only very crude measures of nutrition and since maternal smoking is often confounded with nutrition, I will not comment on these any further.

### 5.1.1 Previous Results

A number of analyses of this data-set have already been carried out. The initial analysis into whether there was any effect of milk supplementation was discussed in Elwood et al. (1981) and Department of Health and Social Security (1981). The main conclusions of these two reports were that there is no effect of milk

supplementation on growth. The trial, which was then considered as a longitudinal study, was also reported in Department of Health and Social Security (1987) and Elwood et al. (1987). These latter papers considered factors, mainly maternal, affecting the growth of children during the first five years of their life. The main factor to be identified as responsible for differences in the measures of growth was the smoking status of the mother. They also attempted to identify other environmental factors which may have been relevant to growth, and concluded that although there were some relationships between variables such as social class and parental height and weight, the most important determinant was maternal smoking.

### 5.1.2 Motivation for Current Analysis & Problems

The two main reasons for re-analysing this data-set are:

- The subjects in this study are currently being followed up to investigate the effects of early life experience on later-life diseases, since there is a large current research interest in 'life-course illnesses'. In particular the interest lies in how early life events, including in-utero events, affect the likelihood of developing diseases like diabetes and coronary heart disease in later life. It is therefore of great interest to ensure that an analysis of factors affecting early life growth is completed so that these factors may be investigated in relation to the later-life diseases.

- The previous analyses nearly always considered the variables in a univariate analysis, i.e. they assess the impact of social class on growth without taking into account other factors. It is therefore of interest to consider a more detailed analysis of the data which will take account of the joint relationships present rather than simply the univariate.

There are a number of problems in attempting to analyse data relating to childhood growth. The process itself is a very complex one that has never been fully understood. The main factors determining a child's ultimate height/weight (and

thus growth) are genetic, but there are a host of social, environmental and psychological factors that can also have an effect on the growth of a child. Therefore in any study of growth, although we can observe most of the external factors, we are not yet able to observe the genetic factors, except to say that we would expect the attained growth of the child to be highly related to the mid-line parental average (the only surrogate measurement of genetic factors available).

## 5.2 Univariate Analysis Of Height and Weight

The main methodology of this analysis is to fit a joint-response chain graph as described on pages 1–2 of this thesis and Chapter 2 of Cox & Wermuth (1996). To fit one of these graphs involves fitting standard statistical models to the data and using the estimated coefficients to form the graph. For this particular data set, we consider relating measurements of growth of the children at birth, age one, age three and age five to the background maternal/paternal variables.

### 5.2.1 Variables To Be Analysed

The two main measures of childhood growth to be considered are the height, or length before age two (in cm), and weight of the children (in kg) at each of the time points. These are the two most indicative measurements of the growth of children, and they are also the most accurately measured. The measurements of these two were regularly monitored by independent observers during the study to ensure their accuracy. It is therefore unlikely that there is appreciable measurement error in these measurements.

As described previously, there are a host of background variables that one might hypothesise as being related to the growth of children. For this analysis, a subset of these were selected by (expert) prior knowledge of those most likely to be involved in determining the growth of children. These are:

- **Maternal Age** A continuous variable indicating the age of the mother (in

years) at the birth of the child.

- **Gestational Weeks** A continuous variable indicating the number of weeks into the pregnancy after which the child was born.

- **Parity** A categorical variable indicating the parity of the child (possible values are 0, 1, 2 and 3+).

- **Smoking During Pregnancy** An ordinal variable with three levels indicating the number of cigarettes smoked (on average) per day during the pregnancy. The levels are none, 1-14 and 15+ cigarettes per day.

- **Maternal Height** A continuous variable representing the height of the mother (in inches) at the birth of the child.

- **Maternal Weight** A continuous variable representing the weight of the mother (in pounds) at 20 weeks into pregnancy.

- **Maternal Weight Gain** A continuous variable measuring the proportionate weight gain of the mother between the 20th and 36th weeks of pregnancy.

- **Social Class** A categorical variable with three levels, indicating a High, Middle and Low Social Class of the father (and therefore of the household) when the child was $4\frac{1}{2}$ years old (unfortunately the social class at birth is not available as the data file has been lost). This variable is coded as 0, 1 and 2 representing high, middle and low social classes respectively.

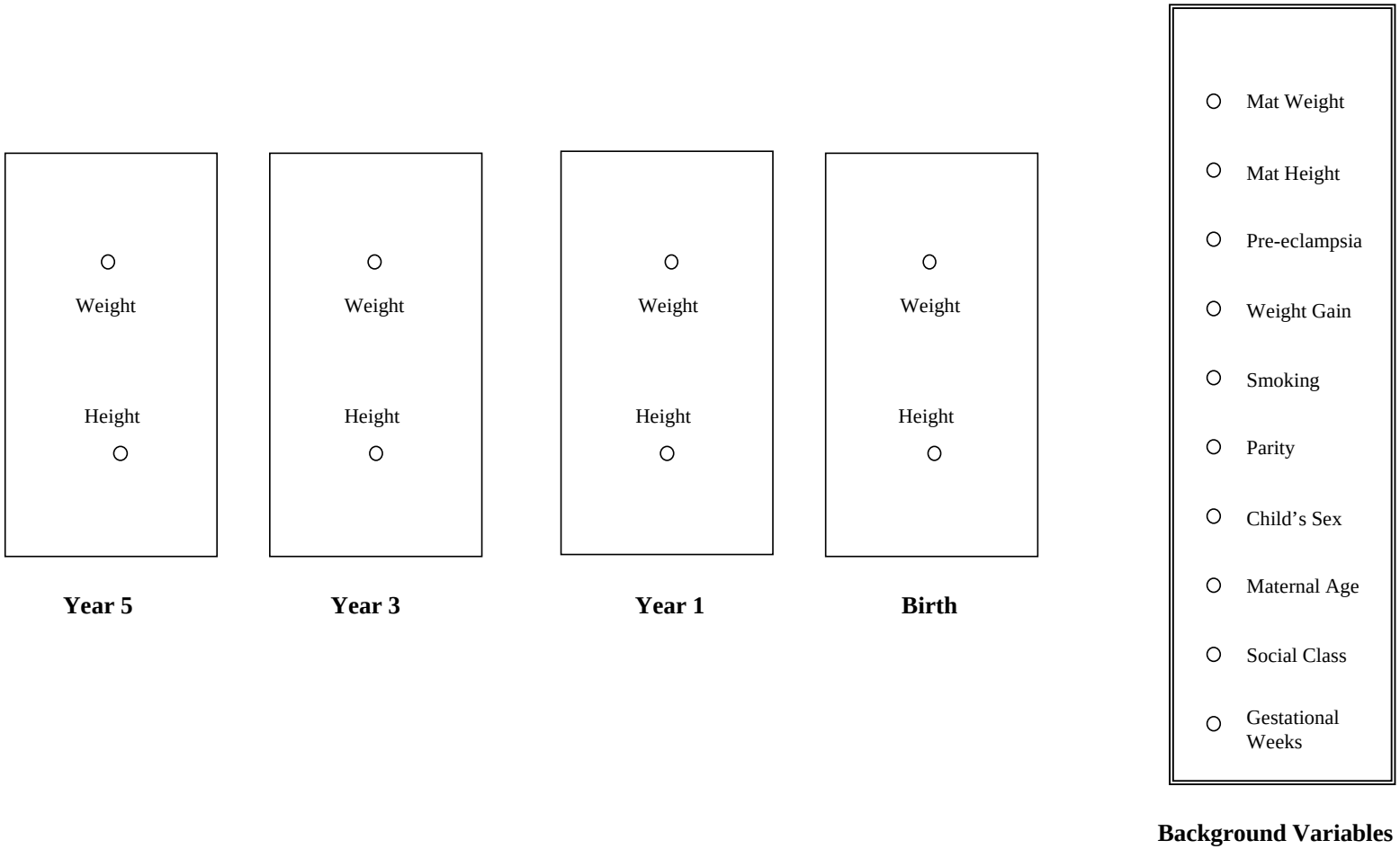
- **Pre-eclampsia** An ordered categorical variable with four levels indicating whether the mother had 0, 1, 2 or all 3 of the triad of symptoms associated with pre-eclampsia. These are oedema, albumin and high blood pressure (this was determined if the diastolic pressure at 36 weeks was above 90 mmHg).

- **Child's Sex** A binary indicator of whether the new born child was male or female (note: males were coded as 0 and females as 1).

Unfortunately at the current time, none of the paternal variables are available for analysis. The computer file on which they were originally recorded has been lost and although there are efforts in progress to have the data re-entered, it is unlikely that this will be completed within the next year.

A first ordering of these variables is given in Figure 5.1 and reflects the temporal ordering present in this data set.

Figure 5.1: A first ordering of the variables for the childhood growth dataset.



In all the analyses that follow we use the logarithms of height and weight data, allowing us to make direct comparisons of regression coefficients even thought the measurement units are different. When considering maternal weight gain, we consider the log of the ratio between the maternal weight at 20 weeks and at 36 weeks.

We consider fitting linear regression models to this data, using either height or weight as the response variable and at each time point, using all of the previous height and weight variables and all of the background variables as possible explanatory variables. This appears a sensible thing to do: if one looks at the distribution of the height or weight of a child at the time points considered in this analysis, they all appear to be normally distributed. The selection of the most parsimonious model was made by fitting all possible explanatory variables and then deleting one by one from the model those whose fitted regression coefficients were not significant. After each deletion, the variables which were not in the model were reinstated and kept in the model if they significantly improved the model fit, thus ensuring that the final model best described the variation in the data set using all of the possible explanatory variables. Finally, the usual linear regression diagnostics were checked for outliers and to ensure a good fitting model. In what follows, the multiple R-Squared and the residual standard error is given for each fitted model, although none of the diagnostic plots are included as there was nothing of significance to note. The main reason for including the R-squared is to confirm that these analyses agree with expert observations (Dunger, 1998) that at birth, one can explain approximately 30% of the variation in children's growth measurements by background variables and this rapidly increases to approximately 70% by the age of five, if one includes a previous value of the children's growth measurement.

It should be noted that there is some missing data here (approximately 5% at each time point), although it would appear that this data is missing at random. This was checked by looking at the distributions of the background variables for those cases whose data was missing at a particular time point and comparing these distributions to those for the cases who had no missing data. On each occasion,

Table 5.1: Summary statistics for background variables measured on a continuous scale.

| Variable Name | Min | Mean | Max | Std Deviation | Missing |
|---|---|---|---|---|---|
| Maternal Age (years) | 15 | 25 | 42 | 4.61 | 4 |
| Log Maternal Weight | 4.45 | 4.91 | 5.44 | 0.154 | 18 |
| Log Maternal Height | 4.04 | 4.14 | 4.24 | 0.0368 | 19 |
| Log Weight Gain | -0.307 | 0.102 | 0.368 | 0.0648 | 64 |
| Gestation (weeks) | 34 | 40.0 | 45 | 1.43 | 47 |

there appeared to be little difference in the distributions between the cases with missing data and those without, and therefore no explicit attempt is made in the modelling procedure to take account of missing data.

### 5.2.2 Analysis of Background Variables

The first analysis provides a summary of the recorded background variables and explores any relationships that exist between them.

A summary of the continuous variables is given in Table 5.1, and of the categorical variables in Table 5.2. It is interesting that in this particular trial (a randomised controlled trial) we have a ratio of males to females of 54:46; upon checking with all those children born into the trial this is not a facet of drop-out as the ratio was identical at birth. It is also of interest to investigate the associations between the background variables. The correlation matrix between these variables is given in Table 5.3: where both of the variables are continuous the standard Pearson correlation is given, otherwise the Spearman rank correlation is given. Also in this table, those correlations which are significantly different from zero are identified by their p-value.

There are a number of associations between the background variables, although for the purposes of this analysis we will only consider those who have a p-value $< 0.01$ as being significantly related. These are:

Table 5.2: Summary of background variables measured on a categorical scale.

| Sex | | Social Class | | Smoking Status | | Parity | |
|---|---|---|---|---|---|---|---|
| Male | 513 | High (I) | 18 | None | 577 | 0 | 360 |
| Female | 438 | Upper Middle (II) | 250 | 1-14 | 201 | 1 | 372 |
| | | Middle (III) | 402 | 15+ | 167 | 2 | 137 |
| | | Low (IV/V) | 163 | Missing | 6 | 3+ | 70 |
| | | Missing | 118 | | | Missing | 12 |

| Pre-eclampsia | |
|---|---|
| No Symptoms | 575 |
| 1 Symptom | 256 |
| 2 Symptoms | 49 |
| 3 Symptoms | 6 |
| Missing | 65 |

- Maternal age and parity.
- Maternal age and maternal height.
- Maternal age and social class.
- Parity and smoking.
- Parity and weight gain.
- Smoking and weight gain.
- Smoking and social class.
- Maternal height and maternal weight.
- Maternal weight and weight gain during pregnancy.

Most of these relationships are to be expected on general grounds, or on the basis of previous studies. The most interesting relationship is that between parity and

smoking, although this is likely to be an artifact of the associations between weight gain and smoking, and parity and weight gain. The association between maternal age and social class would appear to suggest that this study recruited a biased age/social class distribution. The positive association between social class and smoking implies that lower income groups smoke more.

Table 5.3: Estimated correlations between the background variables.

| | Gestweeks (weeks) | Parity | Smoking | Log Mat. Height | Log Mat. Weight | Log Weight Gain | Social Class | Pre-Eclampsia | Sex |
|---|---|---|---|---|---|---|---|---|---|
| Mat Age (years) | −0.078† | 0.453‡ | −0.067† | 0.103‡ | 0.084† | -0.047 | −0.169‡ | 0.013 | 0.042 |
| Gestweeks (weeks) | | -0.020 | 0.019 | 0.072† | 0.069† | 0.069† | 0.056 | -0.044 | 0.031 |
| Parity | | | 0.120‡ | 0.066† | -0.023 | −0.107‡ | 0.084† | -0.016 | -0.030 |
| Smoking | | | | 0.004 | 0.003 | −0.109‡ | 0.187‡ | -0.028 | -0.049 |
| Log Mat. Height | | | | | 0.448‡ | -0.032 | −0.080† | 0.010 | -0.001 |
| Log Mat. Weight | | | | | | −0.402‡ | 0.001 | 0.083† | -0.018 |
| Log Weight Gain | | | | | | | -0.017 | 0.082† | 0.017 |
| Social Class | | | | | | | | 0.018 | -0.006 |
| Pre-Eclampsia | | | | | | | | | 0.059 |

**Note:** A † indicates a p-value of less than 0.05 and ‡ indicates a p-value of less than 0.01 for the hypothesis test that the correlation is zero.

Table 5.4: Summary statistics for the log height measurements.

| Variable Name | Min | Mean | Max | Std Deviation | Missing |
|---|---|---|---|---|---|
| Log Birth Length | 3.80 | 3.95 | 4.07 | 0.0414 | 40 |
| Log Year One Length | 4.19 | 4.32 | 4.43 | 0.0348 | 5 |
| Log Year Three Height | 4.43 | 4.55 | 4.67 | 0.0390 | 54 |
| Log Year Five Height | 4.48 | 4.68 | 4.79 | 0.0402 | 21 |

### 5.2.3 Analysis of Children's Height

In this section, we perform an analysis at each of the time points of the log height of children, and the factors that influence it. We examine these factors in two stages: firstly looking at only the background factors, and secondly looking at background factors and previous height and weight measurements. Initially, Table 5.4 summarises the log height measurements at each of the time points.

#### Birth

If we consider attempting to model the log birth-length of the children using all of the background variables and the selection procedure described above to find the most parsimonious model, we obtain the estimated regression coefficients in Table 5.5.

The multiple R-Squared for this model is 0.260 which, although rather small, is in line with expectation when trying to predict birth length without knowing anything about the genetic components of growth. The residual standard error for this model is 0.0356 on 814 degrees of freedom. Also, the estimated correlation between the residuals from this fitted model and the log birth weight of children is 0.561.

This analysis does not include social class as a possible explanatory variable for differences in the birth length of the children. There could be a number of reasons for this, the main one being that the recording of this variable was done at $4\frac{1}{2}$ years,

Table 5.5: Estimated regression coefficients from model of child's log birth length using the background variables.

| Variable | Estimate | Std. Error |
|---|---|---|
| (Intercept) | 2.83 | 0.144 |
| Gestweeks (weeks) | 0.00846 | 0.00093 |
| Log Mat Height | 0.147 | 0.0382 |
| Log Mat Weight | 0.0363 | 0.0101 |
| Sex | -0.0227 | 0.00249 |
| Log Weight Gain | 0.0567 | 0.0215 |
| Smoke (1-14/day) | -0.0116 | 0.00311 |
| Smoke (15+/day) | -0.0135 | 0.00345 |

and it is possible that this may have changed since the child was born. However, one plausible reason is that maternal smoking during pregnancy is strongly related to social class and thus any differences between levels of social class are confounded by differences in the smoking status of the mother. This confounding could be more enhanced since social class normally indicates differences between amounts of money available to a family, and thus can be considered as a surrogate nutritional indicator. Department of Health and Social Security (1987) and Elwood et al. (1987) also showed that smoking status was related to a number of crude measures of nutritional status and thus the confounding between smoking status, social class and nutritional differences is quite high. The negative effect of maternal smoking is very significant and indicates that smoking during pregnancy (and some of its possible confounders) are responsible for a significant proportion of inter-uteral growth retardation (IUGR). Maternal age and parity are not included in the final model, as maternal age and parity are highly correlated, and both are related to a number of other variables in the final model. Another factor affecting growth is gestational duration: the longer the pregnancy the larger the child. Rather obviously, males tend to be longer than females and thus the effect of sex is as expected. Maternal weight gain is positively associated with the length of the

Table 5.6: Estimated regression coefficients from model of child's log height at age one using their log birth length, log birth weight and all background variables.

| Variable | Estimate | Std. Error |
|---|---|---|
| (Intercept) | 1.99 | 0.125 |
| Sex | -0.0111 | 0.0019 |
| Log Mat Height | 0.183 | 0.0254 |
| Log Birth Length | 0.399 | 0.0233 |

children and this variable sometimes (although not always) is a reflection of inter-uteral growth. It is unusual that pre-eclampsia is not included in the final model as it is commonly seen that children of mothers who suffer pre-eclampsia tend to be both smaller and lighter. However, Table 5.2 shows that there are only six cases of full pre-eclampsia and thus its non-inclusion is therefore not that surprising. Thus all of the factors associated with birth length have coefficients in the directions one would expect.

**Year One**

The model that results from analysing the log length of children at year one, using their log length and log weight at birth and all of the background variables as possible explanatory variables, is given in Table 5.6.

The multiple R-Squared for this model is 0.378 and the residual standard error is 0.0273 on 885 degrees of freedom. The estimated correlation between the residuals from this model and the log weight measurements of the children at age one is 0.526.

The estimates which result from fitting a model to the log length of children at year one using only the background variables as possible explanatory variables is given in Table 5.7.

The multiple R-Squared for this model is 0.211, smaller than that obtained for

Table 5.7: Estimated regression coefficients from model of child's log height at age one using only the background variables.

| Variable | Estimate | Std. Error |
|---|---|---|
| (Intercept) | 3.20 | 0.119 |
| Smoke (1-14/day) | -0.0057 | 0.0026 |
| Smoke (15+/day) | -0.0122 | 0.0028 |
| Log Mat Weight | 0.0206 | 0.0076 |
| Log Mat Height | 0.248 | 0.0316 |
| Sex | -0.0195 | 0.0021 |

predicting log birth length from the background variables. The residual standard error is 0.0313 on 909 degrees of freedom.

It is interesting that in the second model, virtually none of the background variables except maternal height and weight, and maternal smoking are in the model. This implies that the other background variables partly responsible for IUGR no longer have an effect, and therefore that the children have recovered from the effect of these variables on their growth inter-uterally. The estimated regression coefficients for maternal smoking are relatively smaller than those fitted in the model for log birth-length and hence it can be argued that maternal smoking is responsible for less of the difference in heights at this time. In other words, the children are catching up to the height they would have been had their mothers not smoked during pregnancy, but they are still slightly smaller than those children whose mothers did not smoke.

Looking at the first model alone implies that conditionally on the child's birth length (the coefficient for log birth weight is very close to zero), the child's length at age one is conditionally independent of all of the background variables. There are, therefore, some background factors that marginally affect the child's length at year one. However, when one takes into account the child's birth length, the background factors offer no further explanation of the differences in the children's lengths. The child's sex shows less dependence in the first model compared to the

Table 5.8: Estimated regression coefficients from model of child's log height at age three using previous measures of log height and log weight and the background variables.

| Variable | Estimate | Std. Error |
|---|---|---|
| (Intercept) | 0.446 | 0.126 |
| Log Mat Height | 0.141 | 0.0244 |
| Log Year One Length | 0.815 | 0.0258 |

one for birth, which is as one would expect since the birth length encapsulates some information and helps differentiate between males and females.

### Year Three

The model that results from analysing the log height of children at year three, using their previous log lengths and log weights and all of the background variables as possible explanatory variables, is given in Table 5.8.

The multiple R-Squared for this model is 0.618 which, as expected, is becoming larger as we obtain more implicit information about the genetic component of growth as the child becomes older. The residual standard error is 0.0251 on 870 degrees of freedom. The correlation between the residuals from this model and the children's log weight measurements at age three is 0.646.

The model using only the background variables as possible explanatory variables is given in Table 5.9.

The multiple R-squared for this model is 0.150 which is of similar magnitude to those obtained for the models at age one and birth using only the background variables. The residual standard error is 0.0362 on 871 degrees of freedom.

An interesting point arising from the first analysis is that the dependence of a child's height on previous values of the child's height (and weight, although the coefficient in the regression of log height on previous log weight values is always

Table 5.9: Estimated regression coefficients from model of child's log height at age three using only the background variables.

| Variable | Estimate | Std. Error |
|---|---|---|
| (Intercept) | 3.03 | 0.139 |
| Smoke (1-14/day) | -0.0074 | 0.0031 |
| Smoke (15+/day) | -0.0128 | 0.0033 |
| Log Mat Height | 0.367 | 0.0335 |
| Sex | -0.0104 | 0.0025 |

very close to zero) appears to be a first order Markov process. Thus height at age three is conditionally independent of all other variables, except maternal height, given the child's height at age one. We also note that the child's sex is no longer significant, implying that all the information about the difference between males and females is encapsulated in the length data at age one.

It is also interesting that in the second analysis, smoking during pregnancy still has an effect on a child's height, even at age three. Thus, unlike the other background variables, maternal smoking during pregnancy continues to have an effect on the child's height. Maternal height continues to be strongly associated with the child's height in both analyses, reflecting the fact that it is the main implicit indicator of the genetic component of growth.

### Year Five

The model that results from fitting a linear model to the log height of children at age five, using their previous log lengths and log weights and all of the background variables as possible explanatory variables, is given in Table 5.10.

The multiple R-Squared for this model is 0.704, fairly similar to that for the year three model and the residual standard error is 0.0215 on 856 degrees of freedom. The correlation between the residuals from this model and the children's log weight measurements at age five is 0.241.

Table 5.10: Estimated regression coefficients from model of child's log height at age five using the previous log height and log weight values and the background variables.

| Variable | Estimate | Std. Error |
|---|---|---|
| (Intercept) | 0.289 | 0.117 |
| Log Mat Height | 0.109 | 0.0244 |
| Log Year Three Height | 0.866 | 0.0227 |

Table 5.11: Estimated regression coefficients from model of child's log height at age five using only the background variables.

| Variable | Estimate | Std. Error |
|---|---|---|
| (Intercept) | 2.94 | 0.148 |
| Sex | -0.0086 | 0.0026 |
| Smoke (1-14/day) | -0.0057 | 0.0033 |
| Smoke (15+/day) | -0.0143 | 0.0035 |
| Log Mat Height | 0.419 | 0.0357 |

The model for the child's log height at age five using only the background variables is given in Table 5.11.

The multiple R-squared for this model is 0.154, which is again similar to the values from the models for year three, year one and birth length (height) analyses. The residual standard error is 0.0370 on 903 degrees of freedom.

We learn nothing further from this analysis than we knew previously. The same variables that are marginally related in the model for the child's height at year three remain so at year five. Also, conditionally on the previous values of height and weight, year five height is only related to the year three data, thus giving more evidence that this dependence can be captured via a first order Markov process. Maternal height remains significant in both the marginal and conditional model clearly indicating its role as an implicit genetic determinant of childhood growth.

Further, in the marginal model smoking during pregnancy remains significantly associated with height at age five. Smoking during pregnancy therefore not only results in IUGR, but is also associated with a slightly smaller child at age five compared with those children from mothers who didn't smoke during pregnancy.

### Fitted Graphical Model

From these results we can form the graphical model shown in Figure 5.2. Using this graph, we can read off the various independence relationships (using the results in Cox & Wermuth, 1996, Chapter 2): for example, height at age five is conditionally independent of maternal weight gain given height at age three and maternal height. The graph shows the relationships between the variables in this data set; although nearly all the background variables have an effect on the birth length of the child, by the time the child is aged one they no longer have any effect. This is true for all maternal variables except maternal height, which continues to be associated with the child's height throughout this study, confirming the belief that this variable implicitly indicates the genetic growth potential of the child. It would appear from this graph that the dependence of a child's height on previous measurements is a first order Markov process, i.e. a child's height at age five is only dependent on their height at age three and so on.

Figure 5.2: Fitted graphical model for the marginal analysis of the log height of children.

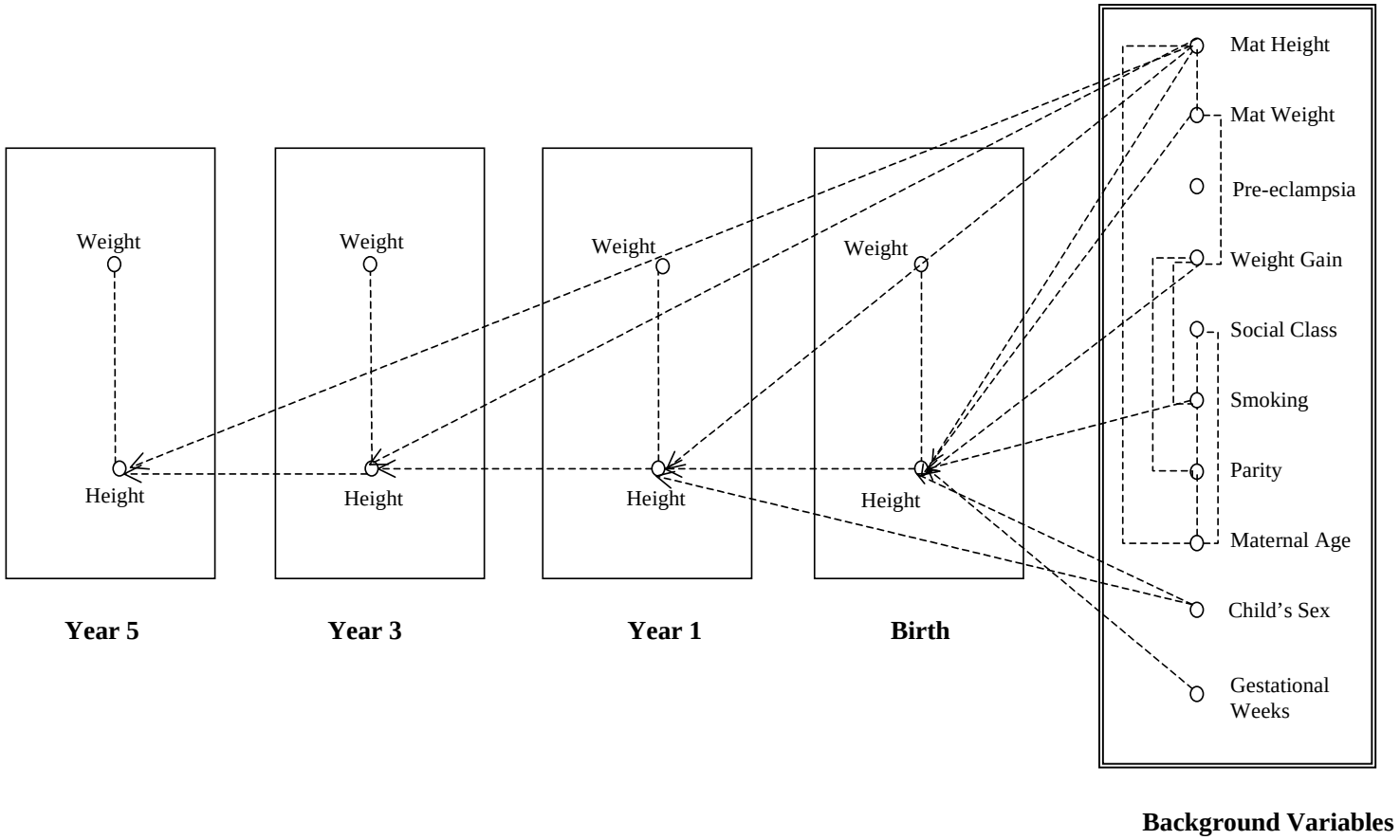


Table 5.12: Summary statistics for the log weight measurements.

| Variable Name | Min | Mean | Max | Std Deviation | Missing |
|---|---|---|---|---|---|
| Log Birth Weight | 0.308 | 1.20 | 1.58 | 0.156 | 0 |
| Log Year One Weight | 1.95 | 2.33 | 2.85 | 0.117 | 5 |
| Log Year Three Weight | 2.33 | 2.69 | 3.11 | 0.114 | 67 |
| Log Year Five Weight | 2.40 | 2.93 | 3.43 | 0.125 | 26 |

### 5.2.4 Analysis of Children's Weight

In this section, we consider the analysis at each of the time points of the log weight of children, and the factors that influence it. Firstly we present in Table 5.12 a number of summary measures of the log weight measurements at each of the time points.

#### Birth

If we model the log birth-weight of the children using all of the background variables described above, the estimated regression coefficients are given in Table 5.13.

The multiple R-Squared for this model is 0.253, which is small, although in line with prior expectation especially when one has a minimal amount of information about the genetic information transmitted between parents and child. The residual standard error is 0.126 on 834 degrees of freedom. The estimated correlation between the residuals from this model and the children's log length at birth is 0.562.

The model that has been fitted for log birth weight, is almost identical to that fitted for log birth length. The only feature to differ between these two models is that maternal height is not included as a potential explanatory variable. Although maternal height will offer information on the genetic component of growth, at this time more of the variation is explained by the maternal weight, and the weight gain of the mother between the 20th and 36th week of pregnancy. It would be

Table 5.13: Estimated regression coefficients from model of child's log birth weight using the background variables.

| Variable | Estimate | Std. Error |
|---|---|---|
| (Intercept) | -1.21 | 0.192 |
| Gestweeks (weeks) | 0.0298 | 0.0033 |
| Log Weight Gain | 0.244 | 0.0748 |
| Smoke (1-14/day) | -0.0477 | 0.0109 |
| Smoke (15+/day) | -0.0879 | 0.0120 |
| Log Mat Weight | 0.254 | 0.0312 |
| Sex | -0.0506 | 0.0087 |

expected that maternal height will become more predictive of a child's weight as the age of the child increases. Other than this one difference, all the other variables have a similar effect on the birth weight of the children as they did on the children's birth length. They have identical signs and the previous discussion applies.

**Year One**

The fitted model for the log weight of the child at age one using the previous values of log height and log weight along with all of the possible explanatory variables is given in Table 5.14.

The multiple R-Squared for this model is 0.278, only slightly greater than that for weight at birth, and the residual standard error is 0.0998 on 875 degrees of freedom. The estimated correlation between the residuals from this model and the log height of children at age one is 0.483.

If we model the log weight of the child at age one using only the background explanatory variables, then the resulting model is given in Table 5.15.

The multiple R-squared for this model is 0.166, which is very similar to that

Table 5.14: Estimated regression coefficients from model of child's log weight at age one using the previous log height and log weight values and the background variables.

| Variable | Estimate | Std. Error |
|---|---|---|
| (Intercept) | -0.892 | 0.560 |
| Log Mat Height | 0.252 | 0.104 |
| Log Mat Weight | 0.0588 | 0.0250 |
| Sex | -0.0547 | 0.0070 |
| Log Birth Weight | 0.201 | 0.0337 |
| Log Birth Length | 0.424 | 0.124 |

Table 5.15: Estimated regression coefficients from model of child's log weight at age one using only the background variables.

| Variable | Estimate | Std. Error |
|---|---|---|
| (Intercept) | 0.0990 | 0.412 |
| Log Weight Gain | 0.183 | 0.0617 |
| Log Mat Height | 0.381 | 0.111 |
| Log Mat Weight | 0.137 | 0.0291 |
| Sex | -0.0745 | 0.0073 |

obtained from the model for log birth weight and the residual standard error is 0.1065 on 861 degrees of freedom.

It can be noted that in both of these models, maternal height is now strongly predictive of the weight of the child. This is to be expected: by the age of one, the weight of the child is following a genetically pre-determined growth curve, and is thus highly related to the implicit genetic variables (i.e. maternal height and weight).

Unlike the fitted model for the year one length data, maternal smoking is no longer marginally associated with the weight of the child at age one. This result concurs with that found by Nafstad et al. (1997) in Norway, who showed that although children born to mothers who smoked were generally lighter than those from non-smoking mothers, by the time these children reached age one they had caught up any losses in weight due to maternal smoking. An identical pattern is observed in this data; the IUGR caused by maternal smoking is completely reversed by the end of the first year of life.

All the other variables and their fitted coefficients in both these models are in the same direction as the fitted models for the length of the children at age one.

### Year Three

The model resulting from fitting a linear model to the log weight data at age three using the previous log height and log weight values and all the possible explanatory variables is given in Table 5.16.

The multiple R-Squared for this model is 0.645, significantly higher than that for year one weight, which is as expected when one has discovered more implicit information about the genetic component of growth as the child becomes older. The residual standard error for this model is 0.0682 on 844 degrees of freedom. The estimated correlation between the residuals from this model and the children's log height at age three is 0.348.

The corresponding fitted model using only the background explanatory variables

Table 5.16: Estimated regression coefficients from model of child's log weight at age three using the previous log height and log weight values and the background variables.

| Variable | Estimate | Std. Error |
|---|---|---|
| (Intercept) | -1.47 | 0.410 |
| Log Mat Height | 0.182 | 0.0741 |
| Log Mat Weight | 0.0351 | 0.0174 |
| Log Year One Weight | 0.676 | 0.0284 |
| Log Year One Height | 0.383 | 0.0972 |

Table 5.17: Estimated regression coefficients from model of child's log weight at age three using only the background variables.

| Variable | Estimate | Std. Error |
|---|---|---|
| (Intercept) | 0.0492 | 0.4284 |
| Log Weight Gain | 0.205 | 0.0636 |
| Log Mat Height | 0.450 | 0.116 |
| Log Mat Weight | 0.158 | 0.0306 |
| Sex | -0.0451 | 0.0075 |

is given in Table 5.17.

The multiple R-squared for this model is 0.106, which is again smaller than the one for the corresponding model for year one weight, but not unexpected when one takes account of the fact that no information on nutrition is included in this model. The residual standard error is 0.107 on 805 degrees of freedom.

It is interesting to note that both maternal height and weight are marginally and conditionally associated with the weight of the child at age three. This is, as it was at age one, related to the increased implicit genetic information conveyed by maternal height and weight. When one considers only the height of the child, maternal weight offers little genetic information; however, when one considers the

Table 5.18: Estimated regression coefficients from model of child's log weight at age five using the previous log height and log weight values and the background variables.

| Variable | Estimate | Std. Error |
|---|---|---|
| (Intercept) | -0.888 | 0.331 |
| Log Mat Weight | 0.0597 | 0.0150 |
| Sex | 0.0150 | 0.0045 |
| Log Year Three Weight | 0.869 | 0.0295 |
| Log Year Three Height | 0.260 | 0.0838 |

child's weight, maternal weight offers explanation as to the inter-uteral growth of the child, whereas maternal height offers more explanation for the genetic component of growth.

The other interesting point to draw from this analysis is that the weight of the child appears, like the height of the child, to follow a first order Markov process. When one considers the first model, only the height and weight of the child at year one is included as a potential explanatory variable. Here, the majority of dependence on the previous measurements is for the child's previous weight, whereas when we looked at the model for the height of the child at age three, the majority of the dependence was on the previous height measurement. This is, however, as one would expect.

### Year Five

The fitted model for the year five log weight data using both the previous log height and log weight measurements and all of the explanatory variables is given in Table 5.10.

The multiple R-Squared for this model is 0.740, which is fairly close to that for year three, and the residual standard error is 0.0637 on 836 degrees of freedom. The estimated correlation between the residuals from this model and the log height of

Table 5.19: Estimated regression coefficients from model of child's log weight at age five using only the background variables.

| Variable | Estimate | Std. Error |
|---|---|---|
| (Intercept) | 0.0607 | 0.463 |
| Log Weight Gain | 0.280 | 0.0695 |
| Sex | -0.0262 | 0.0081 |
| Log Mat Height | 0.440 | 0.1249 |
| Log Mat Weight | 0.210 | 0.0329 |

children at age five is 0.147.

The model for the year five log weight data using only the previous explanatory variables is given in Table 5.19.

The multiple R-squared for this model is 0.117, which is very similar to that obtained from the corresponding model for the year three weight data. The residual standard error is 0.118 on 842 degrees of freedom.

The most interesting point to draw from the first of these models is that maternal height is no longer associated with the child's weight at age five. The most likely explanation is that all the information conveyed by this variable is encapsulated within the previous height and weight measurements of the children. This model also helps to confirm our belief that the dependence of the weight of the children on previous values of the child's height and weight is a first order Markov process.

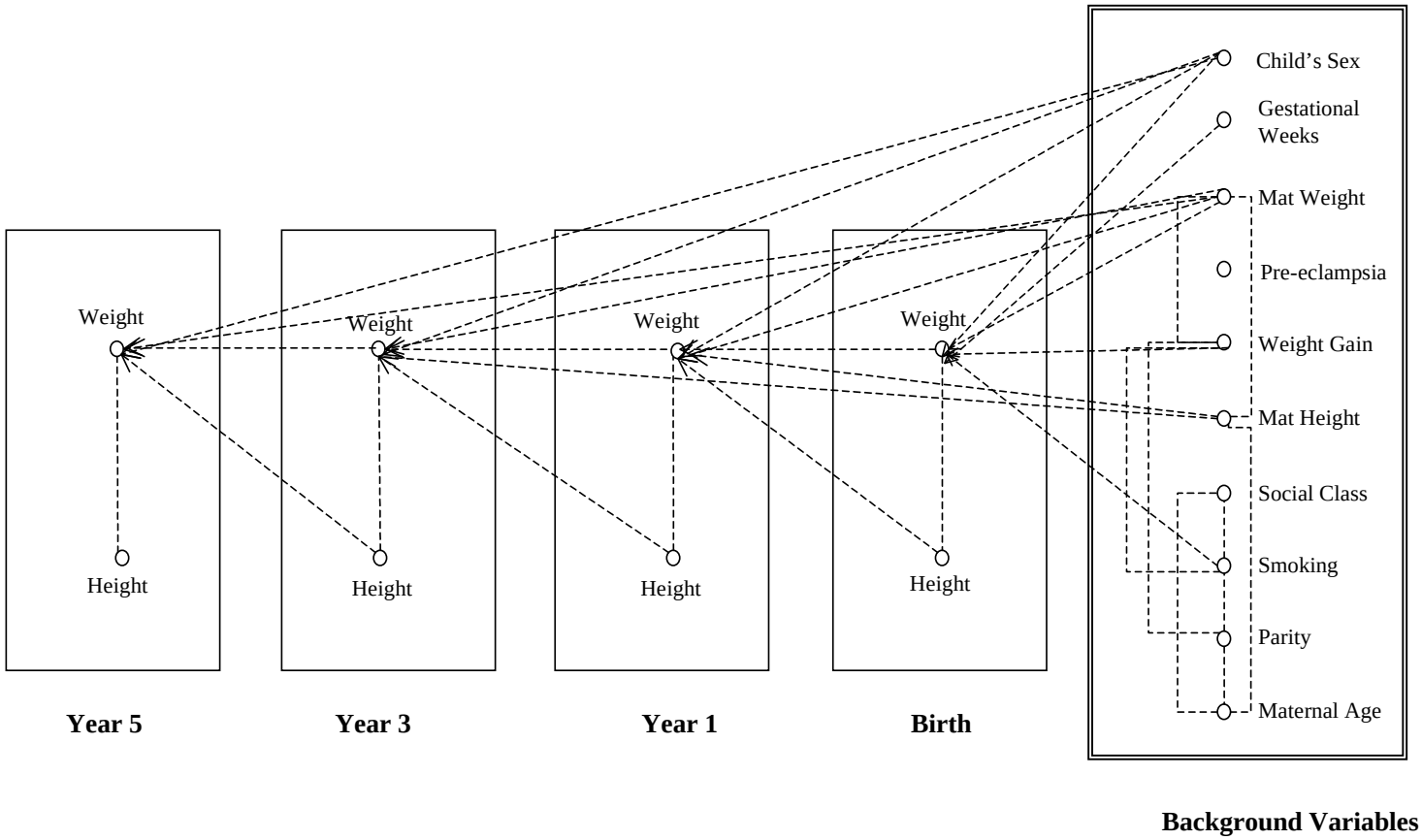
The fitted marginal model includes the same variables with coefficients in the same direction as the corresponding model for the year three weight data and the year one weight data.

### Fitted Graphical Model

With these results we can form the graphical model shown in Figure 5.3. Using this graph, we can read off the various independence relationships: for example,

weight at age five is conditionally independent of maternal weight gain given the height and weight of the child at age three, the child's sex and the maternal weight. This fitted graph gives a clear representation of the relationships between all the variables. It shows quite clearly that the children's weight, like the children's height appears to be governed by a first order Markov process. It also shows that all the maternal variables (except weight and height) that have an effect on the birth weight of the child no longer have any effect after the child reaches the age of one.

Figure 5.3: Fitted graphical model for the marginal analysis of the log weight of children.



### 5.2.5 Discussion

Both these analyses have shown that there are a number of background variables associated with the birth weight and birth length of children. However, as is often the case in observational studies, it has proven difficult to isolate effects of particular variables as there is a degree of confounding between explanatory variables. In this study in particular, it is difficult to isolate the effects of social class, smoking and parity due to high correlations between these variables. It was also shown in this analysis that most of the background variables that influence the weight and height of the children have little effect after the age of one, conditional on previous height and weight measurements. However, it was interesting that maternal smoking continued to marginally associated with the childrens height, although not their weight, even after age one. As discussed above, it could be hypothesised that the measured variables indicate IUGR in the unborn child and thus result in a decreased birth length or weight. During the first year of life the child recovers from the growth retardation and rejoins their pre-programmed growth curve.

It has also been shown that, without a specific measure of the genetic contribution of growth, the use of surrogate measures such as maternal height and weight (and paternal height and weight should those measurements be available), can only go so far in producing a variable that is an indication of the growth potential of the child. These surrogate measures will at best be an indirect measure of genetic potential since they themselves are subject to environmental factors affecting the growth of the parents. As can be seen from the above fitted models, maternal height and weight are strongly associated with the child's height and weight respectively, and are thus indicators of the genetic contribution of growth. The maternal height also appears to be associated at two of the studied time points with the child's weight, suggesting that more information can be gained from both of the maternal measurements than from only one. It would have been interesting to have the paternal height and weight measurements, as Elwood et al. (1987) showed that while maternal height and weight were strongly associated with the

birth measurements, the paternal measurements appeared to be more strongly associated with the growth of the children during the first years of their life. This could explain why maternal height is only associated with some of the weight measurements of the children and why maternal weight is only associated with birth length, and not with any of the other height measurements of the children.

If one compares the graphical models for height and weight, it is interesting to note that, while both the previous values of height and weight are used in forming a model for the child's weight at any time point other than birth, only the previous height measurement is used in the equivalent model for the child's height. This suggests firstly that the growth (both in terms of the child's height and weight) is regulated by a first order Markov process, and also that the child's height is more important in determining the ultimate regulation of the child's growth. In other words, it is more likely that the weight of a child has a greater degree of variability, whereas the child's height is a more stable and informative process for determining ultimate growth.

## 5.3 Analysis of Height and Weight Jointly

There are a number of techniques that can be utilised when investigating the joint relationships between a set of variables: these include the classical techniques of canonical correlation (or canonical regression analysis). More recently, Cox & Wermuth (1992) and Wermuth & Cox (1995) discuss a technique that allows the calculation of derived variables in the analysis of multivariate responses. They consider the multivariate regression of a $p \times 1$ vector $Y$ of random variables on a $q \times 1$ vector of explanatory variables $X$. The interest lies in a set of linear combinations of $Y$ that can form the basis for interpretation in terms of the explanatory variables. The proposed technique produces a set of linear combinations $Y^*$ such that in the multivariate regression of $Y_i^*$ on $X$, only the coefficient of $X_i$ is non-zero.

However, both these papers only considered a vector response on a single occa-

sion $Y$. In many situations, including the setting of the current analysis, we have repeated observations of the response variable at different time points. The interest is then in the set of derived variables that best explain the joint relationship between the responses, in this case the children's height and weight. There are two possibilities that can be considered in this situation: we can hypothesise either that the derived variables are different at each time point, or that they are the same. For the current analysis, the second of these two possibilities seems the most appropriate: there is, for example, no reason to believe that the relationships between a child's height and weight should change over time.

### 5.3.1 Two Time Points

Let us assume that we have two $p \times 1$ variables $Y_n, Y_{n-1}$ and we are looking for combinations $a_i^T \ i = 1, \ldots, p$ such that in the new co-ordinate base, $a_i^T Y_n$ is only dependent on $a_i^T Y_{n-1}$. If we let $Y_n^* = AY_n, Y_{n-1}^* = AY_{n-1}$, this implies that we require the matrix of regression coefficients to be diagonal and equal to $D$. We observe that

$$\begin{aligned}
\text{Cov}(Y_n^*, Y_{n-1}^*) &= \Sigma_{Y_n^*, Y_{n-1}^*} = A\Sigma_{n,n-1}A^T, \\
\text{Cov}(Y_{n-1}^*) &= \Sigma_{Y_{n-1}^*, Y_{n-1}^*} = A\Sigma_{n-1,n-1}A^T,
\end{aligned}$$

where $\Sigma_{n,n-1} = \text{Cov}(Y_n, Y_{n-1})$ and $\Sigma_{n-1,n-1} = \text{Cov}(Y_{n-1}, Y_{n-1})$. Thus we require, assuming that A is non-singular,

$$\begin{aligned}
B_{Y_n^*, Y_{n-1}^*} = (A\Sigma_{n,n-1}A^T)(A\Sigma_{n-1,n-1}A^T)^{-1} &= D, \\
A\Sigma_{n,n-1}\Sigma_{n-1,n-1}^{-1}A^{-1} &= D, \\
AB_{n,n-1} &= DA,
\end{aligned}$$

where $B_{n,n-1}$ is the matrix of regression coefficients of $Y_n$ on $Y_{n-1}$.

That is, the rows of $A$ are the left eigenvectors of $B_{n,n-1}$ with eigenvalues the elements of $D$, $d_j$. Note that since $B_{n,n-1}$ is a matrix of regression coefficients and is in general neither symmetric nor positive definite, there is no guarantee that the eigenvalues will be real. If there is a complex eigenvalue, this implies that it

Table 5.20: Summary of the joint relationships between the height and weight of children between birth and age five.

| | Age Five and Three | Age Three and One | Age One and Birth |
|---|---|---|---|
| First Eigenvalue | 0.94 | 0.85 | 0.47 |
| First Eigenvector[a] | $(0.77, 0.63)^T$ | $(0.98, 0.21)^T$ | $(1.00, 0.07)^T$ |
| Second Eigenvalue | 0.65 | 0.56 | 0.13 |
| Second Eigenvector[a] | $(-0.96, 0.29)^T$ | $(-0.93, 0.37)^T$ | $(-0.93, 0.36)^T$ |

[a]The first component of the vectors represents the log of the child's height and the second component represents the log of the child's weight.

is not possible to represent the dependence relationships in the form specified. If there is a zero eigenvalue, this would imply that a reduced number of variables would be sufficient to define the dependence.

**Example**

Let us now consider applying this technique to the childhood growth data set. Initially, let us consider the heights and weights of children at ages birth, one, three and five. We can then consider applying this technique to each consecutive pair of time points in order to attempt to learn about the joint relationships between height and weight of children over time.

The results from this set of analyses are given in Table 5.20. It would appear from this analysis that between birth and age one nearly all of the first joint relationship is attributed to the height component, whereas the second of the joint relationships is the ratio of weight to height to the power 2.5. The latter of these two relationships is often referred to as the Benn index, a measure of adiposity often used in small children – it is a measure midway between the body mass index (the ratio of weight to height squared) and the ponderal index (the ratio of weight to height cubed). The first of these two components can be considered a measure of size, while the second of the two components can be considered a

measure of shape.

Between the ages of one and three, the joint relationships are something like weight plus height to the power 4.5 along with the Benn index. It is interesting that these analyses do not suggest the more commonly used joint relationship between height and weight – the body mass index – which is often thought to be more appropriate after the child starts walking (between one and two years of age). Between ages three and five, the first component is approximately the geometric mean of height and weight, while the second component is approximately the ponderal index. It remains interesting that none of these analyses suggests using the body mass index as the most appropriate measure of the joint relationship between height and weight. In all cases, the first component of the joint relationships appears to be a measure of size, whereas the second component appears to be a measure of shape.

It appears from these results that the joint relationships do change between very young children and older children; assuming that the joint relationships between a child's height and weight are constant between the first five years of life is, therefore, not appropriate.

### 5.3.2 Three Time Points

If there are three repeated measurements then there are a number of interesting possibilities. Let $Y_n, Y_{n-1}$ and $Y_{n-2}$ represent the repeated measurements. In the current discussion we are going to assume that there are no explanatory variables, although it would be possible to consider extending this discussion to include explanatory variables.

#### Case 1

In the first case, we assume the Markov property, that is in the new co-ordinate base we have that $a_i^T Y_n$ is dependent only on $a_i^T Y_{n-1}$, and that $a_i^T Y_{n-1}$ is dependent only on $a_i^T Y_{n-2}$. In this case, we are essentially in the same setting as the case

of only two time points. For the joint relationships to remain constant over the three time points then we require that the matrices of regression coefficients

$$B_{Y_n,Y_{n-1}} \text{ and } B_{Y_{n-1},Y_{n-2}},$$

must have the same left eigenvectors, although they do not necessarily have to have the same eigenvalues.

**Case 2**

If we no longer assume the Markov property, we simply require that in the new co-ordinate basis $a_i^T Y_n$ can depend on $a_i^T Y_{n-1}$ and $a_i^T Y_{n-2}$, whereas $a_i^T Y_{n-1}$ only depends on $a_i^T Y_{n-2}$. By the required properties of $Y^*_{n-1}$ on $Y^*_{n-2}$ we require that

$$AB_{n-1,n-2} = D_{n-1}A. \tag{5.1}$$

We have to introduce the relationships relating to time $n$, so we let

$$Z^*_{n-1} = \begin{pmatrix} AY_{n-1} \\ AY_{n-2} \end{pmatrix},$$

so that

$$\mathrm{Cov}(Z^*_{n-1}) = \begin{pmatrix} A\Sigma_{n-1,n-1}A^T & A\Sigma_{n-1,n-2}A^T \\ \cdot & A\Sigma_{n-2,n-2}A^T \end{pmatrix}.$$

Thus

$$\mathrm{Cov}(Y^*_n, Z^*_{n-1}) = \begin{pmatrix} A\Sigma_{n,n-1}A^T & A\Sigma_{n,n-2}A^T \end{pmatrix},$$

and

$$B_{Y^*_n,Z^*_{n-1}} = \begin{pmatrix} A\Sigma_{n,n-1}A^T & A\Sigma_{n,n-2}A^T \end{pmatrix} \begin{pmatrix} A\Sigma_{n-1,n-1}A^T & A\Sigma_{n-1,n-2}A^T \\ \cdot & A\Sigma_{n-2,n-2}A^T \end{pmatrix}^{-1}. \tag{5.2}$$

For the derived variables to be consistent we require $B_{Y^*_n,Z^*_{n-1}}$ to have the form $(D_1|D_2)$, where $D_1, D_2$ are diagonal matrices.

To find the derived variables in this situation, therefore, we can find $A$ from equation (5.1) and substitute this into equation (5.2) to check whether the double diagonal form holds.

**Case 3**

We now assume arbitrary dependence between the components of $Y_n^*$ and the components of $Y_{n-1}^*$, but impose the constraint that $a_i^T Y_{n-1}$ only depends on $a_i^T Y_{n-2}$. The discussion follows that of case 2 and we simply inspect the matrix $B_{Y_n^*, Z_{n-1}^*}$, to observe the actual dependence structure between $Y_n^*$ and $Y_{n-1}^*$. A zero entry in this matrix will imply that there is no relationship between the two components, and a non-zero entry will imply that there is a relationship.

In general, the matrix form of equation (5.2) will allow us to assess which of the three cases is the most suitable to describe the evolution of the joint dependence relationships. Case 1 corresponds to the matrix $D_2$ being a zero matrix, and the matrix $D_1$ being diagonal. Case 2 corresponds to the double diagonal form of $(D_1|D_2)$, and case 3 allows the matrix to be of any form.

In all of these cases, it remains unclear how to test the hypothesis that entries of this matrix are zero, or whether the components of the joint relationship have a particular form. However, the method presented here is proposed as a relatively simple way to investigate possible relationships between joint responses measured over time. It is therefore believed that this should be a flexible approach to be used as a tool for investigation and not necessarily as an inferential procedure.

**Example**

Consider the results presented in Table 5.20. If we are to assume the Markovian structure for the three time points, then we require that both sets of eigenvectors are the same. It would appear that for none of the combinations of time points considered here is this an appropriate assumption. Certainly between ages one, three and five, the first component of the joint relationship changes considerably. However, between birth and ages one and three, it may be possible that the set of joint relationships are approximately Markovian. To investigate this assumption, we can calculate the matrix in equation (5.2) for both of these situations. If we

consider birth and ages one and three, then we calculate

$$B_{Y_n^*,Z_{n-1}^*} = \begin{pmatrix} 0.803 & 0.078 & 0.078 & -0.051 \\ -0.004 & 0.544 & -0.003 & 0.052 \end{pmatrix},$$

and for ages one, three and five, we calculate

$$B_{Y_n^*,Z_{n-1}^*} = \begin{pmatrix} 0.900 & 0.248 & 0.038 & -0.189 \\ 0.087 & 0.591 & -0.067 & 0.167 \end{pmatrix}.$$

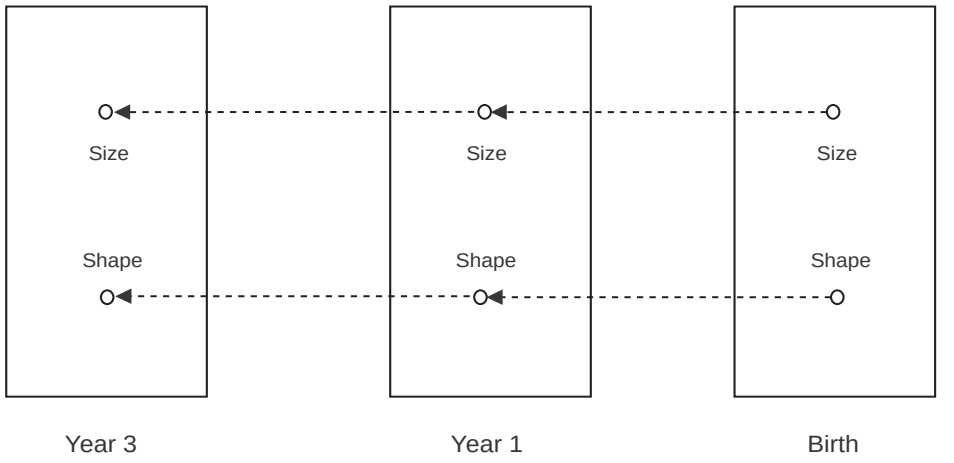
Looking at the first of these two matrices, it seems apparent that the first of the three cases applies; between birth, age one and age three it appears that a Markovian structure is appropriate for the evolution of the joint relationships between the height and weight of children. However, between ages one, three and five the evolution of the joint relationships is quite complex. The first component, which can be considered a size component, is dependent at age five on the previous size component and the previous two shape components, while the second (shape) component is dependent only on the previous shape components. This tends to suggest that a person's size is dependent on their previous size and shape, while a person's shape depends only on their previous shape.

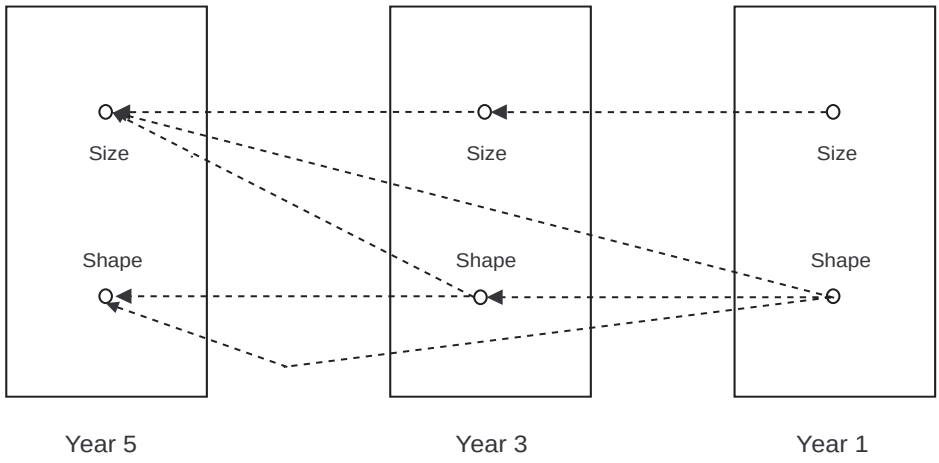
We can represent both of these relationships in the form of graphical models (Figure 5.4), allowing an easier insight into the structure of these joint relationships.

### 5.3.3 Discussion

This section has provided a new method for the analysis of a set of repeated measurements in terms of derived variables, where it was assumed that the same combination of derived variables was appropriate at each time point. The illustration of this method on the childhood growth data has provided some very useful insights into the structure of the growth of children. It has shown that during the first few years of life the growth of children is governed by a Markovian process. However, during the latter part of early childhood, the structure of the process is much more complex. The size component of growth appears to be governed not only by its previous values, but also by the shape component. This implies that a

Figure 5.4: Graphical models for the joint dependence relationships between the log height and log weight of children.



person's size is related to previous values of size and that it is also related to their shape. It is also plausible that this analysis implies that while a person's shape remains relatively constant over the first five years of life, their size changes quite dramatically in relation to this.

Although this section has only considered a maximum of three time points, extending this to more than three time points is direct, we simply redefine the matrices in an appropriate manner. The issue which requires careful thought is how to interpret the results of this analysis as there are a number of interesting possibilities. It would also be possible to investigate the effects that explanatory variables have on the derived variables; this was not considered here as it was of more interest to consider the structure of the growth process, rather than its dependence on explanatory variables. The sort of approach that one may take would be a combination of the original treatment by Cox & Wermuth (1992) and the treatment here.

## 5.4 Acknowledgements

I would like to thank Dr Yoav-Ben Shlomo and Professor George Davey-Smith of the Department of Social Medicine, University of Bristol for providing me with access to the data and also to Dr Jim Holt, Dr Anne Thompson and Dr David Dunger of the John Radcliffe Hospital, Oxford for some interesting and useful discussions regarding the variables and analyses that one may wish to consider applying to data of this type.

# Bibliography


Aitken, A. C. & Gonin, H. T. (1935). Fourfold sampling with and without replacement. *Proceedings of the Royal Society of Edinburgh* **55**, 114.

Aitken, A. C. & Turnbull, H. W. (1931). *Theory of Canonical Matricies.* Blackie, London.

Armitage, P. & Berry, G. (1987). *Statistical Methods in Medical Research.* Blackwell Scientific, Oxford, 2nd edition.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete Multivariate Analysis.* MIT Press, Cambridge, MA.

Bollen, K. A. (1989). *Structural Equations with Latent Variables.* Wiley, Chichester.

Carroll, R. J., Ruppert, D., & Stefanski, L. A. (1995). *Measurement Error in Nonlinear Models.* Chapman & Hall, London.

Cochran, W. G. (1968). Errors of measurement in statistics. *Technometrics* **10**, 637–666.

Cox, D. R. (1958). *Planning of Experiments.* Chapman & Hall, London.

Cox, D. R. & Hinkley, D. V. (1974). *Theoretical Statistics.* Chapman & Hall, London.

Cox, D. R. & Snell, E. J. (1989). *Analysis of Binary Data.* Chapman & Hall, London, 2nd. edition.

Cox, D. R. & Wermuth, N. (1990). An approximation to maximum likelihood estimates in reduced models. *Biometrika* **77**, 747–61.

Cox, D. R. & Wermuth, N. (1991). A simple approximation for bivariate and trivariate normal integrals. *International Statistical Review* **59**, 263–269.

Cox, D. R. & Wermuth, N. (1992). On the calculation of derived variables in the analysis of multivariate responses. *Journal of Multivariate Analysis* **42**, 162–170.

Cox, D. R. & Wermuth, N. (1994). A note on the quadratic exponential binary distribution. *Biometrika* **81**, 403–408.

Cox, D. R. & Wermuth, N. (1996). *Multivariate Dependencies – Models, Analysis and Interpretation.* Chapman & Hall, London.

Dale, J. R. (1984). Local versus global association for bivariate ordered responses. *Biometrika* **71**, 507–514.

Dale, J. R. (1986). Global cross-ratio models for bivariate, discrete, ordered responses. *Biometrics* **42**, 909–917.

Department of Health and Social Security (1981). Growth and development in the first five years of life. In *Medical Aspects of Food Policy, Subcommittee on Nutritional Surveillance: Second Report.* HMSO, London.

Department of Health and Social Security (1987). Growth and development in the first five years of life. In *Medical Aspects of Food Policy, Subcommittee on Nutritional Surveillance: Third Report.* HMSO, London.

Desu, M. M. & Raghavarao, D. (1990). *Sample Size Methodology.* Academic Press, New York.

Devine, O. J. & Smith, J. M. (1998). Estimating sample size for epidemiologic studies: The impact of ignoring exposure measurement uncertainty. *Statistics in Medicine* **17**, 1375–1389.

Dunger, D. (1998). Personal Communication.

Edwards, D. (1995). *Introduction to Graphical Modelling.* Springer-Verlag, Berlin.

Elwood, P. C., Haley, T. J. L., Hughes, S. J., Sweetnam, P. M., Gray, O. P., & Davies, D. P. (1981). Child growth (0-5 years), and the effect of entitlement to a milk supplement. *Archives of Disease in Childhood* **56**, 831–835.

Elwood, P. C., Sweetnam, P. M., Gray, O. P., Davies, D. P., & Wood, P. D. P. (1987). Growth of children from 0-5 years: with special reference to mother's smoking in pregnancy. *Annals of Human Biology* **14**, 543–557.

Eysenck, H. J. (1972). Personality and the maintenance of the smoking habit. In Dunn, W. L., editor, *Smoking Behaviour: Motives and Incentives.* Winston, Washington D.C.

Fitzmaurice, G. M. & Laird, N. M. (1993). A likelihood-based method for analysing longitudinal binary responses. *Biometrika* **80**, 141–151.

Freedman, L. S., Schatzkin, A., & Wax, Y. (1990). The impact of dietary measurement error on planning sample size required in a cohort study. *American Journal of Epidemiology* **132**, 1185–1195.

Fuller, W. A. (1987). *Measurement Error Models.* John Wiley & Sons, New York.

Glonek, G. F. V. (1996). A class of regression models for multivariate categorical responses. *Biometrika* **83**, 15–28.

Glonek, G. F. V. & McCullagh, P. (1995). Multivariate logistic models. *Journal of the Royal Statistical Society, Series B* **57**, 533–546.

Goodman, L. A. (1981). Association models and the bivariate normal for contingency tables with ordered categories. *Biometrika* **68**, 347–355.

Goodman, L. A. (1991). Measures, models, and graphical displays in the analysis of cross-classified data. *Jounal of the American Statistical Association* **86**, 1085–1138.

Greenland, S. (1988). On sample-size and power calculations for studies using confidence intervals. *American Journal of Epidemiology* **128**, 231–237.

Grizzle, J. E., Starmer, C. F., & Koch, G. G. (1969). Analysis of categorical data by linear models. *Biometrics* **25**, 489–504.

Kendall, M. G. (1941). Proof of relations connected with tetrachoric series and its generalization. *Biometrika* **32**, 196–198.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press, Oxford.

Liang, K.-Y. & Zeger, S. L. (1986). Longitudinal data analysis using generalised linear models. *Biometrika* **73**, 13–22.

Liang, K.-Y., Zeger, S. L., & Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society, Series B* **54**, 3–40.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall, London, 2nd. edition.

McFadden, J. A. (1955). Urn models of correlation and a comparison with the multivariate normal integral. *Annals of Mathemtical Statistics* **26**, 478–489.

McFadden, J. A. (1956). An approximation for the symmetric quadrivariate normal integral. *Biometrika* **43**, 206–207.

McKeown-Eyssen, G. E. & Tibshirani, R. (1994). Implications of measurement error in exposure for the sample sizes of case-control studies. *American Journal of Epidemiology* **139**, 415–421.

Nafstad, P., Jaakkola, J. J. K., Hagen, J. A., Pedersen, B. S., Qvigstad, E., Botten, G., & Kongerud, J. (1997). Weight gain during the first year of life in relation to maternal smoking and breast-feeding in Norway. *Journal of Epidemiology and Community Health* **51**, 261–265.

Pearson, K. (1909). On a new method of determining correlation between a measured character $A$, and a character $B$, of which only the percentage of cases

wherein $B$ exceeds (or falls short of) a given intensity is recorded for each grade of $A$. *Biometrika* **7**, 96–105.

Price, W. E., Grizzle, J. E., Postlethwait, R. W., Johnson, W. D., & Grabicki, P. (1970). Results of operation for duodenal ulcer. *Surgery, Gynaecology and Obstetrics* **131**, 233–244.

Reeves, G. K., Cox, D. R., Darby, S. C., & Whitley, E. (1998). Some aspects of measurement error in explanatory variables for continuous and binary regression models. *Statistics in Medicine* **17**, 2157–77.

Richman, N. & Graham, P. J. (1971). A behavioural screening questionnaire to use with three-year-old children. Preliminary findings. *Journal of Child Psychology and Psychiatry and Applied Disciplines* **12**, 5–33.

Schervish, M. J. (1984). Algorithm AS195. Multivariate normal probabilities with error bound. *Applied Statistics* **33**, 81–94.

Sheppard, W. F. (1898). On the application of the theory of error to cases of normal distributions and normal correlations. *Philosophical Transactions of the Royal Society, London Series A.* **192**, 101–167.

Steck, G. P. (1962). Orthant probabilities for the equicorrelated multivariate normal distribution. *Biometrika* **49**, 433–445.

Tiesch, J. M., Sommer, A., Witt, K., Katz, J., & Royall, R. M. (1990). Blindness and visual impairment in an American urban population. *Archives of Opthalmolology* **108**, 286–290.

Ware, J. H., Dockery, D. W., Spiro, A. I., Speizer, F. E., & Ferris, B. G. J. (1984). Passive smoking, gas cooking and respiratory health in children living in six cities. *American Review of Respiratory Disease* **129**, 366–374.

Wermuth, N. & Cox, D. R. (1995). Derived variables calculated from similar joint responses: some characteristics and examples. *Computational Statistics & Data Analysis* **19**, 223–234.

Wermuth, N. & Cox, D. R. (1998). On the application of conditional independence to ordinal data. *International Statistical Review* **66**, 181–199.

Whittaker, J. (1990). *Graphical models in applied multivariate statistics*. Wiley, Chichester.

Williams, O. D. & Grizzle, J. E. (1972). Analysis of contingency tables having ordered response categories. *Journal of The American Statistical Society* **67**, 55–63.

Zeger, S. L. & Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**, 121–130.

Zeger, S. L., Liang, K.-Y., & Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics* **44**, 1049–1060.

Zhao, L. P. & Prentice, R. L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika* **77**, 642–648.

# Appendix

The Mathematica code used to fit the models in Section 3.2.3 is given below. The main function which is used is `awrmv` which takes as its arguments five parameters. These parameters are

- `data` The observed data vector arranged into lexicographical order within each of the groups.
- `x` The $X$ matrix for all the groups arranged into the same lexicographical order as the data vector.
- `dims` A vector representing the dimensions of each of the response variables.
- `type` A vector representing the response variable types (nominal=0 or ordinal=1).
- `rdim` The dimension of the contingency table within each group.

The function call required for the coal miners data example is given at the end of this appendix.

```
<<LinearAlgebra`MatrixManipulation`

awrcord[dim_]:=BlockMatrix[Outer[Times,IdentityMatrix[dim-1],
{{1},{-1}}]]

awrcnom[dim_]:=Join[IdentityMatrix[dim-1], {Table[-1, {dim-1}]}]

awrgenc[dims_, type_]:=Module[{nor, ccur, cprev, ctmp, loop, add1d,
add2d, add3d, add4d, add1, add2},
ccur={{1}};
```

```
nor=Length[dims];
For[loop=1, loop<=nor, loop++, cprev=ccur;
If[type[[loop]]==0, ctmp=awrcnom[dims[[loop]]],
ctmp=awrcord[dims[[loop]]]];
ctmp=BlockMatrix[Outer[Times,cprev,ctmp]];
add1d=Dimensions[ctmp][[2]];
add2d=Dimensions[cprev][[2]];
add3d=Dimensions[cprev][[1]];
add4d=Dimensions[ctmp][[1]];
add1=ZeroMatrix[add3d,add1d];
add2=ZeroMatrix[add4d,add2d];
ccur=BlockMatrix[{{cprev,add1},{add2,ctmp}}];];
ccur]

awrgenl[dims_, type_]:=Module[
{nor,lcur,lprev,loop,ltmp,ltmp1},
lcur={{1}};
nor=Length[dims];
For[loop=1,loop<=nor,loop++, lprev=lcur;
ltmp={Table[1, {dims[[loop]]}]};
If[type[[loop]]==0, ltmp1=IdentityMatrix[dims[[loop]]],
ltmp1=awrlord[dims[[loop]]]];
ltmp=BlockMatrix[Outer[Times,lprev,ltmp]];
ltmp1=BlockMatrix[Outer[Times,lprev,ltmp1]];
lcur=BlockMatrix[{{ltmp},{ltmp1}}];];
lcur]

awrlord[dim_]:=Module[{tmp,tmp1,ind1,ind2},
tmp=LowerDiagonalMatrix[awrtmp, dim];
tmp1=ZeroMatrix[dim]-tmp+1;
res=ZeroMatrix[2 (dim-1), dim];
For[ind1=1; ind2=1, ind1<=(dim-1) 2, ind1+=2, res[[ind1]]=tmp[[ind2]];
If[ind1!=(dim-1) 2, res[[ind1+1]]=tmp1[[ind2]]];
ind2++;];
res]
```

```
awrtmp[x_,y_]:=1

awrest[table_, dimr_]:=Module[{total,loop},
total=Sum[table[[loop]], {loop,1,dimr}];
{table/total, total} ]

awrestimate[rawdata_, dimr_, nor_]:=Module[{lpest,res1,res2,temp,tmp},
For[lpest=1; res1=Null; res2=Null; temp=rawdata, lpest<=nor, lpest++,
tmp=awrest[Take[temp,dimr], dimr];
res1={res1, tmp[[1]]};
res2={res2, tmp[[2]]};
temp=RotateLeft[temp,dimr]];
{Rest[Flatten[res1]], Rest[Flatten[res2]]}]

awrextendx[xmat_]:=Module[{res, temp, size, inat},
size=Dimensions[xmat][[1]];
inat=Dimensions[xmat][[2]];
temp=IdentityMatrix[size];
res=xmat;
For[i=1, i<=size, i++, res=Insert[res, temp[[i]], {i,inat+1}];
res[[i]]=Flatten[res[[i]]];];
res]

awrvarest[estprobs_, totals_, dimr_]:=Module[{estvar},
estvar=estprobs/Flatten[Transpose[Table[totals, {dimr}]]];
DiagonalMatrix[estvar]]

awrmatsolve[extx_, nopar_]:=Module[{extxrr,res,row,col},
row=Dimensions[extx][[1]];
col=Dimensions[extx][[2]];
extxrr=RowReduce[extx];
res=extxrr[[Range[row], Drop[Range[col],nopar]]];
res]

awrcox[varmat_, mles_, nopar_]:=Module[{size},
size=Dimensions[varmat][[1]];
mles[[Range[nopar]]]-varmat[[Range[nopar],Range[nopar+1,size]]].
```

```
Inverse[varmat[[Range[nopar+1,size],Range[nopar+1,size]]]].
mles[[Range[nopar+1,size]]]

awrtfr[par_, dimr_, cm_, lm_]:=Module[{loop,res,temp,nor},
nor=Length[par]/dimr;
For[loop=1; res=Null; temp=par, loop<=nor, loop++,
res={res, Drop[Transpose[cm].Log[lm.temp[[Range[dimr]]]],1]};
temp=RotateLeft[temp,dimr]];
Rest[Flatten[res]]]

awrzero[tfp_, pdim_]:=Module[{swap,pos,zerostart,res},
res=tfp;
pos=Flatten[Position[res,0]];
While[pos[[1]]<=pdim, pos=Drop[pos,1]];
zerostart=Length[tfp]-pdim+1;
While[pos[[1]]<zerostart && Length[pos]<=pdim, swap=pos[[1]]+1;
While[res[[swap]]==0, swap++];
res[[pos[[1]]]]=res[[swap]];
res[[swap]]=0;
pos=Flatten[Position[res,0]];
While[pos[[1]]<=pdim, pos=Drop[pos,1]];];
res]

awrfitted[estparams_, xmat_, nor_, cm_, lm_, rdim_]:=
Module[{lp,res,fitted,totake,topass},
fitted=xmat.estparams;
totake=Dimensions[xmat][[1]]/nor;
For[lp=1; res=Null, lp<=nor, lp++,
topass=Insert[Take[fitted,totake], 0, 1];
res={res, awrinvert[topass, cm, lm, rdim]};
fitted=RotateLeft[fitted, totake];];
Rest[Flatten[res]]]

awrinvert[par_, c_, l_, rdim_]:=
Module[{cur,prev,test,d,tmp1,tmp2,notok,i},
cur=Table[1/rdim, {rdim}];
```

```
cur=N[Log[cur]];
notok=True;
epsilon=0.000001;
While[notok==True, prev=cur;
d=N[Inverse[DiagonalMatrix[1.Exp[prev]]]];
tmp1=N[Inverse[Transpose[c].d.1.DiagonalMatrix[Exp[prev]]]];
tmp2=N[Transpose[c].Log[1.Exp[prev]]-par];
cur=N[prev-tmp1.tmp2];
cur=Exp[cur];
cur=cur/Sum[cur[[i]], {i,1,rdim}];
cur=N[Log[cur]];
test=N[Abs[cur-prev]];
For[i=1; notok=False, i<rdim, i++ If[test[[i]]>epsilon, notok=True]];];
N[Exp[cur]]]

awrtotalinfo[fitpbs_, xm_, tot_, c_, l_, dimr_, nop_, nor_]:=
Module[ {i,pbstmp,xtmp,info,se,totoake},
totake=Dimensions[xm][[1]]/nor;
For[i=1; pbstmp=fitpbs; xtmp=xm; info=0, i<=nor, i++,
info=info+awrinfo[Take[pbstmp, dimr], Take[xtmp, totake],
tot[[i]], nop, dimr, c, l];
pbstmp=RotateLeft[pbstmp, dimr];
xtmp=RotateLeft[xtmp,totake]];
info=Inverse[info];
For[i=1;
se=Null, i<=Dimensions[info][[1]], i++, se={se, info[[i,i]]}];
Sqrt[Rest[Flatten[se]]]]

awrinfo[fitprob_, xmat_, total_, nopar_, dimr_, cm_, lm_]:=
Module[ {tfrpa,jac,dmat,tmpx},
tfrpa=Transpose[cm].Log[lm.fitprob];
tmpx=Insert[xmat, Table[0, {nopar}], 1];
dmat=DiagonalMatrix[lm.fitprob];
jac=Inverse[Transpose[cm].Inverse[dmat].lm].tmpx;
N[total*Transpose[jac].Inverse[DiagonalMatrix[fitprob]].jac]]
```

```
awrloglike[datalike_, problike_, dimdata_, dimresp_]:=
Module[{i,loglike,probstmp,datatmp},
For[i=1; probstmp=problike; datatmp=datalike;
loglike=0, i<=dimresp, i++,
loglike=loglike+datatmp[[Range[dimdata]]].
Log[probstmp[[Range[dimdata]]]];
probstmp=RotateLeft[probstmp,dimdata];
datatmp=RotateLeft[datatmp,dimdata]];
loglike]

awrmv[data_, x_, dims_, type_, rdim_]:=Module[{nor,nop,extended,extend,
totals,params,p,cmat,lmat,dmat,tfrparams,tfrjac,loop, loop1,jac,i,j,
vest,tfrvest,extx,fitted,fittedv,info,like},
nor=Length[data]/rdim;
nop=Dimensions[x][[2]];
extended=awrestimate[data, rdim, nor];
extend=extended[[1]];
totals=extended[[2]];
cmat=awrgenc[dims, type];
lmat=awrgenl[dims, type];
jac=ZeroMatrix[(rdim-1) nor, rdim nor];
For[loop=1, loop<=nor, loop++,
tfrjac=Drop[Transpose[cmat].Inverse[DiagonalMatrix[lmat.
Take[extend, {(loop-1) rdim+1, loop rdim}]]].lmat,1];
For[i=(loop-1) (rdim-1)+1; ijac=1, i<=loop (rdim-1), i++,
For[j=(loop-1) rdim+1; jjac=1, j<=loop rdim, j++,
jac[[i,j]]=tfrjac[[ijac,jjac]];jjac++; ]; ijac++];];
vest=awrvarest[extend, totals, rdim];
tfrvest=jac.vest.Transpose[jac];
extx=awrextendx[x];
tfrparams=N[LinearSolve[extx, awrtfr[extend, rdim, cmat, lmat]]];
tfrparams=awrzero[tfrparams,nop];
vest=awrmatsolve[extx, nop];
tfrvest=N[vest.tfrvest.Transpose[vest]];
```

```
estparams=N[awrcox[tfrvest, tfrparams, nop]];
fitted=awrfitted[estparams, x, nor, cmat, lmat, rdim];
fittedv=N[fitted Flatten[Transpose[Table[totals, {rdim}]]]];
like=N[awrloglike[data, fitted, rdim, nor]];
info=awrtotalinfo[fitted, x, totals, cmat, lmat, rdim, nop, nor];
{estparams, info, like} ]
```

To obtain the approximate maximum likelihood fit to the coal miners example, the following code was used. The other examples were computed in a similar way.

```
data={9,7,95,1841,23,9,105,1654,54,19,177,1863,
121,48,257,2357,169,54,273,1778,269,88,324,1712,
404,117,245,1324,406,152,225,967,372,106,132,526}

xmatrix={{1,-4,0,0,0,0},
{0,0,1,-4,0,0},
{0,0,0,0,1,-4},
{1,-3,0,0,0,0},
{0,0,1,-3,0,0},
{0,0,0,0,1,-3},
{1,-2,0,0,0,0},
{0,0,1,-2,0,0},
{0,0,0,0,1,-2},
{1,-1,0,0,0,0},
{0,0,1,-1,0,0},
{0,0,0,0,1,-1},
{1,0,0,0,0,0},
{0,0,1,0,0,0},
{0,0,0,0,1,0},
{1,1,0,0,0,0},
{0,0,1,1,0,0},
{0,0,0,0,1,1},
{1,2,0,0,0,0},
{0,0,1,2,0,0},
{0,0,0,0,1,2},
```

```
{1,3,0,0,0,0},
{0,0,1,3,0,0},
{0,0,0,0,1,3},
{1,4,0,0,0,0},
{0,0,1,4,0,0},
{0,0,0,0,1,4}}

awrmv[data, xmatrix, {2,2}, {0,0}, 4]
```