HIERARCHICAL MODELS IN MEDICAL RESEARCH

Thesis submitted for the degree of

Doctor of Philosophy

At the University of Leicester

By

Paul Christopher Lambert B.Sc. M.Sc

Department of Epidemiology and Public Health

University of Leicester

November 2000

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

HIERARCHICAL MODELS IN MEDICAL RESEARCH

Paul Christopher Lambert B.Sc. M.Sc

Abstract

This thesis describes and develops the use of hierarchical models in medical research from both a classical and Bayesian perspective. Hierarchical models are appropriate when observations are clustered into larger units within a data set, which is a common occurance in medical research. The use and versatility of hierarchical models is shown through a number of examples, with the aim of developing improved and more appropriate methods of analysis. The examples are real data sets and present real problems in terms of statistical analysis.

The data sets presented include two data sets involved with longitudinal data where repeated measurements are clustered within individuals. One data set has repeated blood pressure measurements taken on pregnant women and the other consists of repeated peak expiratory flow measurements taken on asthmatic children. Bayesian and classical analyses are compared. A number of issues are explored including the modelling of complex mean profiles, interpretation and quantification of variance components and the modelling of heterogeneous within-subject variances. Other data sets are concerned with meta-analysis, where individuals are clustered within studies. The classical and Bayesian frameworks are compared and one data set investigates the potential to combine estimates from different study types in order to estimate the attributable risk. One of the meta-analysis data sets included individual patient data, where there is a substantial amount of missing covariate data. For this data set models that incorporate individuals with incomplete data when modelling survival times for children with Neuroblastoma are developed.

This thesis thus demonstrates that hierarchical models are of great importance in analysing data in medical research. In many situations a Bayesian analysis provides a number of advantages over classical models especially when introducing realistic complexity that would be hard to incorporate using classical methodology.

Acknowledgments

I would like to thank my supervisor Keith Abrams for all his help and guidance throughout my studies. I am grateful to Paul Burton who encouraged me to pursue a Ph.D., all those years ago, and also for a number of helpful discussions. I would like to thank other members of the medical statistics group in the Department of Epidemiology and Public Health, particularly David Jones and Alex Sutton for helpful and thought provoking discussion of my work. I am grateful to Adrian Brooke for clinical discussion of the peak expiratory flow data and similarly to Aidan Halligan for the ABPM data. I would like to thank the International Neuroblastoma Staging System (INSS) group for allowing me access to the Neuroblastoma data. Finally, and most importantly, thanks and love go to Lucy Smith and my beautiful children, Martha and George.

Contents

1 INTRODUCTION

1.1 Aims of the Thesis

In this thesis I describe and develop the use of hierarchical models in medical research from both a classical and Bayesian perspective. I explore the use of such models through a number of different examples that demonstrate the range of applications that hierarchical models cover. All the examples in this thesis are real datasets and real problems and I explore the potential for improved and more appropriate methods of analysis.

1.2 Hierarchical Data and Hierarchical Models

Hierarchical data occurs when different *levels* of information exist in a data set, in that observations are grouped (or clustered) into larger units, with each unit consisting of a number of observations. For example, patients are grouped into GP practices. The existence of such a hierarchy is non-ignorable which has led to the development of hierarchical models. The use of such models has grown extensively in recent years due to advances in methodology and computing resources. To aid the understanding of hierarchical data and hierarchical models, in Chapter 2 I provide background information to hierarchical data and discuss model fitting and estimation from both a classical and Bayesian perspective. As two of the data sets in this thesis are concerned with repeated measurements, I describe hierarchical models in this context. I demonstrate the dangers of ignoring the hierarchical structure of the data and demonstrate the methods described through use of a simple example.

1.3 Repeated Measures

A common form of hierarchically structured data occurs when there are measurements taken on individuals on more than one occasion. In this situation the repeated measurements are clustered within individuals. Generally with repeated measurements, there is greater variation between individuals than within individuals, leading to high correlation of observations recorded on the same individual. In Chapter 3 I use a data set

that investigates repeated blood pressure measurements over a 24-hour period in pregnant women, with the aim of comparing women who subsequently give birth to an infant with intra-utrine growth retardation to those who do not. In Chapter 4 I use a data set that consists of repeated measurements of peak expiratory flow on young children with the aim of comparing atopic and non-atopic children. Both data sets are analysed from both a classical and Bayesian perspective. A number of issues are explored including interpretation and quantification of variance components and heterogeneity of within-subject variances.

1.4 Meta Analysis

Meta-analysis is the quantitative synthesis of results from different studies. Data from meta-analyses exhibit a hierarchical structure as individual patients are grouped within particular studies and there will be both between-study and within-study variation. Often in a meta-analysis data are only available at the study level in the form of a summary measure and a standard error. In Chapter 5 I present the results from two data sets using both classical and Bayesian models. The first investigates the effect of lowering cholesterol on mortality and the second is a more complex example that combines estimates from three different types of study in order to estimate the attributable risk of a history of infertility on perinatal mortality. In Chapter 6 I present a meta-analysis including individual patient-level data, but the main issue of interest in this chapter is missing covariate data.

1.5 Missing Data

Missing data is a common problem in medical and other areas of research. Hierarchical models are of potential use for missing data problems, as missing data is usually a multivariate problem in that a data set may consist of both response variables and covariates, with any of these variables being potentially missing. The variables can be considered to be clustered within individuals and it is likely that there are interdependencies between the variables. In Chapter 6 I present a data set of survival times for children with Neuroblastoma with the aim of exploring the relationship between tumour markers and survival time. I use hierarchical models to model missing covariate data, with

the main aim of including in the analyses individuals who have some covariates missing as well as those with complete information. I demonstrate how in simple cases multiple imputation techniques can be used using classical methods, but for more complex situations Bayesian models offer far greater flexibility.

1.6 Discussion

In this thesis I investigate the use of hierarchical models in medical research, but many of the findings are applicable to other areas of research. I demonstrate that the use of Bayesian models can offer a number of advantages over classical models, especially in more complex situations. In Chapter 7 I discuss the work I have presented and suggest directions for future research.

2 HIERARCHICAL MODELS

2.1 Introduction

In this Chapter I give an introduction to hierarchical data and hierarchical models. I begin in section 2.2 by introducing the concept of hierarchical data and discuss why it needs to be considered in a different way to non-hierarchical data. In section 2.3 I give a brief summary of techniques used in the analysis of a common form of hierarchically structured data, namely repeated measures (or longitudinal) data. Section 2.4 introduces the concept of hierarchical models by describing how a simple repeated measures analysis could be performed. In section 2.5 I describe methods of estimation for hierarchical models, with Iterative Generalized Least Squares (IGLS) described in section 2.5.1 and a Bayesian approach to estimation using Markov Chain Monte Carlo (MCMC) methods described in section 2.5.2. Section 2.6 demonstrates that if the hierarchical structure is ignored then the results can be biased. A simple example is given in section 2.7. Finally, in section 2.8, I give a summary and discuss the issues raised in the chapter.

2.2 Hierarchical Data

In many situations there is a natural hierarchy to data with there being different *levels* of information. Other ways of describing this is to state that there is a clustered or nested structure (Longford 1993; Goldstein, 1995). Often data of this type is encountered when individuals are grouped (or clustered) into larger units, with each unit consisting of a number of individuals. One simple example is where patients are grouped into GP practices. It is important to recognise hierarchically structured data, since individuals (or units) within the same cluster will tend to be *more* similar than individuals (or units) from different clusters. This is perhaps best illustrated by examples. A first example concerns the family unit where offspring are nested within families, where you would expect there to be similarities *within* families due to both genetic and environmental factors. A second example is where patients are nested within wards, which are nested within hospitals, where there may be similarities between patients in the same ward and patients in the same hospital, for example, specialist and non-specialist hospitals.

Two of the chapters in this thesis are concerned with longitudinal data where repeated observations are nested within subjects. One would generally expect repeated observations made on the same subject to be more similar than observations made on different subjects. There is an extensive literature on longitudinal data and it is discussed further in section 2.3. Much of the recent research regarding hierarchical structures and models has been in educational research where pupils are nested within classes, which are nested in schools (Aitkin and Longford, 1986; Goldstein, 1995). Again one would expect some degree of similarity between children in the same class and between children in the same school.

A further example of clustering is in a multi-centre clinical trial where one may expect there to be similarities between subjects within the same centre. Related to this is meta-analysis, where the results of two or more independent studies are statistically combined. Again one would expect similarities between subjects within the same study. In Chapter 5 I show how hierarchical models can be used in meta-analysis. Another example is cluster-randomised trials where randomisation occurs at, for example, the general practice level rather than at the individual level. There is a growing amount of work in this area demonstrating the need to account for similarities between individuals within the same randomisation unit both in terms of design and analysis (Donner, 1998). A final example of hierarchically structured data is when there is a multivariate response. For example, there may be a number of different outcomes recorded on each individual. Again one would expect the outcomes recorded on the same subject to be more similar than outcomes recorded on different subjects, for example systolic and diastolic blood pressure. I use hierarchical models for multivariate structured data for the analyses involving missing covariate data in Chapter 6.

The main reason why it is important to take into account the hierarchical structure when performing an analysis is that units within clusters tend to be more similar than units from different clusters or alternatively there are observable differences between units. In other words, data with a hierarchical structure induces a correlation structure leading to a lack of independence between measurements within the same cluster. Ignoring this lack of independence, sometimes known as naïve pooling (Burton *et al.*, 1998), can lead to the

wrong or misleading inferences being drawn. A well known example in educational research investigates school effectiveness for 907 pupils in 18 schools (Aitkin and Longford, 1986). In this data set a hierarchical model was shown to be more realistic and could potentially lead to different conclusions than models ignoring the grouping of pupils into schools, estimating a separate intercept for each school and using a summary measure for each school (aggregation of pupil data). Generally, if the clustering is ignored then the main problem with the analysis will be with biased standard errors. Cluster level covariates will tend to have too small standard errors whilst within-cluster level covariates will tend to have too larger standard errors. I demonstrate a simple example of how the wrong inferences can be drawn when ignoring the lack of independence in section 2.6.

I shall adopt the definitions of Goldstein (1986) and Byrk and Raudenbush (1992) in that units are grouped at different *levels* in a hierarchy. Level 1 units are at the lowest level of the hierarchy and level 2 are the units in which the level 1 units are grouped. For example, in a cluster randomised trial, where randomisation is by GP practice, the patients are the level 1 units and the GP practices are the level 2 units, as patients are grouped (or nested or clustered) within GP practices. Hierarchical structures can be extended to situations where there are more levels. For example, in educational research data may be analysed at three levels with pupils (level 1) nested within classes (level2) nested within schools (level 3).

As discussed above, the important aspect regarding hierarchical structures is that we expect there to be differences between the units at each level. For example, we expect there to be differences between practices in a cluster randomised trial or in the case of repeated measures we expect there to be variation between patients. Often, we are not interested in estimating the effect of each of the higher level units, but more interested in describing the features of various groups, for example the mean treatment difference in a randomised controlled trial. The important aspect of hierarchical data structures is that *random effects* can be used to quantify and explore the nested structure of the data. Random effects are coefficients that are allowed to vary between units according to some specified distribution (conventionally Normal). These will be discussed in more detail with application to repeated measures problems in section 2.4.

In the rest of this chapter I will concentrate on hierarchical structures with a continuous response, where the response is assumed to have a Normal distribution. However, it is important to realize that other types of data can have a hierarchical structure, for example binary or count data. In Chapter 5 I investigate hierarchical structured data with a combination of Poisson and Binomial responses and in Chapter 6 the response is (censored) survival time.

In this thesis, I shall use the term hierarchical model when modelling hierarchical data. However, there are various other names used in the literature such as multilevel models (Goldstein, 1986), random effects models (Laird and Ware, 1982), mixed effects models (Breslow and Clayton, 1993), random coefficients models (Longford, 1993; Rutter and Elashoff, 1994), and variance component models (Longford, 1987). The Laird and Ware paper has become so well known that hierarchical models in the context of longitudinal data are often referred to as Laird-Ware models. The term hierarchical models was first introduced by Lindley and Smith (1972) in the investigation of Bayesian estimation of linear models. However, hierarchical models were not really used much in practice, mainly due to computational problems, until the introduction of the EM algorithm (Dempster *et al.*, 1977). The EM algorithm was shown to be appropriate for hierarchical data by Dempster *et al.* (1981), and Laird and Ware (1982).

Other estimation procedures have been developed for hierarchical models, a Fisher scoring algorithm (Longford, 1987) and a generalized least squares algorithm (Goldstein, 1986). The latter known as Iterative Generalised Least Squares (IGLS) is one of the approaches I use in this thesis. Methods of estimation are discussed in further detail in section 2.5. It is also worth stating the increase in use of hierarchical models has been largely due the development of specialist statistical software. The most common of these is MLWin (Rasbash *et al.*, 1999) and its predecessors (MLn, ML3 and ML2) developed at the Institute of Education by Harvey Goldstein and colleagues, and HLM developed by Raudenbush and Byrk (1988). Other programs exist such as VARCL and MIXREG as well as implementation in certain larger statistical software packages, e.g. PROC MIXED (Littel *et*

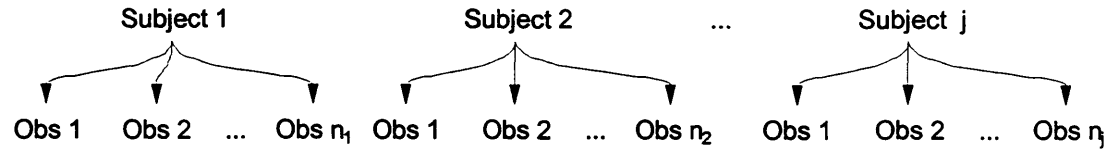

Figure 2.1 Hierarchical structure of repeated measures data.

al., 1996) in SAS and the "lme" function in Splus (Mathsoft 1996). A comparative review of some of the software for fitting hierarchical models can found in Kreft (1994).

2.3 Repeated Measures Data

In this and the following section I will describe hierarchical models for the analysis of repeated measures data, since two of the chapters in this thesis relate to this type of data. However, the analyses and interpretation of hierarchical data is similar across disciplines. For example, in a cluster randomised trial one may interested in the between-general practice variation, while in a repeated measures analysis one is interested in the between-subject variation.

As stated in the previous section it is sensible to think of longitudinal (or repeated measures data) as forming a hierarchy with repeated observations being nested within individual subjects. This hierarchical structure can be seen in Figure 2.1. It has long being realised that when dealing with repeated measures data one explicitly needs to take account of the correlation that the hierarchical structure induces. There is an extensive literature on repeated measures analysis, examples of which are (Crowder and Hand, 1993; Lindsey, 1993; Diggle, Liang, and Zeger, 1994; Hand and Crowder, 1996). The simplest method of analysing repeated measures data is to use summary measures (Matthews *et al.*, 1990). A summary measure is obtained for each individual, which summarises a feature of the response (e.g. mean, slope, time to maximum value etc). In reducing each individual's repeated observations to a single summary measure standard statistical techniques, such as the t-test, can be used to analyse the data. However, although they have the advantage of being relatively simple, there is a potential large decrease in power due to the loss of

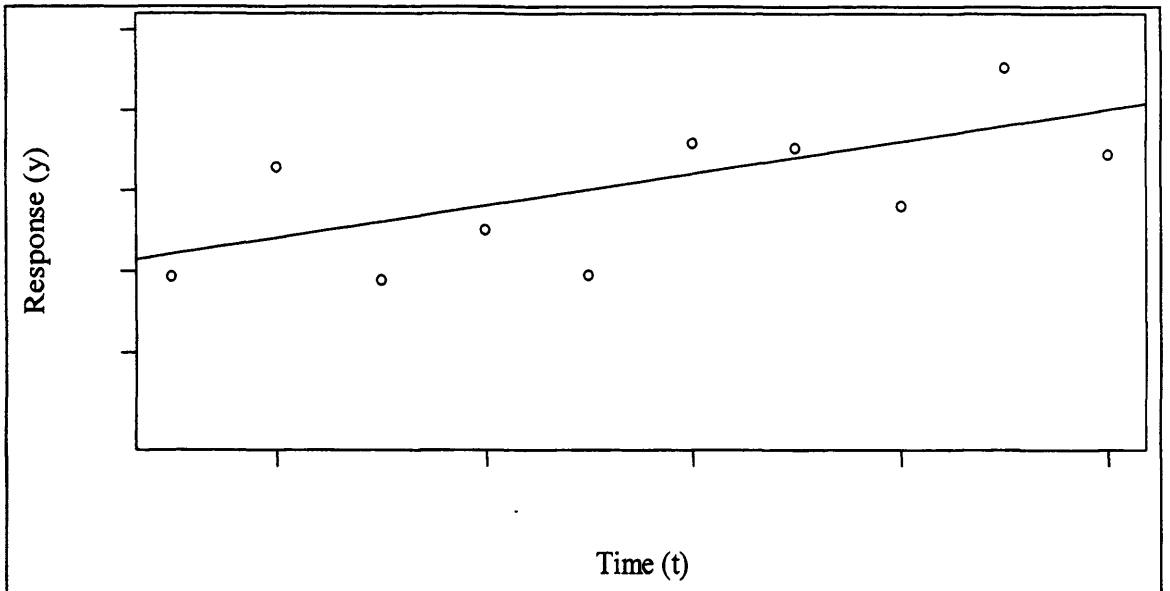

Figure 2.2 Plot showing within-subject variation about a regression line

information (Matthews *et al.*, 1990). Two common techniques for the analysis of repeated measures data that do not require a summary measure to be calculated are the Multivariate Analysis of Variance (MANOVA) and univariate Analysis of Variance (ANOVA) based on the agricultural split plot design (Crowder and Hand, 1993). However, there are a number of problems with these techniques (Diggle, Liang, and Zeger, 1994). Firstly, there is the need to have the same number of repeated observations per individual, measured at the same time points. Secondly, with ANOVA split plot models the assumption of the correlation structure is usually too simplistic (with essentially one variance and one covariance term estimated) and with MANOVA models the correlation structure is over-parameterised (with all possible variances and covariances estimated). With the development of hierarchical random effects models and the ability to model the covariance structure it is likely that the use of these two methods will reduce over time.

2.4 Hierarchical Models for Repeated Measures

This section describes models for repeated measures using hierarchical or multilevel models. The notation is similar to that of Goldstein (1995). Figure 2.2 shows a hypothetical response (y) for an individual where there is a linear increase over time (t). The regression line represents the mean response of y as t increases, while the points scattered about the

regression line show variation about this mean. In this case, such variation can be thought of as *within-subject variation*. A simple regression line can be fitted to data such as this

$$y_i = \beta_0 + \beta_1 t_i + e_i \qquad (2.1)$$

where y_i is the response at the i^{th} time point, t_i is the time at the i^{th} time point, β_0 is the intercept, β_1 is the gradient and e_i is the residual at the i^{th} time point. Generally it is assumed that $e_i \sim N(0, \sigma^2)$ and $cov(e_i, e_j)=0$.

Of course studies are generally not performed on only one subject. If there are a number of subjects then it is likely that there will be *between-subject variability* as it is unlikely that all subjects would have the same response, i.e. some individuals will tend to have higher values of the response. When between-subject variability exists then observations on the same subject will tend to be correlated. For example, it will be seen in Chapter 3 that blood pressures recorded on the same subject are likely to be more similar than blood pressures recorded on different subjects. Thus, with repeated measures on each individual there will be both *between-subject* and *within-subject* variation. This can be explained graphically.

Figure 2.3 shows both between-subject and within-subject variation for a hypothetical response that increases linearly over time, recorded on four subjects on ten occasions. The thick line shows the mean population response over time. The thin lines indicate each of the four subjects' mean responses. It can be seen that these vary about the population mean response. The thin lines therefore demonstrate between-subject variability. There is also variation associated with measurements taken within subjects. This is represented by the symbols that vary about each of the thin lines. It is possible to consider this sort of data as forming a two level hierarchy. The level 1 units are the repeated observations and these are nested within the level 2 units (the subjects). It can be assumed that the mean response for each subject is randomly distributed around the underlying response in the population as a whole. This leads to level 2 or between-subject variation. In a similar way, the repeated observations for each subject are assumed to be randomly distributed around their underlying mean response, thus leading to within-subject or level 1 variability. A model to fit this data in Figure 2.3 can be of the form.

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + u_j + e_{ij} \tag{2.2}$$

where y_{ij} is the response at the i^{th} time point for the j^{th} subject, t_{ij} the is the time at the i^{th} time point for the j^{th} subject, β_0 is the intercept, β_1 is the gradient, u_j is the effect of the j^{th} subject and e_{ij} is the residual at the i^{th} time point, where u_j and e_{ij} are assumed to be normal random variables, with the following parameters

$$u_j \sim N(0,\sigma_u^2), \quad e_{ij} \sim N(0,\sigma^2) \tag{2.3}$$

and $\text{cov}(e_{ij},e_{kj})=0$. Incorporating the random effect u_j leads to the intercept being a random coefficient in that it varies between subjects. This can be seen by rewriting (2.2) as

$$\begin{aligned} y_{ij} &= \beta_{0j} + \beta_1 t_{ij} + e_{ij} \\ \beta_{0j} &= \beta_0 + u_j \end{aligned} \tag{2.4}$$

The model in (2.4) is flexible as not only does it model the fixed parameters (β_0 and β_1), but also the components of variance, in that both between-subject (σ_u^2) and within-subject variation (σ^2) are being estimated. In order to investigate how similar the level 1 units are (in this case the repeated measurements on each subject) the *intraclass correlation coefficient* (ICC) can be estimated (Goldstein, 1995). This is given by the formula

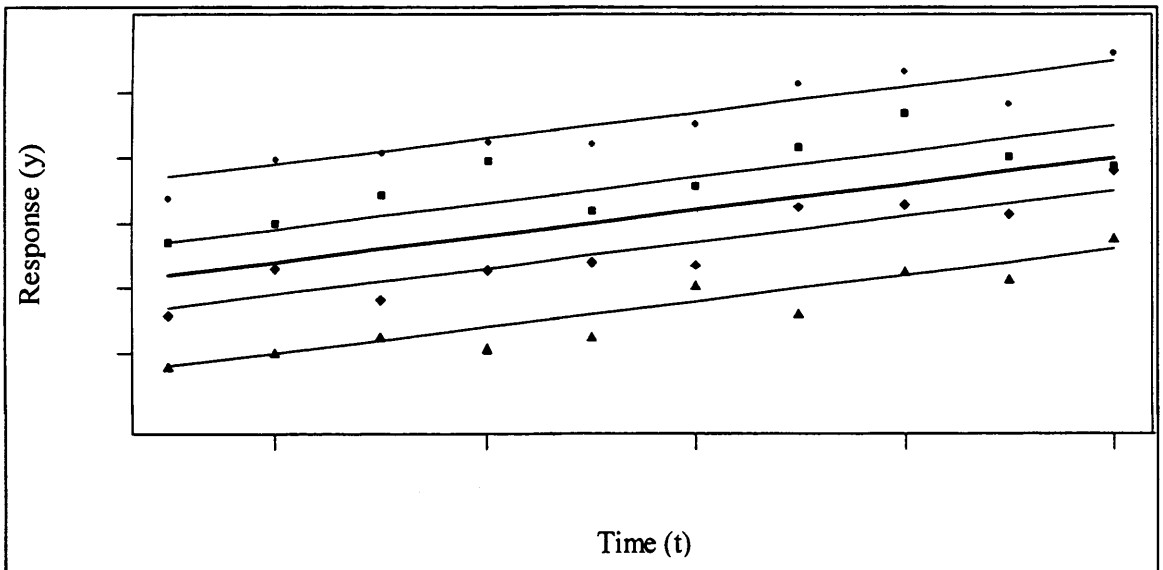

Figure 2.3 Plot showing between and within-subject variation for subjects with a positive linear response over time (different symbols represent different subjects).

$$\rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma^2} \tag{2.5}$$

This is a measure of the proportion of the total variation that is explained by the variation between the level 2 units, in this case subjects, and can be thought of as the correlation between observations made on the same subject. It can be seen that if there is no between-subject variability then ρ=0 and standard statistical methods could be used. However, it must be stressed that this is an unlikely situation with hierarchical data, especially repeated measures data.

It may be of interest not only to describe the mean response of a population over a period of time, but also to quantify how different the subjects are and by how much individuals vary with regard to their responses. If there were only a few subjects then it is possible to treat the u_j's as fixed effects, i.e. obtain an estimate of u_j for each subject. However, if there are a large number of subjects then the number of parameters that need to be estimated will become large. If there is an interaction between subject and time then the number of parameters that need to be estimated will become even larger. The hierarchical model has appeal, as it is not sensible for the regression coefficients to be the same for all subjects and the approach can be considered a sensible compromise between separate models for each subject and models where the coefficients are forced to be equal. The compromise is sensible because a hierarchical model will give more stable estimates than separate models and more interesting parameters than equal coefficients (de Leuuw and Kreft, 1995).

Figure 2.3 may be unrealistic because it assumes that the rate of change in the response over time is the same for all subjects, i.e. the gradients are the same. It may be sensible to consider a response as shown in Figure 2.4, where not only the intercept varies from subject to subject but so too does the gradient. A model to capture this feature can be defined as follows,

$$y_{ij} = \beta_{0j} + \beta_{1j} t_{ij} + e_{ij} \tag{2.6}$$

$$\beta_{0j} = \beta_0 + u_{0j}, \quad \beta_{1j} = \beta_1 + u_{1j}$$

where u_{0j}, u_{1j} and e_{ij} are assumed to be normal random variables, with the following parameters,

$$\underline{u}_j \sim MVN\left[\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}\sigma^2_{u0} & \sigma_{u01}\\ \sigma_{u01} & \sigma^2_{u1}\end{pmatrix}\right], \quad e_{ij} \sim N(0,\sigma^2) \tag{2.7}$$

Where MVN[-,-] denotes a Multivariate Normal distribution.

Using this formulation the variation in intercepts is quantified by σ^2_{u0} and the variation in gradients is quantified by σ^2_{u1}. Note that there is a covariance term (σ_{u01}) for the two between-subject (level 2) random effects. By inspection of Figure 2.4 it becomes apparent why a covariance term may be needed. In the figure the subjects with high intercepts tend to have smaller gradients leading to a negative association between u_{0j} and u_{1j}, and thus a negative covariance term. A positive covariance would indicate that subjects with higher intercepts would also tend to have larger gradients.

The model in (2.6) can be rewritten in the following form,

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + u_{0j} + u_{1j} t_{ij} + e_{ij} \tag{2.8}$$

or of the form

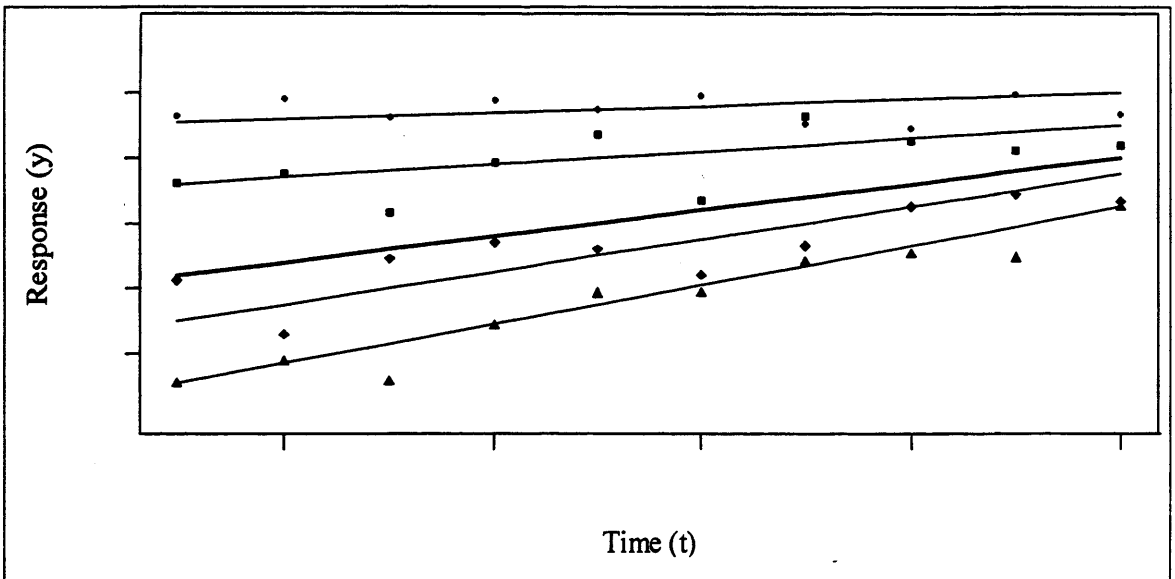

Figure 2.4 Plot where both the gradient and intercept varies between subjects (different symbols represent different subjects)

$$y_{ij} = (\beta_0 + u_{0j}) + (\beta_1 + u_{1j}) t_{ij} + e_{ij} \quad (2.9)$$

In this simplistic example, I have assumed that all subjects were measured at the same time points and there are no missing values. However, a very important aspect about hierarchical models is that level 2 units can have a different number of level 1 units, so in the repeated measure example, different subjects could be measured at different time points and have a different number of observations. These examples are also simplistic in that they assume a linear effect of time and the variation about the mean response is a function of the linear predictor. However, it is possible to extend these models to situations where the mean profile is a more complex function of time and the between-subject variance is also of a more complex form. The models can be extended in the same way as standard linear models to situations where groups of individuals can be compared, e.g. treatment arms in a clinical trial and where adjustments are made for confounding variables. Another aspect I will demonstrate is where the level 1 (within-subject variation) can also be modelled. For example,

$$\sigma_{ij}^2 = \alpha + \beta t_{ij} \quad (2.10)$$

where the within-subject variance is a function of time. This is known as complex level 1 variation (Goldstein, 1995).

2.5 Estimation

It was seen in section 2.4 that when fitting a hierarchical model, both fixed effects and random effects need to be estimated. There are a number of techniques available for estimating the parameters including Iterative Generlized Least Squares (IGLS) (Goldstein, 1986), a Fisher Scoring method (Longford, 1987), and an EM algorithm (Bryk and Raudenbush, 1992). Alternatively, within a Bayesian framework it is possible to use Markov Chain Monte Carlo (MCMC) methods, notably Gibbs Sampling (Zeger and Karim, 1991). In this thesis the program MLn (Rasbash and Woodhouse, 1995) is used for fitting hierarchical models. This program uses Iterative Generalized Least Square (IGLS) which is described in section 2.5.1. I also consider Bayesian methodology for fitting hierarchical models using the BUGS and WinBUGS programs (Spiegelhalter *et al.*, 1996). The

methodology for this is described in section 2.5.2. Descriptions of the other methods can be found in Goldstein (1995) and Longford (1992).

2.5.1 Iterative Generalized Least Squares (IGLS)

For the models described in the previous section it was shown that a hierarchical model consists of both fixed effects and random effects. Furthermore, for a two level model there can be random effects at both level 2 and level 1. Let N be the total number of observations (also the number of level 1 units), M be the number of level 2 units (the number of subjects in a repeated measures analysis) and n_j the number of observations for the j^{th} level 2 unit. Note that

$$\sum_{j=1}^{M} n_j = N \tag{2.11}$$

A general two level model can be written, using the notation similar to that of Goldstein (1988), as

$$Y = X\beta + Z^{(2)}u + Z^{(1)}e \tag{2.12}$$

where Y is the $(N\times 1)$ vector of responses, X is the $(N\times p)$ design matrix for the p fixed effect parameters, β is the $(p\times 1)$ vector of fixed effect parameters, $Z^{(2)}$ is the $(N\times Mq_2)$ design matrix for the q_2 random effect parameters at level 2, $Z^{(1)}$ is the $(N\times Mq_1)$ design matrix for the q_1 random effect parameters at level 1, u consists of M random sub-vectors $u_j, j=1,\ldots,M$, each with q_2 components, and e consists of N random sub-vectors e_{ij}, $i=1,\ldots,N, j=1,\ldots,M$, each with q_1 components.

Since u and e denote the random effects it is assumed that

$$u \sim MVN(0,\Omega_2), \quad e \sim MVN(0,\Omega_1) \tag{2.13}$$

where Ω_2 is the covariance matrix for the level 2 random effects and Ω_1 is the covariance matrix for the level 1 random effects.

The above model can be illustrated by considering the example of the 2 level model described in (2.6)-(2.7) for a linear model with random intercepts and slopes. This can be written in the form of (2.12) as follows;

$$\begin{pmatrix} y_{11} \\ y_{21} \\ \vdots \\ y_{n_jM} \end{pmatrix} = \begin{pmatrix} 1 & t_{11} \\ 1 & t_{21} \\ \vdots & \vdots \\ 1 & t_{n_jM} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \text{diag} \left(\begin{bmatrix} 1 & t_{11} \\ \vdots & \vdots \\ 1 & t_{n_11} \end{bmatrix} \cdots \begin{bmatrix} 1 & t_{1M} \\ \vdots & \vdots \\ 1 & t_{n_mM} \end{bmatrix} \right) \begin{pmatrix} \begin{bmatrix} u_{01} \\ u_{11} \end{bmatrix} \\ \vdots \\ \begin{bmatrix} u_{0M} \\ u_{1M} \end{bmatrix} \end{pmatrix} + I_N \begin{pmatrix} e_{11} \\ e_{21} \\ \vdots \\ e_{n_jM} \end{pmatrix} \qquad (2.14)$$

Note that 'diag' indicates that the matrix is block diagonal, with the blocks denoted by the square brackets. Since at level 1 there is only one variance term (the within-subject variance is assumed constant) the design matrix is the identity matrix.

Also from (2.13) it is assumed that $u \sim MVN(0,\Omega_2)$, $e \sim MVN(0,\Omega_1)$, In this case

$$\Omega_2 = \begin{pmatrix} \sigma^2_{u0} & \sigma_{u01} \\ \sigma_{u01} & \sigma^2_{u1} \end{pmatrix} \quad \textit{and} \quad \Omega_1 = \left(\sigma^2\right) \qquad (2.15)$$

These are the covariance matrices for the random effects and are also known as the *variance components*. In the repeated measures example here, Ω_2 denotes the between-subject variation about the mean slope and Ω_1 the within-subject variation, which is assumed to be constant over time and also equivalent between subjects.

When estimating the parameters in a hierarchical model, the variance of the outcome Y conditional upon the fixed effects is needed.

$$V = Var(Y \mid X\beta) = E[(Y - X\beta)(Y - X\beta)^T] \qquad (2.16)$$

From (2.12) it can be seen that $Y\text{-}X\beta$ is equal to $Z^{(2)}u+Z^{(1)}e$, so

$$V = E[(Z^{(2)}u + Z^{(1)}e)(Z^{(2)}u + Z^{(1)}e)^T] \tag{2.17}$$

$$= Z^{(2)}Var(u)Z^{(2)^T} + Z^{(1)}Var(e)Z^{(1)^T}$$

From (2.13) it is known that Var(u)=Ω_2 and Var(e)=Ω_1, and so

$$V = Z^{(2)}\Omega_2 Z^{(2)^T} + Z^{(1)}\Omega_1 Z^{(1)^T} \tag{2.18}$$

V will block diagonal, which is sensible, as observations made on different level 2 units are considered independent. For the repeated measures linear model example in (2.6)-(2.7) each block would be of the form

$$\begin{pmatrix} \sigma^2_{u0} + 2t_{1j}\sigma_{u01} + t^2_{1j}\sigma^2_{u1} + \sigma^2 & \sigma^2_{u0} + (t_{1j} + t_{2j})\sigma_{u01} + t_{1j}t_{2j}\sigma^2_{u1} & \cdots & \sigma^2_{u0} + (t_{1j} + t_{n_jj})\sigma_{u01} + t_{1j}t_{n_jj}\sigma^2_{u1} \\ \sigma^2_{u0} + (t_{1j} + t_{2j})\sigma_{u01} + t_{1j}t_{2j}\sigma^2_{u1} & \sigma^2_{u0} + 2t_{2j}\sigma_{u01} + t^2_{2j}\sigma^2_{u1} + \sigma^2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^2_{u0} + (t_{1j} + t_{n_jj})\sigma_{u01} + t_{1j}t_{n_jj}\sigma^2_{u1} & \cdots & \cdots & \sigma^2_{u0} + 2t_{n_jj}\sigma_{u01} + t^2_{n_jj}\sigma^2_{u1} + \sigma^2 \end{pmatrix} \tag{2.19}$$

If Ω_1 and Ω_2, and hence V, are known then it is possible to use standard Generalised Least Squares (GLS) (McCullagh and Nelder, 1989) to estimate the fixed effect parameters β, such that

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y \tag{2.20}$$

with covariance matrix

$$V(\hat{\beta}) = (X^T V^{-1} X)^{-1} \tag{2.21}$$

However, in practice V will not be known, and it will have to be estimated from the data. The IGLS method works through the following steps (Goldstein, 1986);

1) Obtain an initial estimate of β (usually assuming all observations are independent, i.e. $V=I_N$)
2) Estimate Ω_1 and Ω_2 using the residuals for the current estimate of β enabling V to be estimated using (2.18).
3) Re-estimate β using the updated estimate of V.
4) Repeat steps 2 and 3 until to convergence, at a pre-specified level of tolerance, of both the fixed effects and the variance components in Ω_1 and Ω_2.

Equation (2.20) shows how step 3 is performed, but obtaining the estimates of Ω_1 and Ω_2 is slightly more complex and uses residuals. The residuals of a hierarchical model are defined as

$$\tilde{Y} = Y - X\hat{\beta} \tag{2.22}$$

Let the cross-products of the residuals be defined by Y^* where

$$Y^* = \tilde{Y}\tilde{Y}^T \tag{2.23}$$

From (2.16) it can be seen that $E(Y^*)=V$. It was seen in (2.18) and (2.19) that V is a linear function of the parameters in Ω_1 and Ω_2. Therefore, if the appropriate design matrix is obtained then the parameters can be estimated using GLS. Y^* is symmetric and so only the lower triangle of the matrix is needed. Thus,

$$Y^{**} = \text{vech}(Y^*) \tag{2.24}$$

Where the function 'vech' means that Y^{**} is formed by stacking the columns of the lower triangle of Y^*. The linear model of the variance components can then be written

$$E(Y^{**}) = Z^*\theta \tag{2.25}$$

where Z^* is the design matrix for the variance components and θ is the vector of parameters in Ω_1 and Ω_2. The GLS estimates of θ, $\hat{\theta}$, can be obtained by

$$\hat{\theta} = (Z^{*T}V^{*-1}Z^*)^{-1}Z^{*T}V^{*-1}Y^{**} \tag{2.26}$$

where V^* is the covariance matrix of Y^{**} and is defined by

$$V^* = V \otimes V \tag{2.27}$$

where $\otimes$ is the Kronecker product which multiplies every element of the left hand matrix by each element of the right hand matrix.

Assuming multivariate normality it is possible to obtain a covariance matrix for $\hat{\theta}$ such that.

$$\text{cov}(\hat{\theta}) = 2(Z^{*T}V^{*-1}Z^*)^{-1} \tag{2.28}$$

However, when estimating the variance components the assumption of multivariate normality may be unrealistic (Goldstein, 1995). Although this should not bias the estimates of the variance components themselves, it can lead to biased estimates of the variances of the variance components. Thus, they should be used with caution.

For the simple repeated measure linear model example in (2.6)-(2.7) where each block of V is shown in (2.19) the variance components are estimated as follows. Y^{**} is formed by stacking the following vectors, y_j^*

$$y_j^* = \begin{pmatrix} \tilde{y}_{11}^2 \\ \tilde{y}_{21}\tilde{y}_{11} \\ \vdots \\ \tilde{y}_{n_j1}\tilde{y}_{11} \\ \tilde{y}_{21}^2 \\ \vdots \\ \tilde{y}_{n_j1}\tilde{y}_{21} \\ \vdots \\ \tilde{y}_{n_j1}^2 \end{pmatrix} \tag{2.29}$$

Z^* is formed by stacking the following matrices z_j^*

$$z_j^* = \begin{pmatrix} 1 & 2t_{1j} & t_{1j}^2 & 1 \\ 1 & t_{2j}+t_{1j} & t_{2j}t_{1j} & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_{n_jj}+t_{1j} & t_{n_jj}t_{1j} & 0 \\ 1 & 2t_{2j} & t_{2j}^2 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_{n_jj}+t_{2j} & t_{n_jj}t_{2j} & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 2t_{n_jj} & t_{n_jj}^2 & 1 \end{pmatrix} \tag{2.30}$$

Finally θ is the vector of parameters to be estimated.

$$\theta^T = \begin{pmatrix} \sigma_{u0}^2 & \sigma_{u01} & \sigma_{u1}^2 & \sigma^2 \end{pmatrix} \tag{2.31}$$

Futher details regarding obtaining the design matrix for the random effects can be seen in Goldstein and Rasbash (1992). Although I have demonstrated the estimation procedure for two levels, the procedure can be easily extended to situation with three or more levels.

Usually the random components (the u_j's) are not of specific interest, but they can be useful for model checking purposes or for showing fitted individual profiles. Unlike a general linear model, there are more than one set of residuals, with the potential for several sets residuals at each of the different levels. For example, for model (2.18) there will be three sets of residuals, two at the between-subject level (the u_{0j}'s and the u_{1j}'s) and one at the within-subject level (the e_{ij}'s). The residuals are not obtained directly in the model estimation, but using the parameter estimates of the hierarchical model it is possible to obtain *shrunken residuals*. A shrunken residual is the expected value of the random components *conditional on* the estimated fixed and random components. For each level of the model, the residuals are estimated by regressing the random component on the overall residual component $\tilde{Y}$. Thus, for a hierarchical model with h levels the level h residuals can be obtained using

$$\hat{u} = R_h^T V^{-1} \tilde{Y} \qquad (2.32)$$

Where R_h is block diagonal, with each block corresponding to a level h unit, and for the j^{th} block is given by

$$Z_j^h \Omega_h \qquad (2.33)$$

Further details of residuals in hierarchical models can be found in Goldstein(1995).

Goldstein (1986) shows that IGLS estimates are equivalent to Maximum Likelihood (ML) estimates. However, it is known that maximum likelihood estimates of the variance components are biased. The reason for this is that the procedure ignores the sampling variability in the estimates of fixed effects. Goldstein (1989) extends his work on IGLS to obtain Restricted Iterative Generalised Least Squares (RIGLS) estimates. He shows that RIGLS is equivalent to restricted maximum likelihood (REML). The choice of IGLS and RIGLS becomes important in small samples. However, the datasets analysed in this thesis

are of a sufficiently large size that use of RIGLS instead if IGLS makes no practical difference. There is an important issue in the definition of small. There may be a large number of level 1 units, but very few level 2 units. In this situation it may be advisable to use RIGLS as there is a small sample size with regard to the level 2 units.

2.5.2 *Bayesian Methodology for Hierarchical Models*

There has been a long debate of the various advantages and disadvantages of Bayesian and Frequentist statistics. However, much of this debate has been regarding philosophical rather than practical aspects of analysis (de Finetti, 1972; Lindley, 1985). There is increasing interest in the use of Bayesian statistical methods in many fields including epidemiology and biostatistics (Kadane, 1995; Berry and Stangl, 1996; Lilford and Braunholtz, 1996; Spiegelhalter *et al.*, 1999). I will not go into detail about the advantages and disadvantages of Bayesian statistics, these can be found in the above references. My adoption of Bayesian methods is pragmatic rather than philosophical, although I do have leanings towards the philosophical arguments. The five main reasons why I believe Bayesian methodology may be useful in the analysis of hierarchical models are as follows;

1) the ability to take appropriate account of all forms of recognised uncertainty;
2) the ability to deal with problems that would be difficult (or impossible) using classical methods;
3) the ability to provide more meaningful interpretations of data;
4) the ability to include pertinent information external to the current study;
5) the ability to predict future/unknown observations.

An important difference between the Frequentist and Bayesian frameworks is in the definition of probability. I will briefly explain the differences between the two frameworks in terms of the comparison of two treatments (a new treatment, A, and the current treatment, B) in a clinical trial where $\underline{x}$ denotes the observed data and θ the parameter of interest (the treatment difference).

The Frequentist approach to inference attempts to demonstrate a treatment difference by assuming that there is not a difference, but then showing that it is unlikely that you would have obtained your data if this were actually the case. This is done by defining a null hypothesis that there is not a treatment difference, choosing an appropriate test statistic and then calculating the probability of observing the test statistic to be as extreme as that obtained given that the null-hypothesis is true. This is, of course the definition of a P-value. Many non-statisticians are confused by this definition and actually interpret P-values as if they were Bayesian probabilities.

The Bayesian approach starts with the observed treatment difference and obtains the probability that, for example, treatment A is better than treatment B. This is of course the question most clinical researchers want answered. Although the interpretation of the Bayesian analysis is simpler, there is a penalty to pay in the form of a prior distribution. One must specify the prior beliefs of the treatment difference before conducting the analysis.

The difference in interpretation between the two frameworks is because P-values are based on a hypothesis testing framework, which is based on an inverse argument, $P(\underline{x} \mid \theta)$, while what most people want to have is $P(\theta \mid \underline{x})$. It is worthwhile stating that in Frequentist inference probabilities refer to the long-run frequencies of repeatable events. In Bayesian statistical inference probabilities are subjective in that a probability can be attached to any event. For example, I could state that "the probability of Ipswich Town not being relegated from the Premier League in the 2000/2001 season is 0.7" (more optimistic than the bookmakers!). A Frequentist could not make this statement, as the event is not repeatable.

A similar problem occurs when obtaining confidence intervals. Many people interpret a Frequentist (95%) confidence interval as if there was a 95% probability that the interval contained the parameter of interest. Since in the Frequentist approach, parameters are considered to be fixed, the parameter is either in the interval or it is not. The correct interpretation of a 95% confidence interval is that it implies that if the experiment were repeated again and again then 95% of the confidence intervals would be expected to contain the true parameter value. Within the Bayesian framework, credible intervals can be

obtained, where a 95% credible interval infers that there is a 95% chance that the parameter of interest lies within the interval.

In the previous section I showed how the parameters $\underline{\theta}$ in a hierarchical model with observed data $\underline{x}$ can be estimated using IGLS which is equivalent to maximising the likelihood, $p(\underline{x}|\underline{\theta})$. In a Bayesian analysis the unknowns, $\underline{\theta}$ need to be given a distribution that reflects the uncertainty about these parameters *before* the data has been observed which may include knowledge based on external evidence. This is known as the *prior distribution* for $\underline{\theta}$. We then want to update our knowledge of $\underline{\theta}$ in light of seeing the data, $\underline{x}$. This is known as the *posterior distribution* and is defined by $p(\underline{\theta}|\underline{x})$. The prior is linked to the posterior distribution through Bayes Thereom.

$$p(\underline{\theta} \mid \underline{x}) = \frac{p(\underline{x} \mid \underline{\theta})p(\underline{\theta})}{\int p(\underline{\theta})p(\underline{x} \mid \underline{\theta})d\underline{\theta}} \tag{2.34}$$

Generally it is not necessary to calculate the denominator so

$$p(\underline{\theta} \mid \underline{x}) \propto p(\underline{x} \mid \underline{\theta})p(\underline{\theta}) \tag{2.35}$$

One of the main criticisms of the Bayesian approach is that the results of an analysis will be dependent on the prior distributions. The choice of prior distribution may vary between individuals and thus these different individuals could potentially reach different conclusions after analysing the same data. Furthermore, in multi-parameter settings, the specification of prior beliefs can be very complex. To overcome these problems it has been suggested that *vague* or *non-informative prior distributions* should be used. With such prior distribution, it assumed that very little is known about the parameters before the analysis and so the data, through the likelihood, dominates the prior distribution. In this thesis I always use vague prior distributions. As I use the WinBUGS software for Bayesian analysis the prior distributions I use are proper, in that they are expressed in the form of a parametric distribution. However, it is relatively easy to make these vague by, for example, using very large variances with a Normal distribution. In addition, one can perform sensitivity analyses in order to investigate how sensitive any parameter estimates are to the choice of prior distributions.

The main reason why Bayesian methods have been little used in practice until the last few years is because of computational problems. This is because although the likelihood will be the product of simple terms, problems occur when making inferences regarding specific parameters as the other parameters need to be 'intergrated out', i.e.

$$p(\theta_k \mid \underline{x}) = \int_k p(\underline{\theta} \mid \underline{x}) d\underline{\theta}_{\backslash k} \tag{2.36}$$

$p(\theta_k \mid \underline{x})$ is the marginal posterior distribution for θ_k with the '$\backslash k$' notation denoting all parameters other than θ_k.

Interest may sometimes lie in the prediction of a future observation. This should be conditional on the data and the prior distributions. Within the Bayesian framework it is possible to obtain such prediction through the use of a *predictive distribution*, $p(y|\underline{x})$, where y is the future observation. A predictive distribution can be defined as

$$p(y \mid \underline{x}) = \int p(y \mid \underline{\theta}) p(\underline{\theta} \mid \underline{x}) \partial \underline{\theta} \tag{2.37}$$

In Chapter 6 I use predictive distributions, not to predict future observations, but to predict missing values for covariates.

When obtaining marginal posterior or predictive distributions it may not be possible (or very complex) to do numerical or analytical integration. For details of these methods see Thisted (1988). However, with the developments of *Markov Chain Monte Carlo* (MCMC) methods it has become relatively simple to obtain *samples* from the *joint posterior distribution* $p(\underline{\theta}|\underline{x})$ (Gilks, Richardson, and Spiegelhalter, 1996). As stated by Draper (1998), "I start out wanting to compute a probability density $p(\underline{\theta}|\underline{x})$, but then I notice after thinking about it, I would be just as happy to have a large sample for as to know its precise form". With a large sample it is possible to approximate marginal statistics of interest from the joint posterior such as means, medians, standard deviations and quantiles. If the sample is large enough then these will be estimated to a high degree of accuracy. The simplest form of MCMC is *Gibbs Sampling*. In Gibbs Sampling after starting values have been chosen or randomly generated for each parameter, then each parameter is then sampled from a distribution conditional on the *current* values of all other parameters and the data. The procedure loops through all parameters many times leading to a large sample from the

joint posterior distribution. If all parameters are contained in the vector $\underline{\theta}$, then estimates are each parameter can be obtained using the following three steps.

1) Choose starting values for $\underline{\theta}$ $(\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_k^{(0)})$

2) Sample $\theta_1^{(1)}$ from $p(\theta_1 \mid \theta_2^{(0)}, \theta_3^{(0)}, \cdots, \theta_k^{(0)}, \underline{x})$

   Sample $\theta_2^{(1)}$ from $p(\theta_2 \mid \theta_1^{(1)}, \theta_3^{(0)}, \cdots, \theta_k^{(0)}, \underline{x})$

   $\vdots$

   Sample $\theta_k^{(1)}$ from $p(\theta_k \mid \theta_1^{(1)}, \theta_3^{(1)}, \cdots, \theta_{k-1}^{(1)}, \underline{x})$

3) Repeat step 2 many 1000's of times and eventually obtain a large sample from $p(\underline{\theta} \mid \underline{x})$.

Note that $\theta_k^{(l)}$ refers to the l^{th} iteration of parameter θ_k.

In the 1000's of repeats of step 2 a sample is obtained from the *full conditional distribution* for each parameter. A full conditional distribution for a parameter is the distribution of that parameter conditional on the data and all other parameters in the model. Due to ergodic theory, the full conditional distribution tends to the marginal posterior distribution (Gelfand and Smith, 1990). Although Gibbs Sampling can be implemented in any statistical programming environment, the use of a user friendly package, such as BUGS (Spiegelhalter *et al.* 1996) or the more recent WinBUGS (Spiegelhalter, Thomas, and Best, 1999), has made it more practical to applied statisticians.

When using Gibbs Sampling to fit a Bayesian hierarchical model *graphical models* (also known as conditional independence models) are useful for breaking down complex models into simple constituent components, for communicating the essential structure of the model and for providing the basis for computation (Whittaker, 1990). Graphical models can be represented by a graph known as a *Directed Acyclic Graph* (DAG) (Best *et al.*, 1996). As an example consider the following model:

$$y_{ij} = \beta_0 + \beta_1(t_{ij} - \bar{t}) + \beta_{0j} + \beta_{1j} + e_{ij} \qquad (2.38)$$

$$\beta_{0j} \sim N(0, \sigma^2_{\beta_0}),\ \beta_{1j} \sim N(0, \sigma^2_{\beta_1}),\ e_{ij} \sim N(0, \sigma^2)$$

Note that no covariance is assumed between β_{0j} and β_{1j} as t_{ij} is centred about its mean. The model can be re-expressed in the following hierarchical form:

$$
\begin{aligned}
y_{ij} &\sim N(\mu_{ij}, \sigma^2) \\
\mu_{ij} &= \beta_{0j} + \beta_{1j}(t_{ij} - \bar{t}) \\
\beta_{0j} &\sim N(\beta_0, \sigma^2_{\beta_0}) \\
\beta_{1j} &\sim N(\beta_1, \sigma^2_{\beta_1})
\end{aligned}
\tag{2.39}
$$

The model is shown graphically as a DAG as in Figure 2.5. Each component of the model appears in the DAG as a *node*. Solid arrows represent probabilistic (or stochastic) dependencies, while dashed arrows represent deterministic (or functional) dependencies. For example y_{ij} is a stochastic node and thus has solid arrows from σ^2 and μ_{ij}. μ_{ij} is a deterministic node and thus, has dashed arrows leading to it.

The important aspect about expressing models in this graphical form is that it can be shown that there are useful properties that can make computation, when using Gibbs sampling, much simpler. Let v be a node in a DAG and V be the set of all nodes. The joint distribution $p(V)$ can be obtained as follows. If a parent of v is any node with an arrow coming from it pointing at v then it can be shown that (Spiegelhalter *et al.*, 1993)

$$
p(V) = \prod_{v \in V} p(v \mid \text{parents}[v]) \tag{2.40}
$$

When defining a parent, deterministic nodes are ignored, so the parents of y_{ij} are β_{0j}, β_{1j} and σ. Thus for the model in Figure 2.5,

$$
\begin{aligned}
p(V) = p(y_{ij} \mid \beta_{0j}, \beta_{1j}, \sigma^2) p(\beta_{0j} \mid \beta_0, \sigma^2_{\beta_0}) p(\beta_{1j} \mid \beta_1, \sigma^2_{\beta_1}) \\
p(\beta_0) p(\beta_1) p(\sigma^2_{\beta_0}) p(\sigma^2_{\beta_1}) p(\sigma^2)
\end{aligned}
\tag{2.41}
$$

It was shown above that when using Gibbs Sampling, samples are taken from full conditional distributions. It can be shown (Spiegelhalter *et al.*, 1993; Gilks *et al.*, 1993) that the full conditional distribution $p(v|V_{-v})$ for node v has the form

$$
\begin{aligned}
p(v \mid V_{-v}) &\propto p(v, V_{-v}) \\
&\propto \text{terms in } p(V) \text{ containing } v
\end{aligned}
\tag{2.42}
$$

For example, the full conditional distributions for β_{0j} is

$$
p(\beta_{0j} \mid \cdot) = p(y_{ij} \mid \beta_{0j}, \beta_{1j}, \sigma^3) p(\beta_{0j} \mid \beta_0, \sigma^2_{\beta_0}) p(\beta_0) \tag{2.43}
$$

and for $\sigma^2_{\beta_0}$ is

$$
p(\sigma^2_{\beta_0} \mid \cdot) = p(\beta_{0j} \mid \beta_0, \sigma^2_{\beta_0}) p(\sigma^2_{\beta_0}) \tag{2.44}
$$

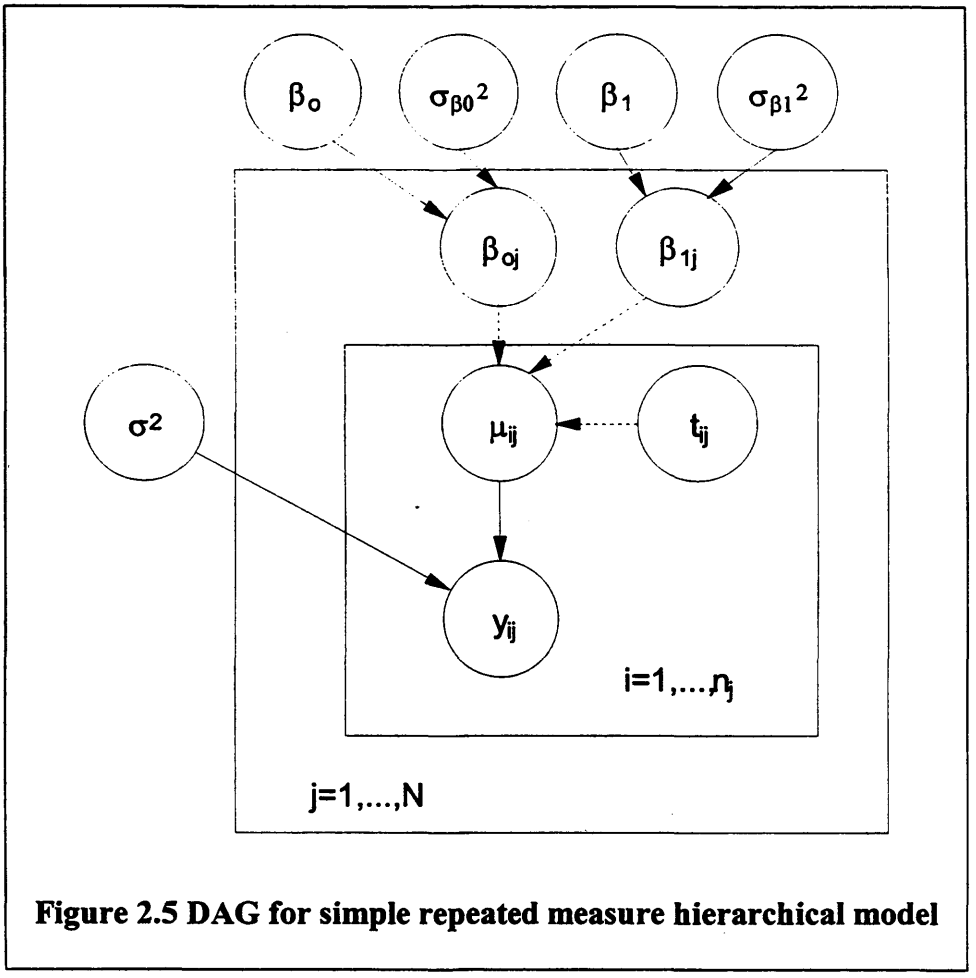

Figure 2.5 DAG for simple repeated measure hierarchical model

At each iteration of the Gibbs Sampler, for each node a value needs to be sampled from its full conditional distribution. The full conditional distribution for any stochastic node is conditional on the *current* values of all other stochastic nodes in the model and, as seen in (2.42), this can be obtained from the DAG. When using WinBUGS, if the full conditional distribution reduces analytically to a known distribution, then the full conditional distribution will be sampled from this. If this is not the case then *adaptive rejection sampling* is used to sample the full conditional distribution (Gilks and Wild, 1992). In adaptive rejection sampling a density $g(y)$ is obtained using (2.42). A function $G(y)$, so that $G(y) \geq g(y)$ for all y, is then chosen. A sample is drawn from the density proportional to $G(y)$ and the point is accepted with probability $g(y)/G(y)$. Accepted points are independent samples from the density proportional to g. A poor choice of G will lead to many rejected points and thus it will take a long time to run the Gibbs Sampler. The adaptive part of adaptive rejection sampling comes from when the sampled point is rejected, G is updated

so that it comes closer to g. Further details of these methods and more complex methods can be found in (Gilks, 1996; Brooks, 1998).

An important advantage of using the Bayesian approach is that one is not restricted to a small number of distributions. For example, rather than assume normality one could assume that between-subject variation could be modelled using a t-distribution with pre-specified degrees of freedom or even estimate the degrees of freedom. With Binomial and Poisson responses it has been shown that classical methods of estimation can be biased (Breslow and Clayton, 1993; Rodriguez and Goldman, 1995) and that Bayesian methods are preferred, especially with small samples (Browne and Draper, 2000).

When using Gibbs Sampling or other MCMC methods a crucial issue is assessment of whether the repeated samples, known as *chains*, have converged to the target distribution. It is important to realise that convergence in this case means convergence to a distribution rather than to a single value. The first issue is to decide on starting values for each parameter. If the starting values are close to the actual values then convergence will be quicker. However, when one is not sure what the actual values are, then one must be careful as there are situations where the joint posterior distribution may not be fully explored, for example, with a bi-modal likelihood. An issue related to the choice of starting value is the length of the 'burn-in'. These are the samples that are discarded while the chain "settles down" to its target distribution. After deciding on the length of burn-in, one must decide on how many samples are needed to obtain reliable estimates. Generally the repeated samples will not be independent, as autocorrelation will exist. With high autocorrelation more samples will be needed. To aid in the choice of burn-in and number of samples a number of convergence diagnostics exist. The most common of these, namely the Geweke statistics and the Gelman and Rubin statistic are outlined below.

Informal convergence assessment for each chain can be performed by inspection of trace plots (a plot of the sampled values vs. iteration number), but there are a number of techniques available to assess convergence more formally. For a comprehensive review of these see Cowles and Carlin (1996). Some of these techniques require more than one chain and assess between and within-chain variation (Gelman and Rubin, 1992; Brooks and

Gelman, 1998). A procedure that uses only one chain is the Geweke Statistic (Geweke, 1992). For each parameter, this looks at an 'early' and a 'late' section of the chain. 'Early' is usually defined as the first 10% of the chain and 'late' as the last 50%. The means (E_{early} and E_{late}) and variances (V_{early} and V_{late}) of both sections of the chain are obtained. Since there may be autocorrelation in the chain the variances are calculated using a spectral density. A test statistic for each parameter is then obtained as follows:

$$Z = \frac{E_{early} - E_{late}}{\sqrt{V_{early} + V_{late}}} \sim N(0,1) \qquad (2.45)$$

Absolute values greater than 2 indicate that there may be problems with convergence. Although it us usually preferable to use more than one chain when using Gibbs Sampling, some models take a long time to run and a large number of iterations are required. In this situation running multiple chains is less attractive due to both time constraints and computing resources. When developing a model using classical estimation and then fitting a Bayesian model, the starting values can be chosen to be equal to those obtained from the classical model (Browne and Draper, 2000). In this situation, one would not expect the parameter estimates to differ dramatically and so detailed convergence assessment is not usually required.

Gelman and Rubin (1992) discuss how there can be problems with just using one chain when assessing convergence. This is because it is possible that the chain will remain in a region that is heavily influenced by the starting values. The problem may be particularly severe if the target distribution is multi-modal. Gelman and Rubin therefore recommend the use of multiple chains with different starting values when assessing convergence.

The Gelman and Rubin convergence diagnostic uses m chains, each with $2n$ iterations and each with different starting values. The starting values should be overdispersed with respect to the target distribution. Convergence is assessed using the components of variance of the multiple sequences and is based on a classical ANOVA calculating the *between-chain variance* and the pooled *within-chain variance*. Convergence is assessed on the last half of the sample, i.e. a sample of size n. However, there is no reason why the proportion of the sample the Gelman and Rubin diagnostic should be half the sample.

The between-chain variance (B) for the parameter of interest θ is calculated as follows,

$$B = \frac{n}{m-1}\sum_{j=1}^{m}\left(\bar{\theta}_{j.} - \bar{\theta}_{..}\right)^2 \tag{2.46}$$

where

$$\bar{\theta}_{j.} = \frac{1}{n}\sum_{t=1}^{n}\theta_{ij} \quad \text{and} \quad \bar{\theta}_{..} = \frac{1}{m}\sum_{j=1}^{m}\bar{\theta}_{j.}$$

$\bar{\theta}_{j.}$ is the mean value of the j^{th} chain and $\bar{\theta}_{..}$ is the overall mean value.

The pooled with chain variance (W) is calculated as follows

$$W = \frac{1}{m(n-1)}\sum_{j=1}^{m}\sum_{t=1}^{n}\left(\theta_{jt} - \bar{\theta}_{j.}\right)^2 \tag{2.47}$$

The total variance of θ, (V) can be estimated by a weighted average of B and W,

$$\hat{V} = \frac{n-1}{n}W + \frac{B}{n} \tag{2.48}$$

It can be seen that if the between-chain variance is small (as one would expect if each chain had converged to the target distribution) the effect of the second term is negligible.

$\hat{V}$ is an unbiased estimate of the true variance if the chain has converged, but will be an overestimate if convergence has not been achieved. For any finite n, the within-chain variance, W, will be an underestimate of the true variance, as the individual chains have not had time to range over all of the target distribution. As $n \to \infty$, both $\hat{V}$ and W approach the true variance from opposite directions.

The Gelman and Rubin diagnostic is obtained by estimating the factor by which the estimated variance of the posterior distribution will be reduced as $n \to \infty$, and is obtained by

$$\sqrt{R} = \frac{\hat{V}}{W} \tag{2.49}$$

It can be seen that if the between-chain variance is very small then $\hat{V}$ will be approximately equal to W and $\sqrt{R} \approx 1$.

Gelman and Rubin (1992) and Brookes and Gelman (1998) give further details of this convergence diagnostic, including procedures to take account of the uncertainty of the estimates of $\bar{\theta}$ and $\hat{V}$ by applying a correction factor (which in practice makes little difference), and advise using the upper 97.5% confidence limit of $\sqrt{R}$. If $\sqrt{R}$ is close to one for all parameters then the m chains have converged to similar distributions. Practical convergence is sometimes defined as $\sqrt{R} < 1.04$ and $\sqrt{R_{97.5\%}} < 1.08$ (Gelman, 1996).

The above method proposes running the chains for $2n$ iterations and then calculating $\sqrt{R}$ for the final n iterations. Brooks and Gelman (1998) suggest using an iterated graphical approach where each of the m chains is divided into batches of length b. $\sqrt{R}$ is then calculated for each of the segments and plotted against the number of iterations. In this way it can be seen how quickly the chains converge and get an idea of how long the 'burn-in' period needs to be.

2.6 The Effect of Ignoring the Hierarchy

In this section I show how standard errors of fixed effect parameters may be wrong if the hierarchical structure of the data is ignored. Consider a situation where there are repeated observations over time for a number of individuals and of interest is the linear relationship between the response (y) and time (t). It will be assumed that the responses for each subject are parallel as in Figure 2.3 and that each subject has the same number of measurements with no missing data. There are N subjects and M time points. Time (t) has been centred about its mean. Let σ_u^2 denote the between-subject variance and σ^2 be the within-subject variance. Thus the model is as follows.

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + u_j + e_{ij} \qquad (2.50)$$

$$u_j \sim N(0, \sigma_u^2), \quad e_{ij} \sim N(0, \sigma^2)$$

If σ_u^2 and σ^2 are assumed known then estimates of the fixed effects (β_0 and β_1) are obtained using equations (2.20) ($\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y$) with variance matrix obtained using (2.21) ($Var(\hat{\beta}) = (X^T V^{-1} X)^{-1}$);

V is block diagonal with blocks V_j of the form

$$V_j = \begin{pmatrix} \sigma_u^2 + \sigma^2 & \sigma^2 & \cdots & \sigma^2 \\ \sigma^2 & \sigma_u^2 + \sigma^2 & & \vdots \\ \vdots & & \ddots & \sigma^2 \\ \sigma^2 & \cdots & \sigma^2 & \sigma_u^2 + \sigma^2 \end{pmatrix} \tag{2.51}$$

V is block diagonal because observations made on different subjects can be considered to be independent. V^{-1} is also block diagonal with each block being the inverse of V_j. Since V_j is symmetric, V^{-1} will also be symmetric and can be solved as follows. Letting

$$a = \sigma_u^2 + \sigma^2 \tag{2.52}$$

and

$$b = \sigma^2 \tag{2.53}$$

The inverse of V can be calculated by solving

$$\begin{pmatrix} a & b & \cdots & b \\ b & a & & \vdots \\ \vdots & & \ddots & b \\ b & \cdots & b & a \end{pmatrix} \begin{pmatrix} p & q & \cdots & q \\ q & p & & \vdots \\ \vdots & & \ddots & q \\ q & \cdots & q & p \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix} \tag{2.54}$$

for p and q.

From (2.54) is is possible to obtain the following equations for p and q.

$$\begin{aligned} ap + (M-1)bq &= 1 \\ bp + (a + (M-2)b)q &= 0 \end{aligned} \tag{2.55}$$

Solving for p and q,

$$\begin{pmatrix} p \\ q \end{pmatrix} = \begin{pmatrix} a & (M-1)b \\ b & a + (M-2)b \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \tag{2.56}$$

The determinent of the square matrix is

$$\begin{aligned} & a(a+(M-2)b)-b(M-1)b \\ = \; & a^2+abM-2ab-b^2M+b^2 \\ = \; & (a-b)^2+M(ab-b^2) \end{aligned} \tag{2.57}$$

Solving for p

$$p=\frac{a+(M-2)b}{(a-b)^2+M(ab-b^2)} \tag{2.58}$$

and for q

$$q=\frac{-b}{(a-b)^2+M(ab-b^2)} \tag{2.59}$$

substituting σ_u^2 and σ^2 back into (2.58) and (2.59) yields V^{-1} as a symmetric matrix with diagonal elements r, where

$$r=\frac{(M-1)\sigma_u^2+\sigma^2}{M\sigma_u^2\sigma^2+(\sigma^2)^2}, \tag{2.60}$$

and off diagonal elements s, where

$$s=\frac{-\sigma_u^2}{M\sigma_u^2\sigma^2+(\sigma^2)^2} \tag{2.61}$$

To obtain Var(β) equation (2.21) is used. The design matrix X is as follows

$$X=\begin{pmatrix} 1 & t_{11} \\ \vdots & t_{21} \\ \vdots & \vdots \\ \vdots & t_{M1} \\ \vdots & t_{i2} \\ \vdots & \vdots \\ 1 & t_{NM} \end{pmatrix} \tag{2.62}$$

and V^{-1} is block diagonal with blocks A, such that

$$\begin{pmatrix} A & 0 & \cdots & 0 \\ 0 & A & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \ldots & 0 & A \end{pmatrix} \tag{2.63}$$

with

$$A = \begin{pmatrix} r & s & \cdots & s \\ s & r & & \vdots \\ \vdots & & \ddots & s \\ s & \cdots & s & r \end{pmatrix} \tag{2.64}$$

The first two terms of equation (2.21), $X^T V^{-1}$, gives

$$X^T V^{-1} = \begin{pmatrix} r+(M-1)s & r+(M-1)s & \cdots & \cdots & \cdots & r+(M-1)s \\ t_{11}r - x_{11}s^* & t_{21}r - t_{21}s & \cdots & t_{M1}r - t_{M1}s & \cdots & t_{MN}r - t_{MN}s \end{pmatrix} \tag{2.65}$$

And multiplying by X, gives

$$X^T V^{-1} X = \begin{pmatrix} MN(r+(M-1)s) & (r+(M-2)s)\sum t \\ (r+(M-2)s)\sum t & (r-s)\sum t^2 \end{pmatrix} \tag{2.66}$$

Note that the off diagonals are zero as t is centred about its mean so $\sum t$ will be zero. To obtain Var(β) the inverse of (2.66) needs to be obtained. The determinent is:

$$\left| X^T V^{-1} X \right| = NM(r+(M-2)s)(r-s)\sum t^2 \tag{2.67}$$

The variance of the intercept β_0 is therefore

$$Var(\beta_0) = \frac{(r-s)\sum t^2}{NM(r+(M-2)s)(r-s)\sum t^2} \tag{2.68}$$

and the variance of the gradient β_1 is

$$Var(\beta_1) = \frac{MN(r+(M-1)s}{MN(r+(M-1)s(r-s)\sum t^2} \tag{2.69}$$

Substituting the terms for r and s *from* (2.60) and (2.61) into (2.68) and (2.69) gives the variance of the intercept,

$$Var(\beta_0) = \frac{M\sigma_u^2 + \sigma^2}{NM}, \tag{2.70}$$

and the variance of the gradient

$$Var(\beta_1) = \frac{\sigma^2}{\sum x^2} \tag{2.71}$$

When ignoring the between-subject variance using ordinary least squares (OLS) the residual variance (σ^2_{OLS}) will be equal to $\sigma_u^2 + \sigma^2$ so that the variance of the intercept will be

$$Var(\beta_0) = \frac{\sigma_u^2 + \sigma^2}{NM}, \tag{2.72}$$

And the variance of the gradient will be

$$Var(\beta_1) = \frac{\sigma_u^2 + \sigma^2}{\sum x^2} \tag{2.73}$$

It can be seen that the variance of the intercept β_0 for the hierarchical model will always be greater than or equal to that obtained from the simple linear model. The degree to which the variance estimates differs will depend on the number of repeated observations *M*, and the size of the between-subject variance, σ_u^2. Of more interest is the variance of the gradient, β_1 for the hierarchical model, and this will always be less than or equal to that obtained from the simple linear model. The amount the variance differs will depend on the size of the between-subject variance σ_u^2. Although this is a special case it is worth noting that the results are applicable generally to the comparison of hierarchical models and ordinary least squares estimates. For example, in a clustered randomised trial where randomisation is at the cluster level, but analysis is at the individual level the treatment effect will have too small a standard error if ordinary least squares is used as treatment is a cluster level covariate.

2.7 Example (Plasma Citrate Concentration)

In order to illustrate the methods described and the effect of ignoring the hierarchy I will use an example from Hand and Crowder (1996). The data consists of 5 repeated measurements of plasma citrate concentration in micromoles per litre on twenty subjects. The data can be seen plotted in Figure 2.6. It appears from the figure that plasma citrate decreases over time. There is also some evidence that the between-subject variation decreases over time, but this will be ignored in this illustrative example. If y_{ij} denotes the plasma citrate concentration on the i^{th} occasion on the j^{th} subject and t_{ij}, centred time then a hierarchical model can be fitted to the data as follows

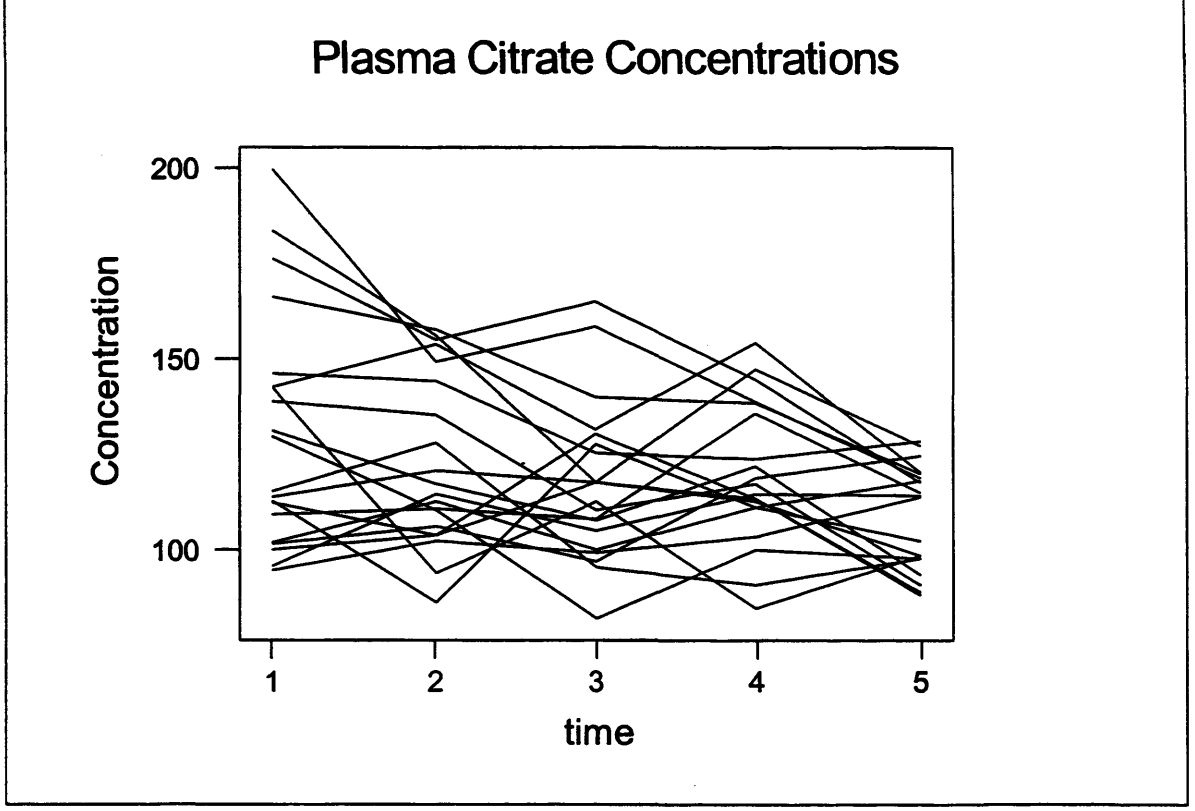

Figure 2.6 Individual profile plots for plasma citrate concentration measured at five time points.

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + u_j + e_{ij}$$
$$\text{where } u_j \sim N(0, \sigma_u^2) \quad \text{and} \quad e_{ij} \sim N(0, \sigma^2) \tag{2.74}$$

If σ_u^2 is constrained to equal 0, then model (2.74) is a simple linear regression model (i.e. it ignores the clustering).

The parameter estimates for the hierarchical linear model and a simple linear regression model can be seen in Table 2.1. It can be seen that the estimates of β_0 and β_1 are the same for both methods of estimation and that $\sigma_{OLS}^2 = \sigma_u^2 + \sigma^2$ as would be expected. The standard error of β_0 for the simple linear model is thus

$$\sqrt{Var(\beta_0)} = \sqrt{\frac{473.75}{20 \times 5}} = 2.18$$

and for the hierarchical linear model,

Paremeter	Hierarchical linear model	Simple linear regression model
β_0	119.9 (3.85)	119.9 (2.18)
β_1	-4.74 (1.05)	-4.74 (1.54)
σ_u^2	252.08	-
σ^2	221.68	-
σ_{OLS}^2	-	473.75

Table 2.1 Parameter estimates for simple linear and hierarchical linear model for plasma citrate concentration data set.

$$\sqrt{Var(\beta_0)} = \sqrt{\frac{5 \times 252.08 + 221.681}{20 \times 5}} = 3.85 .$$

The variance of β_1 for the simple linear model is thus

$$\sqrt{Var(\beta_1)} = \sqrt{\frac{473.75}{200}} = 1.54,$$

and for the hierarchical linear model

$$\sqrt{Var(\beta_1)} = \sqrt{\frac{221.68}{200}} = 1.05$$

Thus, it can be seen that if the hierarchical structure is ignored then in this simple example, the standard error of the intercept is too small, but perhaps more critically the standard error of the gradient is too large. In general when ignoring the hierarchical structure standard errors of subject level covariates will be too small, and standard errors of covariates that change over time will be too large.

The dataset can also be used to illustrate the Bayesian approach to estimation. The model can be written in the following form

$$y_{ij} \sim N(\mu_{ij}, \sigma^2)$$

$$\mu_{ij} = \beta_1 t_{ij} + u_j \qquad (2.75)$$

$$u_j \sim N(\beta_0, \sigma_u^2)$$

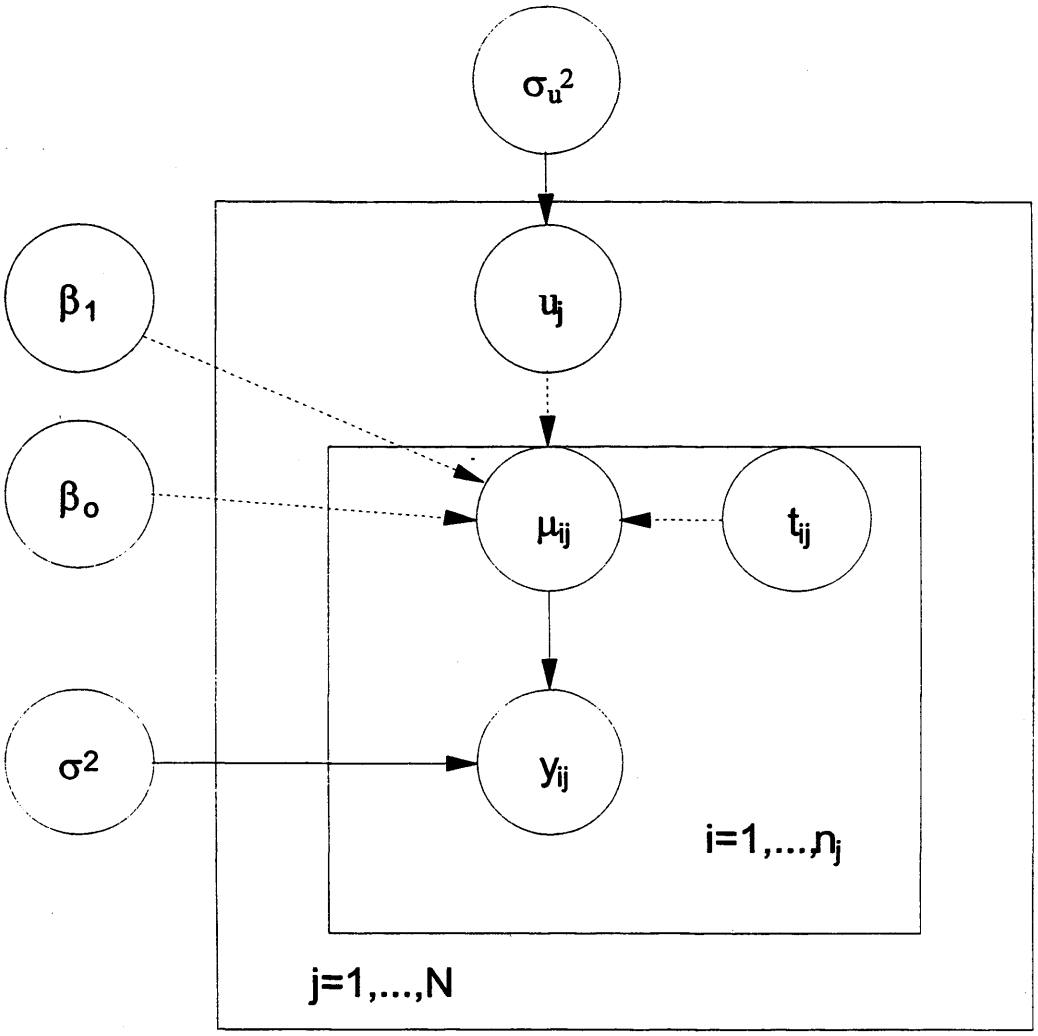

Figure 2.7 DAG for Plasma Citrate Concentration model

As this is a Bayesian model prior distributions need to be specifed for all the unknown parameters. Standard vague prior distributions for the fixed effects are Normal with large variances and Gamma distribution with small parameters for the inverse of the variances, (Spiegelhalter *et al.*, 1996), i.e.

$$\beta_0, \beta_1 \sim N(0,10000)$$

$$\frac{1}{\sigma_u^2}, \frac{1}{\sigma^2} \sim Gamma(0.001,0.001) \qquad (2.76)$$

Further details regarding prior distributions will be discussed in subsequent chapters.

A DAG for the model can be seen in Figure 2.7. This shows how only the intercept is allowed to vary from subject to subject. Although the plot in Figure 2.6 indicates that it may be sensible to allow the slope to vary from subject to subject, I will not include a random effect for this as the example is for illustrative purposes only.

The model was fitted in WinBUGS using 5 chains with each chain having different starting values. These were various combinations of small or large fixed and small or large variances estimates. Each chain was run for 10000 iterations, which took about 10 seconds

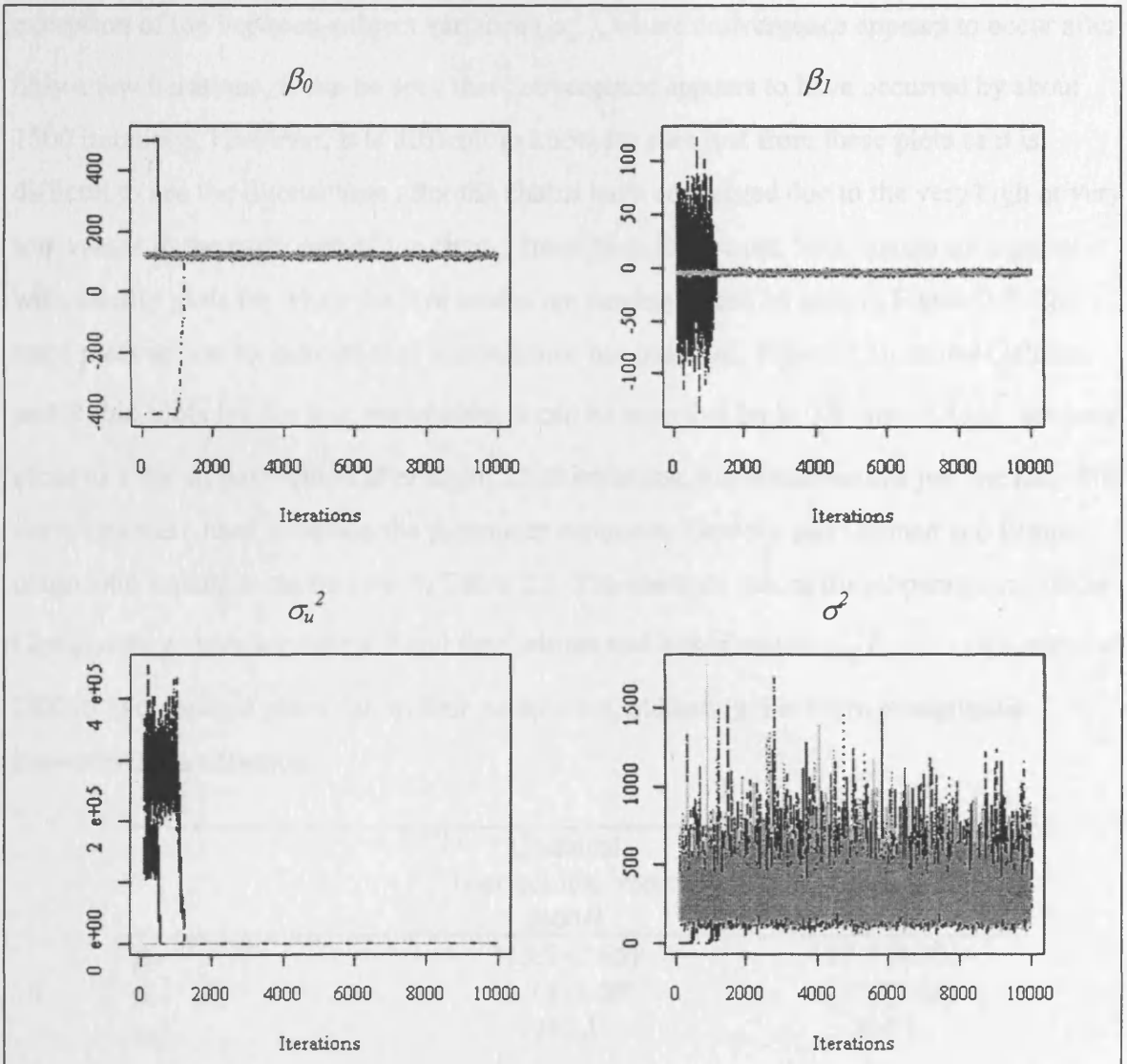

Figure 2.8 Trace plots for the four parameters in the Plasma Citrate Concentration dataset.

Parameter	Geweke Statistic	Gelman and Rubin $\sqrt{R_{97.5\%}}$
β_0	-0.17	1.00
β_1	0.01	1.00
σ_u^2	0.67	1.00
σ^2	-1.13	1.00

Table 2.2 Geweke and Gelman and Rubin diagnostic statistic for the Plasma Citrate Concentration dataset.

on a Pentium II 400Mhz PC. Trace plots for the 10000 iterations for each of the five chains for each of the four parameters in the model can be seen in Figure 2.8. The five chains are overlayed, so it is not always possible to distinguish between the chains. With the exception of the between-subject variance (σ_u^2), where convergence appears to occur after only a few iterations, it can be seen that convergence appears to have occurred by about 1500 iterations. However, it is difficult to know for sure just from these plots as it is difficult to see the fluctuations after the chains have converged due to the very high or very low values in the early part of the chain. Trace plots for the last 5000 iterations together with density plots for when the five chains are combined can be seen in Figure 2.9. The trace plots appear to indicate that convergence has occurred. Figure 2.10 shows Gelman and Rubin plots for the four parameters. It can be seen that both $\sqrt{R}$ and $\sqrt{R_{97.5\%}}$ are very close to 1 for all parameters after about 2500 iterations. For these reasons just the last 5000 iterations were used to obtain the parameter estimates. Geweke and Gelman and Rubin diagnostic statistics can be seen in Table 2.2. The absolute values for all parameters of the Geweke diagnostic are below 2 and the Gelman and Rubin statistic $\sqrt{R_{97.5\%}}$ is estimated at 1.00 to two decimal place for all four parameters, indicating that there is negligible between-chain variation.

	Classical Hierarchical linear model	Bayesian Hierarchical linear model
β_0	119.9 (3.85)	119.9 (4.09)
β_1	-4.74 (1.05)	-4.74 (1.08)
σ_u^2	252.1	295.1
σ^2	221.7	231.9

Table 2.3 Parameter estimates for classical and Bayesian hierarchical linear models for plasma citrate concentration data set.

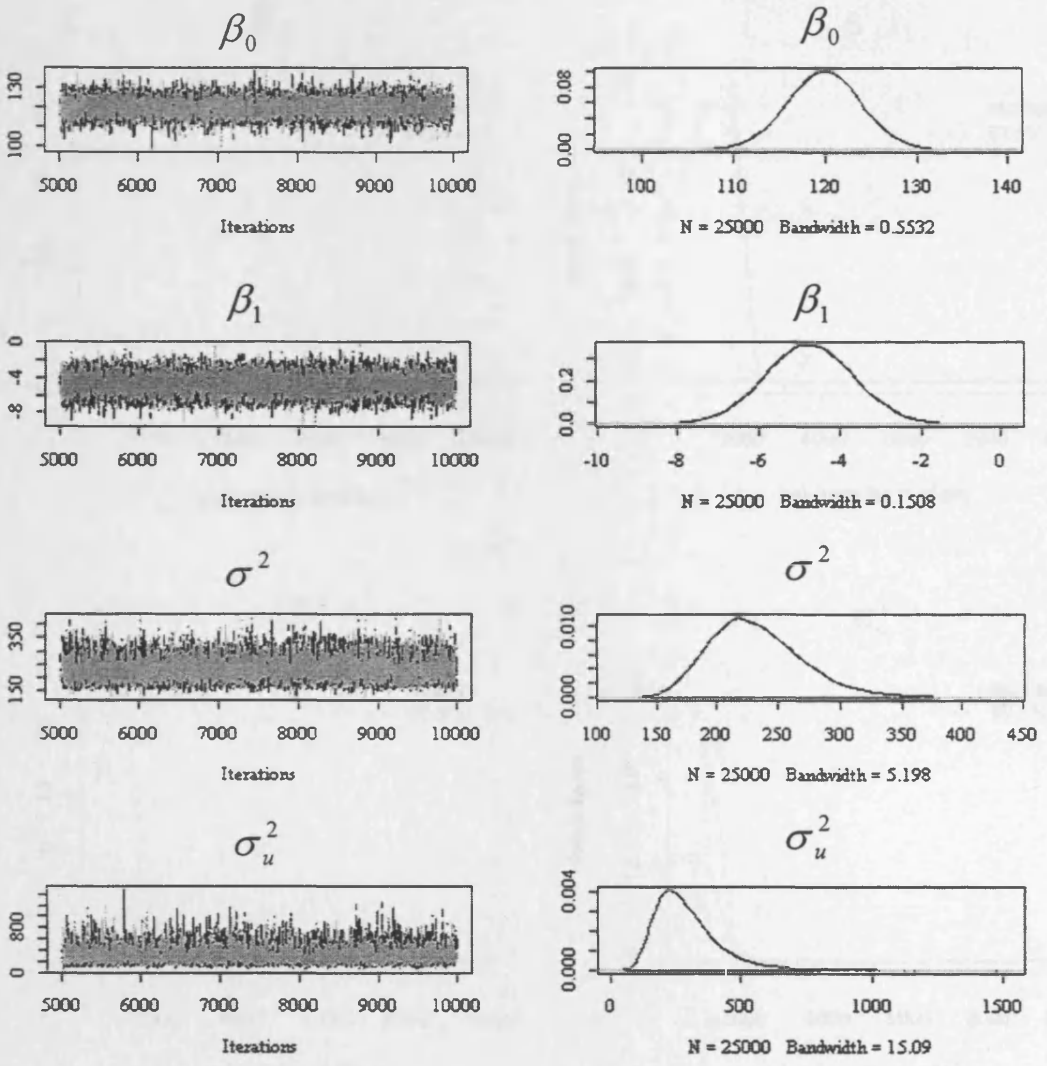

Figure 2.9 Trace and density plots for the parameters in the Plasma Citrate Concentration dataset.

Comparison of the parameter estimates between the classical and Bayesian estimates can be seen in Table 2.3. It can be seen that the fixed effect estimate are identical, but there are slight differences in the both the between and within-subject variances. The standard errors for the fixed effects are slightly larger for the Bayesian model. This is to be expected as the classical model does not take into account the uncertainty associated with the estimates of the between-and within-subject variances when estimating the standard errors. It can be seen from the density plots in Figure 2.9, that there is a considerable amount of uncertainty associated with the variance estimates.

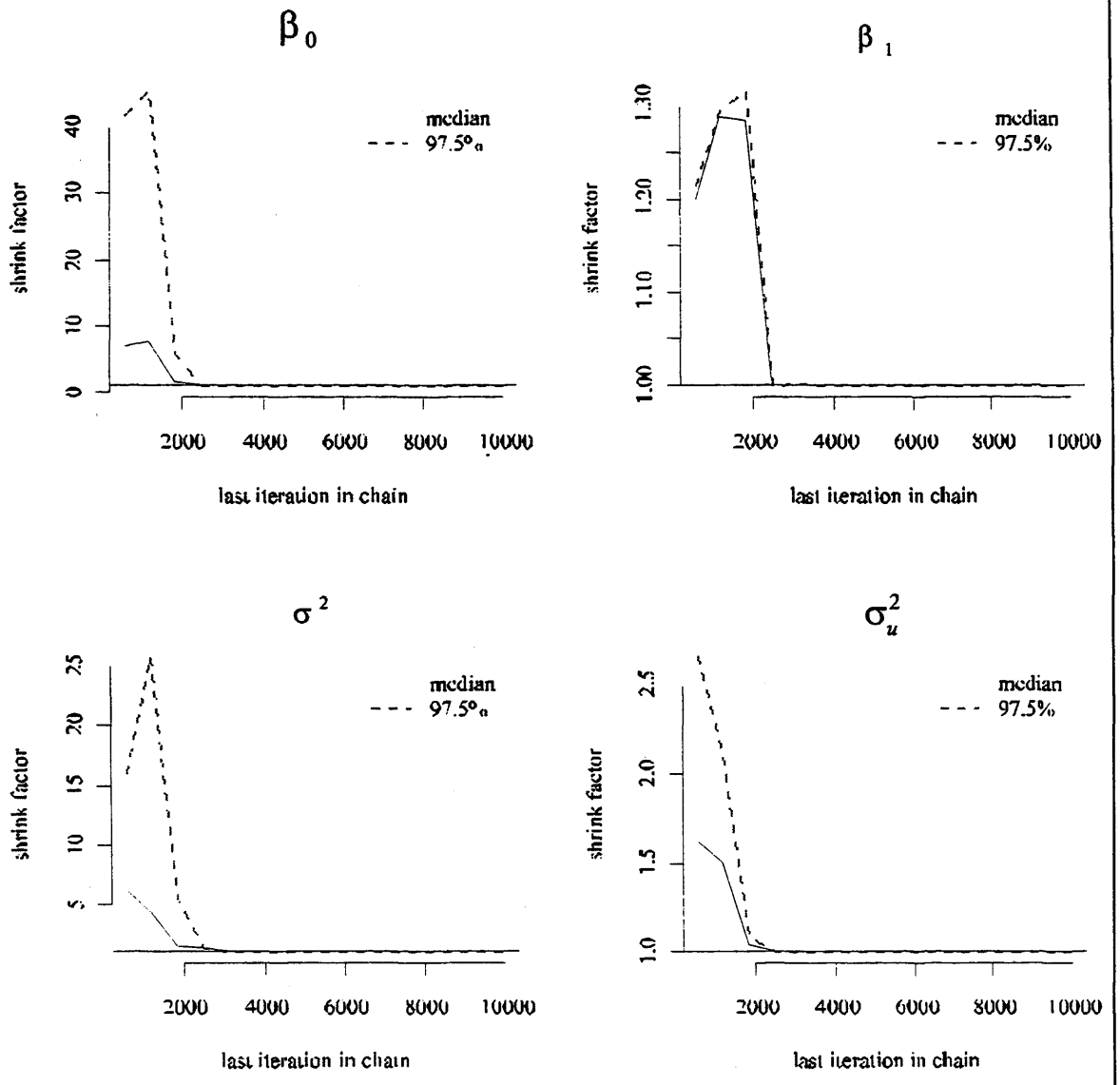

Figure 2.10 Gelman and Rubin plots for Plasma Citrate Concentration data.

2.8 Discussion

In this chapter I have introduced the concept of hierarchical data, discussed the reason why such data needs to be considered differently to non-hierarchical data and shown the effect on the standard errors of the estimates when ignoring the hierarchical structure. I have introduced some of the important concepts through the use of a hypothetical repeated measures example, but it is important to realise that the methodology is applicable to all types of hierarchical data. An important issue is the idea of natural variation between units

(between-subjects in the case of repeated measures data), which can be modelled using random effects. Through the use of random effects, it possible to investigate the source of variation of the response variable at the each of the levels in the model. Thus, as seen in the repeated measures example there is both between and within-subject variation. In chapters 3 and 4 I develop hierarchical models for two different and more complex repeated measures problems.

Although I have concentrated on classical methodology for Gaussian outcomes, there are techniques available for other outcomes, e.g. Binary and Poisson outcomes (Gilks *et al.*, 1993; Breslow and Clayton, 1993). However, in small samples these estimation procedures can be biased (Rodriguez and Goldman, 1995) and it may be better to use Bayesian methods of estimation (Browne and Draper, 2000). In fact the use of Bayesian models offers much more flexibility in terms of model fitting including distributional type and further complexity, as I will demonstrate in subsequent chapters.

3 THE ANALYSIS OF AMBULATORY BLOOD PRESSURE (ABPM) DATA

3.1 Introduction

In this chapter I demonstrate how hierarchical models can be used in the analysis of repeated measures data. I use a two-level model hierarchical model to analyse repeated ambulatory blood pressure monitor measurements with measurements taken every half hour over a 24-hour period. The mean profiles over the 24-hour period exhibit complex curvature and are modelled using restricted cubic splines. I first give a brief background to ambulatory blood pressure monitoring in section 3.2 followed by a description of the data set used in section 3.3. Section 3.4 is concerned with approaches to modelling individual profiles, including conventional summary measures (3.4.1), regression splines (3.4.2), restricted cubic splines (3.4.3), periodic splines (3.4.4), and the choice of the number and location of knots in spline models (3.4.5). In section 3.5 I use restricted cubic splines in a two level hierarchical model. I discuss an initial model in section 3.5 and then perform a number of sensitivity analyses to assess the robustness of the model to various factors including the number of knots (3.5.4), the modelling of the difference in profiles between groups (3.5.5), the choice of between-subject random effects (3.5.6), and the modelling of the within-subject variation (3.5.7) In section 3.5.8 I assess the assumptions of the model. A Bayesian approach is adopted in section 3.6 including a re-analysis of the initial model (3.6.2) and accounting for the within-subject variance heterogeneity (3.6.3). In section 3.6.4 I assess the convergence of the Bayesian model. Finally, in section 3.7, I discuss the techniques I have used and recommend further developments.

3.2 Background to Ambulatory Blood Pressure Monitoring

Blood pressure is one of the most common and important measurements in clinical medicine. For over 100 years (since 1896) mercury sphygmomanometers have been used as the standard method of measuring blood pressure. However, in recent years there has been a growth in the use of Ambulatory (or Automated) Blood Pressure Monitors (ABPM) in

both clinical research and practice (Conway and Coats, 1991; Prasad and Isles, 1996). An APBM is a small device (about the size of a personal stereo) that can be worn by a patient with a belt or a strap over the shoulder. The device measures the patient's blood pressure at pre-set times, usually every 15 or 30 minutes. When using an ABPM it is possible to investigate the blood pressure profile over a 24-hour period or longer. This may be of interest as blood pressure has a circadian rhythm with there being a dipping of blood pressure at night (Seligman, 1971). It has been shown that attenuation of dipping is associated with a number of diseases including pre-eclampsia and chronic renal disease (Pickering, 1990).

The use of ABPM has been advocated for a number of reasons. It may reduce or eliminate some of the problems with conventional sphygmomanometry, including observer error (both terminal digit preference (Patterson, 1984) and systematic under or over reading (Bailey and Bauer, 1993)) and faulty equipment (Burke *et al.*, 1982). Another common problem is that for some individuals blood pressure is increased in the presence of a health professional (Mancia *et al.*, 1983). This is known as 'white coat hypertension' and there is evidence that the use of automated techniques for measuring blood pressure can reduce this problem (Punzi, 1998). An advantage of using ABPM is that it can reduce sampling variation since it involves an increased number of readings compared to sphygmomanometry, where often clinical decisions are based on only one reading (Coats *et al.*, 1992). Another important reason why interest is increasing in the use of automated devices for blood pressure measurement is due to the likelihood that sphygmomanometers will to be banned in the not too distant future due to the toxicity of mercury. In fact they have already been banned in parts of Scandinavia (O'Brien, 1996).

The use of ABPM provides a number of challenges in terms of data analysis, with the large number of repeated observations per subject and changes in blood pressure during the day. However, most approaches to the analysis of ABPM data have tended to simplify the data in that they reduce each subject's ABPM profile to a few summary measures. This results in a loss of information, and it may be preferable to use methods that utilize the whole profile. However, it is important that the results of any analysis are presented in a clinically meaningful manner.

3.3 Description of Data

The data set used to illustrate the methods comes from an observational study comparing the use of ABPM with conventional mercury sphygmomanometry on obstetric outcome for 348 pregnant women with a confirmed clinic BP of at least 140/90 mm Hg and ≥20 weeks gestation (Penny *et al.*, 1998). A number of outcomes were investigated including development of severe hypertension, development of proteinuria, admission to a neonatal intensive care unit, preterm delivery and low birth weight for gestational age. Blood pressure was measured every 30 minutes using a SpaceLabs 90207 ambulatory blood pressure monitor. In the previously reported analysis a day-time mean (10am-8pm) was used as a summary measure. In this chapter I use a reduced data set of 206 women who had at least 10 day-time measurements (10am-8pm) and 5 night-time (12am-6am) measurements. There were 8593 blood pressure recordings in total from the 206 women leading to a mean number of 41.7 blood pressure measurements per women. Analysis was restricted to a 24 hour period starting at 12:00 and finishing at 12:00 the following day. Reasons for exclusion of the 142 women include failure of the monitor, but more usually women removed the monitor, as they were aware that any results would not be used for their clinical management. This was particularly the case at nighttime where the monitor could interrupt their sleep. The question of interest was whether ABPM blood pressure at referral was related to intra-uterine growth retardation (IUGR). This was assessed by investigating whether women who subsequently gave birth to an infant $\leq 10^{th}$ weight centile for gestational age had a different diastolic blood pressure (DBP) profile at referral to those who subsequently gave birth to an infant $>10^{th}$ weight centile for gestational age. This may be important, as previous work has shown little association between maternal blood pressure and birthweight. However, with the greater accuracy of ABPM measurement there is more potential for observing an association (Churchill *et al.*, 1997). Of the 206 women 20(9.7%) subsequently gave birth to an infant $\leq 10^{th}$ centile for gestational age.

A plot of the blood pressure profiles for the first 12 women can be seen in Figure 3.1. Nocturnal dipping can be observed in some of these subjects where blood pressure appears lower at night, but in others it is less clear. The within-subject variation appears to be high indicating problems with making clinical decisions when only taking one blood pressure

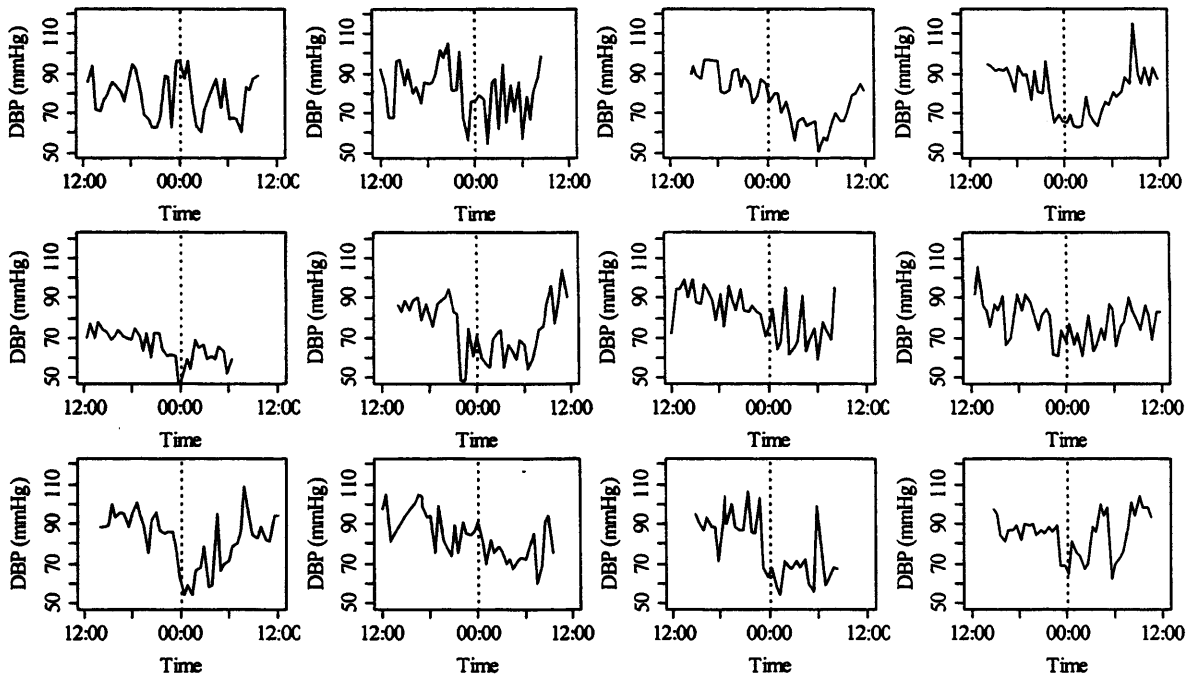

Figure 3.1 Diastolic blood pressure profiles for the first 12 subjects.

measurement. In addition the level of the within-subject variability appears to vary between subjects, i.e. some women have greater variation in their blood pressure measurements.

3.4 Approaches to Modelling Individual Profiles

3.4.1 Summary Measures

With ABPM data there are serial measurements collected over a period of time for each individual. This leads to ABPM exhibiting a natural two-level hierarchical structure with individual blood pressure measurements nested within subjects. Naturally blood pressure measurements made on the same subject will be correlated. Most of the current methods of analysing ABPM data reduce each individual's blood pressure profile to a few summary measures. I will now discuss some of these methods.

The simplest method of reducing an individual's blood pressure profile to a summary measure is to take a mean. This can be a 24 hour mean or more usually a day-time mean and/or a night-time mean (Gatzka and Schmieder, 1995). To overcome the problem of people having different sleep patterns, the use of sleep diaries has been advocated (Peixoto

Filho *et al.*, 1995). However, sleep diaries can be unreliable and so a more simple approach is to define 'narrow bands' for the definition of day and night during which it is expected that most people will be awake/asleep (Peixoto Filho *et al.*, 1995). For example, in previous work on this data set I used 10:00am-8:00pm as a definition of day-time and 12:00am-8:00am as a definition of night-time (Penny *et al.*, 1998).

Although a mean is obviously the simplest method of reducing each individual's blood pressure profile to a summary measure, as discussed in section 3.2, blood pressure is not constant over time and is said to have a circadian rhythm, with blood pressure measurements tending to be lower at night. Much work has looked at fitting curves to individual blood pressure profiles including cosinor analysis (Bingham *et al.*, 1982; Ayala *et al.*, 1997) or its extension fourier analysis (Somes *et al.* 1994). There has also been some recent work on the use of 3rd degree (cubic) polynomials (Corrao *et al.*, 1996).

In cosinor analysis, fourier analysis and the 3rd degree polynomial method, an individuals blood pressure is considered to be a function of time (t) and a model of the following form is fitted to the data:

$$y_i = f(t) + e_i \tag{3.1}$$

where y_i is the i^{th} blood pressure measurement, $f(t)$ is some function of time and e_i is a normal random variable with zero mean and constant variance.

In the 3rd degree (cubic) polynomial approach $f(t)$ is

$$f(t) = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + \beta_3 t_i^3 \tag{3.2}$$

After fitting this model to an individuals blood pressure profile it is possible to obtain a number of summary measures including maximum blood pressure, minimum blood pressure, the blood pressure at the 'flex' point and the times at which the maximum, minimum and the flex point occur.

In cosinor analysis it is assumed that there is one sinusoidal cycle per day (one harmonic period) and $f(t)$ is defined by

$$f(t) = \alpha_0 + A\cos\left(\frac{2\pi t_i}{24} + \phi\right) \tag{3.3}$$

where α_0 is the Mesor (an average blood pressure over the 24 hour period), A is the amplitude (the difference between the Mesor and the highest and lowest parts of the fitted curve) and ϕ is the phase (the time to the peak). Model (3.3) can be written in a linear form as follows (Hallberg, 1969);

$$f(t) = \alpha_0 + \beta_1 \cos\left(\frac{2\pi t_i}{24}\right) + \gamma_1 \sin\left(\frac{2\pi t_i}{24}\right) \tag{3.4}$$

where the amplitude, A, and the phase, ϕ can be calculated as follows;

$$A = \sqrt{\beta_1^2 + \gamma_1^2} \tag{3.5}$$

$$\phi = \arctan\left(\frac{-\gamma_1}{\beta_1}\right) \tag{3.6}$$

Often the amplitude is defined as $2A$ as this represents the difference between the peak and trough of the blood pressure measurements.

Fourier analysis is an extension of cosinor analysis where there are more harmonics, leading to there being j multiple cycles within a day with $f(t)$ taking the form.

$$f(t) = \alpha_0 + \sum_j \left\{\beta_j \cos\left(\frac{2\pi j t_i}{24}\right) + \gamma_j \sin\left(\frac{2\pi j t_i}{24}\right)\right\} \tag{3.7}$$

If j=1 then this reduces to the cosinor model in (3.4). The number of harmonics can be determined by model selection methods (Somes *et al.*, 1994) or can be pre-defined to be the same for all subjects (Germano *et al.*, 1990). Using fourier analysis is possible to obtain fitted maximum and minimum blood pressure measurements as well as Mesor and phase parameters.

Although cosinor techniques are still used extensively, it is known that there are problems with using such methods to analyse ABPM data. For example, cosinor analysis has been criticised for a variety of reasons (Streitberg *et al.*, 1989) including;

(i) It implies an exactly symmetrical behaviour of high and low blood pressure periods, both of the same length, shape and amplitude.

(ii) It fixes the difference between the acrophase (time of peak pressure) and bathyphase (time of lowest pressure) at 12 hours.

(iii) Observed patterns can be multiphasic (several maxima/minima).

(iv) A cosine waves produces either extremely high or extremely low values most of the time, while regions with average values are traversed comparatively fast.

Figure 3.2 shows the raw data for four women together with the fitted values for a cubic polynomial model, a cosinor model and a fourier model with 3 harmonics at 24 hours, 12 hours and 8 hours. It can be seen that the three methods give very different fits to the data, which is obviously problematic. The cubic polynomial and cosinor models do not appear to pick up all peaks and troughs. A potential problem with the cosinor and fourier models is that curvature is being induced in the data leading to parts of the profile where the observed blood pressure appears relatively constant having curvature for the fitted values.

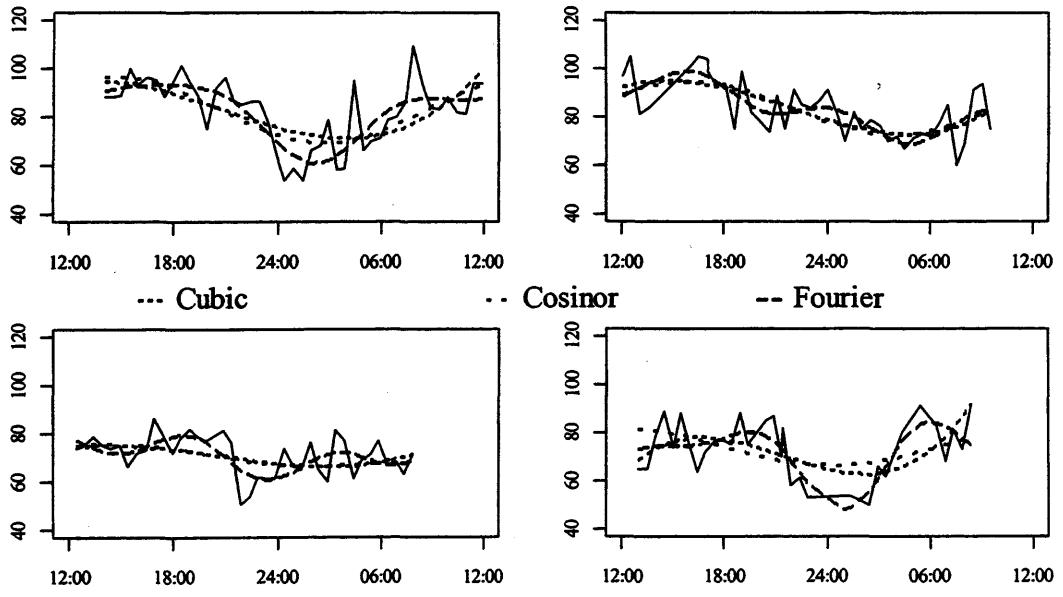

Figure 3.2 Example fit of Cubic Polynomial, Cosinor and Fourier curves for four women in the ABPM study.

Two other procedures used to describe the circadian rhythm of ABPM profiles are cumulative sums (cusums) (Stanton *et al.*, 1992) and square wave fitting (Idema *et al.*, 1992). The cusum method plots the sum of successive deviations from a reference line against time. Various cusum derived statistics are available including the cusum derived crest and trough. The square wave method assumes that there is one period of high blood pressure and one period of low blood pressure and that the change between the two is effectively instantaneous. The duration of the two periods is constrained by the fact that the sum of the two times is equal to 24 hours. These methods have been advocated as they do not rely on pre-defined definition of day and night or require the subjects to keep sleep diaries.

3.4.2 Regression Splines

When the aim of a model is to describe the functional relationship between a response Y and an explanatory variable X, and this relationship is non-linear, a usual approach is to use polynomial regression. For example, in the previous section I demonstrated how third degree polynomials could be used to model ABPM profiles. However, polynomial regression has a number of drawbacks, including the fact that individual observations can exert too much influence on certain parts of the curve (Wold, 1974). This is particularly so for observations towards the minimum or maximum of the explanatory variable X. If the curvature of the data is complex with a number of turning/flex points then the use of polynomials may not be suitable. One possible solution is to use *piecewise polynomials*, which are also known as *regression splines*.

Regression splines can be defined as piecewise polynomials of degree n whose function values and the first $n - 1$ derivatives agree at the points where the polynomials join (De Boor, 1978). The points at which the polynomials join are known as *knots*. Polynomials can be considered as a special case of regression splines, but having no knots, i.e. no joining points. The most common form of regression spline is the *cubic spline*, where the functional relationship is visually smooth with cubic polynomials between each knot, but

with continuous first and second derivatives (Wold, 1974). To repeat a quote of Rice (1969) from given by Wold (1974).

> "spline functions are the most successful approximating functions for practical applications so far discovered. The reader may be unaware of the fact that ordinary polynomials are inadequate in many situations. This is particularly the case when one approximates functions that arise from the physical world rather than from the mathematical world. Functions that express physical relationships are frequently of a disjointed or disassociated nature. That is to say that their behaviour in one region may be totally unrelated to their behaviour in another region. Polynomials, along with most other mathematical functions, have just the opposite property. Namely their behaviour in a small region determines their behaviour everywhere. Splines do not suffer this handicap since they are defined piecewise, yet, for $n \geq 3$, they represent nice, smooth curves in the physical world."

This is an old quotation, but it is a good description of some of the problems with polynomials. However, given that the quote is over 25 years old, one must realise that it does not acknowlege recent developments in curve fitting such as fractional polynomials (Royston and Altman, 1997).

Although there are methods that include both the number and position of knots as unknowns (Gallant and Fuller, 1973; Freedman and Silverman, 1989), when using a fixed number of knots it is possible to obtain parameter estimates using standard linear regression models using standard statistical software packages. However, it must be realized that the fitted model will depend on both the choice of the number of knots and their locations. Before describing a spline model algebraically it is important to understand the "+" function representation (Smith, 1979). Let

$$\begin{aligned} u_+ &= u \quad \text{if} \quad u > 0 \\ u_+ &= 0 \quad \text{if} \quad u \leq 0 \end{aligned} \tag{3.8}$$

In general, with k knots, $t_1<\ldots<t_k$, and the $k+1$ polynomial pieces each of degree n and explanatory variable x, it is possible to write a spline model $S(x)$ with no continuity restrictions as follows

$$S(x) = \sum_{j=0}^{n} \beta_{oj} x^j + \sum_{i=1}^{k} \sum_{j=0}^{n} \beta_{ij} (x - t_i)_+^j + e_i \qquad (3.9)$$

where e_i is a random normal variable with zero mean. If $S^{(j)}$ denotes the j^{th} derivative of $S(x)$ then the presence of a $\beta_{ij}(x-t_i)_+^j$ term allows a discontinuity at t_i for $S^{(j)}$, and its absence forces the continuity of $S^{(j)}$ at t_i. This is perhaps best described by an example. Figure 3.3 shows a blood pressure profile for one subject with four piecewise polynomial models fitted to the data. Each model has two knots, at 20 hours and 28 hours (shown as vertical dashed lines). I am using these graphs to demonstrate the continuity restrictions and not to illustrate the choice of the number and locations of knots. The choice of the number and location of the knots is discussed in section 3.4.5. In Figure 3.3(a) there are no continuity restrictions so that the cubic polynomials do not even join at the knots, clearly giving an unsatisfactory fit to the data. In Figure 3.3(b) the β_{i0}'s from (3.9) are dropped from the model so that the cubic polynomials are forced to join at the knots. Despite the values of the function agreeing at the knots, it can be seen that the resulting fit is not visually smooth, especially before and after the first knot. In Figure 3.3 (c) both the β_{i0}'s and β_{i1}'s from (3.9) are now dropped from the model. This forces both the function values and the first derivative to agree at the knots. It can be seen that the resulting fit is visually smoother than (b). Finally, in Figure 3.3(d) the β_{i0}'s, β_{i1}'s and β_{i2}'s from (3.9) are dropped from the model, thus forcing the function values to join at the knots and have continuous first and second derivatives. It can be seen that the resulting fit is the smoothest of all four models.

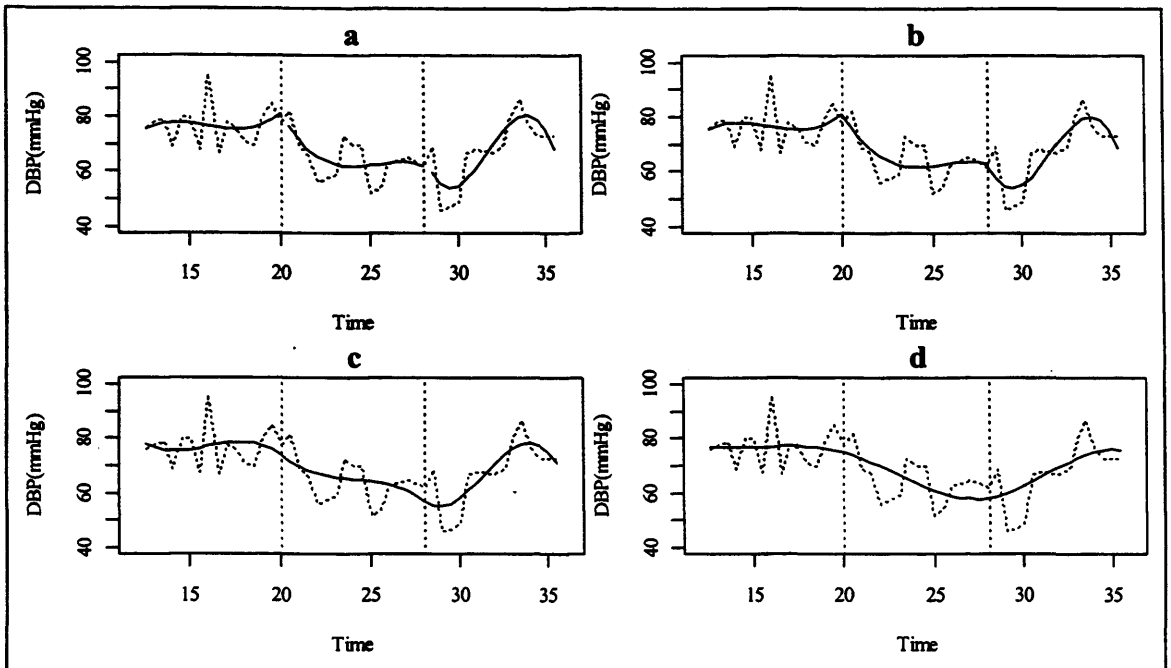

Figure 3.3 **Fitted curves with different continuity restrictions for one subject: (a) no restrictions, (b) forced to join at knots, (c) continuous first derivitive, (d) continuous first and second derivative.**

Inspection of equation (3.9) reveals that there are a number of choices the user has to make when using regression splines (Wold, 1974) . These are (i) the degree of the spline function, n, (ii) the number of knots, k, and (iii) the positions of the knots $(t_1,\ldots,t_k)$.

Cubic splines (n=3) are more generally used due to them being visually smooth as they have continuous first and second derivatives. Higher degree polynomials are generally not needed, since if there was a complicated shape between knots (e.g. with more than 2 turning points) then rather than fitting a higher degree polynomial another knot could be added. When fitting a cubic spline model as in Figure 3.3(d) then (3.9) can be simplified to:

$$S(x) = \sum_{j=0}^{3} \beta_{0j} x^j + \sum_{i=1}^{k} \beta_{i3} (x - t_i)_+^3 + e_i \qquad (3.10)$$

Thus, the number of parameters in a standard regression spline model is k+4.

Regression splines are part of the large family of smoothing methods encompassed by the general term "splines". Sometimes regression splines are expressed using alternative formulations that are more computationally efficient. The most common of these are B-

splines (De Boor, 1978), which are described as being 'nearly orthogonal'. A related non-parametric technique is smoothing splines (Silverman, 1985). The difference between regression splines and smoothing splines is in the choice of knots. When using regression splines a fixed number of knots is set, whereas smoothing splines use each data point as a knot with a penalty term included in the likelihood to control the smoothness of the fitted curve leading to more complex estimation methods particularly when fitting hierarchical models. In this thesis I will not be employing the use of a penalty term to control the smoothness of the curve, but will investigate the sensitivity of the model to the choice of the number and location of knots.

3.4.3 Restricted Cubic Splines

Restricted cubic splines were first suggested by Stone and Koo (1986) and discussed further by Durrleman and Simon (1989) and are a simple extension of regression splines. Restricted cubic splines are forced to be linear in the tails, i.e. before the first knot and after the last knot. There is less information in the tails of the data and over fitting can result if curvature is assumed. If $S(x)$ is to be linear for $x<t_1$ then $\beta_{02} = \beta_{03} = 0$ in (3.10). If $S(x)$ is to be linear for $x>t_k$ then $\sum_{i=0}^{k} \beta_{i3} = 0$ and $\sum_{i=1}^{k} \beta_{i3} t_i = 0$.

It can be shown (Durrleman and Simon, 1989) that using the model formulation in (3.10) that a restricted cubic spline function can be written as follows

$$S(x) = \beta_{00} + \beta_{01} x + \sum_{i=1}^{k-2} \beta_{i3} \left[(x-t_i)_+^3 - \frac{t_k - t_i}{t_k - t_{k-1}} (x-t_{k-1})_+^3 + \frac{t_{k-1} - t_i}{t_k - t_{k-1}} (x-t_k)_+^3 \right] + e_i \quad (3.11)$$

However, this can be simplified as follows. If starting with a time variable, x, then with k knots at locations ($t_1<t_2<\ldots<t_k$), k - 2 new variables are introduced where

$$x_i = (x-t_i)_+^3 - \frac{(x-t_{k-1})_+^3 (t_k - t_i)}{t_k - t_{k-1}} + \frac{(x-t_k)_+^3 (t_{k-1} - t_i)}{t_k - t_{k-1}} i = 1, \ldots, k-2 \quad (3.12)$$

These derived covariates can be easily programmed and as for standard regression splines described in section 3.4.2, the regression coefficients when using restricted cubic splines

can be estimated using a linear model using standard statistical software. For a restricted cubic spline model with k knots there are k regression coefficients to be estimated. Thus, for a standard regression spline model there are 4 extra parameters that need to be estimated when compared to a restriced cubic spline model with the same number of knots.

The fact that the function is forced to be linear in the tails (i.e. before and after the last knot) should not generally be a problem as long as the boundary knots are not too far away from the minimum and maximum values of x and that the function is not changing rapidly at the extremes (Durrleman and Simon, 1989).

Figure 3.4 shows for four subjects an example of fitting a standard cubic regression spline model with 5 knots at (15:00, 20:00, 24:00, 04:00 and 09:00) and a restricted cubic spline model with 5 knots at the same locations. Thus nine parameters need to be estimated for the standard regression spline model compared to five for the restricted cubic spline model. The four plots show that the cubic regression splines pick up more localised peaks and troughs, which may be expected as they have 4 extra parameters. However, interest here lies in the tails and it can be seen that the regression splines appear to be more influential in the tails and start behaving oddly, thus having similar problems in the tails to those encountered when using standard polynomials.

One area where restricted cubic splines have been used is survival analysis where restricted cubic splines have been used for modelling of non-linear continuous covariates (Heinzl and Kaider, 1997) and modelling the baseline hazard function (Herndon and Harrell, 1995). Restricted cubic splines can also be used when testing linearity assumptions in generalized linear models (Harrell *et al.*, 1996).

3.4.4 Periodic Splines

Periodic splines models using robust regression techniques have been suggested for modelling individual ABPM profiles (Streitberg *et al.*, 1989). A periodic spline forces the

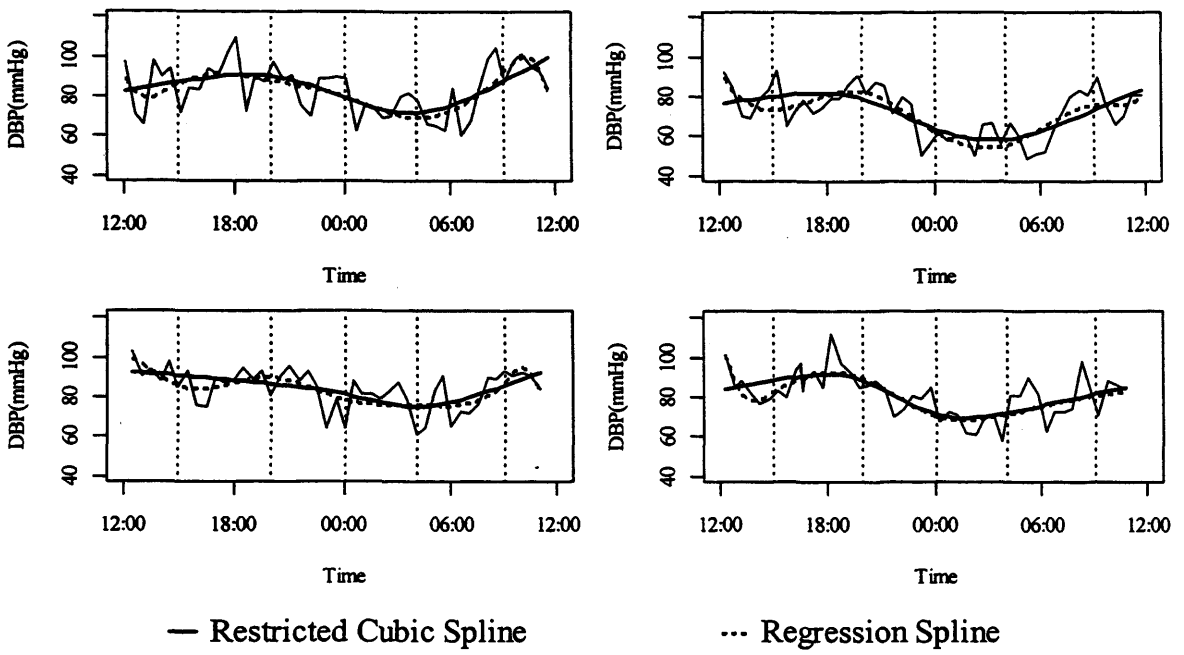

Figure 3.4 Example of regression splines and restricted cubic splines with knots at (15:00, 20:00, 24:00, 04:00 and 09:00) for four women.

fitted ABPM profile to have equal values at the beginning and end of the 24 hour period. With exact repeatability of the 24 hour profiles this may be sensible, but in reality there could be a number of reasons why the blood pressure measurement may be different at the beginning and end of the 24 hour period. For example, when monitoring the effect of blood pressure lowering treatment. Therefore, periodicity in the context of ABPM profiles may be too rigid an assumption (Dickson and Hasford, 1992). Another reason for not assuming periodicity is that the dataset used in this thesis has a number of subjects who do not have complete readings for the 24-hour period. This may lead to complications when forcing the model to agree at the endpoints.

3.4.5 Choice of Number and Location of Knots

As mentioned in section 3.4.2 the number and the location of knots is generally left to the user. The choice can be important, as too many knots will lead to over fitting of the data, while too few will lead to a poor fitting model. In theory, the position and the number of knots can be considered as unknown parameters. Although there are methods that can estimate the number and the location of the knots (Gallant and Fuller, 1973; Denison *et al.*, 1998), it has been argued that using a fixed number of knots and sensibly choosing

their locations may be adequate (Wold, 1974). Wold states that the choice of the locations of the knots corresponds closely with the choice of functional type in an ordinary curve-fitting problem. Generally the choice of function is not considered as a parameter, so the knots in a spline function can be thought of as analogous to this situation. He concludes that the knots should be chosen as to correspond to the overall behaviour of the data (number of points, positions of maxima/minima etc). The curve fitting can then be treated as a standard linear model in terms of the derived covariates. Some rules of thumb for the choice of knots (Wold, 1974) are

(i) Have as few knots as possible (obtaining a balance between overfitting and parsimony).

(ii) Have not more than one maximum or minimum and one inflexion point per interval.

(iii) Have extreme points centred in the intervals.

(iv) Have inflexion points close to the knots.

Other recommendations are (Wegman and Wright, 1983)

(v) Knots should be located at data points.

(vi) A minimum of four or five observations should fall between the knots.

One suggestion is to use 5 knots when using restricted cubic splines, but it has been argued that there is no theoretical basis for this (Durrleman and Simon, 1989). When having 5 knots it has been suggested that it may not be beneficial to have the boundary knots at the extremes due to the potential influence of outliers, but that the boundary knots should not be to far from the extremes due to the restriction that the curve is linear in the tails. Boundary knots at the 5^{th} and 95^{th} percentiles are suggested. A further suggestion is to examine plots of the residuals against time as this may suggest placement of additional knots if curvature is still present. As there is an element of subjectivity when using restricted cubic splines it appears sensible to perform a sensitivity analysis. One would hope that the inferences would not change with a small increase or decrease in the number of knots from an initial model.

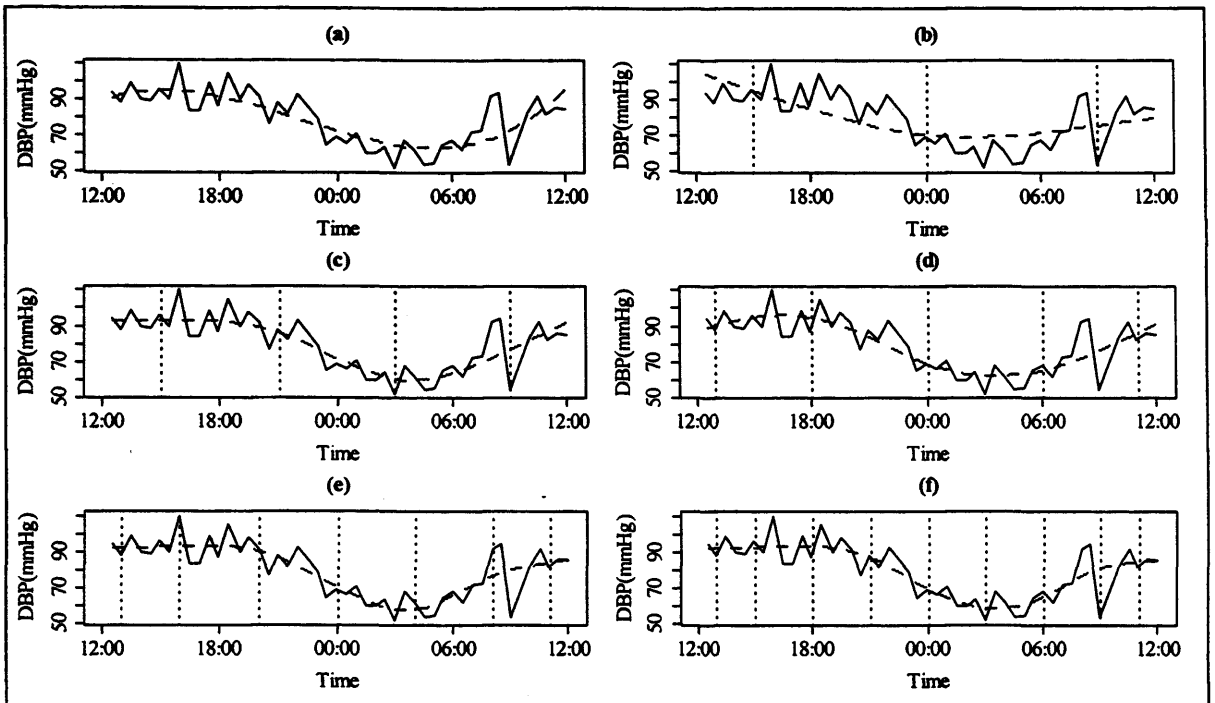

Figure 3.5 **The effect of the number of knots on the fitted curve for a women with 0 knots (a), 3 knots (b), 4 knots (c), 5 knots (d), 7 knots (e), 9 knots (f): Example 1.**

In order to illustrate how the number of knots can affect the shape of the fitted curve Figure 3.5 and Figure 3.6 show the ABPM profiles for two women together with the fitted curves for six different restricted cubic spline models. The six different models have 0 knots (cubic polynomial), and 3, 4, 5, 7 and 9 knots. The locations of the knots are shown in the graphs. For the first example (Figure 3.5), the rate of change in the blood pressure over time is not that great and none of the model fits are totally unacceptable. However, one may have reservations regarding the cubic polynomial (0 knots) and the model with 3 knots. It is hard to differentiate between the models with 4, 5, 7 and knots.

For the second example in Figure 3.6 the blood pressure profile is more complex with a sharp decrease followed by a sharp increase in blood pressure. The cubic polynomial model does not fit the data well with it not picking up some of the large peaks and troughs. These patterns are common and indicates that the third degree polynomial method (Corrao *et al.*, 1996) discussed in section 3.4.1 is not appropriate. Probably worse than this is the restricted cubic spline model with 3 knots. The restriction that the model is linear in the tails leads to there being little flexibility for the rest of the curve. Clearly more knots are

needed. The models with 4 and 5 knots are a slight improvement, but there are still some parts of the curve where the model does not appear to fit well. This can be seen by the fact that at night there are about four hours where the fitted values are below the observed values. The models with 7 and 9 knots appear to pick up most of the observed peaks and troughs, and appear to be a better choice of model for this particular subject.

3.5 Using Restricted Cubic Splines in a Hierarchical Model.

3.5.1 Introduction

In this section I will use restricted cubic splines in a two-level hierarchical model including all 206 ABPM profiles. So far I have just fitted restricted cubic spline models to *individual* profiles, however it is sensible to explore the feasibility of combining all ABPM profiles into one model making use of the hierarchical structure of the data. The structure exists because individual blood pressure measurements are nested within individuals and hence there are two natural units of blood pressure variation, variability between-women and variability within-women. The restricted cubic splines will be used to model the mean

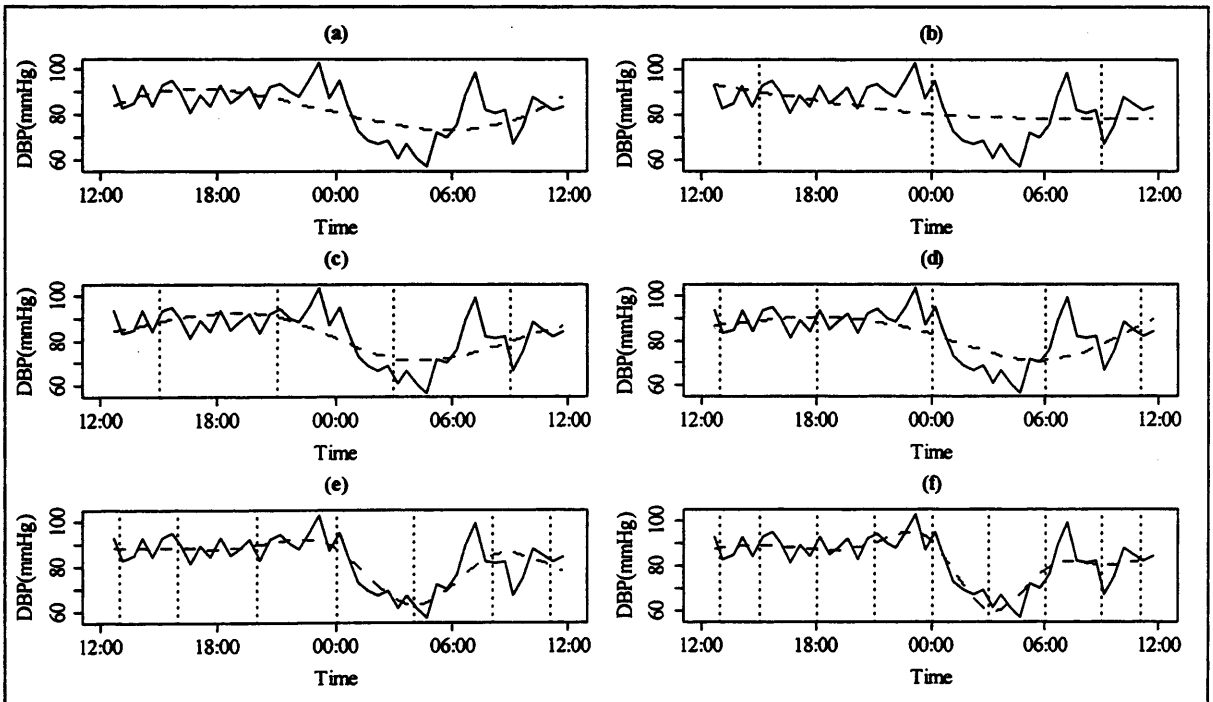

Figure 3.6 The effect of the number of knots on the fitted curve for a women with 0 knots (a), 3 knots (b), 4 knots (c), 5 knots (d), 7 knots (e), 9 knots (f): Example 2.

profiles of the two groups, i.e. those women who gave birth to an infant ≤10th centile for gestational age and those women who gave birth to an infant >10th centile for gestational age. The main question of interest is whether the mean profile differs between the groups, as this will indicate if those women who are going to give birth to a small for gestational age infant can be detected earlier in their pregnancy.

There has been previous work on using splines in the context of longitudinal studies mainly in relation to AIDS patients. One approach used piecewise polynomials (regression splines) (Wang and Taylor, 1995a) when analysing log neopterin values with about 6 observations per subject. In this approach a large fixed number of knots was chosen, but so that the mean curves were smoothed, a penalty term was subtracted from the log likelihood. Thus the method uses penalized maximum likelihood with cross validation methods used to estimate the penalty term. Extensions of this work (Wang and Taylor, 1995b) investigated coverage rates of confidence intervals when using maximum penalized likelihood and the choice of covariance structure. Other work in the area of Aids uses B-splines to model the mean profile of serial CD4 counts (Shi *et al.*, 1996). In this analysis the between-subject random effects are obtained by firstly fitting an overparamerterised model where all fixed effects are also treated as random effects. Principal components analysis is then used to reduce the number of covariance parameters, while still accounting for most of the known between-subject variation observed in the overparameterised model. Another example again in the AIDS area uses semi-parametric model for CD4 counts incorporating non-parametric kernel smoothing to describe the mean profile (Zeger and Diggle, 1994). Cross validation methods are used to provide a smooth mean profile, with a parametric model for covariate adjustment and modelling of the serial correlation.

Restricted cubic splines have been used in growth curve models and were found to perform better than polynomials when predicting future observations (Tian *et al.*, 1994). In a similar type of analysis a hierarchical model using regression splines with 4 knots was used in the analysis of growth data (Pan and Goldstein, 1998). This analysis also incorporated fractional polynomials terms (Royston and Altman, 1997). A further area of application was in the analysis of pulmonary function using B-splines with 10 knots (Wypij *et al.*, 1993). This approach accounted for the within-subject correlations by using generalised

estimating equations (Liang and Zeger, 1986). Because of the large number of observations per subject in this study it has been suggested that there is not need for inclusion of a penalty term (Wang and Taylor, 1995a). This is a similar situation to the ABPM data in this thesis, with up to 48 observations per subject.

There has been little work on the longitudinal analysis of ABPM profiles. One study considers a method that uses weighted least squares to obtain the mean blood pressure at every (fixed) time point for two groups. The weights are obtained by 'smoothing' the variance covariance matrix by assuming that observations set distances apart are equally correlated (Turney *et al.*, 1992). The smoothing assumes that the between-subject variance is constant over the 24-hour period and that the within-subject correlation follows a rather strict pattern in that three separate covariance terms are estimated, one for observations 1 hour apart, one for observations 2-7 hours apart and one for observations >7 hours apart. This method requires rounding the time of each measurement to the nearest hour and estimates the mean blood pressure at each hour for the two treatments. This leads to 48 parameters being estimated in order to describe the 2 blood pressure profiles. Another approach uses a hierarchical model model incorporating a polynomial up to degree 4 to model the mean profiles (Selwyn and Difranco, 1993). This approach has similarities to the approach that I adopt here, but does not incorporate the use of restricted cubic splines. Restricted cubic splines are potentially more flexible than polynomials as discussed in section 3.4.2.

3.5.2 Gram-Schmidt Orthogonalization

Analyses of the ABPM blood pressure profiles when using restricted cubic splines lead to some problems with convergence when fitting some of the more complicated models discussed later in this chapter. This is due to there being strong associations between the derived covariates in the construction of the restricted cubic splines obtained from equation (3.12). There are several ways to formulate spline models, most of which were developed to be more computationally efficient, which is not as important as it was once with advances in computational power. One of the most common methods is B-splines (De Boor, 1978) and these have been used in the analysis of repeated CD4 counts (Shi *et al.*,

1996). However, there are a number of possible formulations that will reduce the association between the derived covariates. Since the derived covariates are not interpretable individually, but need to be considered jointly, it makes sense to orthogonolize the covariates. A simple approach to this that uses linear models is the *Gram-Schmidt* process (Guttman, 1982).

For a restricted cubic spline model with k knots there are k derived covariates $x_0,\ldots,x_{k-1}$, obtained using (3.12), forming a $N\times k$ design matrix X. In order to orthogonalize these covariates a set of new covariates $s_0,\ldots,s_{k-1}$ are computed. Let s_i be the transformation of the i^{th} derived covariate x_i. The s_i are obtained as follows:

$$s_0 = x_0 \tag{3.13}$$

The j^{th} transformed covariate, s_j is obtained by regressing x_j on $s_0,\ldots,s_{j-1}$ and then obtaining the predicted values $\hat{x}_j$. s_j is the vector of residuals of the regression model, thus

$$s_i = x_i - \hat{x}_i \tag{3.14}$$

This is a very simple process and can be easily performed in any statistical package that can fit linear models. For example I wrote a macro in MLn that consisted of about 10 lines of code (see Figure 3.7).

3.5.3 Components of the Model

When fitting a hierarchical model to the ABPM blood pressure profiles there are four aspects of the model that need to be initially considered. These are:

i. The number and location of the knots for the fixed effects.
ii. The choice of random effects to model the between-subject variability.
iii. The modelling of the difference between the groups
iv. The modelling of the within-subject errors.

The choice of the number and location of the knots may be different to that when modelling individual profiles as in section 3.4.5. In the hierarchical model the main interest

```
echo
NOTE ***************************************************************************
NOTE **  This macro uses the Gram-Schmidt process to orthogonalise the derived   **
NOTE **  restricted cubic spline covariates.                                      **
NOTE **                                                                           **
NOTE **  Derived restricted cubic splines need to start in column c80             **
NOTE **  C70 - column of knot locations                                           **
NOTE **                                                                           **
NOTE **  Transformed covariates start in column c110                              **
NOTE ***************************************************************************
count c70 b1
calc b1=b1-1
calc b6=79
calc c79='cons'
calc c110=c79
calc b8=110
loop b4 1 b1
  calc b7=b6+b4
  oreg cb7 b4 c110-cb8 c101
  calc b8=b8+1
  calc cb8=cb7-c101
  calc c109=cb8**2
  sum c109 b11
  calc cb8=cb8/sqrt(b11)
endloop
echo
```

Figure 3.7 MLn macro for Gram-Schmidt Orthogonalisation.

lies in the mean blood pressure profile for each of the two groups. It is probably better to have too many knots rather than too few knots, since with too few knots important aspects of the curve could be missed, while with too many at least these aspects of the curve will be picked up, but with perhaps more 'kinks' due to random variation. I initially chose to have 9 knots at (13:00, 15:00, 18:00, 21:00, 00:00, 03:00, 06:00, 09:00 and 11:00). Note that I have chosen not to have knots at the minimum and maximum times (i.e. 12:00 on both days), but the location of the initial and final knots are close to the boundary to allow for the restriction that they are linear in the tails as discussed in section 3.4.5. With 9 knots there will be 9 fixed effect parameters describing the underlying mean profile. I discuss further the choice of the number of knots in section 3.5.4.

The choice of which covariates are to treated as random effects needs consideration, since these will model the between-subject variation and therefore allow women to have different fitted blood pressure profiles. One option would be to allow all of the coefficients of the derived covariates describing the mean blood pressure profile (i.e the fixed effects) to vary from subject to subject. However, this will probably be unnecessarily complex given the likely variation about the mean profile and would lead to overfitting. With nine knots there will be $k(k+1)/2=45$ unknown parameters in the between-subject (level 2) variance-covariance matrix that requires estimation. At the other extreme, just the intercept

could vary from subject to subject as seen in the simple linear regression model in section 2.4. However, as shown in Figure 2.3, this would force the blood pressure profiles for all individuals to be parallel which is obviously not the case. A sensible compromise may be to use a cubic polynomial for the between-subject random effects, which should allow sufficient variation about the average profile to appropriately define each individual's blood pressure profile (Wang and Taylor, 1995b). It is possible to compare the fitted values for each individual's profile from the hierarchical model to their observed blood pressure, so that the appropriateness of the between-subject random effects can be assessed. This is discussed further in section 3.5.6.

The third aspect of the model to consider is how to model the difference between the two groups. A dichotomous covariate, taking the value 0 for women who subsequently had a child ≤10th centile for gestational age and 1 for women who subsequently had a child >10th centile, can be incorporated into the model to see if there is a 'vertical shift' in the profiles (i.e. parallel profiles, but one group having consistently higher blood pressure throughout the 24 hour period). However, also of interest is whether the blood pressure profiles differ in shape. An interaction between group and the restricted cubic spline parameters is one way of doing this, but will increase the number of parameters by k-1. Assuming that the blood pressure profiles will have a similar shape, it is perhaps more attractive to consider fewer parameters. Fitting a cubic polynomial for the difference between the groups should allow appropriate investigation of any differences in profile shape and avoid overfitting (Beacon *et al.*, 1998). The issue of how to model differences between the two groups is discussed in section 3.5.5.

The final aspect of the model to consider is how to model the within-subject residuals. I will initially assume that the within-subject errors are independently normally distributed with zero mean and constant variance. In doing this I am assuming that the within-subject variance is the same for all women. In section 3.5.7, I investigate modelling of the within-subject variation as a function of time.

I will first fit the following model:

$$BP_{ij} = \sum_{k=0}^{8} \beta_k s_{kij} + \sum_{m=0}^{3} \alpha_{mj} t_{ij}^m + e_{ij} \qquad (3.15)$$

$$\underline{\alpha}_j \sim MVN(0, \Sigma), \qquad e_{ij} \sim N(0, \sigma^2)$$

where BP_{ij} is the i^{th} blood pressure measurement on the j^{th} women, the β's are the fixed effects coefficients associated with the Gram-Schmidt transformed restricted cubic spline covariates, s_{kij}, t_{ij} is the time of the i^{th} blood pressure measurement on the j^{th} woman and the α_{mj}'s are the random effects allowing individual blood pressure profiles to vary about the mean profile. It is assumed that $\underline{\alpha}_j$ has a zero mean vector and variance matrix Σ. This model ignores the group effect and models the between-subject variation as a cubic polynomial. A plot of the mean profile for this model can be seen in Figure 3.8(a). The observed blood pressure profiles of ten randomly chosen women have been added to the plot. These show that there is relatively large between-subject and within-subject variability of blood pressure. It appears from the figure that the mean blood pressure is approximately constant during the day, until about 21:00 when it starts to dip and continues decreasing until just after midnight. The mean blood pressure then appears to be approximately constant during the night until about 6:00, when it starts to increase again to the day-time mean. For the group as a whole there appears to be a reduction in mean blood pressure at night of just over 10 mmHg. A formal assessment of whether the blood pressure profile changes over time can be made by comparing this model with a model that just includes an intercept (i.e. no effect of time on blood pressure) using the likelihood ratio test (Goldstein, 1986). Not surprisingly this yielded a highly significant result ($\chi_8^2 = 828.2$, P<0.001), indicating that there is strong evidence that blood pressure was not constant over the 24-hour period.

The group effect is now added to the model as a dichotomous covariate, bwt_j, which takes the value 1 for women who subsequently had a child ≤10th centile for gestational age and 0 when for women who subsequently had a child >10th centile. Model (3.15) can be extended as follows,

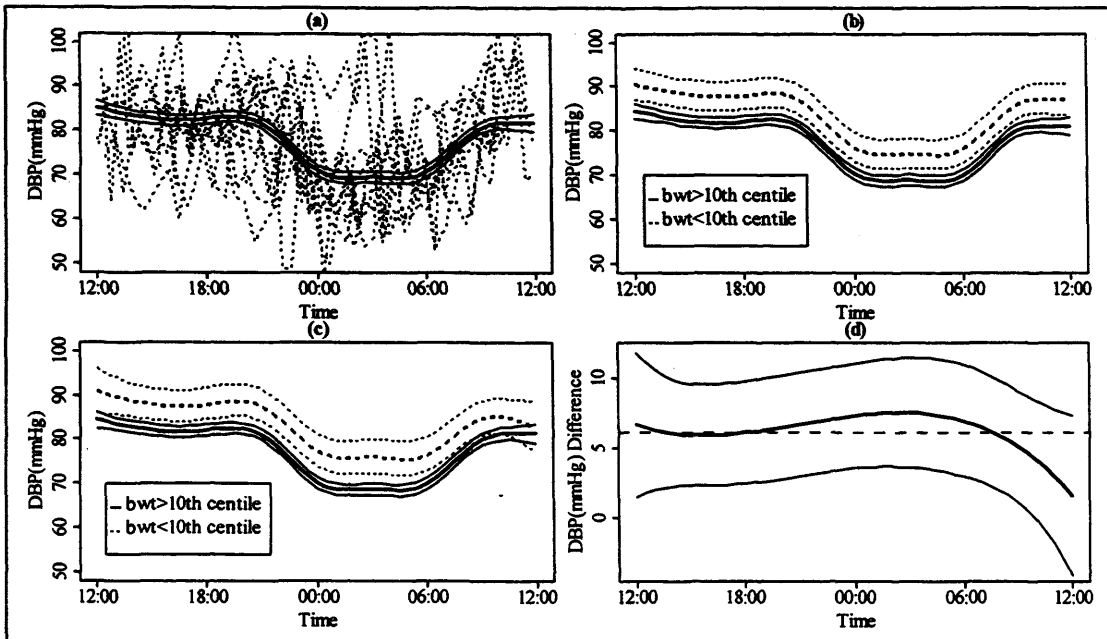

Figure 3.8 Fitted values and confidence intervals for models with (a) no group effect, (b) difference between groups forced to be parallel, (c) difference between groups modelled by cubic polynomial and (d) mean difference between groups when using cubic polynomial.

$$BP_{ij} = \sum_{k=0}^{8} \beta_k s_{kij} + \delta_0 bwt_j + \sum_{m=0}^{3} \alpha_{mj} t_{ij}^m + e_{ij} \qquad (3.16)$$

where δ_0 is the fixed effect associated with bwt_j and represents the mean difference between the groups assuming that the difference is constant throughout the 24 hour period. Figure 3.8(b) shows the fitted mean profiles for this model, where those women who had a child ≤10th centile have a different, but parallel, mean blood pressure profile to those women who had an infant>10th centile. A similar pattern to Figure 3.8(a) can be seen where for both groups blood pressure appears approximately constant during the day and then dips and then is approximately constant during the night. The figure shows that the women with heavier babies for gestational age tended to have higher blood pressure with δ_0 =6.2, which can be interpreted as there being a mean difference of 6.2 mmHg (95% confidence interval 2.8 mmHg to 9.6 mmHg) between the two ABPM mean profiles (Likelihood Ratio test comparing models (3.16) and (3.15): $\chi_1^2 = 12.6$, P<0.001).

The above model may be too restrictive in that it forces the two profiles to be parallel. In the following model the difference between the profiles is modelled by using a third degree (cubic) polynomial:

$$BP_{ij} = \sum_{k=0}^{8} \beta_k s_{kij} + \sum_{n=0}^{3} \delta_n bwt_j \times t_{ij}^n + \sum_{m=0}^{3} \alpha_{mj} t_{ij}^m + e_{ij} \qquad (3.17)$$

Figure 3.8(c) shows the fitted values for this model. Although the two profiles have slightly different shapes they are broadly similar and the difference is non-significant (Likelihood ratio test comparing (3.17) and (3.16): $\chi_3^2 = 4.3$, P=0.23) indicating that there is insufficient evidence to reject the hypothesis that the profiles are in fact parallel. It can be useful to plot the difference in blood pressure between the two groups against time. This is shown in Figure 3.8(d). A line has been added showing the difference observed in the model in (3.16), where the difference was assumed to be constant over time. Although non-significant, it can be seen that the main difference between the profiles is at the end of the 24 hour period. This is where there is less data and where the cubic polynomial modelling

Parameter		Estimate (SE)
FIXED EFFECTS		
β_0	Spline Parameters	76.6 (0.54)
β_1		-276.6 (16.86)
β_2		363.3 (15.71)
β_3		221.4 (13.85)
β_4		28.4 (9.24)
β_5		-118.9 (9.02)
β_6		-128.0 (9.00)
β_7		-6.3 (9.00)
β_8		-73.8 (8.99)
δ_0	Group Comparison	7.4 (1.90)
δ_0		0.163 (0.201)
δ_2		-0.023 (0.014)
δ_3		-0.0026 (0.0020)
RANDOM EFFECTS		
Σ		$\begin{pmatrix} 60.9 & & & \\ 0.74 & 0.47 & & \\ -0.147 & -0.0055 & 0.0020 & \\ -0.0037 & -0.0038 & 0.00005 & 0.00004 \end{pmatrix}$
σ^2		78.5

Table 3.1 Parameter estimates for initial ABPM model.

the difference in the profiles is most likely to be affected by potentially outlying observations. The parameter estimates for this model can be seen in Table 3.1. The parameter estimates of the fixed effects associated with the restricted cubic splines are generally not of interest individually, but it is interesting to note that all bar one are formally significant at the 5% level.

To summarise, the above model appears to indicate that women who had infants ≤10th centile for gestational age had a mean diastolic blood pressure 6.2 mm Hg higher than those women who had infants >10th centile for gestational age. There was little evidence that the two profiles differed in shape. The model defined in equation (3.17) will be referred to as the *initial model*. I will now perform various sensitivity analyses to see how robust the initial model is to changes in the choice of the number of knots, the choice of random effects, how the difference between the profiles is modelled and how the within-subject variance is modelled.

3.5.4 Choice of the Number and Location of Knots

In order to investigate the effect of the choice of the number of knots, six models with differing numbers of knots will be compared. In all models the between-subject variation will be modelled using a cubic polynomial, as in the initial model described in the previous section. Similarly, the within-subject variation is assumed to be independently normally distributed with zero mean and constant variance, and the difference between the mean profiles is the same as the initial model (3.17) where it is modelled as a cubic polynomial. Thus, the only difference between the six models is the number of knots and hence the number of parameters used to describe the underlying mean profile. The six models thus take the form

$$BP_{ij} = \sum_{k=0}^{k} \beta_k s_{kij} + \sum_{n=0}^{3} \delta_n bwt_j \times t_{ij}^n + \sum_{m=0}^{3} \alpha_{mj} t_{ij}^m + e_{ij} \qquad (3.18)$$

$$\underline{\alpha}_j \sim MVN(0,\Sigma) \qquad e_{ij} \sim N(0,\sigma^2)$$

where k is the number of knots. The six models fitted are:

Model A:	0 knots (cubic polynomial)
Model B:	3 knots at (13:00, 00:00, 11:00)
Model C:	5 knots at (13:00, 18:00, 00:00, 06:00, 11:00)
Model D:	7 knots at (13:00, 16:00, 20:00, 00:00, 04:00, 08:00, 11:00)
Model E:	9 knots at (13:00,15:00, 18:00, 21:00, 00:00, 03:00, 06:00, 09:00, 11:00)
Model F:	13 knots at (13:00, 14:00, 16:00, 18:00, 20:00, 22:00, 00:00, 02:00, 04:00, 06:00, 08:00, 10:00, 11:00)

Model E is therefore the same as the initial model in (3.17), with 9 knots. Fitted mean profiles for all six models can be seen in Figure 3.9. The cubic polynomial model (A) in Figure 3.9(a) shows much slower changes when compared to the model with 9 knots (model E). With a cubic polynomial if the minimum and maximum are a long way apart then the curve that joins these two points can only change slowly. Model B with 3 knots shown in Figure 3.9(b) has less parameters than model A and therefore the fit is even worse. Models C and D, shown in Figure 3.9(c) and Figure 3.9(d) with 5 and 7 knots, are an improvement and are picking up the general shape seen in the model with 9 knots, but still show more curvature than model E even though the 9 knot model has the potential to pick up more localised curvature. Model F with 13 knots shown in Figure 3.9(f) shows remarkably similar mean profiles to the model with 9 knots. Visual inspection of the fitted mean profiles for the six models indicates that the models with 9 and 13 knots are probably best. Given that the two models appear to give similar results it would appear preferable to choose the model with 9 knots as it has four less parameters.

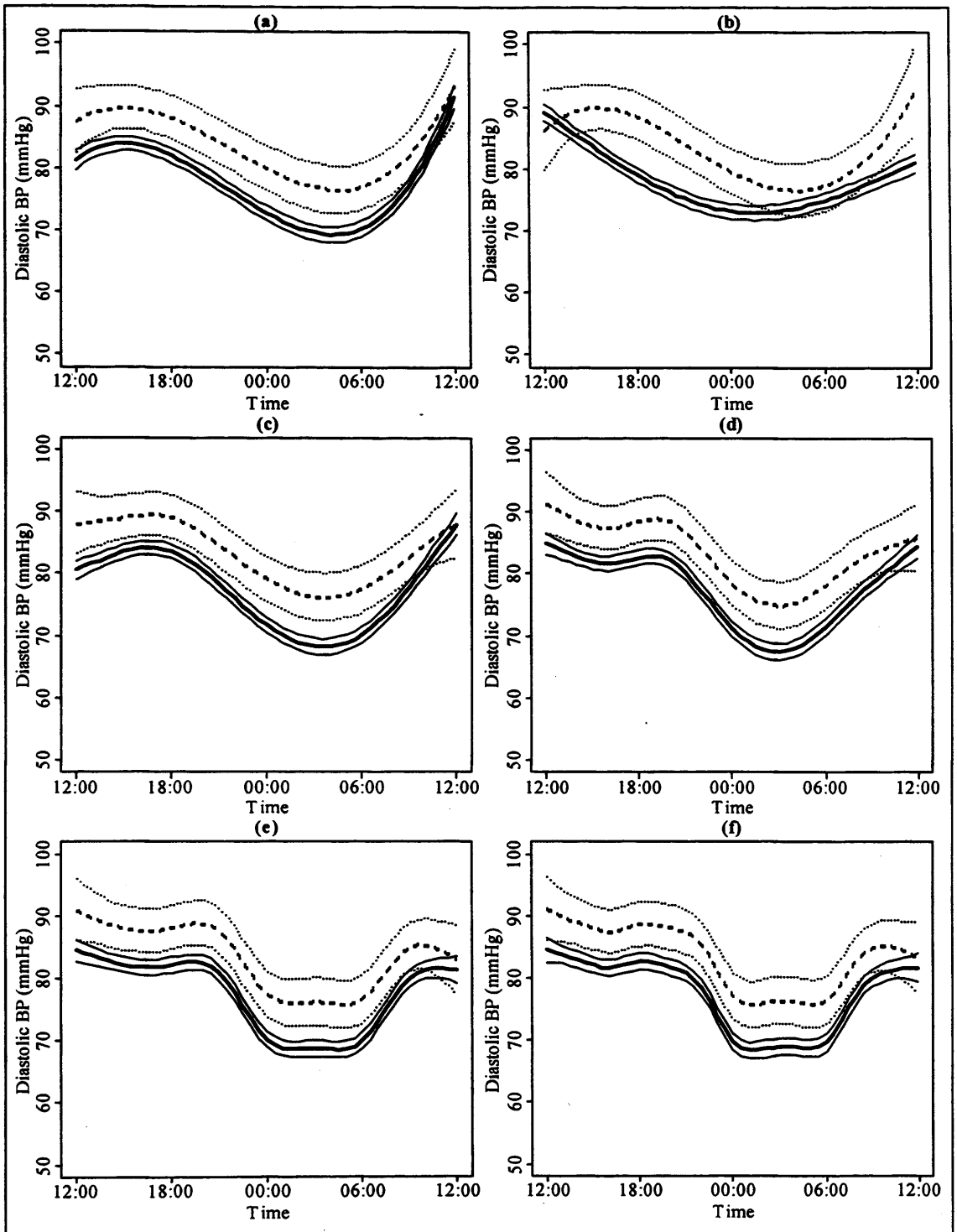

Figure 3.9 Fitted values and 95% confidence intervals for women who subsequently gave birth to an infant >10th centile (—) and those who gave birth to an infant ≤10th centile (----) with (a) 0 knots (cubic polynomial), (b) 3 knots, (c) 5 knots, (d) 7 knots, (e) 9 knots and (f) 13 knots.

An alternative way to compare the models is to formally test whether there is evidence of a difference in the shape of the mean profiles. One would hope that, as long as there were sufficient knots, the different models would give similar results. This can be done by comparing each model to a reduced model where the profiles are forced to be parallel (i.e. for each model removal of the linear(δ_1), quadratic(δ_2) and cubic(δ_3) interactions with bwt_j). Table 3.2 shows the results of this comparison using the likelihood ratio test. It is interesting to note that the models with 7, 9 and 13 knots give almost identical significance levels for a test of a difference in the shapes of the profiles, indicating that perhaps the models are not very sensitive to the choice of the number of knots as long as there are enough of them.

The models with differing numbers of knots can not be compared using the likelihood ratio test as they are not nested. However, it is possible to compare them using other methods which account for the differences in the number of parameters in each model by subtracting a penalty term from the likelihood ratio statistic. I will use two of these, the Bayesian Information Criterion (BIC) (Schwarz, 1978) and the Akaike Information Criterion (AIC) (Akaike, 1980). Let W be the difference in -2×Log likelihood between two models then the BIC is defined by

$$BIC = W - (p_2 - p_1)\log(n) \tag{3.19}$$

where p_i is the number of parameters in model i and n is the total number of observations. The AIC is similarly defined as

$$AIC = W - 2(p_2 - p_1) \tag{3.20}$$

Number of knots	-2×Log likelihood	Change in -2×Log likelihood from Parallel assumption	P-value
A: 0 Knots	63561.8	3.0	0.39
B: 3 Knots	63748.8	12.7	0.0053
C: 5 Knots	63342.3	6.7	0.082
D: 7 Knots	63159.6	4.4	0.22
E: 9 Knots	63085.9	4.3	0.23
F: 13 Knots	63074.8	4.3	0.23

Table 3.2 Formal test of difference in shape of mean profiles for the six models with differing number of knots using the likelihood ratio test.

When comparing models using either the BIC or AIC the "better fitting" models have larger values. Since the BIC and AIC have different penalty terms it is possible they will give different answers. In fact it has been noted that the AIC "keeps too many terms in the model" as it has a smaller penalty term (Carlin and Louis, 1996). In order to compare the six models they all need to be compared to the same initial model. The comparison model I have chosen is a model where only an intercept is fitted as a fixed effect (i.e. no change in blood pressure over time) and a cubic polynomial for the between-subject random effects. The deviance for this model is 63931.0. The results for the AIC and BIC can be seen in Table 3.3. The highest value for the BIC is for the model with 9 knots. This is sensible as it agrees with the fitted value plots in Figure 3.9 in that the model with 13 knots appeared to fit similar mean profiles for the two groups when compared to the model with 9 knots, but had 4 more parameters. The highest value for the AIC is for the model with 13 knots. This could be due to the penalty term for the AIC being smaller than the BIC. I would tend to opt for the results from the BIC as it confirms the initial view that the model with 9 knots gave the "best" fit.

I have chosen to space the knots evenly over the 24-hour period. This does not have to be the case, for example Streitberg (1989) uses clinical opinion to choose the location of the knots. It may be theoretically possible to reduce the number of knots, by changing the location of the knots. For example, having fewer knots where the profile is observed to be flat. However, in a large dataset, such as the ABPM data, the inclusion of one or two extra fixed effect parameters is not a problem. Unless confronted with a small dataset, I would

	Number of Parameters	-2×Log likelihood	BIC	AIC
A: 0 Knots	8	63561.8	305.8	355.2
B: 3 Knots	6	63748.8	136.9	172.2
C: 5 Knots	8	63342.3	525.3	574.7
D: 7 Knots	10	6319.6	689.9	753.4
E: 9 Knots	12	63085.9	745.4	823.1
F: 13 Knots	16	63074.8	720.3	826.2

Table 3.3 BIC and AIC values for the six models with a differing number of knots

prefer to have evenly spaced knots, rather than trying to reduce the number of knots by changing the locations after observing the model fit.

3.5.5 *Modelling the Difference Between Groups*

In the initial model in (3.17) the difference between the two mean profiles was modelled using a cubic polynomial. However, this could lead to problems if the difference between the two mean profiles is more complicated than it is possible to model using a cubic polynomial (for example with more than two turning points). I have also discussed some of the limitations in using polynomials in section 3.4.2. In this section I fit six different models to compare the modelling of the difference in the mean profiles. In order to do this I have standardised other aspects of the model. The underlying mean profile is modelled using 9 knots as in the initial model in (3.17). Similarly the between-subject variation is modelled using a cubic polynomial for the random effects and the within-subject variation is assumed constant. Thus the only difference between the 6 models is how the difference between the two mean profiles is modelled. The six models model the difference in the mean profiles as follows:

Model A:	Intercept (i.e. forces the mean profiles to be parallel)
Model B:	cubic polynomial
Model C:	quartic polynomial
Model D:	5 knots at (13:00, 18:00, 00:00, 06:00 and 11:00)
Model E:	7 knots at (13:00, 16:00, 20:00, 00:00, 04:00, 08:00 and 11:00)
Model F:	9 knots at (13:00, 15:00, 18:00, 21:00, 00:00, 03:00, 06:00, 09:00 and 11:00)

Thus model A is the same model as in (3.16) and model B is the initial model in (3.17). Model F is equivalent to fitting a separate mean profile using 9 knots for each group. Figure 3.10 shows the fitted values for the two mean profiles for the six different models and Figure 3.11 shows the difference in mean profiles together with 95% confidence intervals. The difference between the mean profiles is remarkably similar in all models, except obviously model A where the difference is assumed constant. The difference is

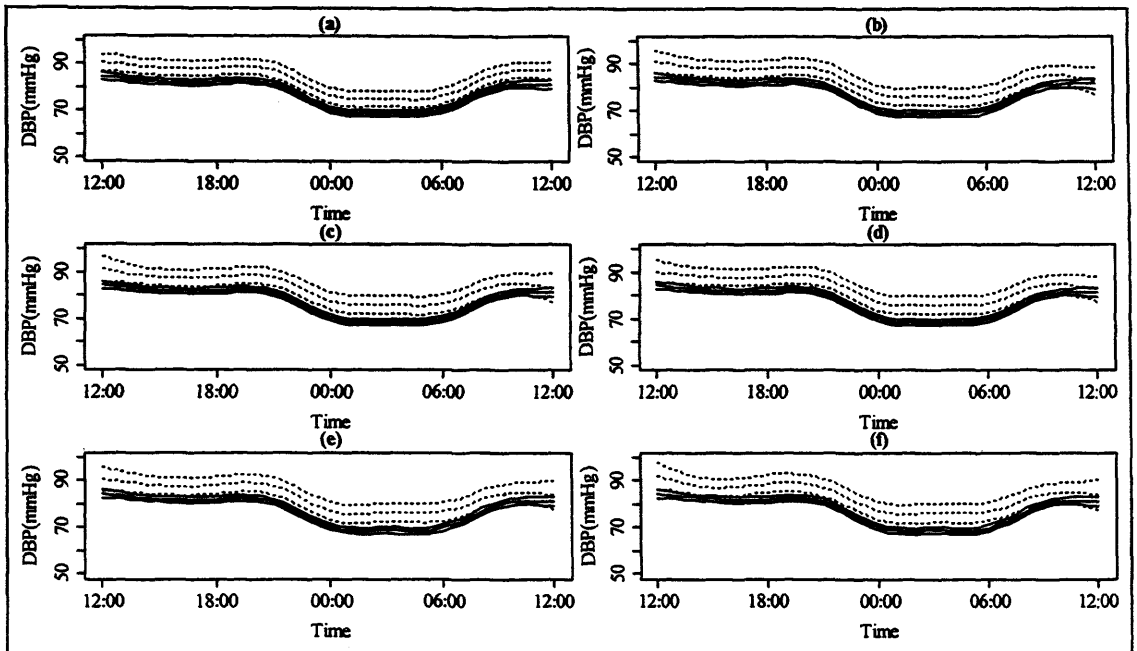

Figure 3.10 Fitted mean profiles and 95% confidence intervals for 6 models with differing interaction terms to model the difference between groups: (a) none, (b) cubic, (c) quartic, (d) 5 knots, (e) 7 knots, (f) 9 knots.

approximately constant at about 6 mm Hg until about 06:00 when it starts to decrease. For the models with more knots and thus more parameters, i.e. the model with 7(E) and 9(F) knots, more local changes are detected, but these are not clinically or statistically important.

Since the main interest lies in whether there is a difference in the mean profiles it is possible to formally compare each of models B-F with model A, where the difference is assumed parallel, using the likelihood ratio test. The results for these analyses can be seen in Table 3.4. All comparisons yield statistically non-significant results so one would conclude in all cases that there was insufficient evidence to state that the mean profiles were not parallel. However, the P-values tend to increase as the number of parameters increases which is to be expected as more parameters are being used to explain a small difference, so the change in likelihood values will decrease by only a small amount while the change in the number of parameters increases.

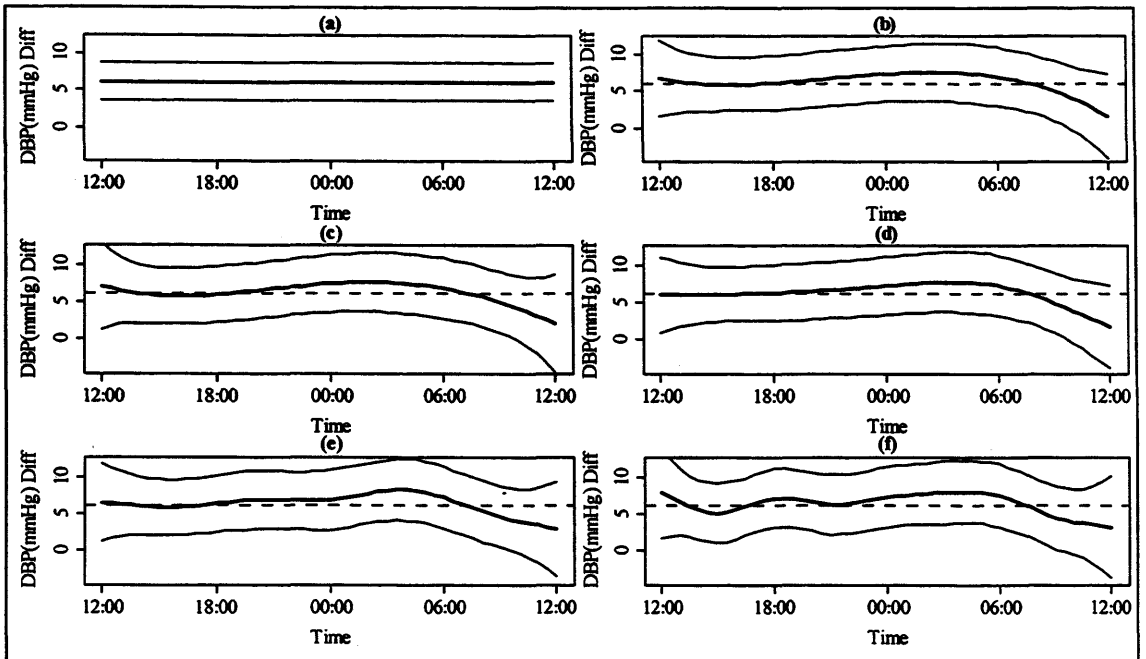

Figure 3.11 Difference in mean profiles for 6 models with differing interaction terms to model the difference between groups: (a) no interaction, (b) cubic, (c) quartic, (d) 5 knots, (e) 7 knots, (f) 9 knots.

3.5.6 Choice of Random Effects

In the initial model the between-subject variation was modelled using a cubic polynomial for the random effects. It was envisaged that this model would allow there to be sufficient variation about the mean profiles to appropriately define each woman's mean blood pressure profile. However, it is of interest to investigate the effect of different choices for the between-subject random effects, as the more random effects there are, the greater the flexibility in the shape of the individual profiles.

	-2×Log likelihood	No. of Parameters for difference	Change in -2×Log likelihood from Parallel assumption	P-value
A: No interaction	63090.2	1	-	-
B: Cubic Polynomial	63085.9	4	4.3	0.23
C: Quartic polynomial	63085.8	5	4.4	0.35
D: 5 Knots	63085.4	5	4.8	0.31
E: 7 Knots	63084.5	7	5.7	0.46
F: 9 Knots	63083.2	9	7.0	0.54

Table 3.4 Formal test of difference in shape of mean profiles for the six models with differing number of knots using likelihood ratio test comparing models B-F with model A.

As for the comparison of the number of knots, and the modelling of the difference between the groups, six models based on the initial model in (3.17) will be compared. Thus, each model will have 9 knots to describe the underlying mean profile, use a cubic polynomial to model the difference between the two profiles and assume a constant within-subject variance. Therefore, the only thing that differs between the six models is the between-subject random effects parameters. The random effects will be modelled in the six models as follow

Model A:	Intercept only
Model B:	Linear polynomial of time
Model C:	Cubic polynomial of time
Model D:	5 knots at (13:00, 18:00, 00:00, 06:00 and 11:00)
Model E:	7 knots at (13:00, 16:00, 20:00, 00:00, 04:00, 08:00 and 11:00)
Model F:	9 knots at (13:00, 15:00, 18:00, 21:00, 00:00, 03:00, 06:00, 09:00 and 11:00)

The number of variance-covariance terms that need to be estimated range from 1 in model A to $9\times(10)/2=45$ in model F. In order to illustrate how the choice of random effects can affect the fitted values for an individual, the fitted values for 5 individuals mean profiles for all six models are shown in Figure 3.12-Figure 3.16. For each of the five women the first model just allows the intercept to vary from women to women and therefore forces the fitted profile for each woman to be parallel to the overall mean profile. It can be seen that this may be adequate for women who have similar profiles to the mean profile, but is clearly not adequate for those women whose profile differs from the mean profile (which will be the case for most women). The linear model B does little better which perhaps one would expect as it only allows the fitted individual profiles to differ from the mean profile by either decreasing or increasing over time and does not allow for a change of direction.

The individual profiles visually appear to have the best fit for the models with more parameters, i.e. models D-F. However, it is unclear which is the "best" model and whether it actually matters if one is not directly interested in the fit of the individual profiles, but rather in the comparison between groups. Thus it is of interest to investigate whether fitting a cubic polynomial, with ten parameters estimated in the variance-covariance matrix, or a

nine knot restricted cubic spline with 45 parameters. As the models are not nested it is not possible to use the likelihood ratio test, but again it is possible to use the BIC and AIC. The deviance, number of parameters in the model, BIC, AIC and the within-subject variance are shown in Table 3.5. The highest value of the BIC is for the model with a restricted cubic spline with seven knots for the between-subject (level 2) random effects, while the AIC chooses the model with nine knots. As stated in section 3.5.4, the AIC tends to lead to over-parameterised models being selected, so perhaps the BIC is more appropriate.

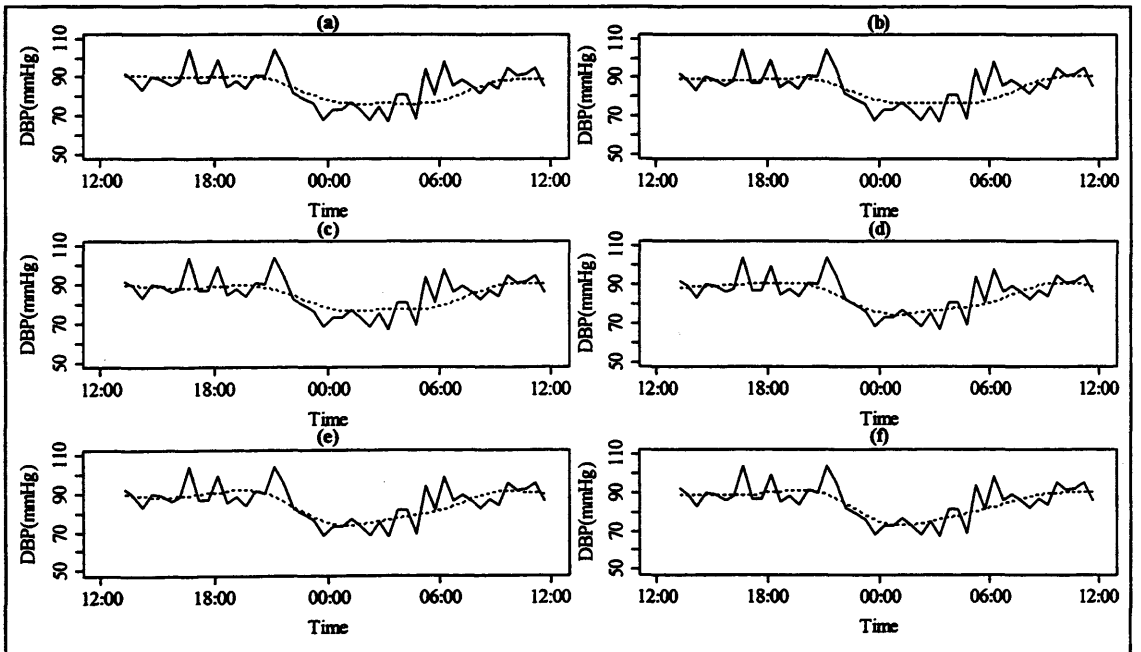

Figure 3.12 **Predicted values for an individual for models with different between-subject random effects: (a) intercept, (b) linear polynomial, (c) cubic polynomial, (d) 5 knots, (e) 7 knots and (f) 9 knots. Example 1.**

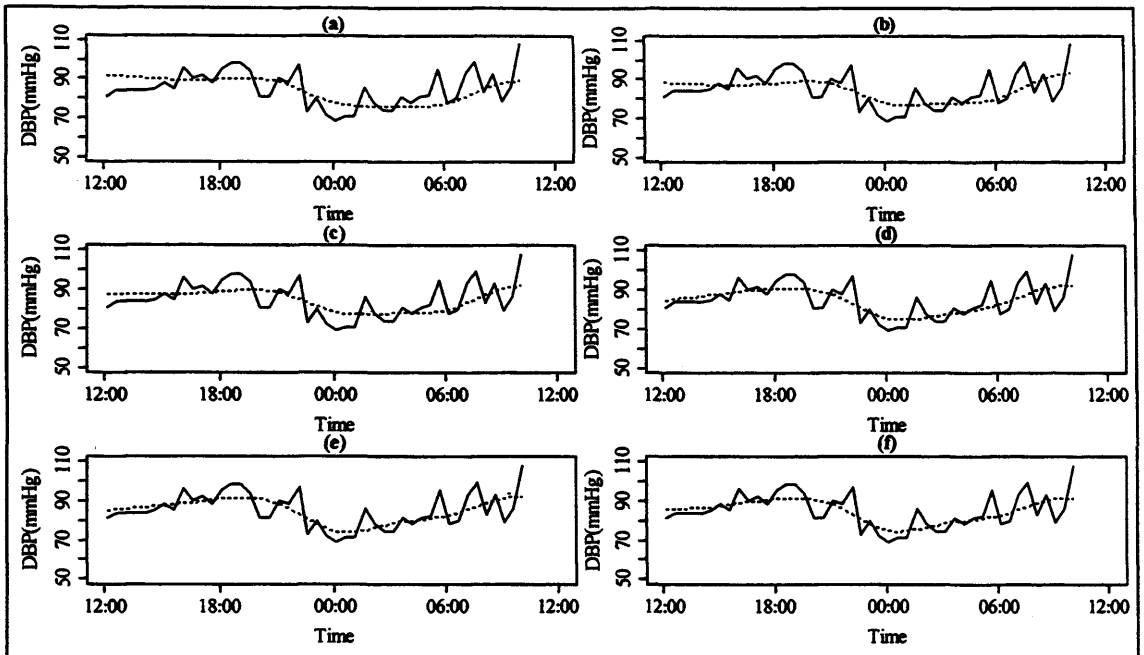

Figure 3.13 **Predicted values for an individual for models with different between-subject random effects: (a) intercept, (b) linear polynomial, (c) cubic polynomial, (d) 5 knots, (e) 7 knots and (f) 9 knots. Example 2.**

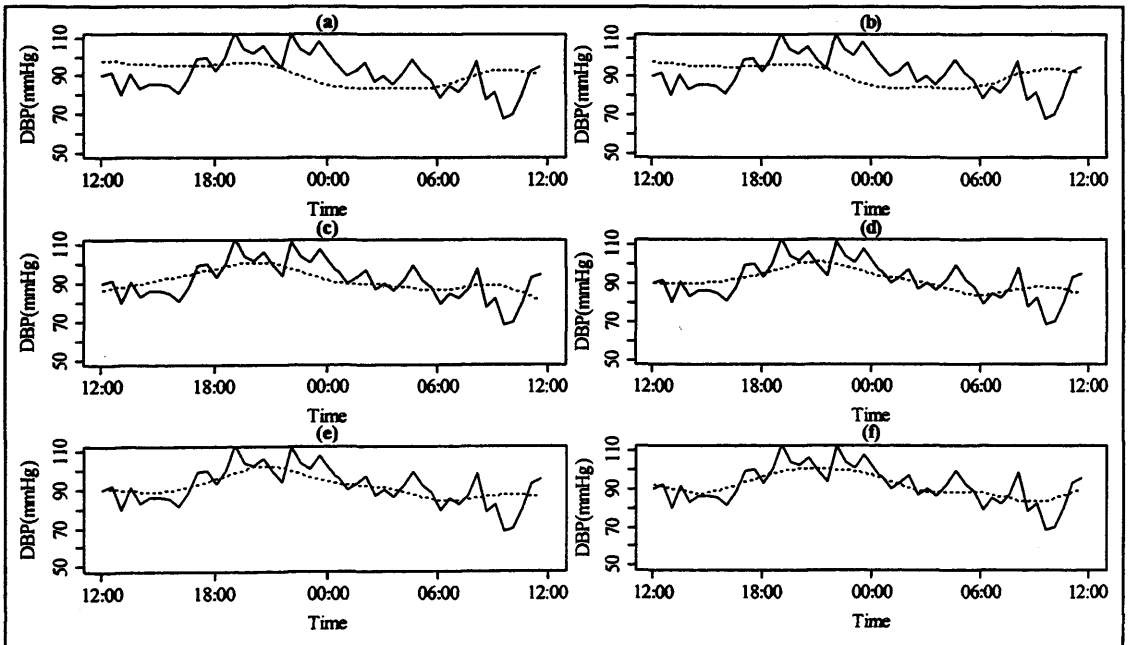

Figure 3.14 **Predicted values for an individual for models with different between-subject random effects: (a) intercept, (b) linear polynomial, (c) cubic polynomial, (d) 5 knots, (e) 7 knots and (f) 9 knots. Example 3.**

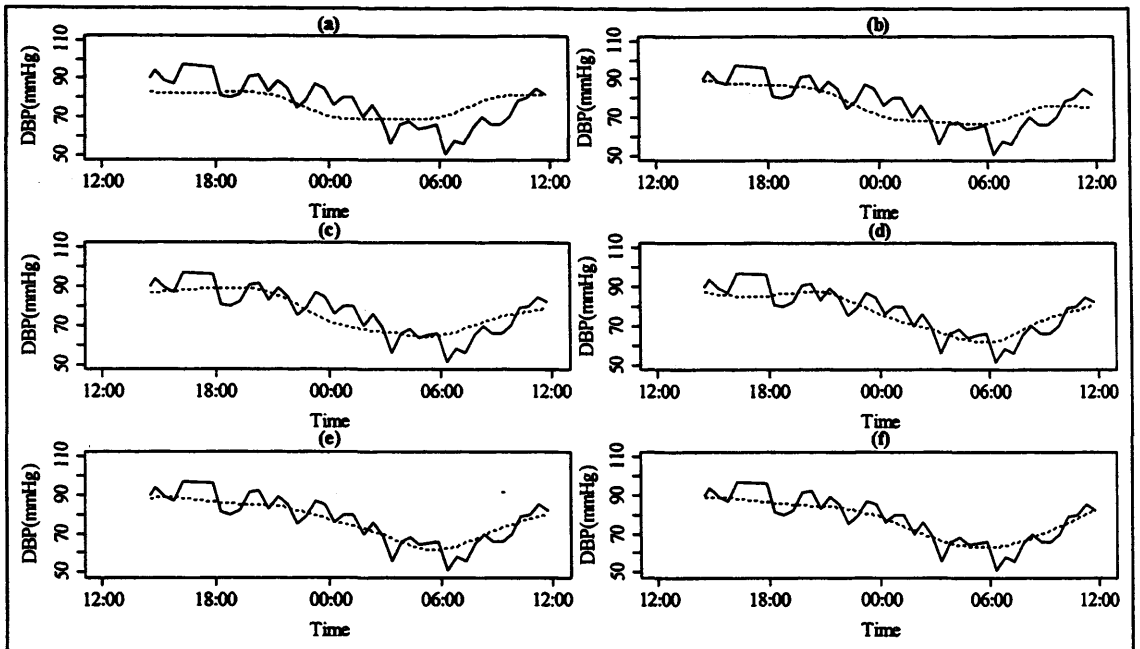

Figure 3.15 **Predicted values for an individual for models with different between subject random effects: (a) intercept, (b) linear polynomial, (c) cubic polynomial, (d) 5 knots, (e) 7 knots and (f) 9 knots. Example 4.**

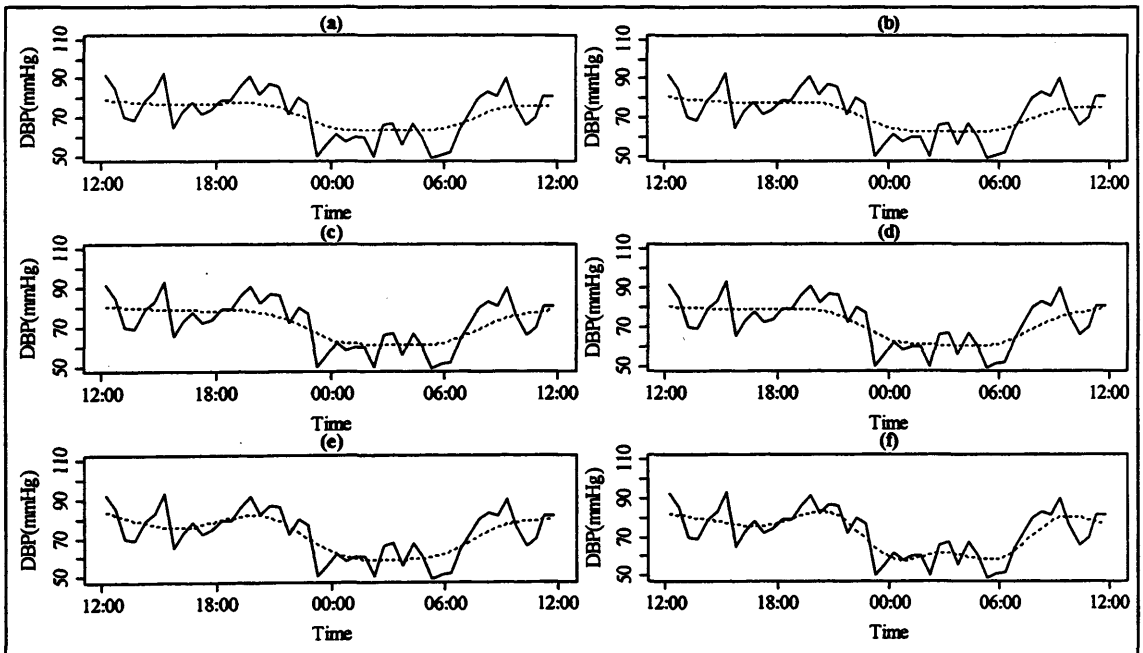

Figure 3.16 **Predicted values for an individual for models with different between-subject random effects: (a) intercept, (b) linear polynomial, (c) cubic polynomial, (d) 5 knots, (e) 7 knots and (f) 9 knots. Example 5.**

Also of interest in the within-subject (level 1) variance for the models with a different number of random effects terms. Figure 3.12-Figure 3.16 show that as the number of random effect parameters increases, the better the fit to each individual's profile. Therefore, one would expect the greater the number of between-subject random effect parameters, the smaller the within-subject variability. The within-subject variability for the six models can be see in Table 3.5. It can be seen that, as expected, the within-subject variance decreases, the greater the number of between-subject random effect parameters. If a large number of extra level between-subject parameters only produced a small reduction in the within-subject variance then this could indicate that the extra parameters are probably not needed.

A crucial question is whether it actually matters how the between-subject variation is modelled when it comes to assessing the differences in the two mean profiles. Table 3.6 shows for each of the six models, the reduction in –2xlog likelihood when testing the null hypothesis that the two mean profiles are parallel. It can be seen that for models A and B there is a significant change in deviance with both models having a P-value of 0.014. However, for the remaining models the change in deviance is non-significant. It is perhaps reassuring that models C and F which could be considered to be realistic in terms of modelling the differences between subjects give very similar results in terms of the change in deviance and hence P-value.

	-2×Log likelihood	No. of Parameters	BIC	AIC	Within-subject Variance
A: Intercept	63452.3	15	2673.8	2758.5	87.2
B: Linear Polynomial	63286.1	17	2821.9	2920.7	83.3
C: Cubic Polynomial	63085.9	23	2967.7	3108.9	78.5
D: 5 Knots	62965.1	28	3043.2	3219.7	76.0
E: 7 Knots	62817.2	41	3073.4	3341.6	72.2
F: 9 Knots	62735.3	58	3001.3	3389.5	69.5

Table 3.5 AIC and BIC values for six different models with differing numbers of between-subject random terms.

	-2×Log likelihood	Change in -2×Log likelihood for testing parallel assumption	P-Value
A: Intercept	63452.3	10.6	0.014
B: Linear Polynomial	63286.1	10.6	0.014
C: Cubic Polynomial	63085.9	4.3	0.23
D: 5 Knots	62965.1	4.4	0.22
E: 7 Knots	62817.2	4.8	0.19
F: 9 Knots	62735.3	4.0	0.26

Table 3.6 Change in deviance in testing assumption of parallel profiles for models with a differing number of random effects using the likelihood ratio test.

3.5.7 Complex Level 1 Variation

Inspection of the plots of the raw data for the first 12 subjects in Figure 3.1 indicated that there may be variation between subjects in terms of the within-subject variances. One advantage of using a hierarchical model is that it is possible to model the within-subject variation in a similar way to the between-subject variation. Goldstein (1995) refers to this as complex level 1 variation. When the within-subject variances are heterogeneous it is good practice to attempt to identify if there are particular types of subjects who tend to have greater or less variation in their response. This is done by modelling the within-subject variance using subject level covariates. It is also possible that the within-subject variance is a function of time. This may be appropriate for this data set as one may believe there to be less variation in blood pressure at night when people tend to be less active. In order to illustrate the modelling of the within-subject variance I will extend the initial model so that the within-subject variation is a function of time for both groups by fitting two models.

The initial model can be extended by adding a subscript *ij* to the within-subject variance term σ^2.

$$BP_{ij} = \sum_{k=0}^{k} \beta_k s_{kij} + \sum_{n=0}^{3} \delta_n bwt_j \times t_{ij}^n + \sum_{m=0}^{3} \alpha_{mj} t_{ij}^m + e_{ij}$$

$$\underline{\alpha}_j \sim MVN(0,\Sigma) \qquad e_{ij} \sim N(0,\sigma_{ij}^2) \tag{3.21}$$

I will model the within-subject variance a cubic polynomial for both groups where

$$\sigma_{ij}^2 = \lambda_0 + \lambda_1 t_{ij} + \lambda_2 t_{ij}^2 + \lambda_3 t_{ij}^3 + \lambda_4 bwt_j + \lambda_5 bwt_j t_{ij} + \lambda_6 bwt_j t_{ij}^2 + \lambda_7 bwt_j t_{ij}^3 \tag{3.22}$$

and as a 5 knot restricted cubic spline function with knots at (13:00,18:00,00:00,06:00 and 11:00) for both groups where

$$\begin{aligned}\sigma_{ij}^2 = {} & \lambda_0 + \lambda_1 s_{1ij} s + \lambda_2 s_{2ij} + \lambda_3 s_{3ij} + \lambda_4 s_{4ij} + \lambda_5 bwt_j + \\ & \lambda_6 bwt_j s_{1ij} + \lambda_7 bwt_j s_{2ij} + \lambda_8 bwt_j s_{3ij} + \lambda_9 bwt_j s_{4ij}\end{aligned} \tag{3.23}$$

Figure 3.17 shows plots of the within-subject variance against time for the two models.

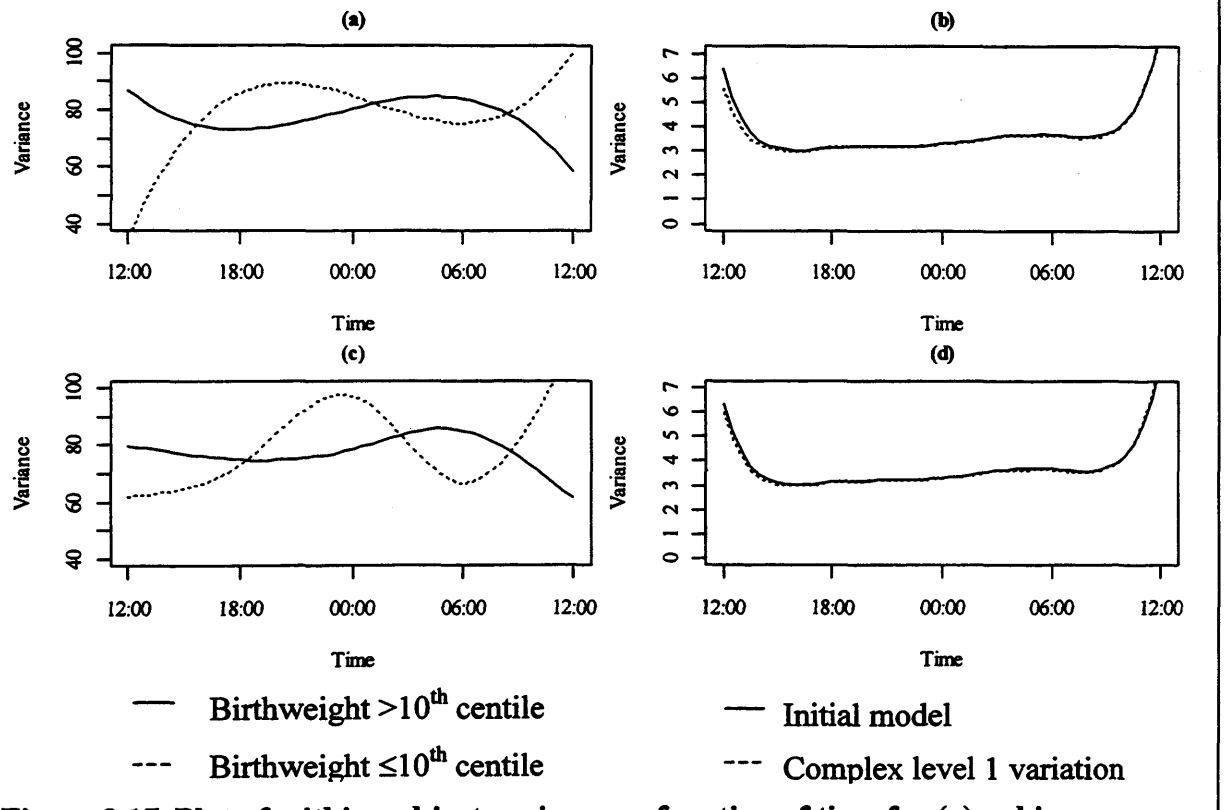

Figure 3.17 Plot of within-subject variance as function of time for (a) cubic polynomial model and (c) 5 knot model. Plot of variance of mean profile for those women who subsequently had an infant >10th centile for (b) cubic polynomial model and (d) 5 knot model compared to initial model.

Also shown is a plot of the variance of the mean profile for those women who had an infant $\leq 10^{th}$ centile for gestational age. The within-subject variance plots show similar patterns for the two models, but with the cubic model giving a slightly flatter profile. Although statistically significant (Likelihood ratio test, cubic polynomial model $\chi^2_7 = 18.0$, P=0.012, 5 knot model $\chi^2_9 = 18.5$, P=0.030), it is difficult to put a clinically meaningful interpretation to the two patterns. However, importantly neither the mean profiles (not shown) or the variances of the mean profiles (Figure 3.17 (b) and (d)) altered to any considerable degree.

3.5.8 Model Checking

I have demonstrated in the previous sections that provided the mean profiles, between-subject random effects and within-subject variances are modelled sensibly the conclusions will not differ qualitatively. However, it is also important to investigate the assumption of normality of the residuals (both within and between-subject residuals) and to investigate whether there are any potential influential or outlying observations. Failure to meet these assumptions could lead to incorrect inferences being drawn.

Figure 3.18 shows histograms and normal probability plots of both the between-subject and within-subject residuals. It appears that the assumption of normality is valid for the within-subject residuals. For each subject there are four between-subject (level 2) residuals (the α_j's). Investigation of the assumption of normality at level 2 is more complicated than at level 1. This is because the between-subject residuals are not actually observed, but as seen in section 2.5.1, estimated from the model and the variance-covariance matrix. However, it is still possible to investigate the assumption of normality on the shrunken residuals. The between-subject residuals are correlated since it is assumed that they have a multivariate normal distribution with q variates (q=4 in this case). However, it has been suggested that in most cases it is sufficient to check for univariate normality for each of the q variates or sometimes to check all the bivariate pairs for bivariate normality (Morrison, 1990). If a formal test of normality is required it has been suggested that use of the Shapiro Wilks W statistic for univariate normality on each of the q variates is often adequate as the test

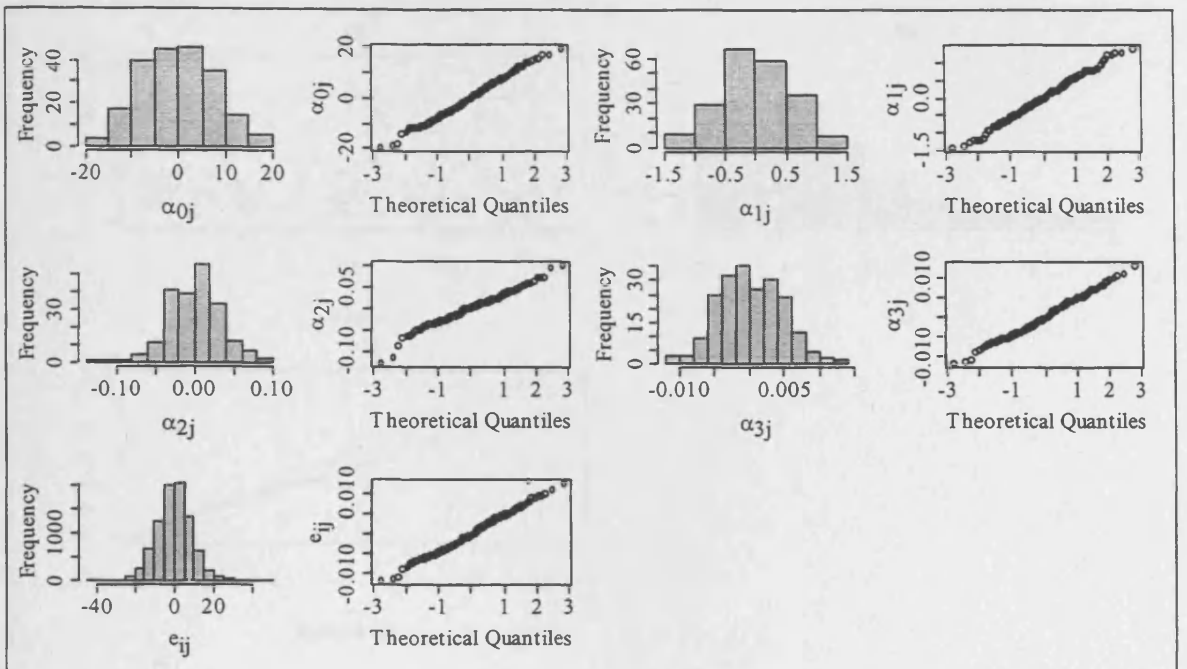

Figure 3.18 Histograms and Normal probability plots for between-subject (α_{0j}, α_{1j}, α_{2j}, α_{3j}) and within-subject residuals (e_{ij}).

statistic generally has low correlation even in the presence of high correlation between the q variates (Royston, 1983). The histograms and normal probability plots indicate that the assumption of univariate normality is valid for all four random effects.

A method of assessing multivariate normality is to obtain the Mahalanobis distance for each subject's level 2 residuals (Bryk and Raudenbush, 1992). It can also be used for investigation of outliers (Morrison, 1990). The Mahalinobis distance is obtained as follows,

$$D_j^2 = u_j' \Sigma^{-1} u_j \qquad (3.24)$$

where u_j is a vector of the subject level residuals for the j^{th} subject and Σ is the variance matrix for the between-subject random coefficients.

Each D_j^2 provides a summary of the degree of departure from multivariate normality of the random effects for each subject. The D_j^2's have a large sample χ^2 distribution with q degrees of freedom, where q is the number of random effects (four in this case). Since the residuals are shrunken, it is expected that in practice they will be less dispersed than a χ_q^2 distribution. The distribution of the Mahalanobis distances for the initial model can be seen

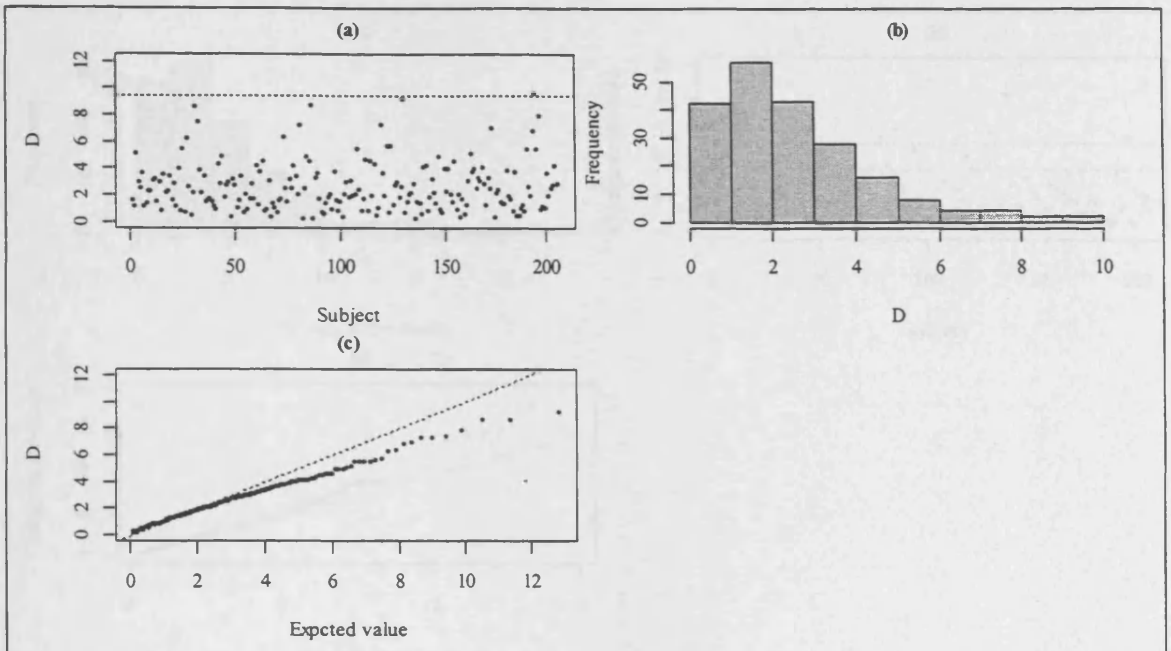

Figure 3.19 Plots of Mahalanobis Distance: (a) against subject id, (b) histogram and (c) quantile plot.

in Figure 3.19 together with a plot vs subject id and a quantile plot. This first plot has a line added to represent significance at the 5% level. It can be seen that only one value exceeds this value indicating that the assumption of normality is valid. The quantile plot shows the ordered, observed Mahalanobis distances versus the expected quantiles from a χ^2 distribution with 4 degrees of freedom. The plot appears to show a fairly straight line again indicating that the assumption of multivariate normality is valid, but shows some departure at higher values due to the use of shrunken residuals.

Recent work in the detection of outliers in multilevel models (Langford and Lewis, 1998) suggests using the deviance to assess whether a subject appears to belong to the multivariate normal distribution describing the between-subject variability. The approach is simple in that a single subject is removed from the random component of the model and the associated effects are included as parameters in the fixed part of the model. Thus if starting with the initial model, where a cubic polynomial was used to model the between-subject variation then one subject can be removed from the fixed part as follows

$$BP_{ij} = \sum_{k=0}^{8} \beta_k s_{kij} + \sum_{n=0}^{3} \delta_n bwt_j \times t_{ij}^n + \sum_{l=0}^{3} \phi_l h_j t_{ij}^l + \sum_{m=0}^{3} \alpha_{mj} (1-h_j) t_{ij}^m + e_{ij} \qquad (3.25)$$

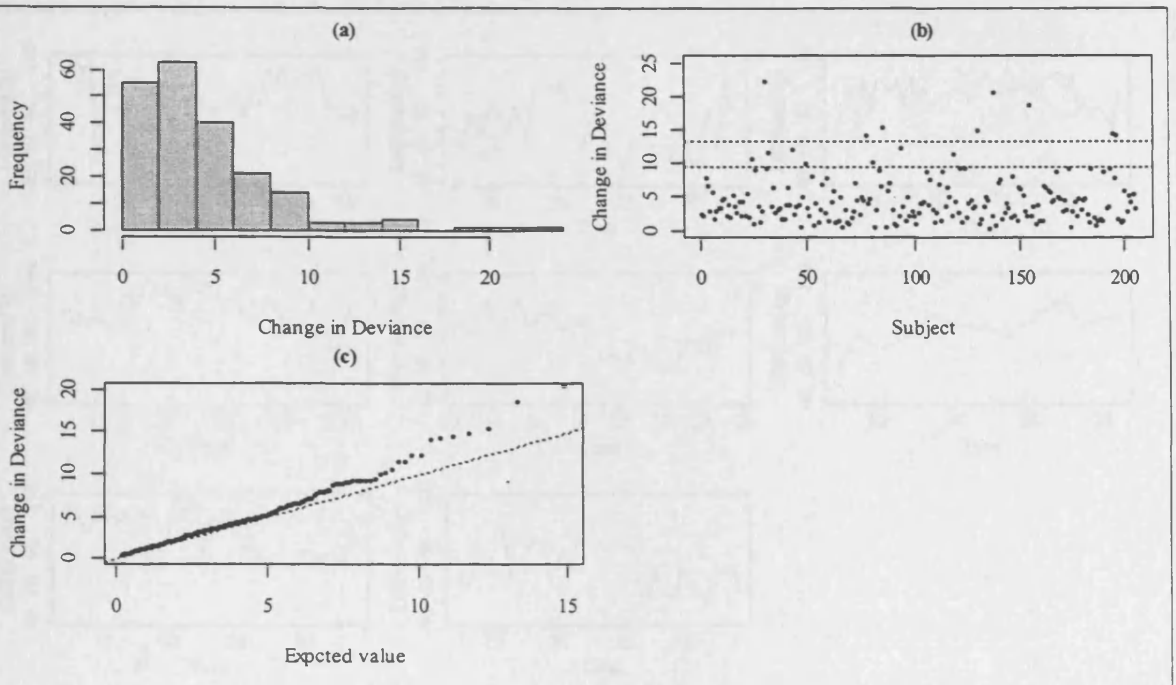

Figure 3.20 Change in Deviance plots: (a) Histogram, (b) vs subject id, (c) quantile plot.

where h_j=1 for the subject of interest and 0 otherwise. This model can then be compared with the initial model by investigation of the change in deviance. This can be repeated for all subjects. Thus, in the case of the dataset used here there will be 206 change in deviance values to be evaluated. A histogram of the change in deviances can be seen in Figure 3.20 together with a plot against subject number. The plot shows the distribution of the change in deviance for the 206 subjects. In the plot vs subject number, lines showing the 5% and 1% significance levels have been added. Given the multiple testing I will only investigate those subjects formally significant at the 1% level. This comprises of 8 subjects whose raw data is shown in Figure 3.21. There appears nothing particularly unusual about these eight subjects and removing them from the analysis makes very little difference to the parameter estimates or the mean profiles. Also shown is a quantile plot for the change in deviance. The line appears relatively straight, but perhaps shows some departure at the upper end.

The change in deviance approach is both time and computer intensive. In this case, 206 different models had to be fitted to the data. The time taken was about 2 hours using a macro written for MLn. However, it has been shown (Langford and Lewis, 1998) that if

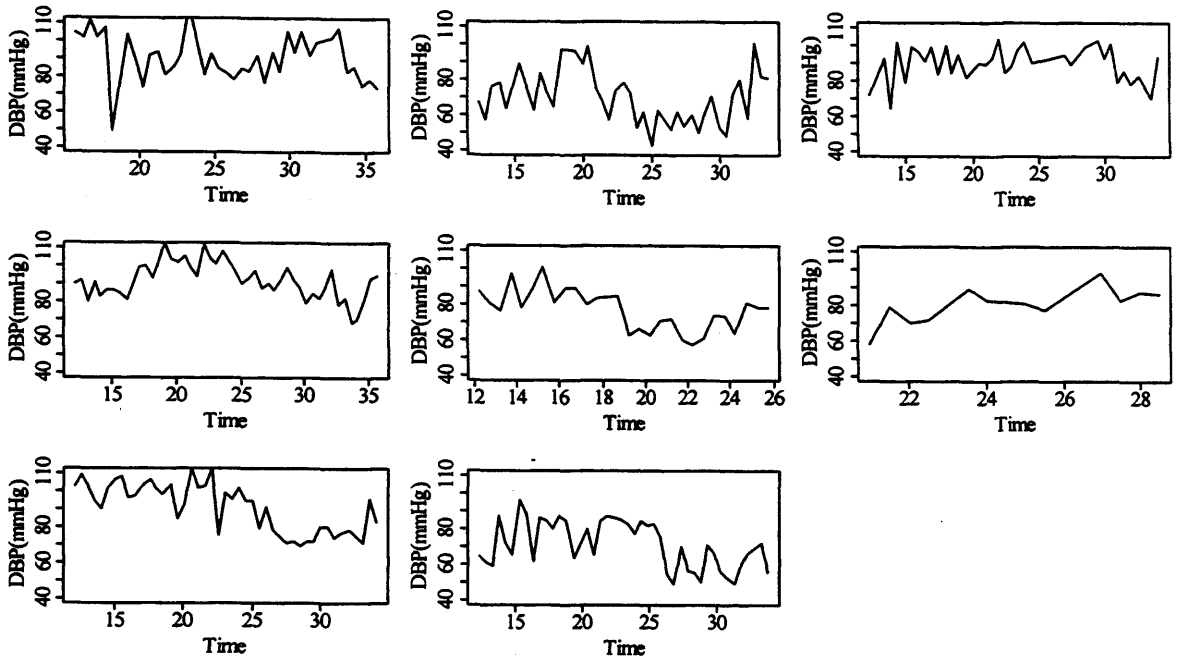

Figure 3.21 Blood pressure profiles for eight subjects identified as potential outliers using change in deviance method.

starting with the converged values in the initial model, then a one-step approximation can be used where only a single iteration is used when fitting each model.

A further assumption is that the within-subject variance is homogeneous across subjects. One way this can be investigated is by calculating the variance (S_j^2) of the within-subject residuals for each of the J subjects. A histogram of these variances is shown in Figure 3.22, which shows a positively skewed distribution. The median variance was 70.7 with a minimum of 24.2 and a maximum of 240.4. A formal test of Level 1 variance heterogeneity (Bryk and Raudenbush, 1992) uses a standardised measure of dispersion for each subject:

$$d_j = \frac{\ln(S_j^2) - \left[\sum (n_j - 1)\ln(S_j^2) / \sum n_j - 1\right]}{\sqrt{2/(n_j - 1)}} \tag{3.26}$$

The test statistic for level 1 variance homogeneity is

$$H = \sum_{i=1}^{J} d_j^2 \tag{3.27}$$

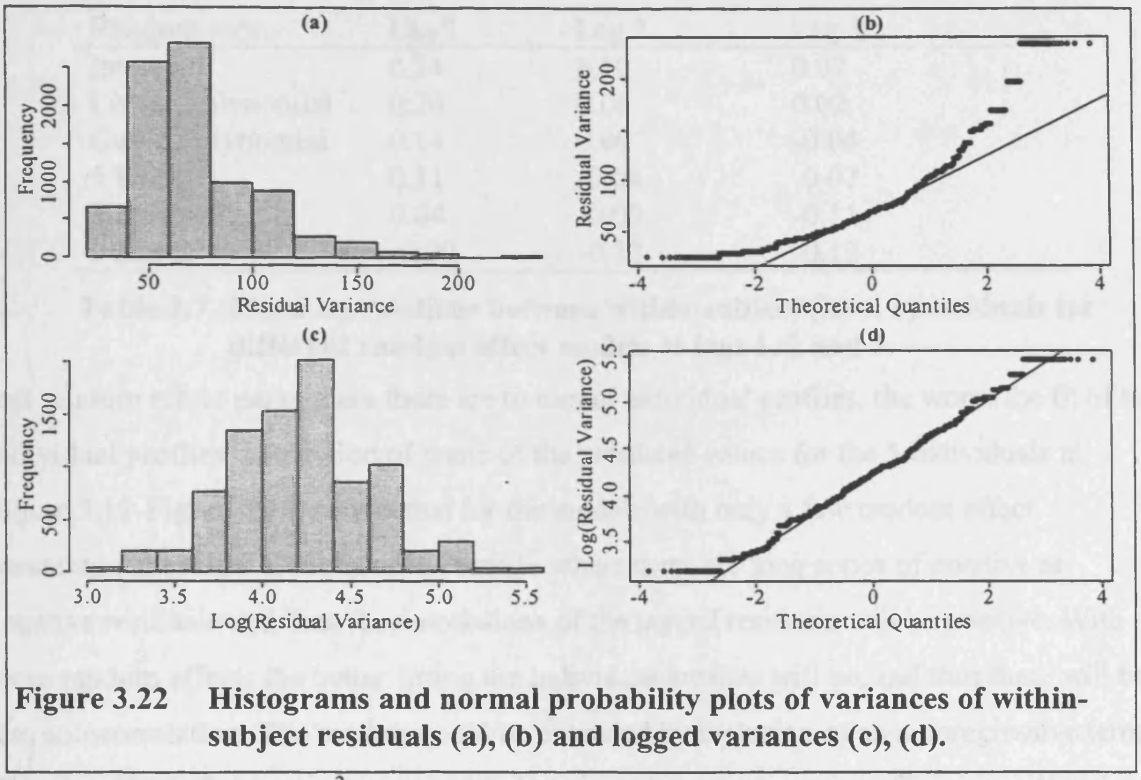

Figure 3.22 Histograms and normal probability plots of variances of within-subject residuals (a), (b) and logged variances (c), (d).

which has a large sample χ^2 distribution with J-1 degrees of freedom. Using (3.26) and (3.27) on the initial model yields H=673.2 with 205 degrees of freedom which gives a P value <0.0001 indicating that heterogeneity of within-subject variances exists.

Another method of formally assessing heterogeneity of the within-subject variances is to fit a separate level 1 variance, σ_j^2, for each subject (Bryk and Raudenbush, 1992). Thus instead of just one parameter being estimated for the level 1 variance, there will be 206 parameters estimated, one for each subject. The two models can then be compared by using the likelihood ratio test. However, this model could not be implemented in either MLn or SAS PROC MIXED, due to memory limitations. In section 3.6 a Bayesian method is developed which allows the within-subject (level 1) variance to vary between subjects.

One final assumption is that the within-subject residuals are uncorrelated. It is possible that the within-subject residuals have an autoregressive structure. In order to investigate this, the correlations between residuals at lags 1, 2 and 3 were calculated for each subject for models with differing number of between-subject random effects.. The mean correlations are shown in Table 3.7. It can be seen that for the models with less between-subject random effects parameters, there appears to be greater autocorrelation. This makes sense since the

Random terms	Lag 1	Lag 2	Lag 3
Intercept	0.24	0.10	0.07
Linear Polynomial	0.20	0.06	0.02
Cubic Polynomial	0.14	0.00	-0.04
5 Knots	0.11	-0.04	-0.07
7 Knots	0.04	-0.09	-0.11
9 Knots	-0.00	-0.13	-0.12

Table 3.7 Mean correlations between within-subject (level 1) residuals for different random effect models at lags 1, 2 and 3.

less random effect parameters there are to model individual profiles, the worse the fit of the individual profiles. Inspection of some of the predicted values for the 5 individuals in Figure 3.12-Figure 3.16, shows that for the models with only a few random effect parameters there are sections of the profile where there are long series of positive or negative residuals and thus the correlations of the lagged residuals will be positive. With more random effects the better fitting the individual profiles will be, and thus there will be less autocorrelation. The models could be extended by inclusion of an autoregressive term (Diggle, 1988; Goldstein *et al.*, 1994), but since the autoregressive effect is small and there are a large number of repeated observations per individual this is unlikely to have a great effect on the estimates of the fixed effects or their standard errors (Wade and Ades, 1998).

In summary, the models I have presented clearly show that there is the expected dipping of blood pressure at night for both groups. There is strong evidence that those women who subsequantely had a baby with a low birthweight for gestational age had higher blood pressure. However, the shape of the two mean profiles was not different between the two groups.

3.6 Bayesian Analysis

3.6.1 Introduction

In this section I consider the use of a Bayesian approach to the analysis of ABPM profiles described above. One reason for this is that it has been suggested that models should be developed using classical methods of estimation, but the final model should be checked using Bayesian estimation as interval estimates are improved (Browne and Draper, 2000).

A further advantage of the Bayesian framework is that it is relatively easy to extend the model to more complex, but more realistic scenarios (Best *et al.*, 1996). In addition, if there was prior information from previous studies about the magnitude of the difference between the two groups, then this could be incorporated through the use of prior distributions. In section 3.5.8, I demonstrated that there was heterogeneity of the within-subject (level 1) variances. I will develop the model so that the within-subject variances are allowed to vary from subject to subject. This is shown in section 3.6.3, but I will first fit the initial model, but now using Bayesian estimation of the parameters.

3.6.2 Initial Model

The initial model used a 9 knot restricted cubic spline for the underlying profile, a cubic polynomial for the difference between the profiles and a cubic polynomial for the between-subject random effects. This can be written in a slightly different but equivalent hierarchical form, as follows:

$$BP_{ij} \sim N(\mu_{ij}, \sigma^2)$$

$$\mu_{ij} = \sum_{k=0}^{8} \beta_k s_{kij} + \sum_{n=0}^{3} \delta_n bwt_j \times t_{ij}^n + \sum_{m=0}^{3} \alpha_{mj} t_{ij}^m \qquad (3.28)$$

where the β's and the δ are the fixed effects and the α's are the random effects. In the initial model it was assumed that the random effects had a multivariate normal distribution with a zero mean vector and variance matrix Σ. When specifying a prior distribution for Σ a Wishart distribution is often used (Carlin, 1996). In Chapter 4 I demonstrate the use a Wishart distribution for this purpose. However, the use of the Wishart distribution is not always intuitive and it is possible to formulate a multivariate normal distribution using a series of linear models (Spiegelhalter *et al.*, 1997; Spiegelhalter, 1998). This is known as the *product normal formulation* of the multivariate normal distribution. For the four between-subject random effect terms, multivariate normality can be expressed as follows

$$\begin{aligned}
\alpha_{0j} &\sim N(0,\sigma^2_{\alpha 0}) \\
\alpha_{1j} &\sim N(\gamma_0\alpha_{0j},\sigma^2_{\alpha 1}) \\
\alpha_{2j} &\sim N(\gamma_1\alpha_{0j}+\gamma_2\alpha_{1j},\sigma^2_{\alpha 2}) \\
\alpha_{3j} &\sim N(\gamma_3\alpha_{0j}+\gamma_4\alpha_{1j}+\gamma_5\alpha_{2j},\sigma^2_{\alpha 3})
\end{aligned} \tag{3.29}$$

The advantage of using this formulation is that instead of using the Wishart distribution as a prior distribution, where a prior is given for the covariance matrix as a whole, univariate prior distributions for $\gamma_0,\cdots,\gamma_5$ and $\sigma^2_{\alpha 1},\cdots,\sigma^2_{\alpha 3}$ can be specified instead. Expressing the relationships between the random effects in this way is also more intuitive. If one considers a simple linear multilevel model where the intercept and gradient vary from subject to

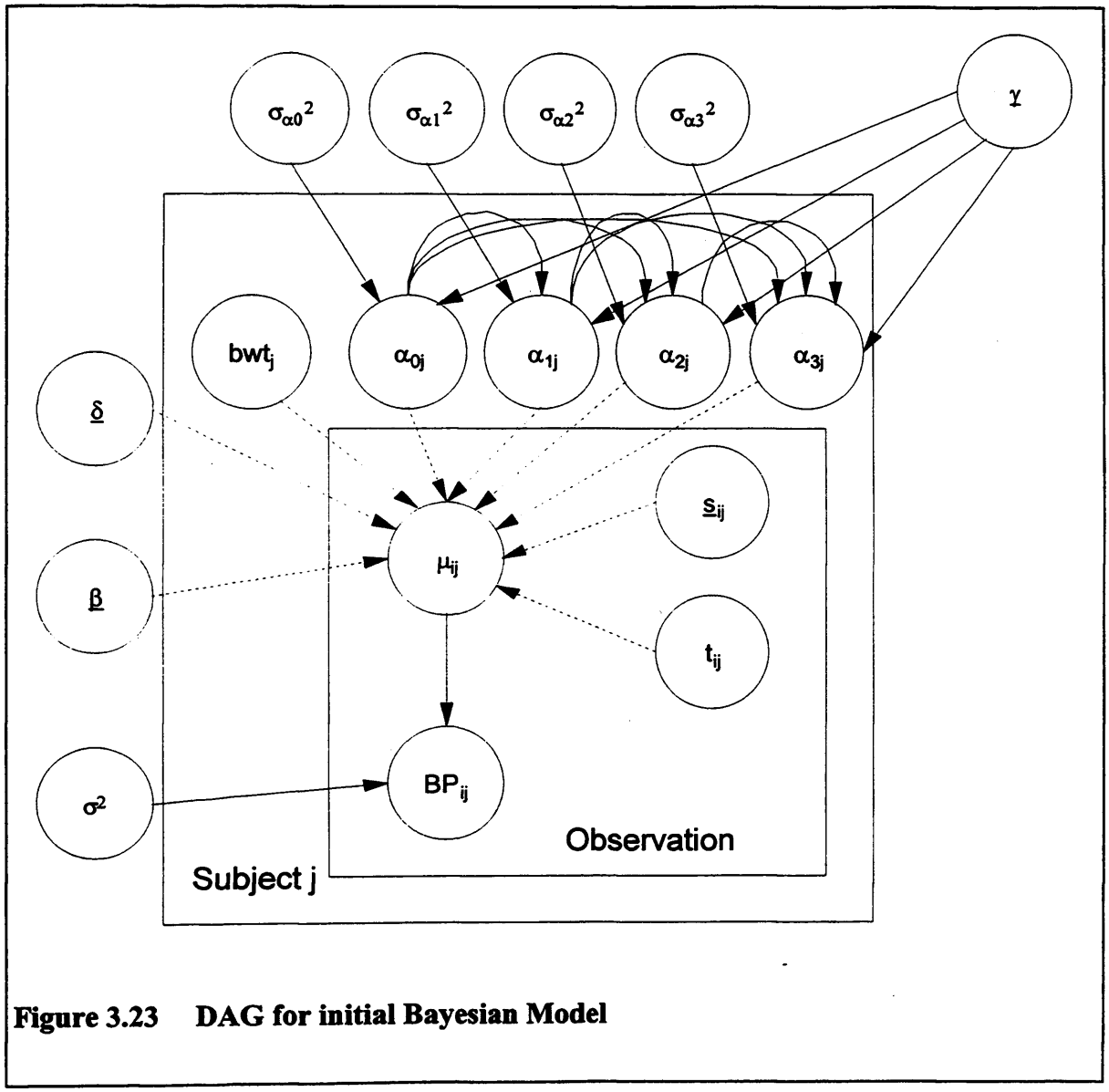

Figure 3.23 DAG for initial Bayesian Model

subject, then using the top two lines in (3.29), it can be seen that the size of the random effect associated with the gradient is related to the intercept through γ_0.

The model can also be expressed as a Directed Acyclic Graph (DAG), which is shown in Figure 3.23. As mentioned in Chapter 2, the DAG is a convenient method of obtaining the factorisation of the joint posterior distribution and, in particular, obtaining the full conditional distribution for each unknown parameter (Best *et al.*, 1996). In the DAG each quantity in the model is shown as a node with the arrows showing direct dependence. Solid arrows represent probabilistic dependencies, while dashed arrows represent functional (deterministic) dependencies. The DAG shows how the mean blood pressure on the i^{th} occasion for the j^{th} subject depends on the fixed effects and the subject level random effects. The interdependencies between the 4 between-subject random effects can be clearly seen.

Since I am now going to adopt a Bayesian approach, prior distributions need to be specified for all unknown parameters. Since I want to compare this model to the classical model I will use non-informative prior distributions. The prior distributions used are

$$\begin{aligned} &\beta's \sim N(0,1000000) \\ &\delta's \sim N(0,1000000) \\ &\gamma's \sim N(0,1000000) \\ &\frac{1}{\sigma_\alpha^2's} \sim Gamma(0.001,0.001) \\ &\frac{1}{\sigma^2} \sim Gamma(0.001,0.001) \end{aligned} \qquad (3.30)$$

Note that it is common in Bayesian models to give the precision (the reciprocal of the variance) a prior rather than the variance.

In order to fit the model WinBUGS was used using overdispersed starting values. A 'burn in' of 5000 iterations was used followed by 50000 samples. This took about eight hours on a Pentium II 400 Mhz PC.

Parameter	Classical Model	Bayesian Model
β_0	76.6 (0.54)	76.7 (0.55)
β_1	-276.6 (16.86)	-276.1 (22.59)
β_2	363.3 (15.71)	373.5 (16.89)
β_3	221.4 (13.85)	221.8 (16.34)
β_4	28.4 (9.24)	33.3 (9.57)
β_5	-118.9 (9.02)	-115.8 (9.11)
β_6	-128.0 (9.00)	-125.3 (9.09)
β_7	-6.3 (9.00)	-3.6 (9.04)
β_8	-73.8 (8.99)	-71.6 (9.02)
δ_0	7.4 (1.90)	7.30 (1.93)
δ_1	0.163 (0.201)	0.162 (0.201)
δ_2	-0.023 (0.014)	-0.022 (0.015)
δ_3	-0.0026 (0.0020)	-0.0026 (0.0024)

Table 3.8 Comparison of fixed effects for the initial Classical and Bayesian Models.

The parameter estimates of the fixed effects for the classical model and the initial Bayesian model are reported in Table 3.8. It can be seen that the parameter estimates are broadly similar. The standard errors are also similar, but tend to be slightly larger for the Bayesian model. However, each parameter has little meaning individually and it is of more interest to investigate the mean profiles in each group. A plot of the predicted values for the classical and Bayesian models together with 95% confidence/credible intervals for each of the two groups is shown in Figure 3.24. The fitted values and 95% credible intervals for the Bayesian model appear to be very similar to the fitted values and 95% confidence intervals for the Classical model. This is perhaps not surprising, as the models are essentially identical and as I am using non-informative priors virtually all the information in the Bayesian model is in the likelihood. If there were less subjects then perhaps one would expect to see differences between the two models as when estimating the fixed effects and their standard errors, the classical model assumes that the random effect variances and covariances are *known* while the Bayesian model treats them as unknown, and appropriately takes the uncertainty associated with them into account. The main difference between the two methods of estimation are in the first and last two to three hours where the Bayesian model has slightly wider intervals. This can also be seen in the plot of the

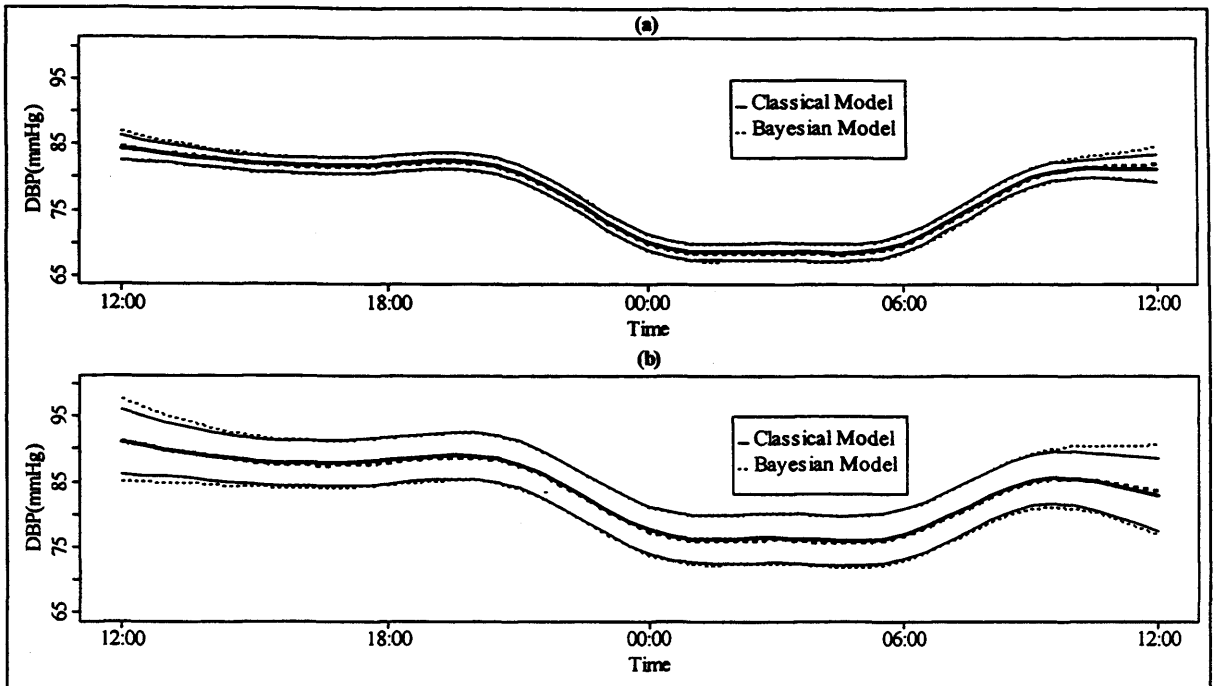

Figure 3.24 Mean profiles with 95% confidence/credible intervals of Classical and Bayesian models for (a) birth weight ≤ 10th centile and (b) birth weight > 10th centile.

difference between the two mean profiles in Figure 3.25(a). The reason for this is that this is where there is less data and hence more uncertainty associated with the random effects.

3.6.3 Heterogeneity of Within- Subject Variation.

The main advantage of the Bayesian approach is its flexibility. Since the within-subject variance appeared not to be constant I will now show how this variation can be incorporated into the Bayesian model. Extending (3.28) gives

$$BP_{ij} \sim N(\mu_{ij}, \sigma_j^2) \tag{3.31}$$

$$\mu_{ij} = \sum_{k=0}^{8} \beta_k s_{kij} + \sum_{n=0}^{3} \delta_n bwt_j \times t_j^n + \sum_{m=0}^{4} \alpha_{mj} t_{ij}^n$$

A subscript j term has been added to the within-subject variance so that now instead of estimating one within-subject variance (or precision), there is a separate within-subject variance for every subject. I will fit three different models for the within-subject variation and compare them to the initial Bayesian model. The first model considers each within-subject variance as a *separate* parameter, so 206 within-subject variances need to be

estimated, one for each subject. A prior is needed for all 206 within-subject variances and can take the following vague form

$$\sigma_j^2 \sim Gamma(0.001,0.001) \tag{3.32}$$

This is the Bayesian equivalent of the model described in section 3.5.8, which can be used to test whether there is evidence of heterogeneous within-subject variation. However, it was not possible to fit these models using Classical methodology.

The above model is probably over parameterised and it may be more sensible to allow the within-subject variance to vary from subject to subject by treating it as a random effect and choosing an appropriate distribution. Figure 3.22 shows that the observed within-subject variances appear to have a skewed distribution, but using a log transformation appears to show approximate normality. Thus, in the next model the within-subject variances are modelled in the following way

$$\log(\sigma_j^2) = \theta_j$$
$$\theta_j \sim N(\mu_\theta, \sigma_\theta^2) \tag{3.33}$$

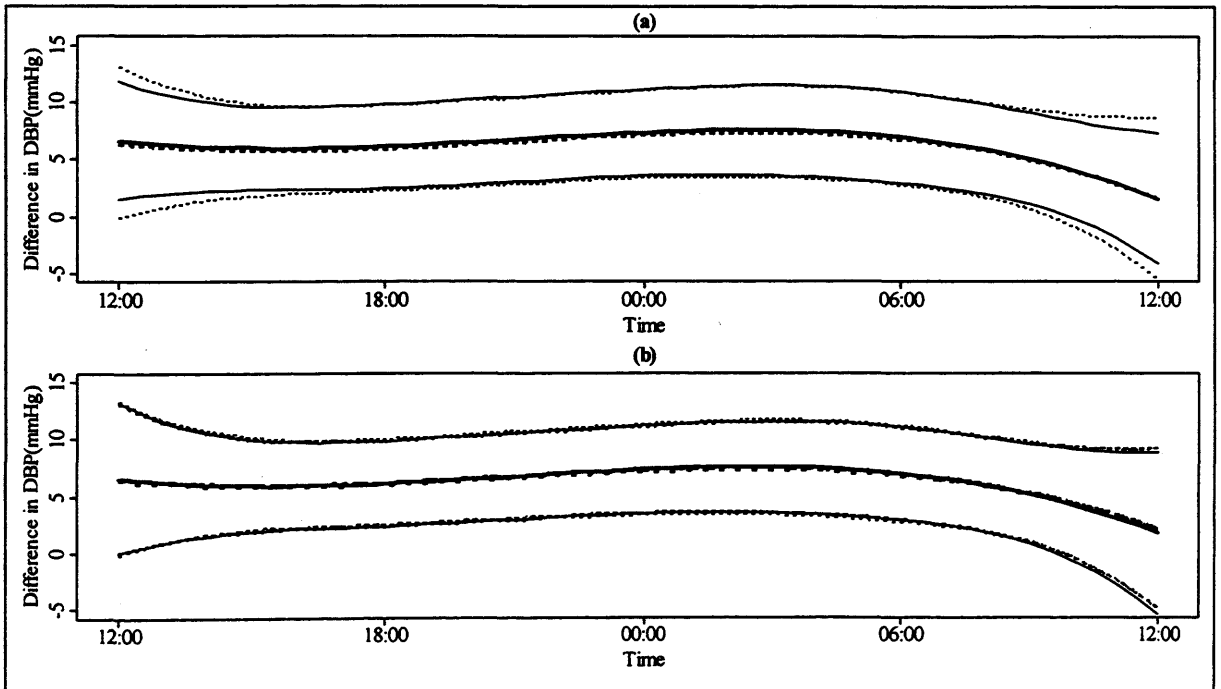

Figure 3.25 Differences between mean profiles for (a) Classical and Bayesian models and (b) the four Bayesian models.

Priors distributions are needed for μ_θ and σ_θ^2 and as with the rest of the prior distributions of the model are relatively non-informative

$$\mu_\theta \sim N(0,1000000) \tag{3.34}$$

$$\frac{1}{\sigma_\theta^2} \sim Gamma(0.001,0.001)$$

It may be preferable not to transform the variances, so the final Bayesian model considers using a Gamma distribution to model the within-subject variances as follows

$$\sigma_j^2 \sim Gamma(a,b) \tag{3.35}$$

Again relatively non-informative prior distributions are used, i.e.

$$\begin{aligned} a &\sim Gamma(0.001,0.001) \\ b &\sim Gamma(0.001,0.001) \end{aligned} \tag{3.36}$$

As with the initial Bayesian model a 'burn in' of 5000 iterations was used for the three extended models with a further 50000 iterations used to form a sample. These models took longer to run than the initial Bayesian model, with the model using the Gamma distribution to model the variation in the within-subject variance taking about 14 hours. The estimates of the fixed effects for the four Bayesian models can be seen in Table 3.9. It can be seen that there is little difference between the four models in either the parameter estimates or

Parameter	Initial Model	Separate Variances	Log-Normal Variances	Gamma Variances
β_0	76.7 (0.55)	76.7 (0.55)	76.6 (0.55)	76.6 (0.55)
β_1	-276.1 (22.59)	-275.4 (22.39)	-275.3(22.48)	-275.7(22.47)
β_2	373.5 (16.89)	373.6 (16.35)	373.3 (16.64)	373.0 (16.55)
β_3	221.8 (16.34)	220.5 (15.86)	220.1 (15.77)	219.7 (16.06)
β_4	33.3 (9.57)	28.4 (9.37)	29.9 (9.44)	29.8 (9.44)
β_5	-115.8 (9.11)	-115.1 (8.79)	-116.2 (8.83)	-115.9 (8.85)
β_6	-125.3 (9.09)	-127.0 (8.84)	-126.8 (8.85)	-126.6 (8.82)
β_7	-3.6 (9.04)	3.0 (8.69)	0.8 (8.83)	0.9 (8.79)
β_8	-71.6 (9.02)	-76.2 (8.69)	-75.2 (8.79)	-74.9 (8.78)
δ_0	7.30 (1.93)	7.22 (1.94)	7.43 (1.96)	7.40 (1.90)
δ_1	0.162 (0.201)	0.142 (0.199)	0.148 (0.200)	0.142 (0.198)
δ_2	-0.022 (0.015)	-0.020 (0.015)	-0.021 (0.015)	-0.021 (0.015)
δ_3	-0.0026 (0.0024)	-0.0022 (0.0024)	-0.0024 (0.0024)	-0.0023 (0.0024)

Table 3.9 Comparison of fixed effect estimates (standard errors) from the four Bayesian Models.

the standard errors. However, it is again more sensible to compare the predicted values. Figure 3.25 shows the mean difference between the two profiles for the four Bayesian models. There is very little difference between the four models both in terms of the mean difference and 95% credible intervals which almost completely overlap.

Figure 3.26 shows a histogram of the observed within-subject variances obtained using the mean values from the Bayesian model estimating a separate variance for each subject. The posterior densities for the two models where the within-subject variance was considered to be a random effect has been superimposed. Both models appear to fit the observed distribution adequately.

3.6.4 Assessment of Convergence

Values of the Geweke test statistic and autocorrelations at lag 10 can be seen in Table 3.10. There are three parameters that give values greater than two for the Geweke statistic, β_0, . β_3, and β_9. It is interesting to note that these are the three parameters with the highest

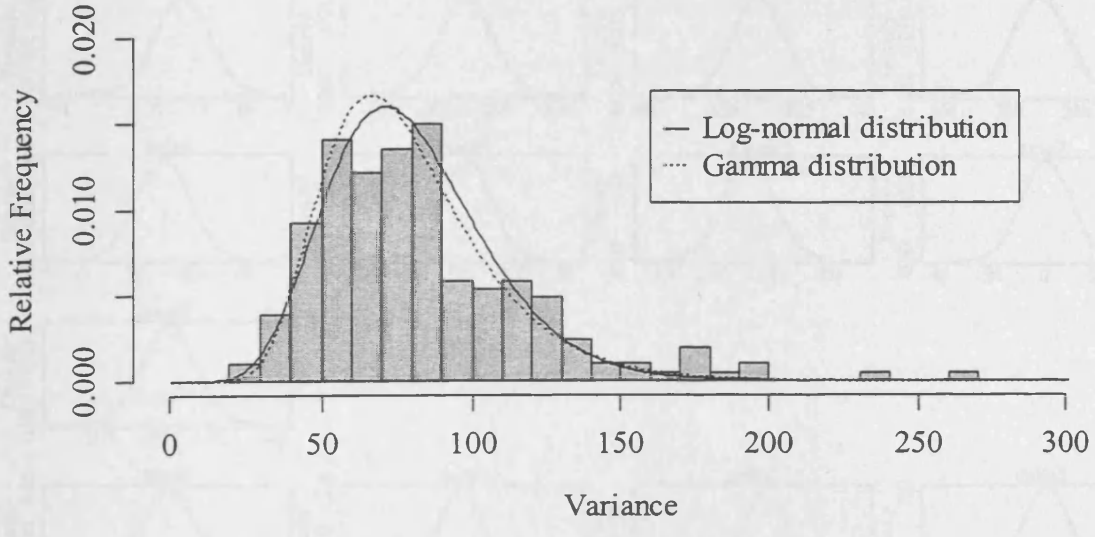

Figure 3.26 Distribution of within-subject variances when each modelled separately together with fitted densities for models where the within-subject variance was modelled using a Log-normal distribution and a Gamma distribution.

autocorrelations. It is well known that high autocorrelations can cause problems in terms of slower convergence. With high autocorrelation it is more likely to have high or low values over a section of the chain leading to a greater chance that the sections will differ.

A problem with the Geweke statistic is that it is only a comparison of the means in the two sections of the chain. In addition it is sometimes more useful to explore the data graphically in order to assess visually how different the two sections are. Figure 3.27 shows density plots for the first 10% and last 50% of the chain for the fixed effects. These provide a graphical alternative to the Geweke statistic and can show potential discrepancies for both the locations and shape of the two sections of the chain. It can be seen that for most of the parameters there is almost complete overlap. For the parameters with high values for the Geweke statistic, i.e. β_0, . β_3, and β_9, although one can see a slight difference one can see that the differences in densities are very small and unlikely have a great deal of influence.

3.7 Discussion

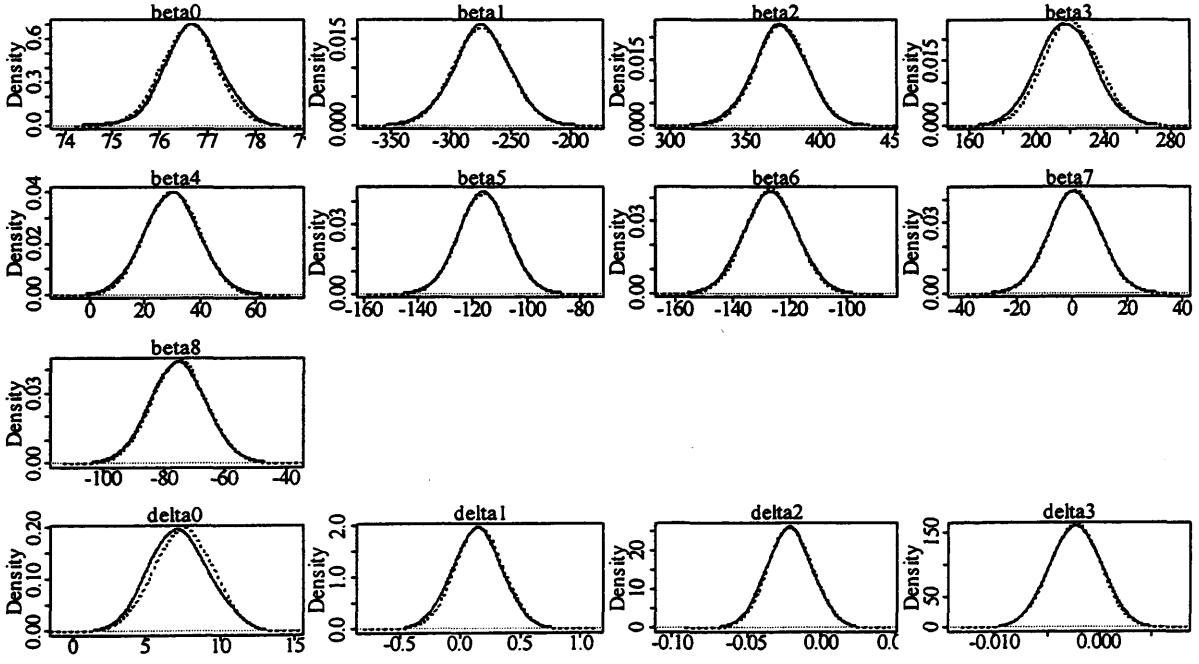

Figure 3.27 Density plots for first 10% (solid line) and last 50% (dotted line) of chains for each fixed effect parameter.

Fixed Effect Parameters			Random Effect Parameters		
	Geweke Z value	Autocorrelation at lag 10		Geweke Z value	Autocorrelation at lag 10
β_0	2.13	0.74	γ_0	-0.27	0.01
β_1	0.54	0.17	γ_1	0.42	0.00
β_2	-1.30	0.09	γ_2	0.82	0.02
β_3	-2.81	0.31	γ_3	0.20	0.00
β_4	0.02	0.01	γ_4	1.20	0.01
β_5	0.42	0.01	γ_5	1.18	0.00
β_6	1.12	0.00	$\sigma^2_{\alpha 0}$	1.07	0.00
β_7	2.07	0.00	$\sigma^2_{\alpha 1}$	-1.11	0.00
β_8	-1.44	0.01	$\sigma^2_{\alpha 2}$	-0.19	0.01
δ_0	-2.05	0.68	$\sigma^2_{\alpha 3}$	0.10	0.00
δ_1	-1.63	0.30	μ_δ	0.04	0.00
δ_2	-0.89	0.06	σ^2_δ	0.36	0.00
δ_3	1.10	0.29			

Table 3.10 Z scores for Geweke assessment of non-convergence and autocorrelations at lag 10.

In this chapter I have demonstrated how repeated ambulatory blood pressures can be modelled using a two level hierarchical model. Use of the two level model allows modelling of mean profiles, shrunken estimates of individual profiles and within-subject variation. Use of Bayesian estimation makes little practical difference to the fitted values or the standard errors of the mean profiles when fitting the same model as in the classical analysis. The Bayesian approach allows greater flexibility in the modelling of the within-subject variance. However, modelling the within-subject variance heterogeneity also makes little difference to the fitted values and standard errors.

The use of restricted cubic splines appears a very useful tool for curve fitting, not necessarily just for hierarchical models. They are relatively simple to use and it takes very little time to fit a model, and can be useful for exploratory purposes when investigating the functional form of a relationship. In the example presented here, it appears that as long as there are a sufficient number of knots, the choice of the exact number is not that crucial in terms of model inferences. However, it would seem sensible to recommend that sensitivity to the choice of the number and location of knots should, at least briefly, be carried out. An

alternative would be to treat the number and location of the knots as unknowns and estimate them in a model. An example of this for non-correlated data from a Bayesian perspective can be seen in Denison *et al.* (1998). However, in this case, given that the model appears robust to changes in the number of knots, this would probably add an unacceptable amount of time to fitting the Bayesian model, which at present takes about 10 hours on a Pentium II 400 Mhz.

The choice of the between-subject random effect parameters will depend on the research question. In most cases the main interest is in the mean profiles. However, in some instances, interest may also lie in modelling individual profiles. When the interest lies in the mean profile I have shown that the choice of random effects variables makes little difference to the results as long as there are sufficient terms to model the between-subject variation sensibly. This concurs with previous work by Taylor and Law (1998) who found that the choice of covariance structure made little practical difference to the mean profiles or their standard errors. However, if one is interested in accurate estimation of individual profiles or prediction of future observations, they found that the choice of covariance structure was much more important. However, since the blood pressure here is recorded over a 24 hour interval prediction of future observations is unlikely to prove that useful.

Using Bayesian estimation procedures made very little difference to the mean profiles or their standard errors. This is not that surprising due to the large number of individuals and observations in the study. However, I would agree with the recommendations of Browne and Draper (2000) that after developing the model using classical methodology one should check parameter estimates and their standard errors using Bayesian estimation. In fact the software program MLWin allows IGLS, RIGLS and MCMC methods of estimation.

One advantage of using a Bayesian analysis is that it is relatively simple to extend the analysis to more complex scenarios. For example, in this case, by incorporating a between-subject random effect for the within-subject variance. Despite strong evidence of within-subject variance heterogeneity there was very little difference in the estimate of the mean profiles. If interest lies in the individual profiles then the modelling of the within-subject heterogeneity will obviously be more crucial (Lin *et al.*, 1997).

The models presented in this chapter could be extended in a number if ways. For example, I have assumed Normality for the response (blood pressures) and the between-subject random effects. Although this is often reasonable it may be possible to explore situations in which the distributions have a different form. For example, it may be possible in some situations to assume that the response has a t-distribution with either pre-specified or estimated degrees of freedom. This allows for the response to have heavier tails than those in a Normal distribution. The use of such methods would be relatively simple using a Bayesian model.

I have assumed vague priors for all parameters in the Bayesian models. In some situations it may be of interest to include more informative prior distributions on the fixed components, the variance components or both. However, since the coefficients associated with the spline variables do not have a sensible interpretation individually , this would be complicated, so it would only realistically be possible to use informative prior distributions for the treatment difference. I have used inverse gamma distributions as prior distributions for the various variance components. Recent work has shown that the use of Pareto distributions for the variance components are often more appropriate than Gamma distributions (Burton *et al.*, 1999). However, the main problem with the use of Gamma distributions is with small variances or when there is only a small amount of data, which is not the situation in the models presented in this chapter.

4 THE ANALYSIS OF PEAK EXPIRATORY FLOW DATA

4.1 Introduction

In this chapter I use hierarchical models to analyse repeated measures of Peak Expiratory Flow (PEF). I extend the two level models of chapter 3 by using a three level model that enables quantification of both the mean level of PEF as well as the within-subject variability of PEF. The latter is broken down into between and within-day variation and has a useful clinical interpretation. I give a brief background to PEF in section 4.2, followed by a description of the data I use in section 4.3. Section 4.4 reviews a number of summary measure techniques for the analysis of repeated measures of PEF. The three level model is described and developed in section 4.5, with the model built up from a simple variance components model in section 4.5.2 through to modelling the mean profiles and between and within-subject variability in section 4.5.5. Section 4.6 investigates the potential use of a Bayesian approach. Finally, section 4.7 discusses the techniques I have used and discuss possible extensions of my work.

4.2 Peak Expiratory Flow

Peak Expiratory Flow is the maximum airflow achieved during a forced expiration from total lung capacity (Ayres and Turpin, 1997). PEF is commonly used in both clinical practice and research. In clinical practice it is often used for monitoring and diagnosing patients with asthma. It is used in the self management of asthma where patients can modify their treatment or seek medical advice accordingly. In clinical research PEF can be used to monitor effectiveness of new treatments in clinical trials (Enright *et al.*, 1994; Toogood *et al.*, 1996). In epidemiological studies it is sometimes used to define asthma cases and to assess severity of asthma (Lebowitz *et al.*, 1987; Toelle *et al.*, 1992).

One of the advantages of PEF is that it is possible for individuals to measure their PEF at home using a *peak flow meter*. Such a device can be seen in Figure 4.1. After the patient has being instructed on how to use the device no clinical supervision is necessary as there are no risks associated with its use. The devices are inexpensive and so measurement of peak flow is relatively cheap. In addition PEF is usually recorded in the form of diaries. In

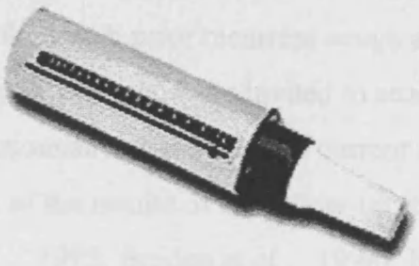

Figure 4.1 Peak flow meter

the research field these are often for a two week period with peak flow recorded twice or more per day.

When investigating a patient's PEF, not only is the actual level of interest, but also the variability in PEF measurements as asthmatics tend to have more variable airways and thus more variable PEF when compared to non-asthmatics. Of particular interest is diurnal variation in PEF (the variation between measurements made on the same day), as this has been shown to be increased in people with asthma (Hetzel and Clark, 1980). In fact some of the current asthma guidelines state that in clinical practice diurnal variability should be calculated when diagnosing asthma and assessing its severity (Bethesda 1995; British Thoratic Society, 1996).

4.3 Description of Data

In 1990 a prevalence study was performed in order to estimate the prevalence of wheeze, doctor diagnosed asthma, and recurrent cough in pre-school children in Leicestershire. This has been reported in detail elsewhere (Luyt *et al.*, 1993). Briefly, the parents of 1650 white caucasian children born between January 1985 and January 1990 were randomly sampled using the Leicestershire Child Health Register as a sampling frame. Parents received a postal questionnaire concerning respiratory symptoms, family history, and environmental/social conditions.

Between March 1992 and March 1994 all children in whom prior wheeze had been reported, together with all children with prior recurrent cough and a random sample of children who had been initially asymptotic, were invited to attend the Leicester Royal Infirmary for a follow-up questionnaire, assessment of current symptoms, and physiological measures. Some of the results of the follow-up study have been reported in detail elsewhere (Brooke *et al.*, 1995; Brooke *et al.*, 1996). In this chapter, I will investigate those who were *current wheezers* in the 1990 study. Current wheeze was defined as wheezing at least once in the year prior to the original study. In addition to the children receiving the follow-up questionnaire, all children who were thought to be able to make repeatable PEF recordings were issued with a mini-Wright peak flow meter and were asked to make recordings at 8:00am (morning measurement), 4:00pm (afternoon measurement) and 8:00pm (evening measurement) for fourteen days. The mini-Wright peak flow meter is one of the most popular devices for measuring PEF and has been shown to be repeatable and correlates well with other measures of lung function (Wright, 1978; Kotses *et al.*, 1984). The importance of adhering to these times was emphasised and parents were asked to omit inaccurately timed readings. Proficiency in using the peak flow meter was checked by a research nurse in a home visit during the two weeks. Completed diaries and meters were returned to the investigators using a freepost system.

The subset of children analysed here consists of 90 children aged between 5 and 8. The atopic status of the children was assessed by skin prick testing (Pepys, 1975). Four aeroallergens: cat hair, dog dander, house dust mite and mixed grass pollen were used. Children who had at least one positive response to any of these four allergens were deemed atopic. The aim of the analysis was to compare PEF for atopic and non-atopic children.

PEF is expressed as percentage predicted (PPEF) using separate predictive equations for boys and girls (Wille and Svensson, 1989), with the equations being a function of age and height. This is standard when analysing PEF, with PPEF being calculated by dividing the observed PEF by the predicted PEF and multiplying by 100. Hence a value greater than 100 indicates that a child had a higher PEF than predicted and a value less than 100 indicates that a child had a PEF less than that predicted. To be included in the analysis children had

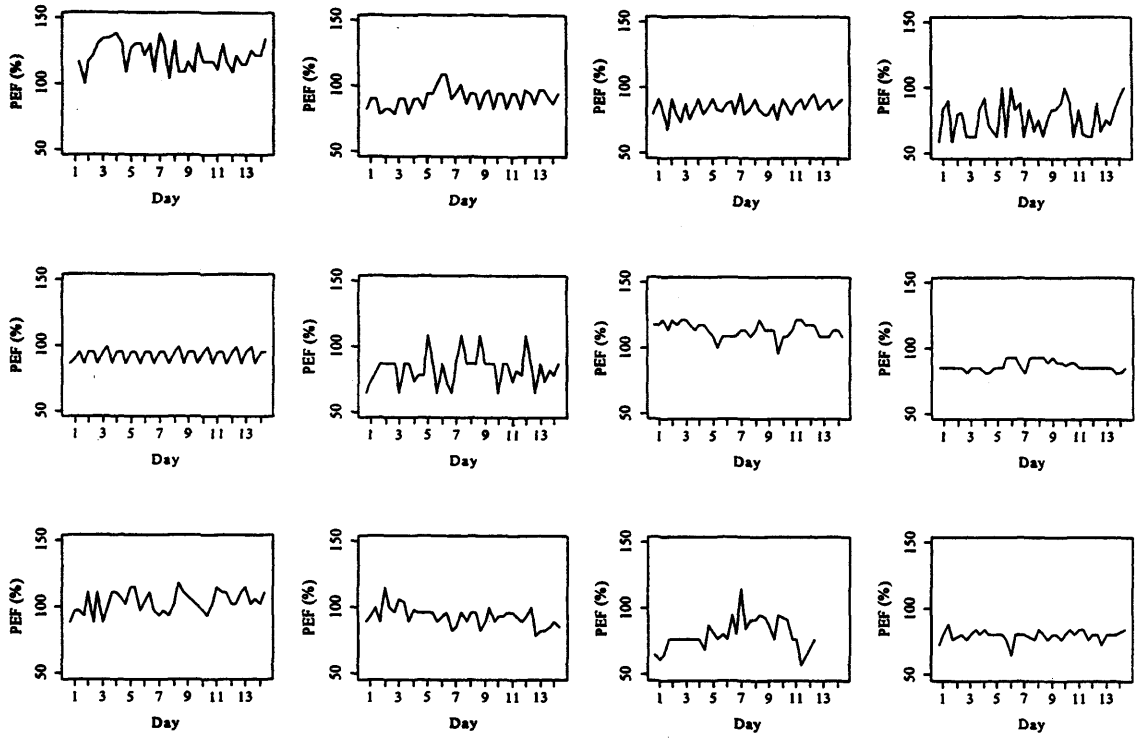

Figure 4.2 Serial plot of PPEF for the first 12 subjects.

to complete at least half of the potential 42 observations in the PEF diary. Of the 90 children, 38 were non-atopic and 52 were atopic with a combined total of 3153 PEF's. A serial plot of the first 10 subjects can be seen in Figure 4.2. These plots demonstrate that there is both, between and within-subject variability. It is of interest to note that within-subject variability appears to vary between subjects

4.4 Standard Approaches to the Analysis of Peak Expiratory Flow

In the research setting, comparison of PEF diaries between groups is of interest. In virtually all cases in the literature a summary measure of the PEF diaries is obtained, with standard procedures such as the t-test used for formal comparison between groups. If one is interested in the level of PEF or PPEF then a mean value can be taken. As it is well known that PEF varies during the day, with lower values in the morning, the values averaged should be measured at similar times of the day.

In the previous section I discussed how not only the level of PPEF, but also the variability of PPEF is of interest. Therefore, a number of summary measures that have been used attempt to measure this. For example, Siersted (1994) uses nine different summary measures for PEF variability. Two of the most commonly used measures are Amplitude percent mean (Amp%Mean) where

$$\text{Amp\%mean} = \text{mean}\left(\frac{\text{daily highest reading} - \text{daily lowest reading}}{\text{daily mean}}\right) \qquad (4.1)$$

And the Standard Deviation percent mean (SD%Mean)

$$\text{SD\%mean} = \frac{\text{SD of PPEF measurements}}{\text{mean of PPEF measurements}} \qquad (4.2)$$

A previous study has shown that Amp%mean is the "best index of separation" between asthmatics and non-asthmatics (Higgins *et al.*, 1992). It has been argued that these two measures are too complex to use in clinical practice as they take too long to calculate in a standard medical consultation and that an even simpler measure should be used, namely 'Lowest PPEF as % of personal best' (Low%best) (Reddel *et al.*, 1999).

$$\text{Low\%best} = \frac{\text{Lowest PPEF}}{\text{Highest PPEF}} \times 100 \qquad (4.3)$$

These three PPEF variability summary measures were obtained using the data described in section 4.3. In addition the mean of the morning measures (AM), the afternoon measures (PM) and the evening measures (EVE) was obtained for each child. Children who had fewer than half (seven) measurements when obtaining this means were excluded from the analysis.

The results of obtaining these six summary measures and a formal comparison of atopic and non-atopic children using the t-test can be seen in Table 4.1. For the analysis of SD % mean, the data was logged as it was positively skewed. It can be seen that the number of children with evening measures of PPEF is less than those measured in the morning. This is because some of the children would be in bed by the time the evening measure was to be made (8:00pm). The table shows that atopic children tend to have lower mean PPEF as percent predicted at all three measurement times. However, only in the morning is the difference formally statistically significant (at the 5% level). For the comparison of PPEF

	Non Atopic		Atopic		Difference in means
	N	Mean (SE)	N	Mean (SE)	(95% CI)
Morning measure	52	97.5 (2.3)	37	90.2 (2.6)	7.3 (0.2 to 14.3)
Afternoon measure	52	101.4 (2.3)	38	95.4 (2.4)	6.0 (-0.7 to 12.7)
Evening measure	43	99.2 (2.5)	31	94.1 (2.5)	5.0 (-2.2 to 12.2)
Amp % mean	52	9.5 (0.62)	38	12.3 (1.08)	-2.7 (-5.1 to –0.4)
Low % best	52	71.7 (1.3)	38	66.9 (1.5)	4.7 (0.7 to 8.8)
Log SD % mean	52	2.02 (0.05)	38	2.22 (0.07)	-0.20 (-0.37 to –0.02)

Table 4.1 Comparison of atopic and non-atopic children using summary measures.

variability, it can be seen that atopic children appear to have greater variability for all three measures with all being significant at the 5% level.

There are a number of potential problems with the use of these summary measures. Firstly, there is the problem of missing data as there are a number of children who do not have complete PEF diaries. For example, in the comparison of PPEF in the morning while a number of children will have recorded all 14 measures, some children will have fewer recordings. Strictly, one should apply weights, as the children with all measurements should contributes more information to the analysis as they will have smaller variances (Matthews, 1993). However, it has been argued that in doing this the attractiveness and the simplicity of summary measurements has been lost and one could therefore use a more powerful method of analysis that does not reduce the data to summary measures (Hand and Crowder, 1996). Another problem is in the choice of which summary measures to use. I have demonstrated six here, but there are numerous others that could have been used. This is especially so for the summary measures estimating PEF variability. The three measures here measure different aspects of variability with Amp%mean measuring mean diurnal variation, SD%mean measuring the variability of all PEF recordings and Low%best investigating the extremes which makes it potentially highly influenced by outlying observations.

4.5 A Three Level Model

4.5.1 Introduction

As I discussed in Chapter 2, with repeated measures data one can normally consider there to be a two level hierarchical structure, with individual measurements nested within subjects. In the case of the PEF data presented here, the PPEF measurements (level 1) are nested within children (level 2). One can also consider the data as having a three level structure and there to be variation at each of these three levels. Firstly, there will be *between-subject* (or in this case between-child) variability in that one would expect PEF's recorded on the same subject to be more similar than PEF's recorded on different subjects (i.e. some children would have consistently higher PEF when compared to other children). There will also be *within-subject* variation. However, due to the way the data was recorded it is possible to break down the within-subject variation into two components. Firstly, one can expect there to be variation on a day to day basis, i.e. *between-day within-subject* variation. In addition there will be variation in PPEF within each day. This is the residual variation but can also be thought of as *within-day within-subject* variation. The three level structure can be seen graphically in Figure 4.3. It can be seen that individual PPEF observations (level 1) are nested within days (level 2) which are nested within subjects (level 3).

There are two reasons for analysing the data as a three level hierarchical model. Firstly, it is of clinical interest to break down the variation in this way as clinicians are generally interested in both aspects of within-subject variability, but mainly concentrate on within-

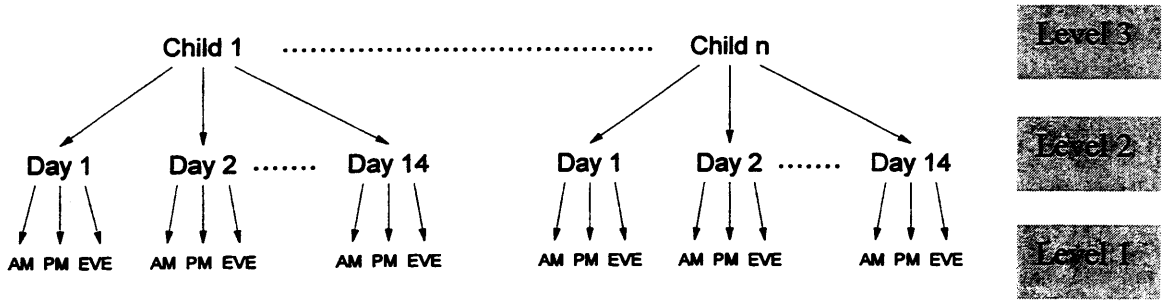

Figure 4.3 Hierarchical structure of the asthma dataset

day variation (Reddel *et al.*, 1999). The second reason for adopting the three level model is that it will induce a slightly more complex correlation structure which may be more realistic. Not only are observations made on the same subject correlated, but by including day as a level in the model, the PPEF recorded on a particular day will be more similar than those recorded on different days for the same subject. This could allow for factors affecting specific days, for example pollution, pollen or humidity levels.

4.5.2 Variance Components Model

The model will have the response percent predicted peak expiratory flow ($PPEF_{ijk}$) which denotes the i^{th} observation (i=1,...,3) on the j^{th} day (j=1,...,14) for the k^{th} subject (k=1,...,90) and will consist of both fixed effects and random effects as follows.

$$PPEF_{ijk} = Fixed\ Effects + Random\ Effects \tag{4.4}$$

The fixed effects will model the mean level of PPEF and the random effects will model the variability of PPEF at each of the three levels. I will start with a simple model, namely a variance components model as follows,

$$PPEF_{ijk} = \beta_0 + \alpha_{0k} + \delta_{0jk} + e_{ijk}$$

$$\alpha_{0k} \sim N(0, \sigma^2_{\alpha_0}) \quad \text{(Between - Subject Variation)} \tag{4.5}$$

$$\delta_{0jk} \sim N(0, \sigma^2_{\delta_0}) \quad \text{(Between - Day Within - Subject Variation)}$$

$$e_{ijk} \sim N(0, \sigma^2) \quad \text{(Within - Day Within - Subject Variation)}$$

The model has one fixed effect (β_0) which is the overall mean PPEF, and three random effect parameters which need to be estimated. The nested random effects give the following covariance structure for PPEF's measured on the same subject.

Day	Time							
1	*AM*	$\sigma^2_{a_0}+\sigma^2_{\delta_0}+\sigma^2$	$\sigma^2_{a_0}+\sigma^2_{\delta_0}$	$\sigma^2_{a_0}+\sigma^2_{\delta_0}$	$\sigma^2_{a_0}$	$\sigma^2_{a_0}$	$\sigma^2_{a_0}$	...
1	*PM*	$\sigma^2_{a_0}+\sigma^2_{\delta_0}$	$\sigma^2_{a_0}+\sigma^2_{\delta_0}+\sigma^2$	$\sigma^2_{a_0}+\sigma^2_{\delta_0}$	$\sigma^2_{a_0}$	$\sigma^2_{a_0}$	$\sigma^2_{a_0}$	...
1	*EVE*	$\sigma^2_{a_0}+\sigma^2_{\delta_0}$	$\sigma^2_{a_0}+\sigma^2_{\delta_0}$	$\sigma^2_{a_0}+\sigma^2_{\delta_0}+\sigma^2$	$\sigma^2_{a_0}$	$\sigma^2_{a_0}$	$\sigma^2_{a_0}$	...
2	*AM*	$\sigma^2_{a_0}$	$\sigma^2_{a_0}$	$\sigma^2_{a_0}$	$\sigma^2_{a_0}+\sigma^2_{\delta_0}+\sigma^2$	$\sigma^2_{a_0}+\sigma^2_{\delta_0}$	$\sigma^2_{a_0}+\sigma^2_{\delta_0}$	...
2	*PM*	$\sigma^2_{a_0}$	$\sigma^2_{a_0}$	$\sigma^2_{a_0}$	$\sigma^2_{a_0}+\sigma^2_{\delta_0}$	$\sigma^2_{a_0}+\sigma^2_{\delta_0}+\sigma^2$	$\sigma^2_{a_0}+\sigma^2_{\delta_0}$	...
2	*EVE*	$\sigma^2_{a_0}$	$\sigma^2_{a_0}$	$\sigma^2_{a_0}$	$\sigma^2_{a_0}+\sigma^2_{\delta_0}$	$\sigma^2_{a_0}+\sigma^2_{\delta_0}$	$\sigma^2_{a_0}+\sigma^2_{\delta_0}+\sigma^2$	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱
14	*EVE*							

(4.6)

The matrix shows the variances and covariances of the first six PPEF's measured on the first two days. The full matrix for an individual will be of dimensions 42x42.

This model was fitted in MLn using IGLS with the results presented in Table 4.2. It can be seen that the overall mean PPEF is 96.9%. The reason why it less than 100 could be due to the fact that the population of children have previously been defined as wheezers. However, it is well known that populations can differ from those from which the predictive equations were obtained and so absolute values should be treated with a degree of caution. It can be seen that the greatest amount of variation is between subjects, which is sensible as one expects children to have different underlying PEF's. Of the within-subject variation the greater proportion is within-days.

Using the variance component parameters it is possible to measure the correlation between observations made on the subject using the *intra class correlation* (Armitage and Berry, 1987). This can be done by converting the covariance matrix in (4.6) to a correlation matrix. Because of the 3 level structure there will be two such measures or correlation, which are obtained as follows

Parameter	Estimate (Standard Error)
FIXED EFFECTS	
β_0 (Overall mean)	96.9 (1.68)
RANDOM EFFECTS	
$\sigma^2_{a_0}$ (Between-Subject Variation)	250.6 (37.8)
$\sigma^2_{\delta_0}$ (Between-Day Within-Subject Variation)	17.6 (1.8)
σ^2 (Within-Day Within-Subject Variation)	58.5 (1.9)

Table 4.2 Parameter estimates of variance components model

$$\begin{aligned} \text{Intra - Subject Correlation} &= \frac{\sigma^2_{\alpha_0}}{\sigma^2_{\alpha_0} + \sigma^2_{\delta_0} + \sigma^2} = 0.77 \\ \text{Intra - Day Correlation} &= \frac{\sigma^2_{\alpha_0} + \sigma^2_{\delta_0}}{\sigma^2_{\alpha_0} + \sigma^2_{\delta_0} + \sigma^2} = 0.82 \end{aligned} \qquad (4.7)$$

The intra-subject correlation is a measure of the similarity between observations measured on the same subject on *different* days and the intra-day correlation is a measure of the similarity between observations measured on the same subject on the *same* day. By definition the inter-day correlation will be greater than or equal to the intra-subject correlation. They will be the same if the between-day within-subject variance is equal to zero. It can be seen that the standard error of the between-day within-subject variance is small compared to the parameter estimate indicating that it is worthwhile including day as a level in the model. However, as I have previously discussed one must be cautious when using standard errors of variance components, although in this case the estimate of the variance is ten times that of the standard error.

4.5.3 *Including the effect of time of day.*

It is well known that PEF tends to vary during the day, with it generally being lowest in the morning and then increasing during the day, until reaching its peak in late afternoon. Since there are three measurements per day at distinct times (8am, 4pm and 8pm) it seems sensible to incorporate this information in the model. PM and EVE are two dichotomous covariates taking the value 1 if the recording was made in the afternoon or evening respectively and 0 otherwise. Model (4.5) can thus be extended as follows

$$\begin{aligned} PPEF_{ijk} &= \beta_0 + \beta_1 PM_{ijk} + \beta_2 EVE_{ijk} + \alpha_{0k} + \alpha_{1k} PM_{ijk} + \alpha_{2k} EVE_{ijk} + \delta_{0jk} + e_{ijk} \\ \underline{\alpha}_k &\sim MVN(\underline{0}, \Sigma_\alpha) \\ \delta_{0jk} &\sim N(0, \sigma^2_{\delta_0}) \\ e_{ijk} &\sim N(0, \sigma^2) \end{aligned} \qquad (4.8)$$

where

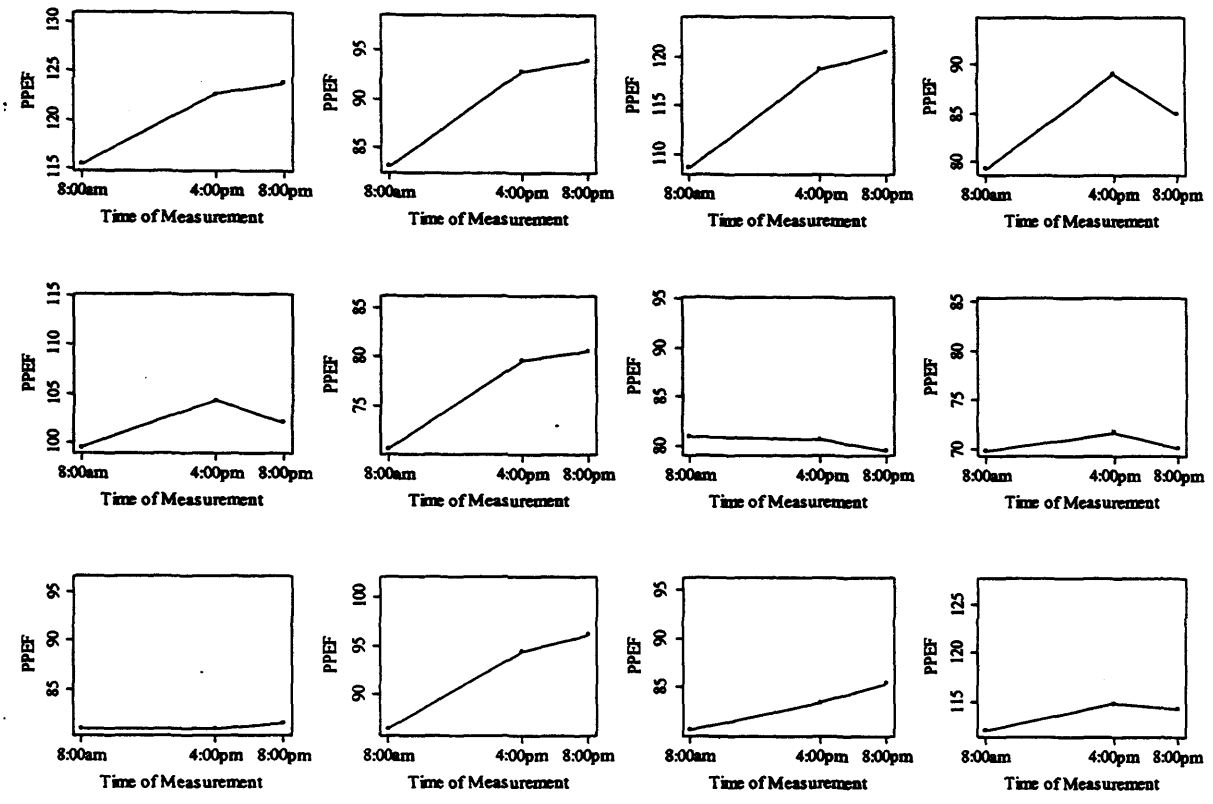

Figure 4.4 Effect of time of measurement for first 12 children.

$$\Sigma_\alpha = \begin{pmatrix} \sigma^2_{\alpha_0} & & \\ \sigma_{\alpha_{01}} & \sigma^2_{\alpha_1} & \\ \sigma_{\alpha_{02}} & \sigma_{\alpha_{12}} & \sigma^2_{\alpha_2} \end{pmatrix} \qquad (4.9)$$

The model now estimates the mean PPEF in the morning, the afternoon and the evening. The effects of PM and EVE are also included as between-subject random effects. This is because the effect of time of day is unlikely to be the same for all subjects with perhaps some subjects having increased or decreased morning dipping when compared to the other subjects. This is investigated in Figure 4.4 where the mean value for each time of day measurement is shown for the first 12 subjects. The figure clearly shows that the effect of the time of day varies from subject to subject with some children having very little mean change in their PPEF during the day while others have increases of over 10%.

I have assumed that the three between-subject random effects come from a multivariate normal distribution with a zero mean vector and covariance matrix Σ_α, as it is likely that

the random effects are correlated. The two sources of within-subject variation are defined in the same way as the previous model (4.8). The results of fitting this model can be seen in Table 4.3. It can be seen that PPEF is lower on average in the morning at 94.4%, increases by about 4.4% in the afternoon and by 3.5% in the evening. It can be seen that these effects are clearly statistically significant as the standard errors are small compared to the parameter estimates. In addition the change in –2×Log likelihood for this model when compared to model (4.5) gives $\chi^2_5 = 298.5$, P<0.001 which clearly indicates that there is an effect of time of day.

The covariance matrix associated with the between-subject variation is also shown in Table 4.3. From this covariance matrix it appears that there is between-subject variation in the effects of PM and EVE, i.e. the effect of time of day does vary from subject to subject. Predictive intervals can be constructed for this matrix. The variance of 274.2 for α_0 indicates that most subjects will be within $2 \times \sqrt{274.2} = 33.1\%$ of the morning mean of 94.4. The variance of 7.2 for α_1 indicates that most subjects will have a mean change in PPEF from morning to afternoon of between $4.4 \pm 2 \times \sqrt{7.2}$. =(-1.0%, 9.8%). The corresponding interval for evening is wider (–4.8%,11.8%).

Parameter	Estimate (Standard Error)
FIXED EFFECTS	
β_0 (Intercept)	94.4 (1.76)
β_1 (PM effect)	4.4 (0.41)
β_2 (EVE effect)	3.5 (0.55)
RANDOM EFFECTS	
Σ_α (Between-Subject Variation)	$\begin{pmatrix} 274.2(41.7) & & \\ -15.6(7.2) & 7.2(2.3) & \\ -28.8(9.8) & 10.3(2.6) & 17.1(3.9) \end{pmatrix}$
$\sigma^2_{\delta_0}$ (Between-Day Within-Subject Variation)	22.5 (1.8)
σ^2 (Within-Day Within-Subject Variation)	47.4 (1.6)

Table 4.3 Including the effect of time of day as both fixed and random effects

Converting the covariance matrix to a correlation matrix gives

$$\begin{pmatrix} 1 & & \\ -0.35 & 1 & \\ -0.42 & 0.93 & 1 \end{pmatrix}$$

The negative correlations for Intercept / PM and Intercept / EVE indicate that those subjects who tend to have higher intercepts (high PPEF in the morning) do not tend to have as large increases in PPEF later in the day when compared to subjects with smaller intercepts. The high positive correlation for PM / EVE indicates that those subjects who have high increases in PPEF in the afternoon when compared to the morning also have a high increase in PPEF in the evening when compared to the morning. This would appear sensible as the PM and EVE measures are likely to be more similar than either of these measures and the morning measure. These correlations indicate that it is necessary to estimate the covariance components of matrix Σ_α as the three between-subject random effects are not independent.

The between-day within-subject variation has increased slightly. In addition the within-day within-subject variation has decreased. This is because the effect of time of day as both fixed and random effects has been incorporated in the model and these are within-day within-subject (level 1) covariates, and thus will explain a proportion of the variation. In other words, by incorporating information about time of day one is explaining part of the within-subject variation.

4.5.4 Including the effect of atopic status.

The model can be further extended by incorporation of the information on atopic status. ATP is a dichotomous covariate taking the value 0 if the child is non-atopic and 1 if the child is atopic. Model (4.8) can be extended to investigate the effects of atopy on mean PPEF as follows

$$PPEF_{ijk} = \beta_0 + \beta_1 PM_{ijk} + \beta_2 EVE_{ijk} + \beta_3 ATP_k + \beta_4 ATP_k \cdot PM_{ijk} + \beta_5 ATP_k \cdot EVE_{ijk}$$
$$+ \alpha_{0k} + \alpha_{1k} PM_{ijk} + \alpha_{2k} EVE_{ijk} + \delta_{0jk} + e_{ijk} \quad (4.10)$$

$$\underline{\alpha}_k \sim MVN(\underline{0}, \Sigma_\alpha)$$

$$\delta_{0jk} \sim N(0, \sigma^2_{\delta_0})$$

$$e_{ijk} \sim N(0, \sigma^2)$$

In this model only the fixed effects of ATP and its interactions with PM and EVE have been included as additions to the model. The definition of the random effects has remained the same as the previous model. The results of fitting this model can be seen in Table 4.4. It can be seen that there appears to be an effect of atopy in that atopic children tend to have lower PPEF. The greatest difference is in the morning at 7.5%. The difference in the afternoon and evening is slightly less, but the interaction in not formally significant (Likelihood ratio test, $\chi^2_2 = 3.5$, P=0.17). Figure 4.5 shows a plot of the mean PPEF at the

Parameter	Estimate (Standard Error)
FIXED EFFECTS	
β_0 (Intercept)	97.6 (2.3)
β_1 (PM effect)	3.8 (0.53)
β_2 (EVE effect)	2.8 (0.71)
β_3 (ATP effect)	-7.5 (3.48)
β_4 (ATP.PM interaction)	1.6 (0.82)
β_5 (ATP.EVE interaction)	1.5 (1.10)
RANDOM EFFECTS	
Σ_α (Between-Subject Variation)	$\begin{pmatrix} 260.7(39.7) & & \\ -12.9(6.8) & 6.5(2.2) & \\ -26.0(9.4) & 9.7(2.5) & 16.5(3.9) \end{pmatrix}$
$\sigma^2_{\delta_0}$ (Between-Day Within-Subject Variation)	22.5 (1.8)
σ^2 (Within-Day Within-Subject Variation)	47.4 (1.6)

Table 4.4 Including the effect of atopy as fixed effects.

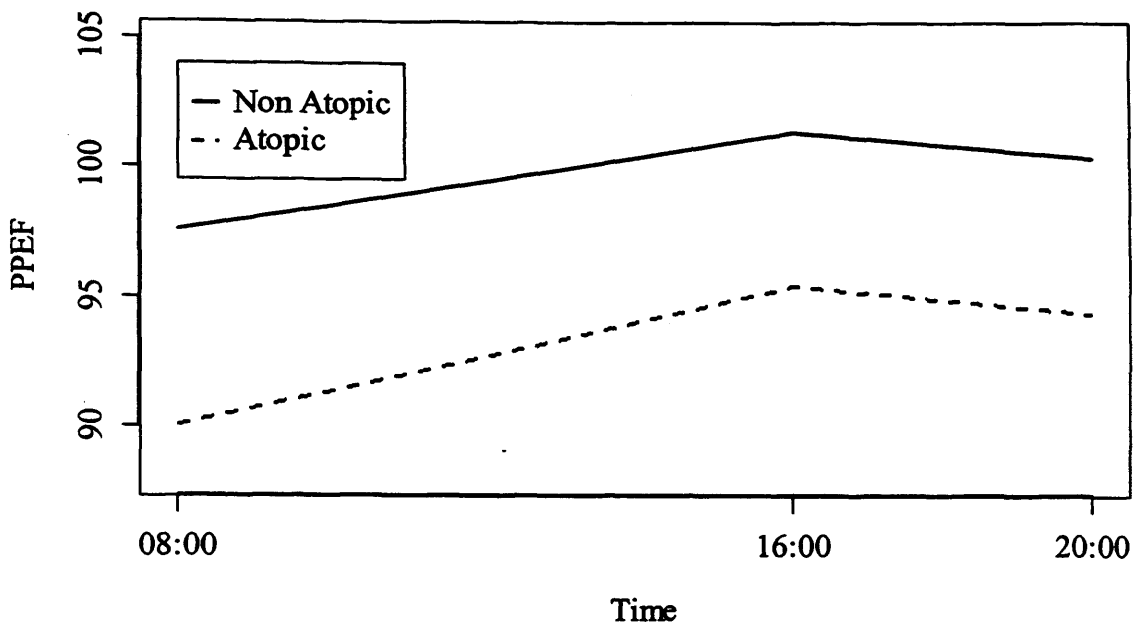

Figure 4.5 Mean PPEF for non-atopic and atopic children measured at the three times of day.

three times of day for non-atopic and atopic children. Visually one can see that the effect of time of day appears to be similar for atopic and non-atopic children.

The random effects have changed very little from the previous model in Table 4.3. There is a slight reduction in the between-subject variances for the three between-subject random effects. This is to be expected as a subject level covariate (ATP) has been added to the model thus explaining part of the between-subject variation.

4.5.5 Modelling the within-subject variance.

As was stated earlier in this chapter, the variability of PPEF is often of interest. It has been shown previously that atopic children have greater between and within-day variability (Clough *et al.* 1991). One of the advantages of fitting a three hierarchical model to this data is that the effect of covariates on the variation at each level can be investigated. Thus, it is possible to add terms to the model that allow for different between-day within-subject variances and different within-day within-subject variances for atopic and non-atopic children. The model is defined as follows

$$PPEF_{ijk} = \beta_0 + \beta_1 PM_{ijk} + \beta_2 EVE_{ijk} + \beta_3 ATP_k + \beta_4 ATP_k \cdot PM_{ijk} + \beta_5 ATP_k \cdot EVE_{ijk} + \alpha_{0k} + \alpha_{1k} PM_{ijk} + \alpha_{2k} EVE_{ijk} + \delta_{jk} + e_{ijk}$$

$$\underline{\alpha}_k \sim MVN(\underline{0}, \Sigma_\alpha)$$

$$\delta_{jk} \sim \begin{cases} N(0, \sigma^2_{\delta_1}) & \text{Non Atopic} \\ N(0, \sigma^2_{\delta_2}) & \text{Atopic} \end{cases}$$

$$e_{ijk} \sim \begin{cases} N(0, \sigma^2_{e_1}) & \text{Non Atopic} \\ N(0, \sigma^2_{e_2}) & \text{Atopic} \end{cases} \qquad (4.11)$$

In comparison to the previous model, there are now two between-day within-subject variances, $\sigma^2_{\delta_1}$ and $\sigma^2_{\delta_2}$, and two within-day with subject variances, $\sigma^2_{e_1}$ and $\sigma^2_{e_2}$. The results of fitting this model can be seen in Table 4.5. The fixed effects and the between-subject random effects have changed very little from the previous model. This is perhaps not really surprising as the additional terms only change the within-subject variability. It

Parameter	Estimate (Standard Error)
FIXED EFFECTS	
β_0 (Intercept)	97.6 (2.3)
β_1 (PM effect)	3.8 (0.51)
β_2 (EVE effect)	2.8 (0.69)
β_3 (ATP effect)	-7.5 (3.48)
β_4 (ATP.PM interaction)	1.5 (0.84)
β_5 (ATP.EVE interaction)	1.5 (1.11)
RANDOM EFFECTS	
Σ_α (Between-Subject Variation)	$\begin{pmatrix} 260.1(39.6) & & \\ -12.6(6.8) & 6.0(2.2) & \\ -25.4(9.3) & 9.7(2.5) & 16.4(3.8) \end{pmatrix}$
$\sigma^2_{\delta_1}$ (Non Atopic)	23.7 (2.3)
$\sigma^2_{\delta_2}$ (Atopic)	20.4 (3.1)
$\sigma^2_{e_1}$ (Non Atopic)	39.8 (1.7)
$\sigma^2_{e_2}$ (Atopic)	58.2 (3.0)

Table 4.5 Allowing separate within-subject variances for non atopic and atopic children.

Group	Variance	90% Confidence Interval
Non Atopic	39.8	±10.3
Atopic	58.2	±12.5

Table 4.6 Quantifying within-day within-subject variation

can be seen that the two between-day within-subject variances are fairly similar indicating that atopic and non-atopic children have fairly similar variation between days. However, there does appear to be a difference in the within-day within-subject variance between non-atopic and atopic children. The difference between the two groups is quantified in Table 4.6. In this table I have taken the square root of the variance to obtain the within-day within-subject standard deviation. I have then constructed a 90% predictive interval. For the non-atopic group the majority of an individual's observations will be within about 10% of that predicted after the fixed effects, between-subject and between-day random effects have been taken into account. For the atopic group this value is about 12.5%.

The interpretation here is important. The fixed effects indicate that atopic children tend to have lower PPEF than non-atopic children, but there is not sufficient evidence to suggest that morning dipping is increased in atopic children (i.e. on average the reduction in PPEF is the same for all three time of day measurements). Incorporation of the between-subject random effects essentially gives each child their own underlying mean morning, afternoon and evening PPEF. The between-day within-subject random effects shifts these three mean measures up or down for each of the 14 days separately for each child. There is then greater variation about these mean values for atopic children when compared to non-atopic children. It is important to realise this is in addition to any differences observed in the fixed effects, i.e slightly increased morning dipping.

The model can be extended further as it is possible that the within-subject within-day variances differ at different times of the day. It thus seems sensible to allow different within-subject within-day variances for the three time of day measurements. Model (4.11) can thus be extended to,

Parameter	Estimate (Standard Error)
FIXED EFFECTS	
β_0 (Intercept)	97.6 (2.3)
β_1 (PM effect)	3.8 (0.51)
β_2 (EVE effect)	2.8 (0.69)
β_3 (ATP effect)	-7.5 (3.48)
β_4 (ATP.PM interaction)	1.5 (0.84)
β_5 (ATP.EVE interaction)	1.5 (1.11)
RANDOM EFFECTS	
Σ_α (Between-Subject Variation)	$\begin{pmatrix} 260.0(39.7) & & \\ -12.2(6.8) & 6.0(2.2) & \\ -25.1(9.3) & 9.3(2.5) & 16.7(3.8) \end{pmatrix}$
$\sigma^2_{\delta_1}$ (Non Atopic)	23.2 (2.3)
$\sigma^2_{\delta_2}$ (Atopic)	20.4 (3.1)
$\sigma^2_{e_1}$ (Non Atopic AM)	45.7 (3.3)
$\sigma^2_{e_2}$ (Non Atopic PM)	41.8 (3.2)
$\sigma^2_{e_3}$ (Non Atopic EVE)	31.4 (2.8)
$\sigma^2_{e_4}$ (Atopic AM)	63.0 (5.3)
$\sigma^2_{e_5}$ (Atopic PM)	60.5 (5.2)
$\sigma^2_{e_6}$ (Atopic EVE)	51.2 (4.9)

Table 4.7 Allowing separate within-subject variances for non atopic and atopic children and separate times of the day.

Group	Variance	90% Confidence Interval
Non Atopic AM	45.7	±11.1
Non Atopic PM	41.8	±10.6
Non Atopic EVE	31.4	±9.2
Atopic AM	63.0	±13.0
Atopic PM	60.5	±12.8
Atopic EVE	51.2	±11.7

Table 4.8 Quantifying within-day within-subject variation

$$PPEF_{ijk} = \beta_0 + \beta_1 PM_{ijk} + \beta_2 EVE_{ijk} + \beta_3 ATP_k + \beta_4 ATP_k \cdot PM_{ijk} + \beta_5 ATP_k \cdot EVE_{ijk}$$

$$+ \alpha_{0k} + \alpha_{1k} PM_{ijk} + \alpha_{2k} EVE_{ijk} + \delta_{jk} + e_{ijk}$$

$$\underline{\alpha}_k \sim MVN(\underline{0}, \Sigma_\alpha)$$

$$\delta_{jk} \sim \begin{cases} N(0, \sigma^2_{\delta_1}) & \text{Non Atopic} \\ N(0, \sigma^2_{\delta_2}) & \text{Atopic} \end{cases} \qquad (4.12)$$

$$e_{ijk} \sim \begin{cases} N(0, \sigma^2_{e_1}) & \text{Non Atopic AM} \\ N(0, \sigma^2_{e_2}) & \text{Non Atopic PM} \\ N(0, \sigma^2_{e_3}) & \text{Non Atopic EVE} \\ N(0, \sigma^2_{e_4}) & \text{Atopic AM} \\ N(0, \sigma^2_{e_5}) & \text{Atopic PM} \\ N(0, \sigma^2_{e_6}) & \text{Atopic EVE} \end{cases}$$

This model allows a separate within-day within-subject variance to be estimated for each time of day for both atopic and non-atopic children, i.e. there now six within-day within-subject variance terms rather than the two in the previous model. The results of fitting this

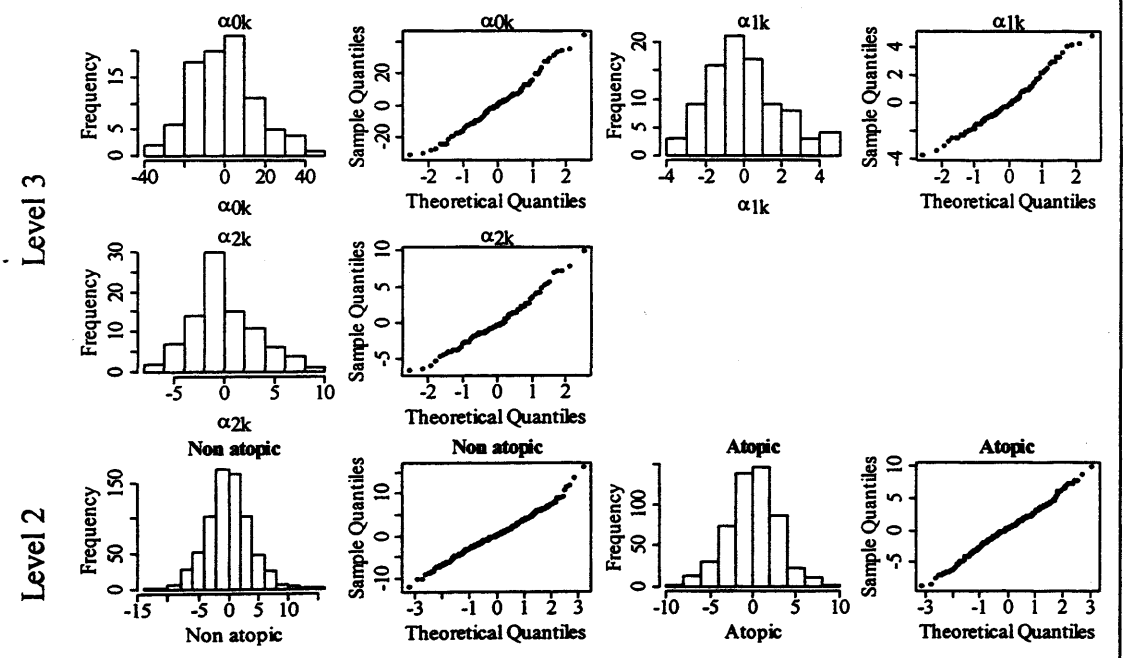

Figure 4.6 Histograms and normal probability plots for level 3 (between-subject) and level 2 (between-day within-subject) residuals.

model can be seen in Table 4.7. The fixed effects and the between-subject and between-day subject within-subject random effects parameters have changed very little from the previous model. For the within-day within-subject variation it can be seen that atopic children have greater variances at all times and that for both groups of children the variances are lower in the evening.

This can be further quantified by calculating 90% prediction intervals as before. The results of this can be seen in Table 4.8. As an example, these values can be interpreted as follows; the PPEF for non atopic children in the morning will be within 11.1% of that child's mean after the random effect of day has been taken into account.

4.5.6 Normality Assumption

Figure 4.6 shows histograms and normal probability plots for the level 3 and level 2 residuals, while Figure 4.7 shows the histograms and normal probability plots for the level 1 residuals. It can be seen that these plots show that the assumption of normality appears valid in all cases.

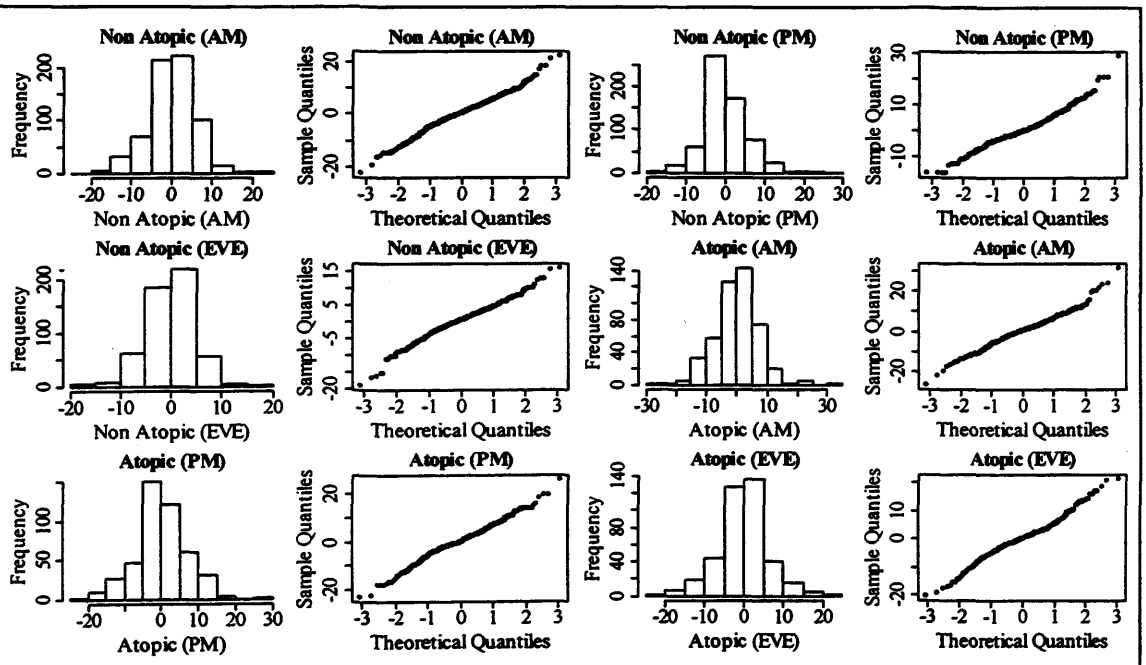

Figure 4.7 Histograms and normal probability plots for level 1 (within-day within-subject residuals.

4.6 A Bayesian Approach

4.6.1 Introduction

In this section I will demonstrate how a Bayesian approach can be adopted for the PPEF model. There are two main advantages for adopting a Bayesian model. Firstly, the IGLS model does not allow for the uncertainty in the estimates of the variance components when estimating the fixed effects or their standard errors. With a three level model it is unclear how estimating the many random effects may affect the uncertainty in the estimates of the fixed effect parameters. A second advantage is that in the IGLS model there are problems in making inferences about the variance components, due to problems regarding the assumption of multivariate normality, when obtaining the parameter estimates of the variances and covariances. Although it is possible to obtain standard errors for the variances, they can be unreliable as discussed in section 2.5.1. I will show in this section how it is possible to obtain credible intervals and density plots for the random effect variances or standard deviations.

4.6.2 Initial estimation of the Bayesian model

I initially attempted to fit a model identical to (4.12) in WinBUGS, i.e.

$$PPEF_{ijk} = \beta_0 + \beta_1 PM_{ijk} + \beta_2 EVE_{ijk} + \beta_3 ATP_k + \beta_4 ATP_k \cdot PM_{ijk} + \beta_5 ATP_k \cdot EVE_{ijk} + \alpha_{0k} + \alpha_{1k} PM_{ijk} + \alpha_{2k} EVE_{ijk} + \delta_{jk} + e_{ijk}$$

$$\underline{\alpha}_k \sim MVN(\underline{0}, \Sigma_\alpha)$$

$$\delta_{jk} \sim \begin{cases} N(0, \sigma^2_{\delta_1}) & \text{Non Atopic} \\ N(0, \sigma^2_{\delta_2}) & \text{Atopic} \end{cases} \qquad (4.13)$$

$$e_{ijk} \sim \begin{cases} N(0, \sigma^2_{e_1}) & \text{Non Atopic AM} \\ N(0, \sigma^2_{e_2}) & \text{Non Atopic PM} \\ N(0, \sigma^2_{e_3}) & \text{Non Atopic EVE} \\ N(0, \sigma^2_{e_4}) & \text{Atopic AM} \\ N(0, \sigma^2_{e_5}) & \text{Atopic PM} \\ N(0, \sigma^2_{e_6}) & \text{Atopic EVE} \end{cases}$$

with univariate non-informative prior distributions for the fixed effects and the within-subject random effect variances as follows

$$\beta_0,\ldots,\beta_5 \sim N(0,1000000)$$
$$\frac{1}{\sigma^2_{\delta_1}},\frac{1}{\sigma^2_{\delta_2}},\frac{1}{\sigma^2_{e_1}},\cdots,\frac{1}{\sigma^2_{e_6}} \sim Gamma(0.001,0.001) \qquad (4.14)$$

Specifying a prior distribution for Σ is slightly more complex. In Chapter 3 I demonstrated how the product normal formulation can be used for multivariate normality. However, another common method for specifying covariance matrices is to use the Wishart distribution as a prior for Σ^{-1}. The Wishart distribution can be defined for a *pxp* symmetric

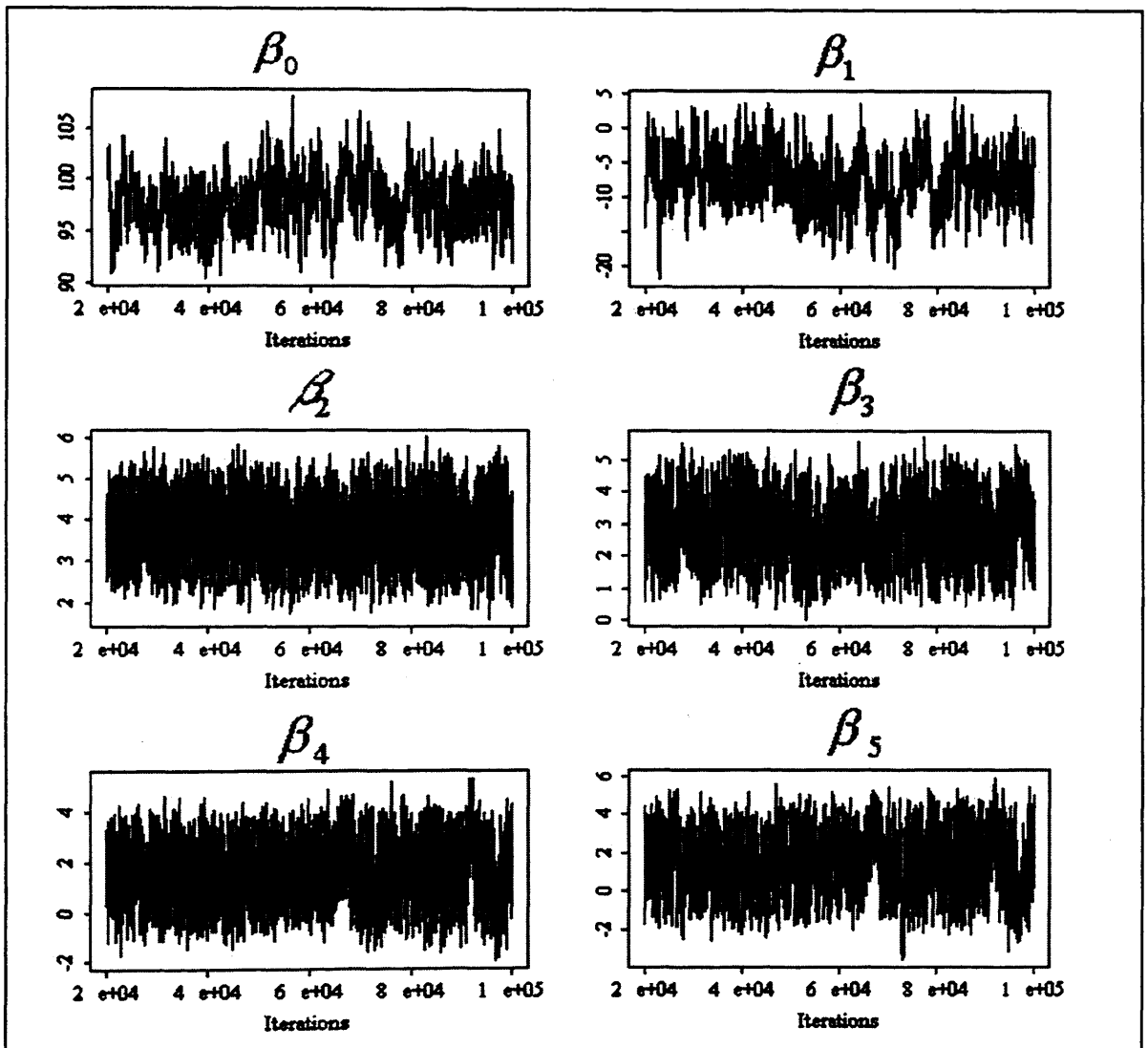

Figure 4.8 Trace plots for fixed effects for initial Bayesian PEF model defined in equation (4.13).

positive definite matrix $\mathbf{x}$ as follows

$$f(\mathbf{x} \mid \mathbf{R}, k) \propto |\mathbf{R}|^{\frac{k}{2}} |\mathbf{x}|^{\frac{k-p-1}{2}} \exp\left[-\tfrac{1}{2}\text{trace}(\mathbf{R}\mathbf{x})\right] \qquad (4.15)$$

where $k \geq p$ are the degrees of freedom, and $\mathbf{R}$ is a $p \times p$ symmetric non-singular matrix. In fact, the Wishart distribution is a multivariate extension of the χ^2 distribution (De Groot 1970). If k is small then Σ should have essentially a non-informative prior distribution. $\mathbf{R}$ can be considered as an estimate of the 'order of magnitude' of the covariance matrix. If k is small then the choice of $\mathbf{R}$ is not crucial and it should lead to a non-informative prior. Thus for model (4.12) the prior distribution for Σ is assumed to be

$$\Sigma^{-1} \sim Wishart\left[\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, 3\right] \qquad (4.16)$$

When fitting this model it immediately became clear that there were severe problems with autocorrelation in the MCMC chains and so I chose a burn-in of 20000 iterations and then drew samples from a further 80000 iterations. The total of 100000 iterations took about 70 minutes on a Pentium II 400Mhz PC. The autocorrelation was severe as can be seen in the trace plots for the fixed effects in Figure 4.8 and the corresponding autocorrelation plots in Figure 4.9. Even though there are 80000 sampled values for each parameter it is still possible to notice patterns in the trace plots, notably for β_0 and β_1. This is confirmed by calculating the autocorrelations up to lag 200, which are shown in Table 4.9. It can be seen that the autocorrelations even after 200 lags are still about 0.5 for β_0 and β_1, and still higher than desired for the other paramaters.

Parameter	Lag 1	Lag 50	Lag 100	Lag 200
β_0 (Intercept)	0.99	0.82	0.69	0.51
β_1 (PM)	0.99	0.82	0.67	0.49
β_2 (EVE)	0.45	0.21	0.17	0.13
β_3 (ATOPY)	0.68	0.35	0.29	0.22
β_4 (ATOPY.PM)	0.51	0.29	0.25	0.21
β_5 (ATOPY.EVE)	0.73	0.44	0.38	0.31

Table 4.9 Autocorrelations for fixed effects at lags of 1, 50, 100 and 200 for intial Bayesian PEF model.

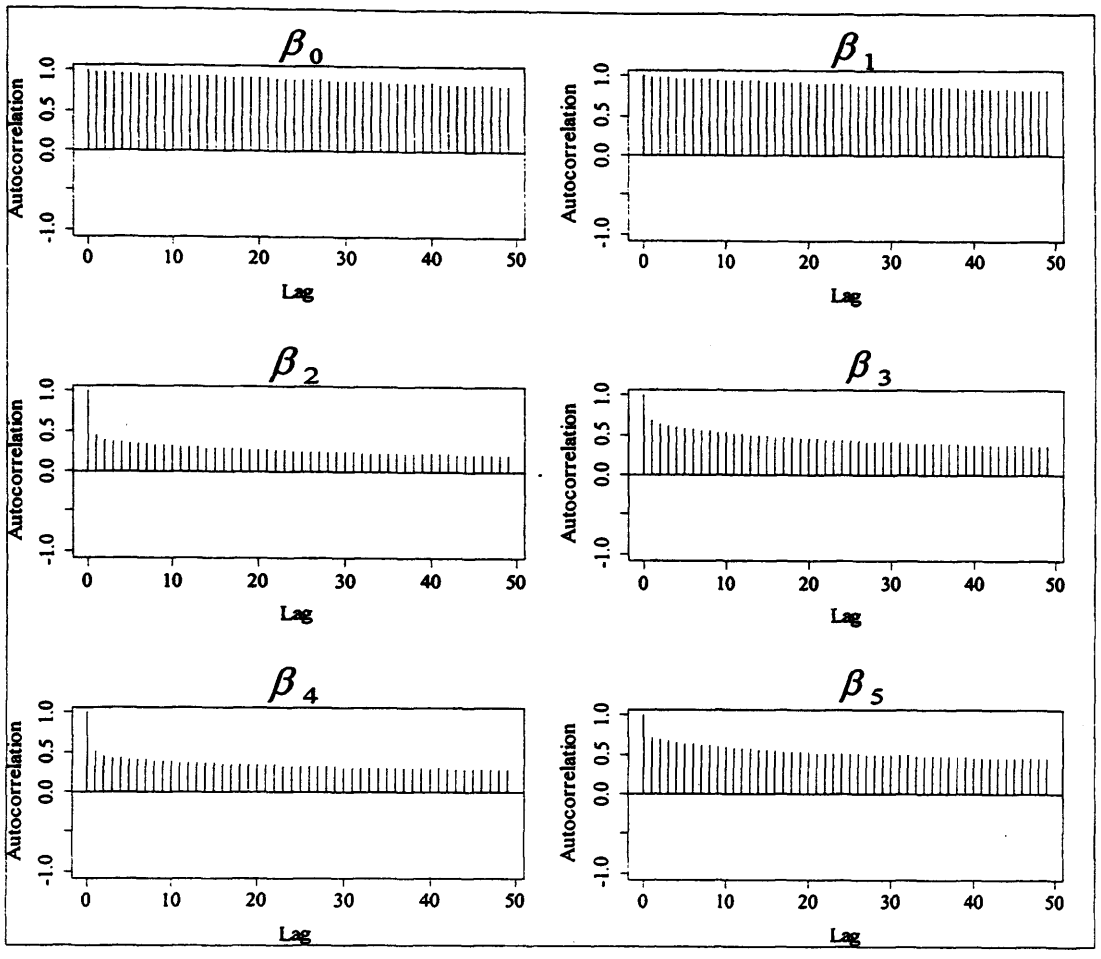

Figure 4.9 Autocorrelation for fixed effects for initial Bayesian PEF model

The Geweke convergence diagnostic (Geweke, 1992) was obtained for the fixed effects with the results presented in Table 4.10. Three of the parameters have absolute Z-scores greater than 2, indicating that the chains still may still have not achieved convergence.

Parameter	Geweke Z score
β_0 (Intercept)	-5.12
β_1 (PM)	7.18
β_2 (EVE)	3.31
β_3 (ATOPY)	1.84
β_4 (ATOPY.PM)	0.46
β_5 (ATOPY.EVE)	1.36

Table 4.10 Geweke convergence diagnostic for initial Bayesian PEF model

The potential lack of convergence is a serious problem. The model in its current form would have to run for much longer in order to achieve convergence, with the main problem being the high autocorrelation. This will lead to problems with computer storage with the current output file of samples being over 25 megabytes with the file having to be split so that the MCMC diagnostic software CODA (Best, Cowles, and Vines, 1995) or BOA (Smith, 2000) could be used. Another problem with the high autocorrelation is that the model will be sensitive to the choice of initial values. If initial values are a long way from the true value, then it will take an extremely large number of samples until the initial value "is forgotten" thus requiring a very long 'burn in'. Such high autocorrelation, which although in theory should not be a problem as long as the chains are run for long enough, is a major problem in practice for the reasons mentioned above and also the problem of being certain that the chains has converged.

4.6.3 *Hierarchical Centring*

One potential method of reducing the autocorrelation and hence reduce the number of iterations required is to reparameterise the model. A method particularly useful for random effects models is to use *hierarchical centring* (Gelfand *et al.*, 1995b). Gelfand showed that in random effects models there can be large posterior correlations between the fixed effects and random effects and between the random effects themselves. It was shown that by reparameterising the model these correlations can be reduced. If one considers the variance components model (4.5)

$$\begin{aligned} PPEF_{ijk} &= \beta_0 + \alpha_k + \delta_{jk} + e_{ijk} \\ \alpha_k &\sim N(0, \sigma^2_{\alpha_0}) \\ \delta_{jk} &\sim N(0, \sigma^2_{\delta_0}) \\ e_{ijk} &\sim N(0, \sigma^2) \end{aligned} \qquad (4.17)$$

This can be rewritten in hierarchically centred form as follows

$$\begin{aligned} PPEF_{ijk} &\sim N(\delta_{jk}, \sigma^2) \\ \delta_{jk} &\sim N(\alpha_k, \sigma^2_\delta) \\ \alpha_k &\sim N(\beta_0, \sigma^2_\alpha) \end{aligned} \qquad (4.18)$$

When writing the model in this form, Gelfand showed that the posterior correlations between parameters are much reduced. In the presence of covariates, it may not be possible to fully centre the parameterisation, but it may be possible to partially centre the model (Gelfand *et al.*, 1995a).

Using this approach model (4.13) can be rewritten as follows

$$\begin{aligned} PPEF_{ijk} &\sim N(\phi_{ijk}, \sigma^2_{e_r}) \quad r = 1,\cdots,6 \\ \phi_{ijk} &= \delta_{jk} + \alpha_{0k} + \alpha_{1k} PM_{ijk} + \alpha_{2k} EVE_{ijk} \\ \delta_{jk} &\sim N(0, \sigma^2_{\delta_s}) \quad s = 1,2 \\ \underline{\alpha}_k &\sim MVN(\underline{\mu}_\alpha, \Sigma) \end{aligned} \tag{4.19}$$

$$\mu_\alpha = \begin{pmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \end{pmatrix} \text{ for non - atopics,} \quad \mu_\alpha = \begin{pmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \end{pmatrix} \text{ for atopics}$$

This is the same model as (4.13), but just parameterised in a different way. The model in (4.19) used partial centring with the α parameters centred, but not the δ parameters. The parameterisation of the fixed effects is slightly different to (4.12), but the γ's can be transformed back to the original parameterisation as follows,

$$\begin{aligned} &\beta_0 = \gamma_0, \quad \beta_1 = \gamma_2, \quad \beta_2 = \gamma_4 \\ &\beta_3 = \gamma_1 - \gamma_0, \quad \beta_4 = \gamma_3 - \gamma_2, \quad \beta_5 = \gamma_5 - \gamma_4 \end{aligned} \tag{4.20}$$

One of the advantages of using MCMC methods is that although the β's are not modelled directly, they can be obtained as functions other parameters. They can be estimated at each iteration of the Gibbs Sampler and monitored as one would monitor any model parameter.

4.6.4 Results of the hierarchically centred model

Five chains were run using WinBUGS, in itially with 5000 iterations each. Various combinations of very high or very low variances and very high and very low fixed effect estimates as well as differing random effects were chosen as starting values. Figure 4.10 shows the Gelman and Rubin plots, as described in section 2.5.2, for the fixed effects. It can be seen that the values of $\sqrt{R}$ and $\sqrt{R_{97.5\%}}$ are close to one after about 1000 to 2000 iterations which indicates that the between-chain variation appears to be very small indicating that the chains have converged. The values of $\sqrt{R}$ and $\sqrt{R_{97.5\%}}$ calculated for the final 50% of the iterations for all parameters can be seen in Table 4.11. It can be seen that with just looking at the final 2500 iterations of the five chains it appears to indicate that the between-chain variance is small and one can be fairly confident that the chains have converged. The plots for the other parameters were very similar and are not shown. It is possible to combine the results from the 5 chains, but often it is simpler to deal with one chain. Therefore I just analysed the samples from one of the chains, but ran it for longer with a 5000 iterations 'burn in' with samples taken from a further 50000 iterations.

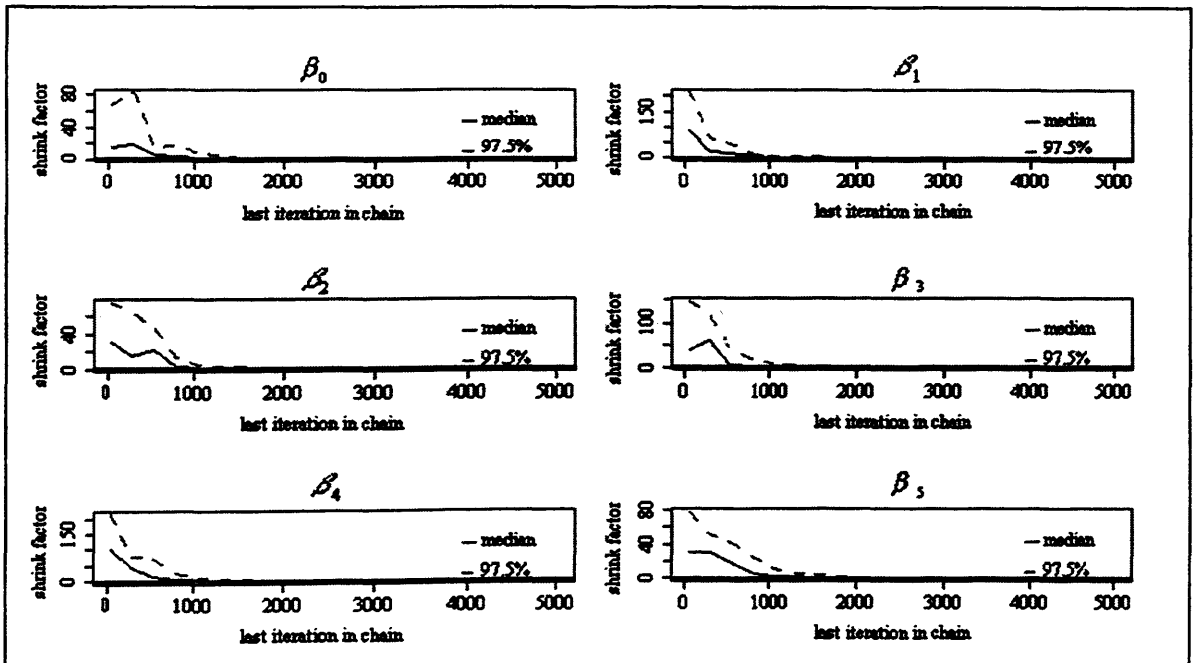

Figure 4.10 Gelman and Rubin plots for the fixed effects for hierarchically centred PEF model

Fixed Effects			Between-subject random effects			Within-subject random effects		
	$\sqrt{R}$	$\sqrt{R_{97.5\%}}$		$\sqrt{R}$	$\sqrt{R_{97.5\%}}$		$\sqrt{R}$	$\sqrt{R_{97.5\%}}$
β_0	1.00	1.01	Σ_{11}	1.00	1.00	$\sigma^2_{e_1}$	1.00	1.00
β_1	1.04	1.08	Σ_{12}	1.01	1.02	$\sigma^2_{e_2}$	1.00	1.00
β_2	1.03	1.07	Σ_{13}	1.00	1.00	$\sigma^2_{e_3}$	1.00	1.01
β_3	1.00	1.01	Σ_{22}	1.01	1.02	$\sigma^2_{e_4}$	1.00	1.00
β_4	1.01	1.04	Σ_{23}	1.00	1.01	$\sigma^2_{e_5}$	1.00	1.00
β_5	1.01	1.04	Σ_{33}	1.00	1.00	$\sigma^2_{e_6}$	1.00	1.00
						$\sigma^2_{\delta_1}$	1.00	1.01
						$\sigma^2_{\delta_2}$	1.00	1.00

Table 4.11 Gelman and Rubin diagnostic for hierarchically centred PEF model.

Although it appeared that the convergence had been achieved by about 2000 iterations, when the models do not take long to run, it seems sensible to be cautious and run the 'burn in' for longer. In addition the 50000 iterations used to draw samples from could also be considered to be larger than needed and over cautious. However, as I will demonstrate shortly there is still evidence of autocorrelation, and in general having too large a sample is preferable to having a sample that is thought to be just about large enough.

The results of fitting the model defined in (4.19) can be seen in Table 4.12 where the parameter estimates are compared to the IGLS model. The parameter estimates of the fixed effects for the two models are very similar. However, perhaps more importantly the standard errors are also very similar. This indicates that the uncertainty associated with the variances of the random effects does not appear to have a large impact on the fixed effects parameter estimates or their standard errors. This is probably due to the fairly large sample size. With a smaller sample size one would expect the standard errors of the fixed effect parameters to be larger for the Bayesian model. The parameter estimates of the between-subject random effect variances and the within-subject variances are also very similar for the two methods of estimation. The standard errors of these parameters are also similar.

Parameter	IGLS Estimate (Standard Error)	Bayesian Estimate (Standard Error)
FIXED EFFECTS		
β_0	97.6 (2.3)	97.6 (2.3)
β_1	3.8 (0.51)	3.7 (0.50)
β_2	2.8 (0.69)	2.8 (0.70)
β_3	-7.5 (3.48)	-7.5 (3.58)
β_4	1.5 (0.84)	1.5 (0.83)
β_5	1.5 (1.11)	1.5 (1.16)
RANDOM EFFECTS		
Σ_α	$\begin{pmatrix} 260.0(39.7) & & \\ -12.2(6.8) & 6.0(2.2) & \\ -25.1(9.3) & 9.3(2.5) & 16.7(3.8) \end{pmatrix}$	$\begin{pmatrix} 268.8(42.3) & & \\ -12.1(7.0) & 5.5(2.0) & \\ -25.6(9.7) & 8.9(2.6) & 16.6(4.0) \end{pmatrix}$
$\sigma^2_{\delta_1}$	23.2 (2.3)	23.2 (2.4)
$\sigma^2_{\delta_2}$	20.4 (3.1)	19.6 (3.3)
$\sigma^2_{e_1}$	45.7 (3.3)	46.1 (3.5)
$\sigma^2_{e_2}$	41.8 (3.2)	42.4 (3.3)
$\sigma^2_{e_3}$	31.4 (2.8)	31.7 (2.9)
$\sigma^2_{e_4}$	63.0 (5.3)	63.8 (5.5)
$\sigma^2_{e_5}$	60.5 (5.2)	61.4 (5.4)
$\sigma^2_{e_6}$	51.2 (4.9)	52.0 (4.9)

Table 4.12 Comparison of estimates from IGLS and Bayesian PEF models

This is despite known problems with using classical methods to obtain standard errors for the variances parameters. One of the advantages of using the Bayesian model is that credible intervals can be constructed using the quantiles of the samples chains. This is especially advantageous if the distributions of the variance parameters are skewed.

Figure 4.11 shows a trace plot for the fixed effect parameters and Figure 4.12 shows the corresponding autocorrelation plots. The trace plots are hard to interpret as they are so dense. However, they are clearly an improvement on the trace plots for the initial hierarchical model in Figure 4.8. Although there is still some autocorrelation it is much reduced from the non-hierarchically centred model described in equation (4.13). This can

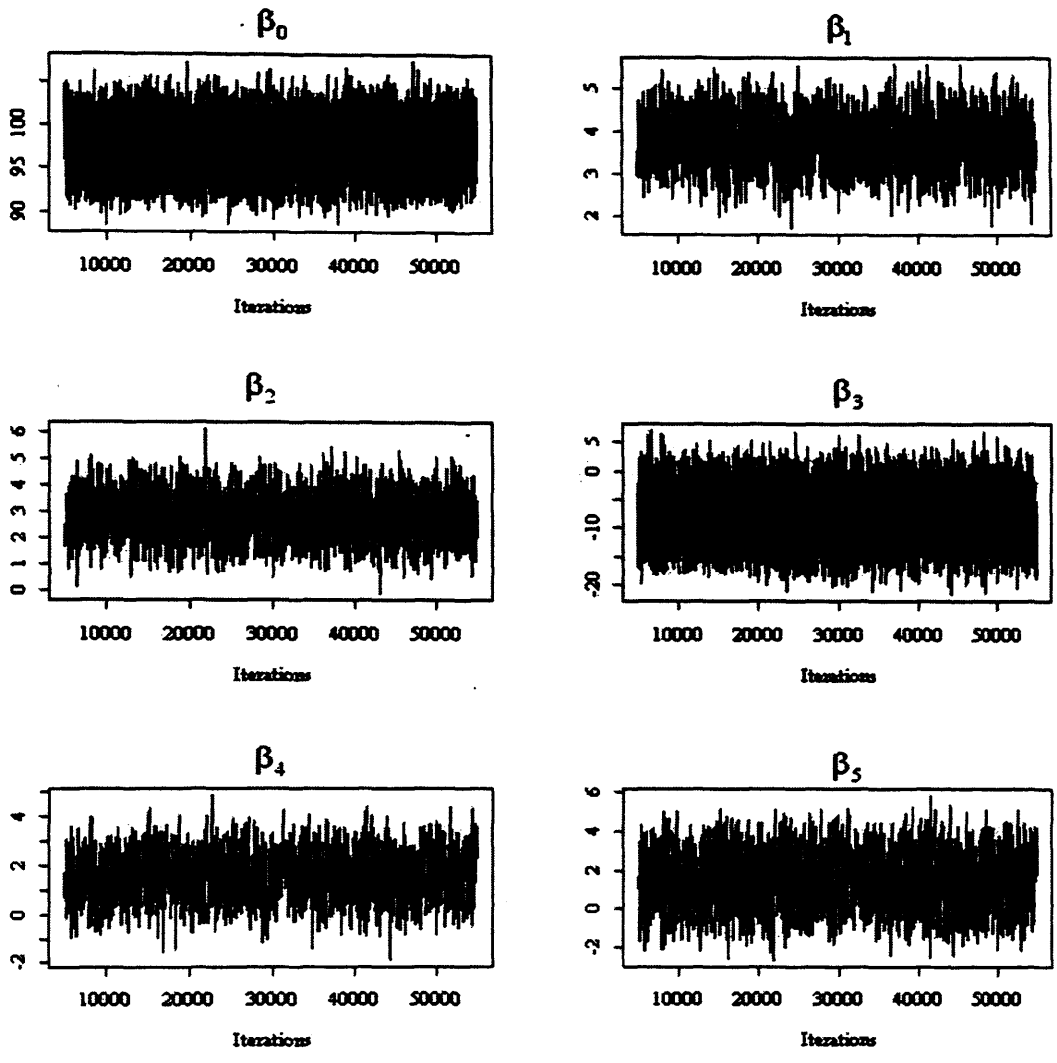

Figure 4.11 Trace plots for fixed effects for hierarchically centred model.

be seen by comparison of the autocorrelation plots in Figure 4.9. The autocorrelation for the random effect variances was smaller, disappearing by about lag 10.

Figure 4.13 shows density plots for the first 10% and last 50% of the chains for each model parameter. As demonstrated in the last chapter, these are a graphical equivalent of the Geweke test. It can be seen that the two densities for most parameters overlap. There are some minor differences in the densities of the fixed effects (the β's), which had the highest autocorrelation, but the differences are very small.

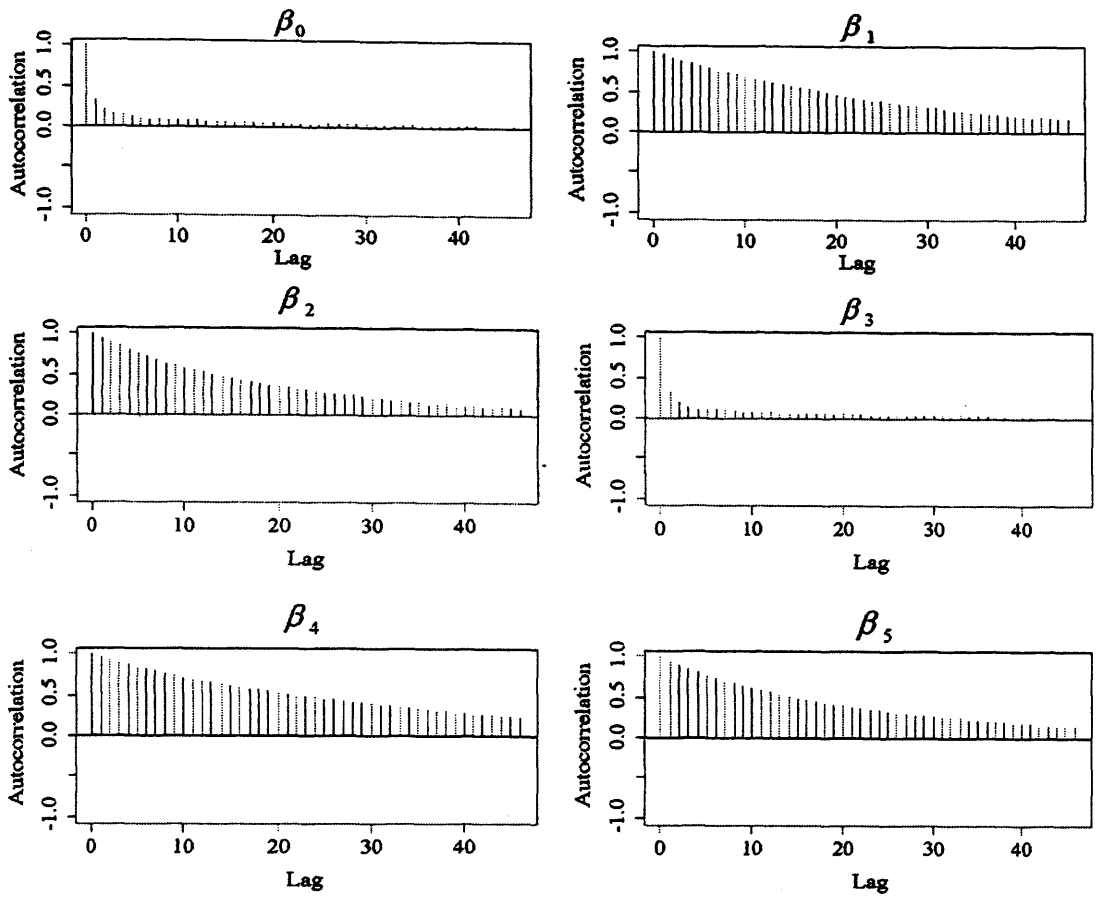

Figure 4.12 Autocorrelation of beta's from hierarchically centred Bayesian model

4.6.5 Interpretation of random effect variances

When using the IGLS model there are problems with making inferences and obtaining confidence intervals for the random effect variances. This is because of problems in estimating the standard errors (Goldstein, 1995). A clear advantage of the Bayesian model is that any function of the model parameters can be obtained and quantiles used to obtain

	Non Atopic	Atopic
AM	6.8 (6.3,7.3)	7.9 (7.3,8.7)
PM	6.5 (6.0,7.0)	7.8 (7.2,8.5)
EVE	5.6 (5.2,6.2)	7.2 (6.6,7.9)

Table 4.13 Within-day within-subject standard deviations with 95% credible intervals for hierarchically centred Bayesian PEF model.

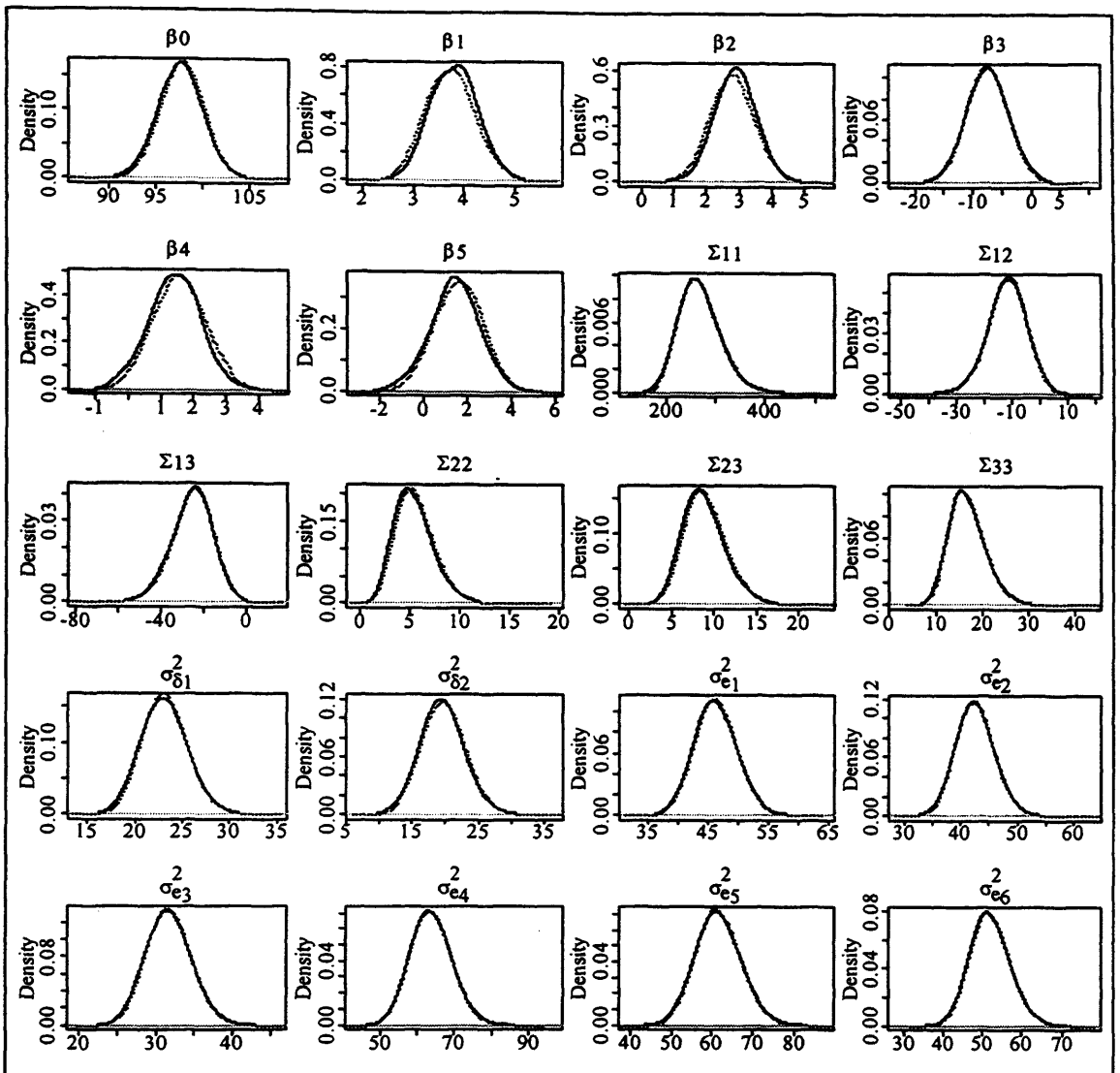

Figure 4.13 Density plot for first 10% (solid line) and last 50% (dotted line) of chains for each model parameter

credible intervals. Since it is easier to interpret standard deviations rather than variances it makes sense to present these. Figure 4.14 displays density plots for the six within-day within-subject standard deviations. These clearly show that at all three times, the standard deviation for the atopic children is greater than that of the non-atopic children and that there is very little overlap between the densities. It is also possible to calculate 95% credible intervals for the standard deviations. These are shown in Table 4.13.

An alternative way of looking at the variances is to obtain the difference in within-day within-subject standard deviations. The densities are shown in Figure 4.15 with the 95% credible intervals in Table 4.14. In addition the probability that the atopic standard

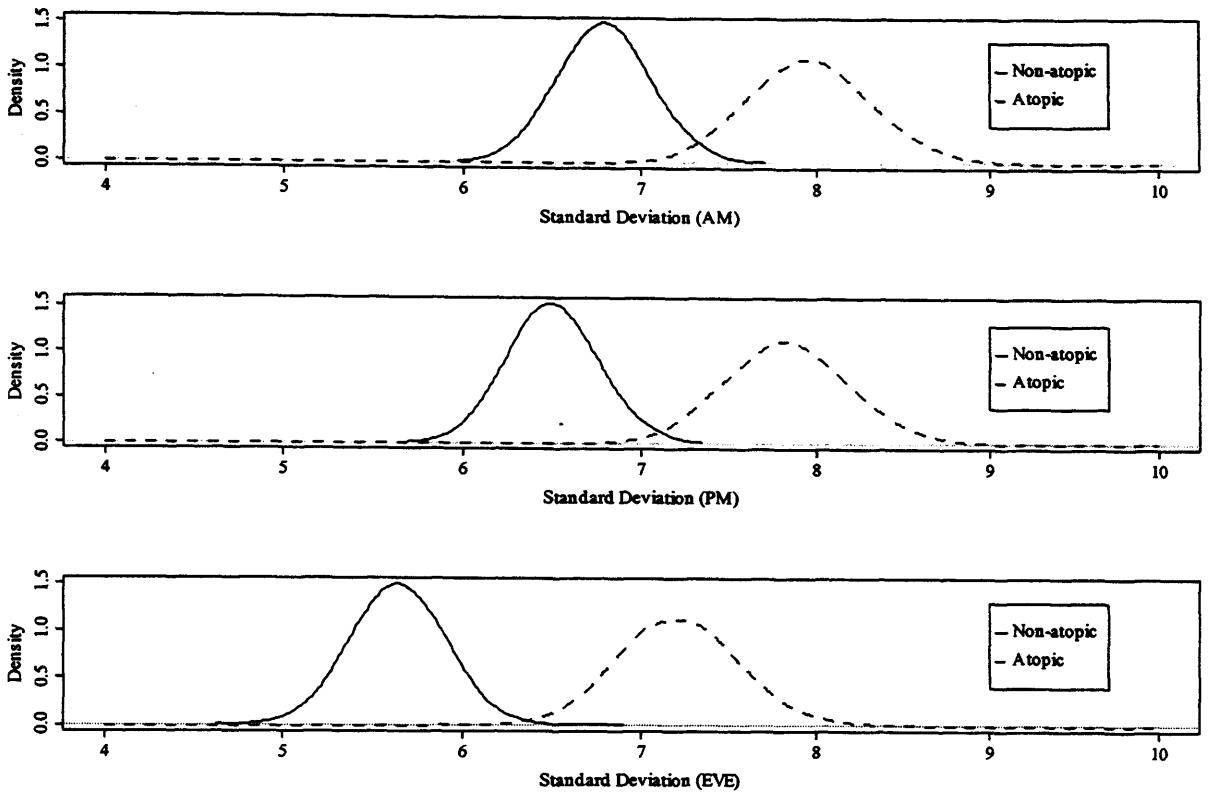

Figure 4.14 Density plots for within-subject within-day standard deviations for Bayesian hierarchically centred PEF model.

deviation is greater than the non- atopic standard deviation has been added to the table. This probability is obtained by observing the proportion of the 50000 iterations the atopic standard deviation is greater than the non-atopic standard deviation (i.e. the difference is <0).

For all time of day measures there is very strong evidence that the within-day within-subject standard deviation is greater for atopic children. This can be seen by inspection of the difference in the standard deviations, their credible intervals and by looking at the probabilities. The probability of one arises from the fact that for none of the 50000 samples

	Difference in SD	P(Atopy SD>Non-atopy SD)
AM	1.2 (0.4,2.1)	0.9982
PM	1.3 (0.5,2.2)	0.9995
EVE	1.6 (0.7,2.4)	1.0

Table 4.14 Difference in within-day within-subject standard deviation for hierarchically centred Bayesian PEF model.

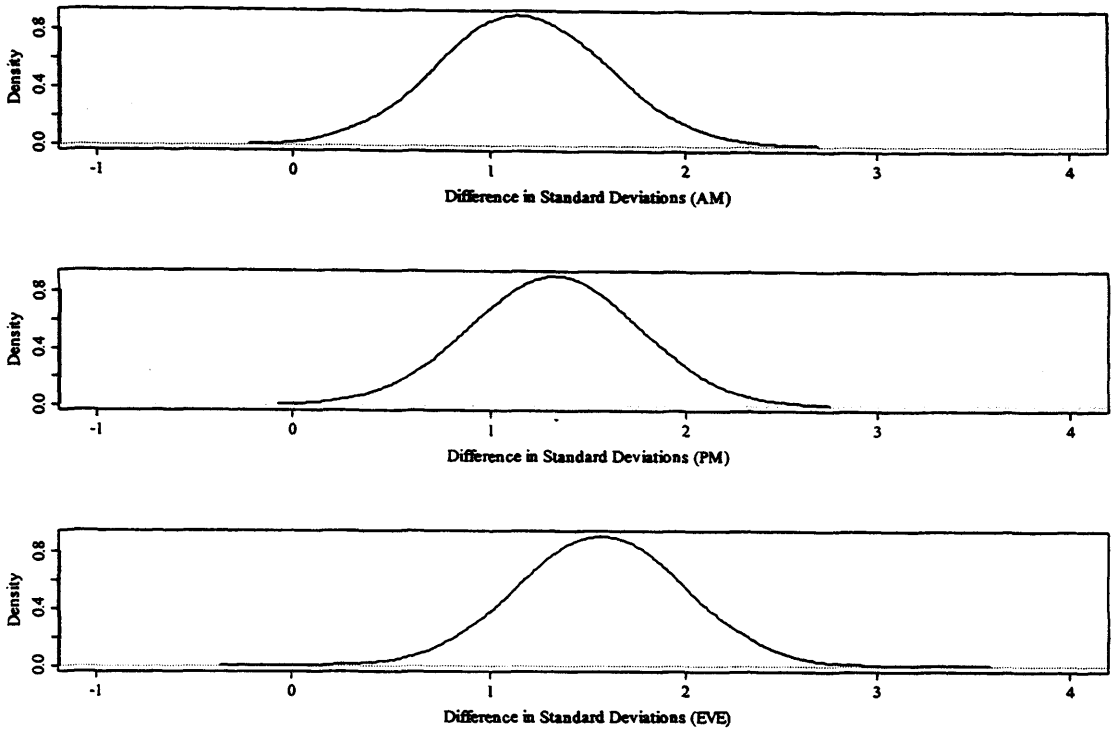

Figure 4.15 Difference between within-subject standard deviations for non atopic and atopic children for hierarchically centred Bayesian PEF model.

was the within-subject standard deviation larger for the non-atopic group. Thus, there is very strong evidence that the atopic children had greater within-day variation.

4.7 Discussion

In this chapter I have demonstrated how data obtained from PEF diaries can be analysed using a hierarchical model. The nature of the recording of the diary with measurements nested within-days, which in turn are nested within-subjects, makes it possible to analyse the data using a three level hierarchical structure rather than the more standard two level structure usually used in the analysis of repeated measures data. Although the model is more complicated than the standard summary measures used in the analysis of PEF data, it is fairly simple to interpret. Perhaps most crucial of all, it provides evidence on a scale that is clinically meaningful and in way which is relatively easy to interpret.

When performing a complex statistical analysis it is important to think about whether the complexity is necessary. If the same inferences can be obtained using simpler techniques, such as summary measures, then generally these should be reported as the results will then be easier for non-statisticians to understand. In the area of PEF there are many summary measures that have been used that attempt to measure some aspect of 'within-subject variability'. However, the many different summary measures all measure different aspects of the variability as I discussed in section 4.4. The lack of consensus and the apparent confusion on which summary measure to use leads one to think that some of the analyses previously performed may be inappropriate. The most important aspect regarding the three level hierarchical model is that it *simultaneously* investigates the level of PEF (as percent predicted) and the three levels of variability, without the need to reduce the data to summary measures. An important aspect of the model is the ability to break down the within-subject variability into two components, namely between-day and within-day variation. Interpretation of the fixed effects is simple as they represent mean values. Interpretation of the random effect variances is more complex, However, these can be converted to standard deviations or prediction intervals which most clinical researchers should be able to understand.

I have presented both a classical and Bayesian analysis of the data. I have shown that the parameter estimates and standard errors were very similar for the two methods of estimation. This is encouraging as the same model was fitted using both methods of estimation. One might expect the estimates and standard errors to differ with smaller sample sizes, as the estimates in the classical model do not take into account the uncertainty associated with the estimates of the variance components. With the Bayesian model there were initial problems with very high autocorrelation. I demonstrated that expressing the model in a slightly different way, using hierarchical centring, decreases the autocorrelation. It is still higher than desired leading to a larger number of iterations being required in order to obtain the parameter estimates. However, given that the model runs in under an hour and that the model was developed using classical methodology, it is probably not of great practical importance. For more complex models that take a long time to run it may be worthwhile exploring other techniques of reducing the autocorrelation (Gilks and Roberts, 1996). Perhaps the most appealing aspect of the Bayesian model is the ability to

make inferences regarding the random effect variances. With the classical model one must be careful when using the standard errors of the variance estimates to obtain confidence intervals, as they can be unreliable. The Bayesian model does not have this problem and it is simple to obtain densities/credible intervals for standard deviations, which are easier to interpret. An alternative way of looking at the data is to obtain the difference in standard deviations between atopics and non-atopics together with credible intervals. I feel that presenting the data in the form of either Table 4.13 or Table 4.14 or Figure 4.14 or Figure 4.15 make interpretation much more intuitive, especially for non-statisticians.

The models presented in this chapter could be extended in a number of ways. I have analysed PEF, but there are a number of other measures of lung function that could also be analysed using a three level hierarchical model. In fact there are known inaccuracies with the use of peak flow meters, especially in children (Sly *et al.*, 1994). With the recent introduction of small, electronic, portable devices that record peak flow and other measures of lung function, such as Forced Expiratory Volume in 1 second (FEV1), there is potential for more accurate measures (Hamid *et al.*, 1998). One possible extension is to consider a four-level model which would enable investigation of the correlation structure between different types of outcome measure (Beacon and Thompson, 1996)

The Bayesian model could be extended in the same way I extended the Bayesian models of ABPM in the previous chapter, where I allowed the within-subject variance to vary between subjects. In addition, it is plausible that PEF has longer tails than those defined by normality, so it may be of interest to explore the use of alternative distributions.

For the Bayesian model presented in this chapter I have used the Wishart distribution as a prior when I have assumed multivariate normality. In chapter 3 I used the product normal formulation in the analysis if ABPM data. In general I find the latter more intuitive than the Wishart distribution and so could be used for the models presented in this chapter.

Although the models are relatively easy to interpret, in a practical setting where a doctor is concerned with an individual patient the models are clearly of less use. However, because the models quantify how and when different groups of patients may differ in terms of the

PEF, the models could be used in construction or justifying suitable summary measures. For example, in the dataset I present, there is little difference in the between-day variation between atopic and non-atopic children. If one wants to differentiate between these two groups, then use of a summary measure that only assess within-day variation will be more powerful.

5 META-ANALYSIS USING HIERARCHICAL MODELS

5.1 Introduction

In this chapter I demonstrate how hierarchical models can be used for meta-analysis. The use of hierarchical models is slightly different to that in the previous two chapters, as often in meta-analysis information is only available in aggregate form, i.e. a study summary statistic with a standard error. In section 5.2 I give a brief introduction to meta-analysis. In section 5.3 I describe how hierarchical models can be used for meta-analysis by giving a fairly standard example that investigates the effect of lowering cholesterol on mortality of coronary heart disease, using both a classical analysis (section 5.3.1) and a Bayesian analysis (section 5.3.2). In section 5.4 I extend the methods shown in section 5.3 and show how meta-analysis can be applied when interest lies in estimation of the attributable risk. Section 5.4.1 introduces an example where interest lies in the effect of a history of infertility on perinatal mortality. Section 5.4.2 gives a classical analysis, while section 5.4.3 gives a Bayesian analysis that has a number of potential advantages including the synthesis of data from different types of studies. Finally in section 5.5, I discuss the models I have used and possible further research.

5.2 Meta-Analysis

Over the last 10-15 years or so the use of meta-analysis in medical research has grown extensively. Meta analysis can be defined as the quantitative synthesis of results from different studies, with the term first used by Glass (1976). The aim of a meta-analysis is to obtain a pooled estimate of an effect size that is more precise than that achieved by any individual study. One of the reasons why the use of meta-analysis has grown and is likely to continue to grow is due to the increasing emphasis on evidence-based-medicine (Sackett, 1996). Hand in hand with the growth of the use of meta-analysis has been the attention on analytical methods for meta-analysis. A comprehensive review of these methods can be seen in Sutton *et al.* (1998).

At the simplest level, meta-analysis assumes a fixed effects model in which each study is assumed to estimate an unknown overall population effect (Fleiss, 1993). In a fixed effects model it is assumed that all studies are estimating a single underlying effect size and that there is no *heterogeneity* between study results other than sampling error. However, it is often observed that there is considerable heterogeneity *between* individual studies with respect to their effect sizes. Reasons for this include, differences between studies due to dosage differences, different inclusion/exclusion criteria etc (Fleiss, 1993). Heterogeneity can be formally assessed in a number of similar ways (Dickersin *et al.*, 1992), with the method of Cochran (1954) the most commonly used. This tests whether the variation between studies is greater than that expected due to sampling error alone. A problem with assessing heterogeneity is that the statistical power of the test is low, due to the small number of studies often included in the meta-analysis (L'Abbe *et al.*, 1987). In the presence of heterogeneity a random effects model has been advocated (Dersimonian and Laird, 1986). In these models each study is assumed to be estimating its *own* unknown study effect. There is considerable controversy between the use of fixed and random effects models for meta-analysis. For example, see (Peto, 1987; Thompson and Pocock, 1991; Thompson, 1993). More important than just allowing for the heterogeneity between studies, is the practice of exploring reasons for its existence, which leads to the use of mixed effect models (Raudenbush, 1994; Thompson, 1994). These extend the random effects meta-analysis by incorporating fixed covariates that attempt to explain the differences between treatment effects, with the remaining variation modelled using a random component.

Meta-analysis fits naturally into a hierarchical structure in that individuals can be considered to be nested within-studies. It is therefore sensible to think about variation at both the individual (level 1) and the study (level 2) levels. Raudenbush and Byrk (1992) give five reasons why hierarchical models are useful for the analysis of meta-analysis data. These are:

1. to estimate the average effect size across a group of studies;
2. to estimate the variance of the effect size parameters;
3. to pose and test a series of linear models to explain variation in the effect size parameters;

4. to estimate the residual variance of the effect size parameters for each linear model; and
5. to use information from all studies to derive empirical Bayes estimates of each studies effect.

A characteristic of meta-analysis data is that often information is only available at the study level in the form of a summary measure with an associated standard error, for example an an odds ratio. This is not always the case and in chapter 6 I show how a meta-analysis can be performed using individual patient data (IPD). In this chapter I shall present two examples of meta-analysis, from both a classical and Bayesian perspective. The first introduces some of the basic concepts of meta-analysis and investigates the effect of lowering cholesterol on mortality. The second example investigates attributable risk of a history of infertility on perinatal mortality and introduces the problem of combining evidence from different types of studies. Both examples are analysed from a classical and Bayesian perspective.

5.3 Meta-Analysis of Cholesterol Data

An example of a meta-analysis is given by Davey-Smith, Song and Sheldon (1993) in which the effects of lowering blood serum cholesterol levels on mortality (both all-cause and cardiac) were assessed in 35 different randomised controlled trials. Various study-level covariates were collected, amongst the most important were thought to be baseline-risk, i.e. the cardiac mortality rate in the control group. The data can be seen in Table 5.1. I have used this data previously to demonstrate the application of hierarchical models in meta-analysis using MLn (Lambert and Abrams, 1995).

Study	Died/Total		Risk Group	Odds Ratio*	Log OR	Standard error (Log OR)
	Treatment	Control				
1	25/204	45/202	High Risk	0.494505	-0.7042	0.270396
2	62/285	35/147	High Risk	0.890172	-0.11634	0.239859
3	34/156	39/119	High Risk	0.578684	-0.547	0.273656
4	2/88	2/30	High Risk	0.333333	-1.09861	0.920135
5	0/30	2/33	High Risk	0.213559	-1.54384	1.569883
6	47/279	73/276	High Risk	0.568093	-0.56547	0.209459
7	37/206	50/206	High Risk	0.689691	-0.37151	0.242401
8	17/123	20/129	High Risk	0.886025	-0.12101	0.352768
9	97/1018	97/1015	High Risk	0.997827	-0.00218	0.150621
10	71/427	23/143	High Risk	1.031301	0.030822	0.260083
11	25/244	44/253	Medium Risk	0.54943	-0.59887	0.26649
12	13/50	10/50	Medium Risk	1.426614	0.355304	0.469755
13	13/47	5/48	Medium Risk	3.187246	1.159157	0.554857
14	0/30	4/60	Medium Risk	0.212806	-1.54737	1.50755
15	826/5552	632/2789	Medium Risk	0.596603	-0.5165	0.058873
16	41/424	50/422	Medium Risk	0.800298	-0.22277	0.221789
17	25/199	25/194	Medium Risk	0.976945	-0.02332	0.300103
18	34/350	35/367	Medium Risk	1.024196	0.023908	0.251638
19	2/79	4/78	Medium Risk	0.541031	-0.61428	0.805325
20	19/1149	31/1129	Medium Risk	0.602057	-0.5074	0.291244
21	35/221	26/237	Medium Risk	1.527386	0.423558	0.275672
22	8/54	1/26	Medium Risk	3.175824	1.155567	0.919258
23	5/71	6/72	Medium Risk	0.859072	-0.1519	0.604764
24	61/4541	54/4516	Medium Risk	1.124158	0.117034	0.187233
25	32/421	44/417	Medium Risk	0.702139	-0.35362	0.241839
26	0/94	1/94	Low Risk	0.333333	-1.09861	1.639495
27	17/311	8/317	Low Risk	2.171059	0.775215	0.42593
28	32/1906	44/1900	Low Risk	0.72371	-0.32336	0.233052
29	14/2051	19/2030	Low Risk	0.734461	-0.30862	0.348189
30	28/6582	3/1663	Low Risk	2.063205	0.724261	0.56706
31	91/5331	77/5296	Low Risk	1.176138	0.162237	0.15561
32	0/48	0/49	Low Risk	1.042105	0.041243	2.010179
33	1/94	0/52	Low Risk	1.702703	0.532217	1.642075
34	1/23	0/29	Low Risk	4.116279	1.41495	1.656807

*0.5 added to each cell when calculating odds ratios and corresponding standard errors.

Table 5.1 Cholesterol Data for use in Meta-Analysis

5.3.1 Classical Analysis

A simple random effects meta-analysis model (not allowing for study level covariates) can be defined as follows. Let y_i be the observed log odds ratio in the i^{th} study and s_i its observed standard error. The model can be written as

$$y_i = \beta_0 + \delta_i + e_i \tag{5.1}$$

where

β_0 is the estimate of the pooled log odds ratio

δ_i is the effect of the i^{th} study and is distributed $\delta_i \sim N(0, \sigma_\delta^2)$

e_i is the error associated with the i^{th} study where $E(e_i)=0$ and $Var(e_i)=\sigma_i^2$ which is estimated by s_i^2

The pooled estimate can be obtained using $\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y$ (see equation (2.20) in section 2.5.1), where the design matrix X will be a vector of 1's, Y a vector of the log-odds ratios and V a diagonal matrix with diagonal elements $\sigma_\delta^2 + \sigma_i^2$. The pooled estimate is thus

$$\hat{\beta}_0 = \left(\sum_{i=1}^{n} \frac{1}{\sigma_\delta^2 + \sigma_i^2} \right)^{-1} \sum_{i=1}^{n} \frac{y_i}{\sigma_\delta^2 + \sigma_i^2} \tag{5.2}$$

which is equivalent to equations (2) and (3) in Dersomonian and Laird (1986).

The variance of the pooled estimate can be obtained using $V(\hat{\beta}) = (X^T V^{-1} X)^{-1}$ (see equation (2.21)). The variance of the pooled estimate is thus

$$Var(\hat{\beta}_0) = \sum_{i=1}^{n} \frac{1}{\sigma_\delta^2 + \sigma_i^2} \tag{5.3}$$

which is equivalent to equation (4) in DerSimonian and Laird (1986).

To obtain model based estimates of the log-odds ratio the level 2 residuals need to be obtained using $R_2^T V^{-1} \tilde{Y}$ (see equation (2.33)) where R_2 is a diagonal matrix with identical elements σ_δ^2 and $\tilde{Y}$ is a vector of the differences between the observed log odds ratios and

Parameter	Estimate (SE)	OR (95% CI)
β_0 (pooled estimate)	-0.118 (0.061)	0.89 (0.79,1.00)
σ_s^2 (between-study varince)	0.0434	

Table 5.2 Parameter estimates for pooled estimate for meta-analysis of cholesterol data.

the pooled log odds ratio, $\hat{\beta}_0$. Thus the model based estimate ($\hat{\delta}_i$) for the log-odds ratio in the i^{th} study is given by subtracting these residuals from the pooled log-odds ratio ($\hat{\beta}_0$).

$$\hat{\delta}_i = \frac{\hat{\sigma}_s^2 y_i + \hat{\sigma}_i^2 \hat{\beta}_0}{\hat{\sigma}_s^2 + \hat{\sigma}_i^2} \tag{5.4}$$

The standard error of the model based estimate can be obtained by $R_2^T V^{-1} R_2$ and is thus

$$se(\hat{\delta}_i) = \sqrt{\frac{\hat{\sigma}_\delta^2 \hat{\sigma}_i^2}{\hat{\sigma}_\delta^2 + \hat{\sigma}_i^2} + \left(\sum_{i=1}^{n} \frac{1}{\hat{\sigma}_\delta^2 + \hat{\sigma}_i^2} \right)^{-1}} \tag{5.5}$$

It should be noted that like all classical hierarchical models, these standard errors take account of the uncertainty in the estimate of the fixed effects ($\hat{\beta}_0$), but not the random effects ($\hat{\sigma}_\delta^2$). Accommodating the uncertainty with regard to $\hat{\sigma}_\delta^2$ can be performed using classical methods (Hardy and Thompson, 1996; Biggerstaff and Tweedie, 1997), but can also be performed using a Bayesian analysis which will be demonstrated in section 5.3.2.

The results of fitting this model can be seen in Table 5.2. This shows that the estimate of the odds ratio is 0.89 (95% confidence interval 0.79 to 1.00) indicating a benefit in lowering cholesterol, but with some uncertainty of the benefit as the upper bound of the 95% confidence interval is 1. The between–study variance is estimated to be 0.0434.

The model can be extended to incorporate information on baseline risk. Baseline risk is defined as the number of deaths from coronary heart disease per 1000 person years for control subjects. Baseline risk was divided into three categories with three corresponding dummy covariates created. The three categories were,

HR - High Risk (>50 deaths)

MR - Medium Risk (>10-50 deaths)

LR - Low Risk (<10 deaths).

Model (5.1) can therefore be extended to a mixed effect meta-analysis by inclusion of the risk groups as fixed effect covariates.

$$y_i = \beta_1 HR_i + \beta_2 MR_i + \beta_3 LR_i + \delta_i + e_i \qquad (5.6)$$

where

β_1, β_2, β_3 are the estimates of the pooled log odds ratio in the three groups

δ_i is the effect of the i^{th} study and is distributed $\delta_i \sim N(0, \sigma_\delta^2)$

e_i is the error associated with the i^{th} study where $E(e_i)=0$ and $Var(e_i)=\sigma_i^2$ which is estimated by s_i^2

The results of fitting this model can be seen in Table 5.3 There appears to be some benefit in lowering cholesterol in the high risk group, less in the medium risk group and a detrimental effect in the low risk group. However, the confidence intervals in the latter two categories include one. Note that the between-study variance has reduced, which is to be expected as some of the between-study heterogeneity is being explained by including information on baseline risk.

Figure 5.1 shows the observed and model based estimates of the log odds ratios and 95% confidence intervals for each study together with the pooled estimates. It can be seen that the model based estimates are shrunk towards their corresponding overall group estimate.

Parameter	Estimate (SE)	OR (95%CI)
β_1 (low risk group)	0.162 (0.121)	1.18 (0.93 to 1.49)
β_2 (medium risk group)	-0.130 (0.077)	0.88 (0.76 to 1.02)
β_3 (high risk group)	-0.284 (0.099)	0.75 (0.62 to 0.91)
σ_s^2 (between-study variance)	0.027	

Table 5.3 Parameter estimates for mixed model meta-analysis of cholesterol data.

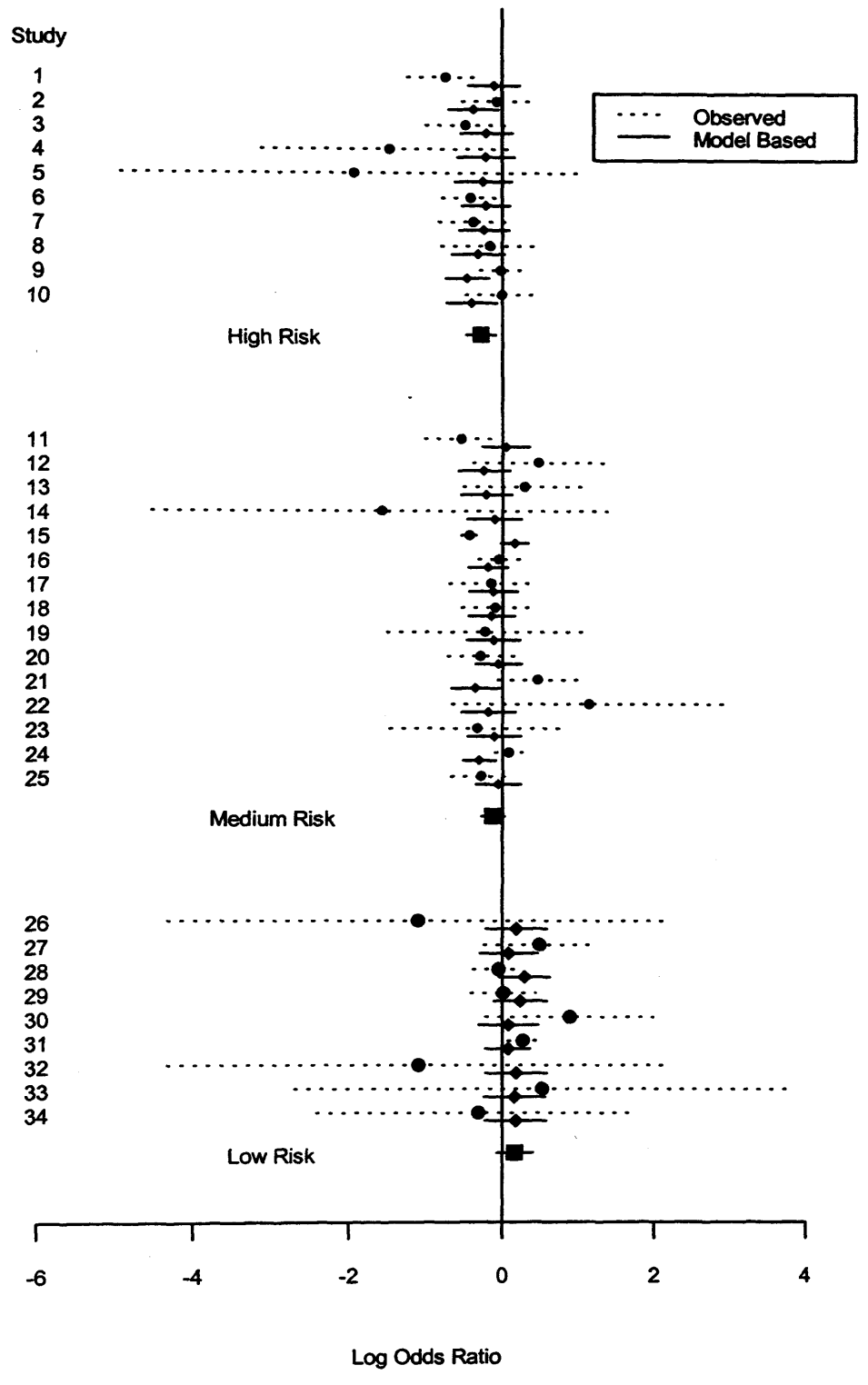

Figure 5.1 **Observed and model based estimates for log odds ratios with 95% confidence intervals for the cholesterol data.**

5.3.2 Bayesian Analysis

As I have mentioned previously, one advantage of using a Bayesian approach for analysing hierarchically structured data is that the uncertainty in the estimates of the variance components are automatically taken into account when estimating the fixed effects and their standard errors. With large datasets this is generally not important, as seen in the repeated measures examples in chapters 3 and 4. However, a meta-analysis is often performed with under ten studies and the issue becomes much more important (DuMouchel and Harris, 1983).

The simple classical model in equation (5.1) can be written as follows,

$$\begin{aligned} y_i &\sim N(\delta_i, s_i^2) \\ \delta_i &\sim N(\beta_0, \sigma_\delta^2) \end{aligned} \tag{5.7}$$

Prior distributions need to be specified for β_0 and σ_δ^2. These are generally vague and I have defined these as follows

$$\begin{aligned} \beta_0 &\sim N(0,10000) \\ \frac{1}{\sigma_\delta^2} &\sim Gamma(0.001,0.001) \end{aligned} \tag{5.8}$$

The above model has assumed that the outcome (log odds ratio) is normally distributed. However, the raw data shown in Table 5.1 gives the number of events and the total number of subjects in the control and treatment arms. Therefore, as an alternative to modelling the calculated log odds ratios, it has been suggested that it may be more sensible to model the raw data (Skene and Wakefield, 1990; Smith *et al.*, 1995). In this case a Binomial distribution can be assumed, with the log odds ratio obtained through the use of a logit transformation. Let r_{Ci} and r_{Ti} denote the observed number of events in the control and treatment arms for the i^{th} study, and n_{Ci} and n_{Ti} the corresponding total number of individuals in each arm.

The model can then be defined a follows.

$$\begin{aligned} r_{Ci} &\sim Binomial(\pi_{Ci}, n_{Ci}) \qquad r_{Ti} \sim Binomial(\pi_{Ti}, n_{Ti}) \\ \text{logit}(\pi_{Ci}) &= \mu_i \qquad \text{logit}(\pi_{Ti}) = \mu_i + \delta_i \\ \delta_i &\sim N(\beta_0, \sigma_\delta^2) \end{aligned} \tag{5.9}$$

Thus, μ_i is the log odds in the control arm of the i^{th} study and δ_i is the treatment effect in the i^{th} study. Prior distributions can be specified as follows,

$$\begin{aligned} \mu_i &\sim N(0,10000) \\ \beta_0 &\sim N(0,10000) \\ \frac{1}{\sigma_\delta^2} &\sim Gamma(0.001,0.001) \end{aligned} \tag{5.10}$$

Using this formulation 0.5 does not have to be added when there are zero events as was done for both the classical and Bayesian models of the calculated odds ratios. This can be important for rare events and/or small studies.

The results of these two Bayesian models can be seen in Table 5.4, where the estimates are compared with those from the classical analysis.

It can be seen that the estimates of the pooled log odds ratios are broadly similar, with the standard errors for the Bayesian models being slightly larger as expected. The estimate of the standard error for the Bayesian model using the Binomial formulation is larger than using the normal formulation. One reason for this is due to the estimate of the between-study variance being larger using the Binomial formulation.

Both models (5.7) and (5.9) can be extended to incorporate information on baseline risk as in model (5.6). Thus model (5.7) becomes

$$\begin{aligned} y_i &\sim N(\delta_i, s_i^2) \\ \delta_i &\sim N(\beta_1 HR_i + \beta_2 MR_i + \beta_3 LR_i, \sigma_\delta^2) \end{aligned} \tag{5.11}$$

Parameter	Classical	Bayesian Normal formulation	Bayesian Binomial Formulation
β_0	-0.118 (0.061)	-0.121 (0.064)	-0.160 (0.080)
σ_δ^2	0.0434 (0.0245)	0.0540	0.089

Table 5.4 Comparison of estimates of classical and two Bayesian models for meta-analysis of cholesterol data.

and model (5.9) becomes

$$\begin{aligned} r_{Ci} &\sim Binomial(\pi_{Ci}, n_{Ci}) \qquad & r_{Ti} &\sim Binomial(\pi_{Ti}, n_{Ti}) \\ \text{logit}(\pi_{Ci}) &= \mu_i & \text{logit}(\pi_{Ti}) &= \mu_i + \delta_i \\ & & \delta_i &\sim N(\beta_1 HR_i + \beta_2 MR_i + \beta_3 HR_i, \sigma_\delta^2) \end{aligned} \tag{5.12}$$

In both of these model the prior distribution for the β's are

$$\beta_1, \beta_2, \beta_3 \sim N(0,10000) \tag{5.13}$$

The results of fitting the two Bayesian models incorporating baseline risk can be seen in Table 5.5. It can be seen that as in the previous table the Bayesian models give larger standard errors for the pooled log odds ratios, with the Binomial formulation giving the largest standard errors. Again the between-study variance is larger using the Binomial formulation.

5.3.3 *Potential bias in the use of baseline risk in meta-analysis*

Since I originally analysed this data (1995) there has been further work on the use of baseline risk in exploring heterogeneity in meta-analysis (Walter, 1997; Thompson *et al.*, 1997). These two papers demonstrate that simplistic use of baseline risk as a covariate can lead to bias. In order to explain these biases and solutions to them I will reformulate the model by treating baseline risk as continuous covariate rather than categorising it into risk groups. Thus, the classical model in (5.6) becomes

Parameter	Classical	Bayesian Normal formulation	Bayesian Binomial Formulation
β_1 (High Risk)	-0.284 (0.099)	-0.287 (0.101)	-0.323 (0.135)
β_2 (Medium Risk)	-0.130 (0.077)	-0.122 (0.087)	-0.143 (0.118)
β_3 (Low Risk)	0.162 (0.121)	0.167 (0.133)	0.111 (0.189)
σ_δ^2	0.027	0.040	0.086

Table 5.5 Comparison of estimates classical and two Bayesian models for meta-analysis of cholesterol data incorporating baseline risk.

$$y_i = \beta_0 + \beta_1 x_i + \delta_i + e_i \qquad (5.14)$$

Where

y_i is the log odds ratio in the i^{th} study.

x_i is the log odds of the event rate in the control group.

β_0 is the estimate of the intercept

β_1 is the estimate of the gradient

δ_i is the effect of the i^{th} study and is distributed $\delta_i \sim N(0, \sigma_\delta^2)$

e_i is the error associated with the i^{th} study where $E(e_i)=0$ and $Var(e_i)=\sigma_i^2$ which is estimated by s_i^2

The results of fitting this model classically can be seen in Table 5.6. It can be seen that the absolute value of the gradient estimate is substantially larger than its standard error indicating strong evidence of relationship between baseline risk (on the log odds scale) and the log odds ratio. The between study variation is smaller than that found in the model categorising the baseline risk into three groups (0.027) indicating that categorisation is less efficient. The relationship between the odds ratio and baseline risk is best examined graphically. Figure 5.2 shows a plot of the log odds ratio against the log odds in the control group with the fitted line obtained from (5.14). The size of the plotting symbol for each study is inversely proportional to the standard error of the log odds ratio. It can be seen how the odds ratio decreases as the event rate in the control group increases. It is also apparent that the fitted line will be dominated by a few large studies.

The model defined above can lead to bias for the following reasons. Firstly, the model fails to take into account the fact that the explanatory covariate, the event rate log odds in the control group, is subject to sampling error. Secondly, the log odds in the control group is

Parameter	Estimate (SE)
β_0 (Intercept)	-0.508 (0.087)
β_1 (Gradient)	-0.153 (0.031)
σ_s^2 (between-study variance)	0.006

Table 5.6 Parameter estimates of classical basline risk model for cholesterol data.

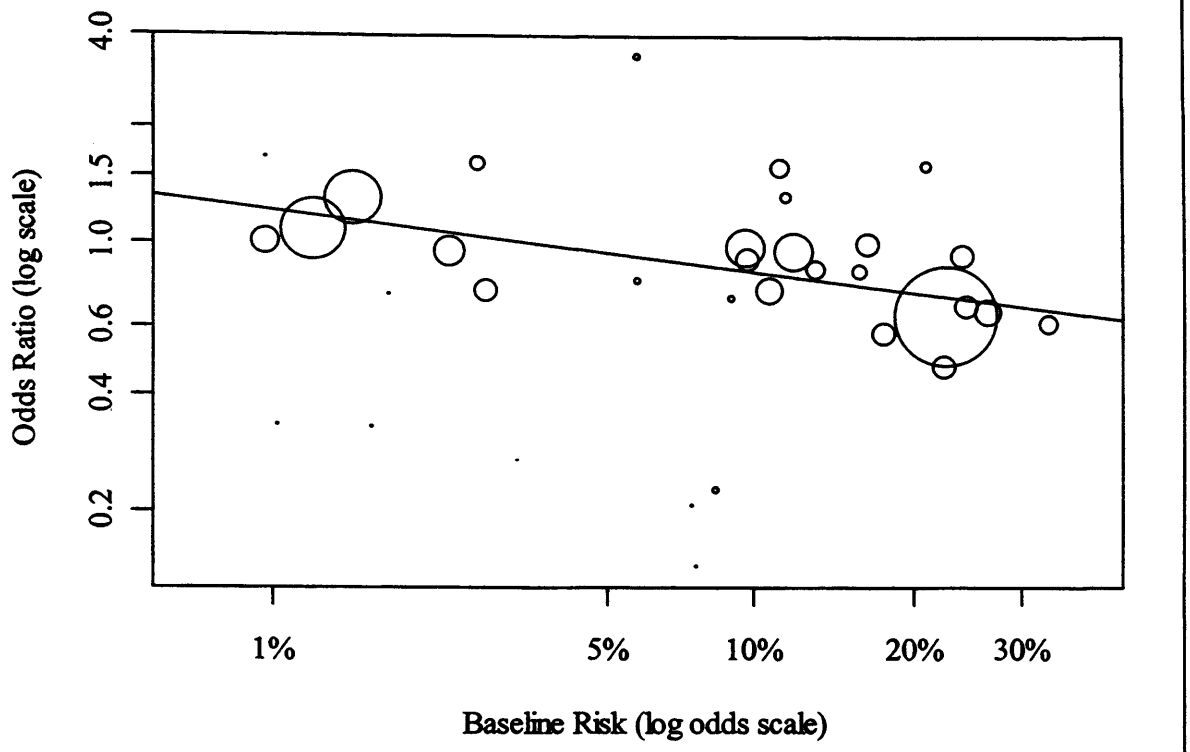

Figure 5.2 Plot of log odds ratio against control group log odds for cholesterol data.

part of the function used to derive the log odds ratio. In fact Thompson *et al.* (1997) show that one would expect there to be a relationship between the log odds ratio and the log odds of the event rate in the control group even if there was no real effect.

In order to circumvent these problems Thompson *et al.* (1997) suggested an extension to the Bayesian model described in (5.12)

$$\begin{aligned} r_{Ci} &\sim Binomial(\pi_{Ci}, n_{Ci}) & r_{Ti} &\sim Binomial(\pi_{Ti}, n_{Ti}) \\ \text{logit}(\pi_{Ci}) &= \mu_i & \text{logit}(\pi_{Ti}) &= \mu_i + \delta_i \\ & & \delta_i &= \delta_i^{'} + \beta(\mu_i - \hat{\mu}) \\ & & \delta_i^{'} &\sim N(\delta, \sigma_\delta^2) \end{aligned} \qquad (5.15)$$

In the above model the treatment effect in the i^{th} study, δ_i, depends on the log odds of the event rate in the control group, μ_i, through a linear relationship. The subtraction of the mean for the gradient effect (β) is to improve convergence by reducing autocorrelation.

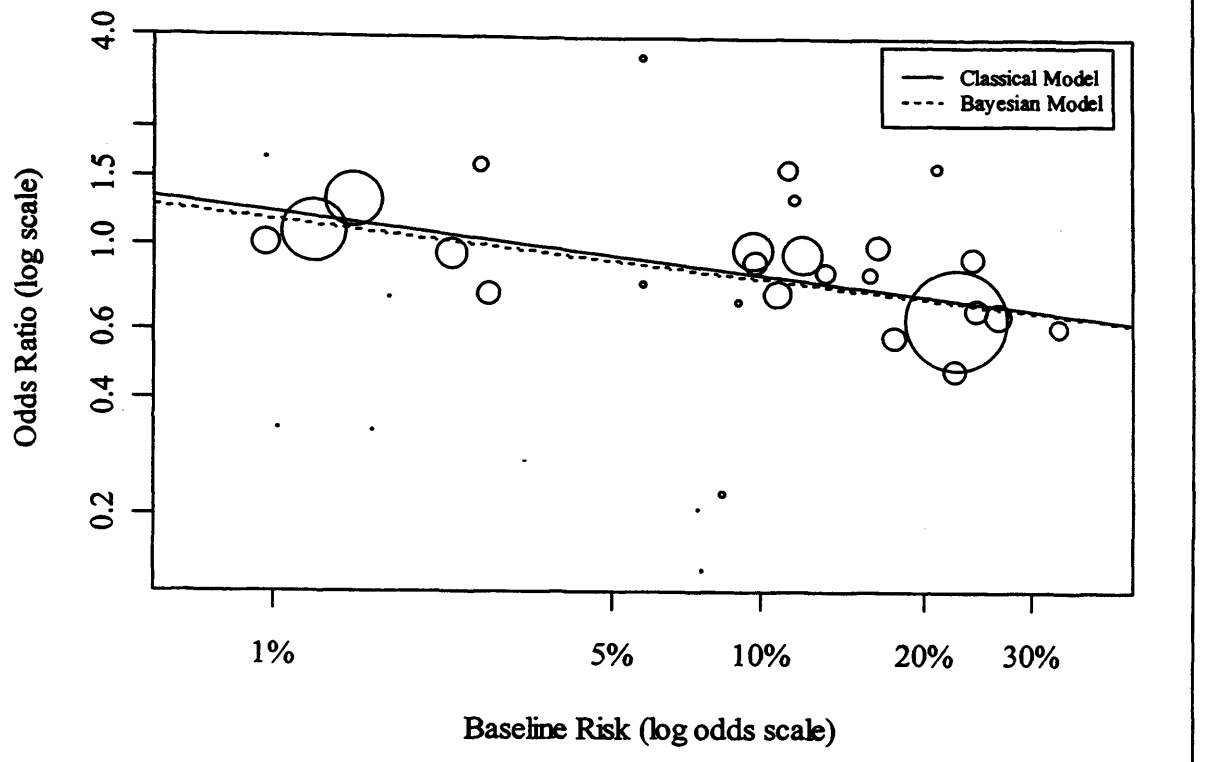

Figure 5.3 Plot of log odds ratio against control group log odds for cholesterol data for both Classical and Bayesian models.

Caclulating the mean of a stochastic variable in WinBUGS slows down the iteration procedure, so I subtracted a scaler (-3.5), that approximated the mean of the μ_i's

Vague priors are specified as follows

$$\begin{aligned} \mu_i &\sim N(0,10000) \\ \beta, \delta &\sim N(0,10000) \\ \frac{1}{\sigma_\delta^2} &\sim Gamma(0.001,0.001) \end{aligned} \qquad (5.16)$$

The results of fitting this model compared to the classical model can be seen in Table 5.7. The estimate of gradient is slightly reduced in the Bayesian analysis, but the standard error

Parameter	Classical Analysis Estimate (SE)	Bayesian analysis Estimate (SE)
β_1 (Gradient)	-0.153 (0.031)	-0.144 (0.0481)
σ_s^2 (between-study variance)	0.006	0.045 (0.039)

Table 5.7 Comparison of slope and between study variance for Classical and Bayesian models incorporating baseline risk.

is over 50% larger. This is mainly due to the fact that the uncertainty in the estimate of the explanatory covariate is now taken into account.

The relationship between the odds ratio and baseline risk can again be examined graphically. Figure 5.3 shows a plot of the log odds ratio against the log odds in the control group with the fitted line obtained from both the classical and Bayesian models. It can be seen how the odds ratio decreases as the event rate in the control group increases, and that the fitted lined are similar for both estimation methods. Thompson *et al.* (1997) found relatively large differences between his Bayesian model and the naïve classical model. The reason for the similarity in this case is that there is considerable variability in the baseline risk in the control groups and that the fitted regression lines are dominated by a few large trials.

5.4 Meta-analysis of Attributable Risk

In epidemiology a commonly used measure is the relative risk, which measures the excess risk when exposed to a particular factor. However, the full implications of excess risk will depend not only on the magnitude of the relative risk, but also on the proportion of the population who are exposed to the factor of interest. A moderate relative risk with a high level of exposure may produce more cases than a high relative risk with a low level of exposure. An example given by Walter (1976) is where the risk of lung cancer can be about 40 to 50 times higher in industrial workers exposed to certain types of chemicals. However, although smoking has a lower relative risk for lung cancer (between 5 and 10), it causes many more cases of disease, as smoking is a much more common exposure and thus has a larger impact on the population as a whole.

A measure of association that takes into account both the magnitude of the relative risk and the proportion of the population exposed is the attributable risk (Levin, 1958), here defined as λ.. It can be defined as

$$\lambda = \frac{\theta(\phi - 1)}{1 + \theta(\phi - 1)} \qquad (5.17)$$

where θ is the proportion of the population exposed to the exposure of interest and ϕ is the relative risk. The attributable risk is also known as the population attributable risk, aetiologic fraction and attributable fraction.

When it is possible to infer causation for a particular exposure, the AR can be interpreted as the proportion of cases in a population that are attributable to the risk factor in question. Knowledge of the effect on the population of a particular exposure makes the attributable risk a potentially very useful measure for developing public health and health service provision policy. For example, in 1992 the results of a case-control study that investigated risk factors for Sudden Infant Death Syndrome (SIDS) was published (Mitchell *et al.*, 1992). The most important risk factor observed was prone sleeping position (on baby's front), where an estimate of the relative risk was 3.70 and an estimate of the prevalence was 33%. This leads to an estimate of the attributable risk of 0.47 indicating that if the prone sleeping position is a causal risk factor one could expect almost a 50% reduction in the incidence of sudden infant death syndrome if babies were no longer placed to sleep on their front. The findings in this study prompted the introduction of 'The Back to Sleep' campaign in the UK with similar campaigns in New Zealand and the rest of the world. In 1990, before the campaign began, the incidence of SIDS in the UK was about 2 deaths per 1000 births and in New Zealand about 4 deaths per 1000 births. In 1996, after the campaigns had been running, the incidence was reduced to about 0.7 deaths per 1000 births in the UK and 1.9 deaths per 1000 births in New Zealand, a reduction of over 50%.

The attributable risk can be estimated from case-control studies, where the odds ratio is used to estimate the relative risk and the proportion of the population exposed to risk factor estimated from the control group (Breslow and Day, 1980). This assumes that the controls are representative of the population of interest, which may not always be the case, for example if they are hospital controls. It is of course necessary to quantify the uncertainty of the estimate of the attributable risk and when calculating its variance from a case-control study it is important to account for the covariance between the odds ratio and the prevalence. Details of different methods of calculating variances of attributable risk obtained from case-control studies can be found in Whittemore (1983).

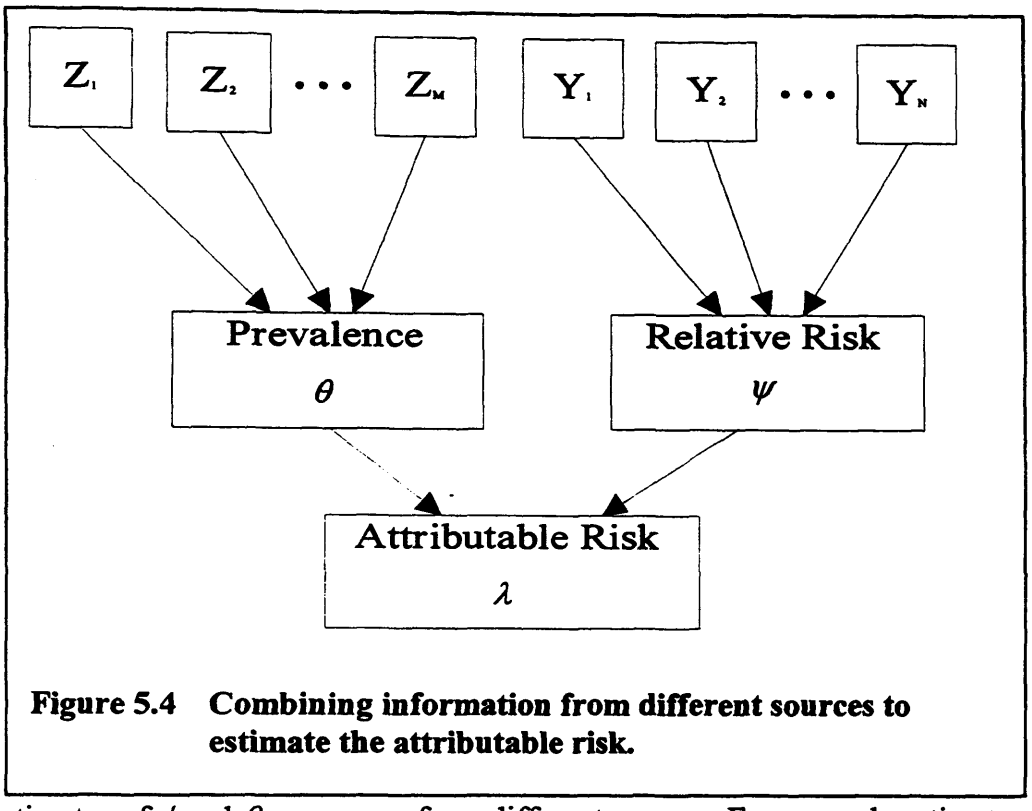

Figure 5.4 Combining information from different sources to estimate the attributable risk.

Often estimates of ϕ and θ may come from different sources. For example estimates of ϕ may come from clinical trials, case-control studies or cohort studies and the estimate of θ may come from prevalence studies or the control group of a case-control study. It may not always be appropriate to use the control group from a case-control study to estimate θ as they may be specially selected (e.g. hospital controls) or from a different geographical area than that to which the estimate of the AR will apply. Figure 5.4 shows how the estimate of the AR may come from a number of different sources with $Z_1,\ldots,Z_M$ representing M independent studies that estimate the prevalence, θ, of the exposure of interest and $Y_1,\ldots,Y_N$ represent N independent studies that estimate the relative risk, ψ. Both the prevalence studies and the relative risk studies will generally provide point estimates and standard errors. From Figure 5.4, it can be seen that there are two independent meta-analyses where the two pooled estimates are then combined to obtain an estimate of the attributable risk. When estimating the attributable risk it is important to account for the uncertainty associated with both the estimates of the prevalence *and* the relative risk. I will show how the pooled estimates of the two meta-analyses can be combined from both a classical and Bayesian perspective and then develop the Bayesian model further to situations where case-control studies can contribute to both the estimate of the prevalence and the relative risk.

	Year	Infertile Group	Fertile Group	RR (95% CI)
Cohort Studies				
Maresh *et al.*	1982	3/123	45/3568	1.94 (0.59,6.38)
Varma *et al.*	1988	5/464	75/7348	1.06 (0.42,2.67)
Beral *et al.*	1990	43/1581	6463/659490	2.78 (2.04,3.77)
Rizk *et al.*	1991	22/961	6463/659490	2.34 (1.53,3.59)
Venn *et al.*	1993	49/1465	680/61253	3.01 (2.23,4.04)
Case-Control Studies		Cases	Controls	
Draper *et al.*	1999	65/567	34/972	3.57 (2.32,5.48)

Table 5.8 Cohort studies and case-control study investigating relationship between infertility and perinatal mortality.

5.4.1 Example

In the early 1980's research was undertaken to explore the possibility that a history of infertility may lead to increased risk of perinatal mortality, i.e. a still birth or death in the first week of life. Infertility is a biological condition that leads to the social condition of childlessness. There are problems with the definitions of infertility, but a common definition is "failure to conceive a clinically recognised pregnancy by a couple having regular sexual intercourse for at least a year without the use of contraception" (Hammond, 1994). Some women who have a history of infertility will eventually become pregnant and it is these women who it is hypothesised have a higher risk of perinatal mortality. Table 5.8 shows six studies that have estimated the relative risk perinatal mortality comparing women who had a history of infertility and those that were fertile. The six studies consisted of five cohort studies and one case-control study. Given in the table is the relative risk estimates with 95% confidence intervals, with a relative risk>1 indicating the increased risk of perinatal mortality associated with infertility.

The relative risks and confidence intervals are shown in a forest plot in Figure 5.5. In addition two pooled estimates are shown using both a Maximum Likelihood (ML) random effects model as in section 5.3.1, and a Bayesian hierarchical model as in section 5.3.2. The pooled estimates shown have been transformed back from the log relative risk scale. For the Bayesian model the same vague priors as in equation (5.8) were used. It can be seen that the pooled estimates are similar with the Bayesian model having wider credible

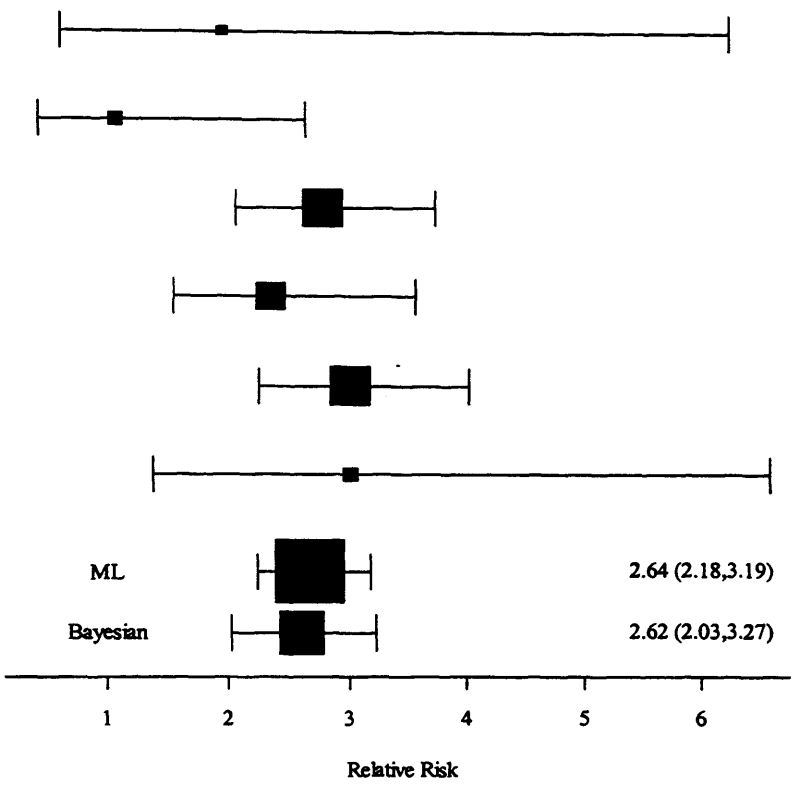

Figure 5.5 Forest plot for Relative Risks of perinatal mortality with pooled estimates obtained from maximum likelihood and Bayesian models.

intervals. This is sensible as there are only six studies and the Bayesian model appropriately takes into account the uncertainty associated with the estimate of the between-study variance. The pooled relative risks indicate that having a history of infertility leads to over a doubling of the risk of perinatal mortality.

The prevalence of infertility may vary from region to region and country to country. When estimating the prevalence in calculating an attributable risk, it is important that it applies to the population of interest. In this case the population of interest is that in Leicestershire. However, it appears sensible to assume that the prevalence in Leicestershire is similar to that in the rest of the UK. Table 5.9 shows the estimated prevalence of infertility in six UK studies together with 95% confidence intervals.

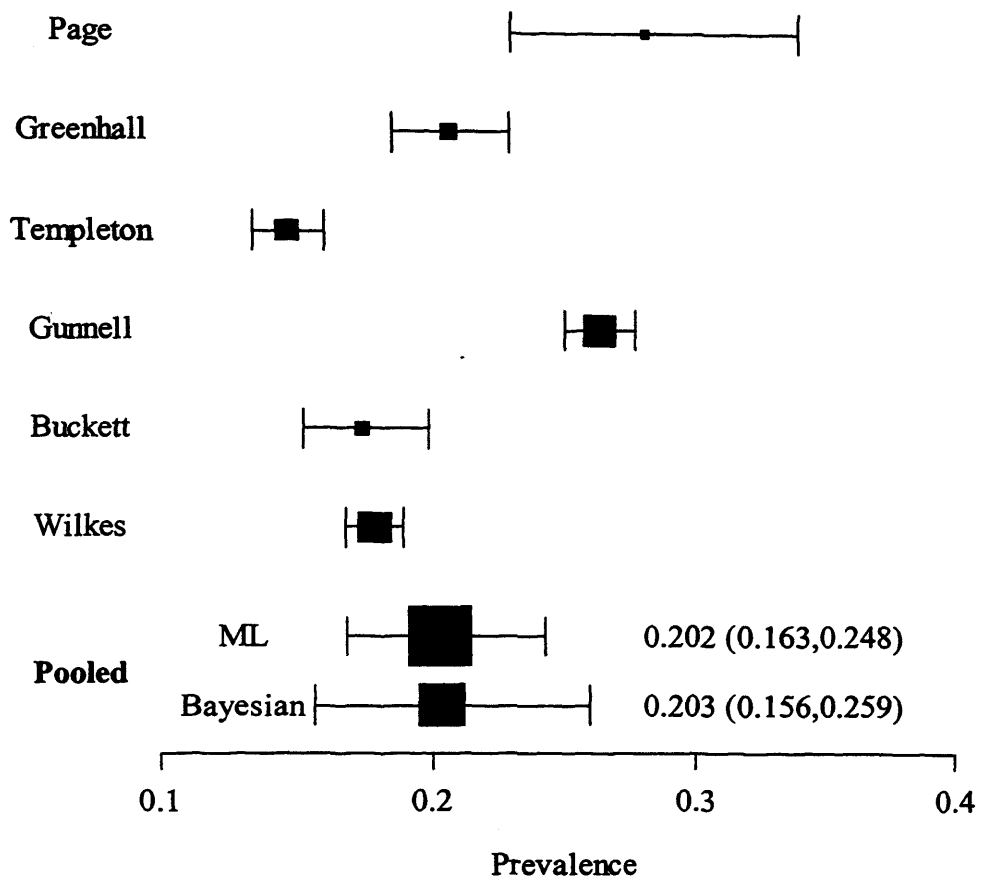

Figure 5.6 Forest plot for prevalence estimates of infertility with pooled estimates obtained from maximum likelihood and Bayeisan models.

Figure 5.6 shows a forest plot for the prevalence estimates. In addition to the prevalence estimates from the six studies the figure shows two pooled estimates that have been transformed back from the logit scale. As with the relative risks these are for a Maximum Likelihood (ML) random effects model and a Bayesian hierarchical model. Both pooled

Author	Year	r/n	Risk (95% CI)
Page	1989	43/153	28.0% (21.0%,35.0%)
Greenhall *et al.*	1990	179/872	20.5% (18.0%,23.0%)
Templeton *et al.*	1991	293/2008	14.6% (13.1%,16.1%)
Gunnell *et al.*	1994	628/2377	26.4% (24.6%,28.2%)
Buckett *et al.*	1995	126/728	17.3% (14.6%,20.0%)
Wilkes *et al.*	1995	623/3500	17.8% (16.5%,19.1%)

Table 5.9 Estimates of prevalence of infertility in six UK studies.

estimates are in broad agreement with again the Bayesian credible interval being wider. What is of interest is the large amount of variation between studies. The two largest studies have confidence intervals that do not overlap. It can be seen that because of the variation between studies, the pooled estimates have a relatively large amount of uncertainty associated with them, with the Bayesian credible interval being wider than any individual study. If a fixed effects meta-analysis were used, which assumes no heterogeneity between studies, the pooled estimate would be 0.203 (95% CI 0.188 to 0.220). It can be seen that the confidence interval is much narrower than when using a random effects model. The reasons for the variation in the estimation of the prevalence of infertility is unclear as all use similar, if not exact, definitions of infertility. It is clear that there is a great deal of uncertainty in what the true estimate of the prevalence of infertility is. However, is important that this uncertainty is appropriately taken into account when estimating the attributable risk.

5.4.2 Classical Analysis

In order to obtain an estimate of λ, the attributable risk, it is possible to substitute the pooled estimates of the relative risk and the prevalence into (5.17). To obtain the standard error, and hence a confidence interval, it is possible to use the delta method (Cox and Hinkley, 1974). The delta method can be used to obtain the variance of a function of two random variables. Initially I will not use the controls from the case-control study (Draper *et al.*, 1999) to estimate the prevalence as this leads to complications in a simple analysis, as there will be a covariance term between the estimate of the relative risk and the prevalence from this study. In section 5.4.3 I will show how the prevalence estimate from this case-control can be used in the pooled estimate of the prevalence as well as the relative risk. The delta method is as follows,

$$Var(f(x,y)) = \left(\frac{df(x,y)}{dx}\right)^2 Var(x) + \left(\frac{df(x,y)}{dy}\right)^2 Var(y) + \left(\frac{df(x,y)}{dx}\right)\left(\frac{df(x,y)}{dx}\right) Cov(x, \quad (5.18)$$

In the calculation of the variance of the attributable risk, the covariance term can be ignored if the estimates of the prevalence and the relative risk come from independent studies. I will assume this is the case in my initial analysis.

Differentiating equation (5.17) with respect to both θ and ψ gives

$$\frac{d\lambda}{d\theta} = \frac{\phi - 1}{(\theta(\phi-1)+1)^2} \qquad \frac{d\lambda}{d\phi} = \frac{\theta}{(\theta(\phi-1)+1)^2} \tag{5.19}$$

substituting into equation (5.18) gives

$$Var(\lambda) = \frac{(\phi-1)^2 Var(\theta) + \theta^2 Var(\phi)}{(\theta(\phi-1)+1)^4} \tag{5.20}$$

It is more convenient to express the above in terms of Var(log(ψ)). Again using the delta method,

$$Var(\psi) = Var\left(e^{\ln(\phi)}\right) = \phi^2 Var(\ln(\phi)) \tag{5.21}$$

and

$$Var(\theta) = Var\left(\frac{e^{\text{logit}(\theta)}}{1+e^{\text{logit}(\theta)}}\right) = Var\left(\left[1+e^{-\text{logit}(\theta)}\right]^{-1}\right) = \theta^2(1-\theta)^2 Var(\text{logit}(\theta)) \tag{5.22}$$

giving

$$Var(\lambda) = \frac{(\phi-1)^2\theta^2(1-\theta)^2 Var(\text{logit}(\theta)) + \theta^2\phi^2 Var(\ln(\phi))}{(\theta(\phi-1)+1)^4} \tag{5.23}$$

Using the pooled ML estimates together with their standard errors given in Figure 5.5 and Figure 5.6 gives an estimate of λ of 0.248 with variance 0.00104 and hence 95% confidence interval 0.185 to 0.311. Thus, about 25% of perinatal deaths are associated with having a history of infertility.

5.4.3 Bayesian Analysis

All that is occurring in the analysis above are two simultaneous meta-analyses, where the two pooled estimates are combined in a function to obtain the estimate of the attributable risk. It is possible to adopt a Bayesian approach to estimation of the attributable risk in a similar way to how the classical meta-analysis was extended in section 5.3.2. Let Z_i be the log-odds of the prevalence of infertility in the i^{th} prevalence study and Y_j be the log relative risk in the j^{th} relative risk study. A hierarchical model can be defined as follows,

$$Z_i \sim N(\delta_i, \sigma^2_{Z,i}) \qquad Y_j \sim N(\gamma_j, \sigma^2_{Y,i})$$

$$\delta_i \sim N(\rho, \sigma^2_\delta) \qquad \gamma_j \sim N(\psi, \sigma^2_\gamma)$$

$$\theta = \frac{e^\rho}{1+e^\rho} \qquad \phi = e^\psi \qquad (5.24)$$

$$\lambda = \frac{\theta(\phi-1)}{1+\theta(\phi-1)}$$

As with the more standard meta-analyses vague prior distributions are used, so

$$\rho \text{ and } \psi \sim N(0,1000000)$$

$$\frac{1}{\sigma^2_\delta} \text{ and } \frac{1}{\sigma^2_\gamma} \sim \text{Gamma}(0.001,0.001) \qquad (5.25)$$

The DAG for this model can be seen in Figure 5.7. This clearly shows that there are two independent meta-analyses where the two pooled estimates are then combined. The model specified in (5.24) and (5.25) was fitted using WinBUGS with a 'burn in' of 1000 iterations and a sample of 5000 iterations. The results of this analysis can be seen in Table 5.10

It can be seen that the estimates of the pooled relative risk and the pooled prevalence are identical to that obtained in the two separate analyses shown in Figure 5.5 and Figure 5.6. The estimate of λ is 0.24, which is similar that obtained in the ML analysis. However, the credible intervals are wider when compared to the ML confidence intervals, which is to be expected as the Bayesian model takes account of all uncertainty specified in the model including the two between-study random effects.

Parameter	Estimate (median)	95% Credible Interval
θ (Prevalence)	0.202	0.154 to 0.260
ϕ (Relative Risk)	2.583	1.929 to 3.274
λ (Attributable Risk)	0.240	0.150 to 0.332

Table 5.10 Parameter estimates for attributable risk hierarchical model

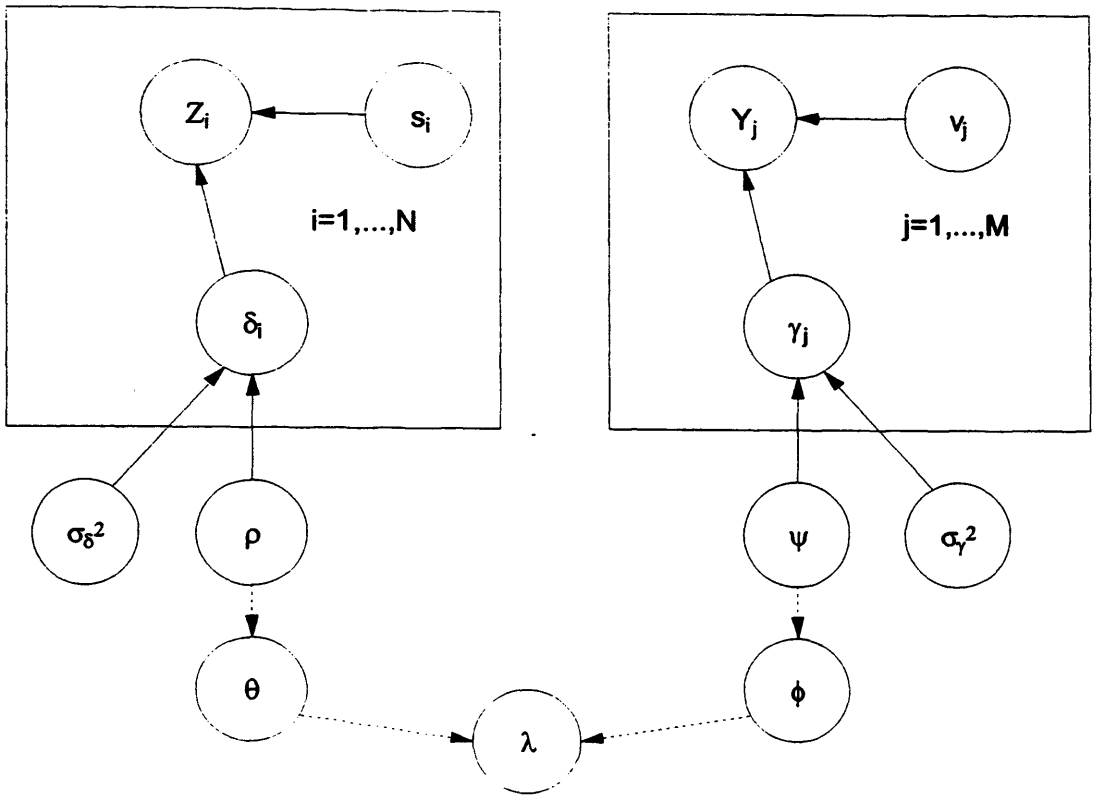

Figure 5.7 DAG for Hierarchical Model for calculation of attributable risk.

The above Bayesian analysis is essentially a simple extension of the classical analysis, but importantly does allow for the uncertainty associated with the estimates of the between-study variances. However, a further advantage of adopting a Bayesian model is that further complexity can be incorporated relatively easily. In this case, instead of modelling the summary measures, it is possible to model the observed frequencies directly. In doing this it is possible to use both the relative risk estimate and the prevalence estimate from the case-control study in the estimation of the attributable risk. For the cohort studies the numbers having a perinatal death can be modelled using a Poisson distribution in each of the two groups with the total number in each group as an offset. For this part of the model let $y_{E,i}$, $n_{E,i}$ and $y_{\underline{E},i}$, $n_{\underline{E},i}$ denote the number of perinatal deaths and the total number of women in the exposed and non-exposed groups respectively, for the i^{th} cohort study. Similarly prevalence studies can be modelled using a binomial distribution with $r_{P,j}$ denoting the number of infertile women out of $n_{P,j}$ for the j^{th} prevalence study. The case-

control study can be modelled with both the case and control groups being assigned a binomial distribution with $r_{D,k}$, $n_{D,k}$ and . $r_{\underline{D},k}$, $n_{\underline{D},k}$ denoting the number of those exposed and the total number in the case (diseased) and control (non-diseased) groups respectively for the k^{th} study (here k=1). When using the frequencies, each study type will need to be modelled separately, but with the case-control study linking to the cohort studies through the estimate of the relative risk and to the prevalence studies through the prevalence estimate from the controls. The model can be defined in the general case where there are M (j=1,...,M) cohort studies, L (k=1,...,L) case-control studies and N (j=1,...,N) prevalence studies as follows,

$$
\begin{array}{lll}
y_{E,i} \sim P(\alpha_{E,i}) & r_{\mathrm{D},k} \sim Bin(\pi_{\mathrm{D},k}, n_{\mathrm{D},k}) & r_{\mathrm{P},j} \sim Bin(\omega_j, n_{\mathrm{P},j}) \\
y_{\underline{E},i} \sim P(\alpha_{\underline{E},i}) & r_{\underline{\mathrm{D}},k} \sim Bin(\pi_{\underline{\mathrm{D}},k}, n_{\underline{\mathrm{D}},k}) & \\
\log(\alpha_{E,i}) = \log(n_{E,i}) + \mu_i + \tfrac{1}{2}\gamma_i & \mathrm{logit}(\pi_{D,k}) = \delta_{k+N} + \gamma_{k+M} & \mathrm{logit}(\omega_j) = \delta_j \\
\log(\alpha_{\underline{\mathrm{E}},i}) = \log(n_{\underline{E},i}) + \mu_i - \tfrac{1}{2}\gamma_i & \mathrm{logit}(\pi_{\underline{\mathrm{D}},k}) = \delta_{k+N} & \\
 & \delta_{k+N} \sim N(\rho, \sigma_\delta^2) & \delta_j \sim N(\rho, \sigma_\delta^2) \\
\gamma_i \sim N(\psi, \sigma_\gamma^2) & \gamma_{k+M} \sim N(\psi, \sigma_\gamma^2) &
\end{array}
\qquad (5.26)
$$

$$\phi = \exp(\psi) \qquad \theta = \frac{\exp(\rho)}{1+\exp(\rho)}$$

$$\lambda = \frac{\theta(\phi-1)}{1+\theta(\phi-1)}$$

This model when written down appears complex, and it is in situations such as these when DAG becomes very useful in order to simplify the explanation of the model. The DAG for the model can be seen in Figure 5.8. It can be seen that there are now three meta-analyses for the three different study types. What is important to observe about the model and the DAG is that both the cohort studies and case-control studies contribute to γ, the between-study random effect associated with the log relative risk, and that both the prevalence studies and the case-control studies contribute to δ, the between-study random effect

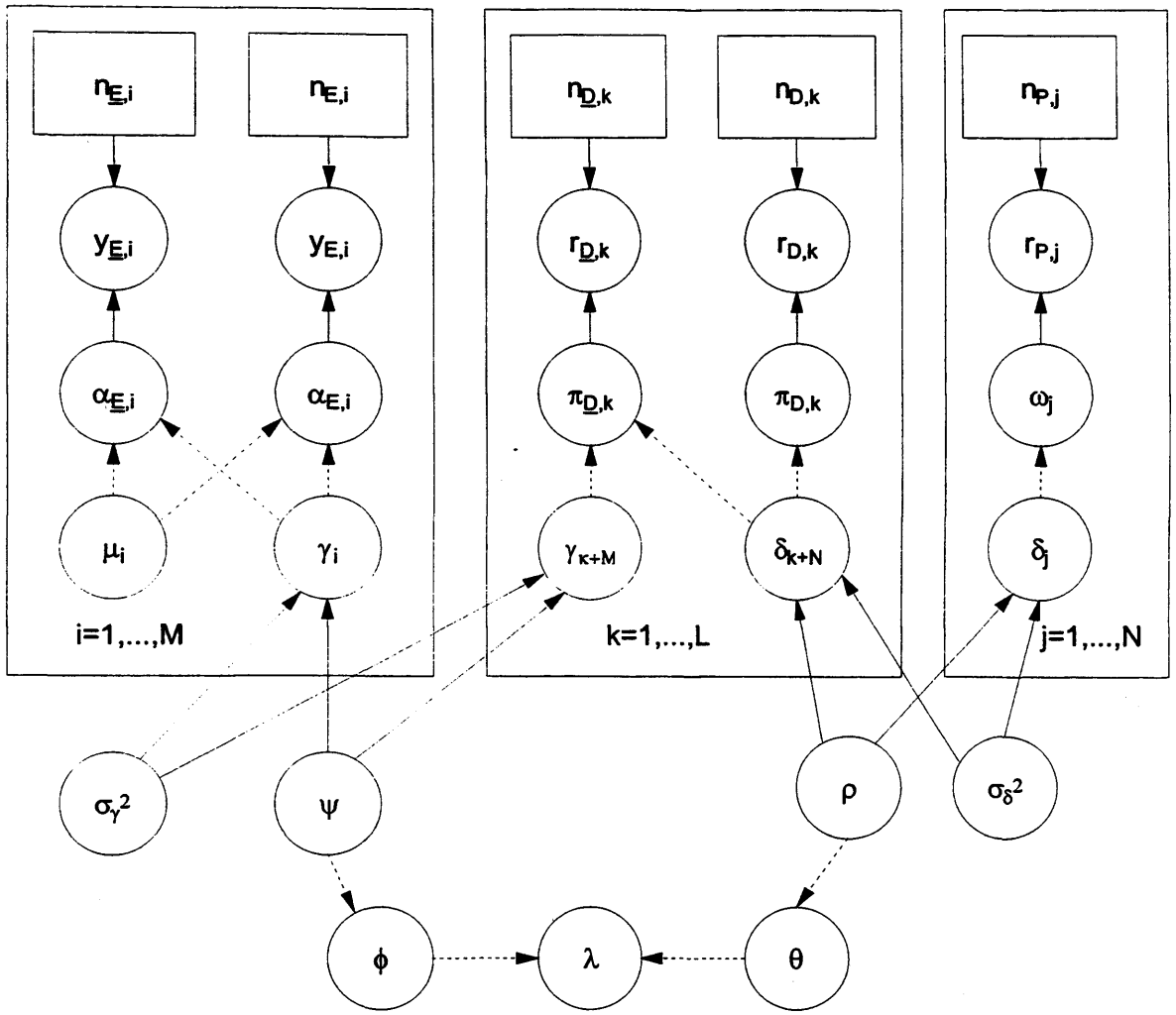

Figure 5.8 DAG for combined hierarchical model for calculation of attributable risk.

associated with the logit of the prevalence. The estimates of ϕ and θ are combined in the same way as before to obtain an estimate of λ, the attributable risk. Prior distributions need to be specified for the model parameters. These were chosen to be vague.

$$\mu_i's \text{ and } \rho \sim N(0,1000000)$$

$$\frac{1}{\sigma_\gamma^2} \text{ and } \frac{1}{\sigma_\delta^2} \sim Gamma(0.001,0.001) \tag{5.27}$$

The model was fitted in WinBUGS using a burn-in of 1000 iterations with a further sample of 5000 iterations. The results of this analysis can be seen in Table 5.11. It is of interest to note the pooled prevalence estimate has decreased from 0.20 to 0.16. This is because the control group from the case-control study is now being incorporated in the estimate of the prevalence, and in that particular study only 34/972 (3.5%) were reported to have a history

Parameter	Estimate (median)	95% Credible Interval
θ (Prevalence)	0.164	0.090 to 0.284
ϕ (Relative Risk)	2.615	1.876 to 3.229
λ (Attributable Risk)	0.206	0.101 to 0.333

Table 5.11 Results of combined hierarchical model for calculation of attributable risk.

of infertility. Thus, further adding to the heterogeneity between the prevalence studies. The pooled estimate of the relative risk of 2.62 is similar to that estimated in the previous Bayesian model, which is not surprising as the case-control study adds relatively little information to the pooled value. The attributable risk estimate of 0.21 is reduced from the previous Bayesian model, which is to be expected, as the prevalence estimate is lower. The confidence interval for the attributable risk is wider than previous Bayesian analysis, which reflects the greater uncertainty regarding the prevalence estimate because of the inclusion of the controls from the case-control study.

5.5 Discussion

In this chapter I have shown how hierarchical models can be used in meta-analysis of summary data. I have presented a standard meta-analysis, using the cholesterol data and a more complex problem where interest lies in estimating the attributable risk of a history of infertility on perinatal mortality.

With the cholesterol data I demonstrated the importance of including study level covariates in the model, with the aim of reducing the observed heterogeneity between studies. An advantage of the Bayesian approach was that the models explicitly allow for the uncertainty in the between-study variance. I have also demonstrated how a Bayesian model can be developed to include the observed frequencies of deaths using the Binomial distribution, rather than having to assume that a summary measure is Normally distributed and using the estimated standard error. The use of baseline risk as a covariate can lead to bias due to both the uncertainty in the estimate of the baseline risk and the fact that that baseline risk is functionally related to the odds ratio. I demonstrated that the method of Thompson *et al.* (1997) can be used to overcome this problem, although in this case there was very little

difference in the parameter estimateof the slope, but there was an increase in its associated standard error.

For the infertility data I have shown how meta-analyses of relative risk and prevalence can be combined to obtain a pooled estimate of the attributable risk. For the infertility dataset there are only a small number of studies and the advantage of the Bayesian appoach is clear in allowing for the uncertainty in the estimate of the between-study variance. The Bayesian approach offers clear advantages when modelling the observed counts in each study through the use of Poisson and Binomial distributions, rather than summary measures. This enables case-control studies to contribute to both the prevalence estimate and the relative risk estimate. Although the model in the attributable risk example is sensible, perhaps it is not the best dataset to use it on. This is because the attributable risk is most beneficial for modifiable risk factors and infertility is a risk factor that is hard to modify. However, knowledge of the attributable risk may be useful in the planning of neonatal intensive care services. Further research should be conducted using alternative datasets.

The model I have used in obtaining a pooled eatimate of the attributable risk combines data from different study types and is therefore related to cross-design synthesis methods (Droitcour *et al.*, 1993). In cross-design synthesis involves combining results from different study types, usually randomised trials and observational studies. Their use has been advocated in situations where randomised controlled trials may be difficult to perform for ethical or other reasons (Abrams and Jones, 1995).

6 USE OF HIERARCHICAL MODELS IN THE ANALYSIS OF MISSING COVARIATE DATA WITH A CENSORED RESPONSE

6.1 Introduction

In the previous chapter I demonstrated how meta-analysis could be performed when there are summary data for each study. It is generally accepted that if individual patient data is available from each study then it should be used in the meta-analysis (Stewart and Clarke, 1995). This type of analysis is known as an Individual Patient Data (IPD) meta-analysis. There has been some previous work comparing IPD meta-analyses with meta-analyses using aggregated data. Some of this work has found discrepancies in the results between the methods (Pignon and Arriagada, 1993; Stewart and Parmar, 1993; Jeng *et al.*, 1995). However, little work has been undertaken to assess why these differences exist (Oxman *et al.*, 1995). One of the main advantages of using IPD meta-analysis is that a fuller exploration of the data can be undertaken, such as sub-group analyses or adjustment for specific covariates. For survival data it has be said that an IPD analysis is the only satisfactory way of combining the data (Hunink and Wong, 1994). With survival data, interest often lies in development of a prognostic model that identifies individuals most at risk of poor outcome (Simon and Altman, 1994). In order to do this properly a substantial amount of data is needed at the individual level. However, with any IPD analysis, a potential problem when combining data from different sources is *missing covariate data*. For example, when developing a prognostic model certain disease markers may or may not have been recorded in every study. The missing data makes the combining of the data into a sensible prognostic model difficult. Although methods exist for analysing data with missing data, such as multiple imputation, there tend to be problems when it comes to analysing censored data.

In this chapter I develop Bayesian models for analysing censored data, which has missing data for a dichotomous and a continuous covariate. I give a brief overview of methods for dealing with missing data in section 6.2. In section 6.3 I use a simple multiple regression example to demonstrate the methodology, using a hierarchical model to generate multiple imputation data sets in section 6.3.1 and a fully Bayesian analysis in section 6.3.2. In

section 6.4 I use simulation techniques to investigate potential bias in the different ways the models can be formulated. Section 6.5 gives an introduction to the Neuroblastoma data, which is used in later sections. Section 6.6 presents a complete case analysis of the data with sections 6.7 and 6.8 extending this model to deal with the missing data. In section 6.9 I demonstrate how extra covariates can be included in the model and in section 6.10, I discuss the techniques I have used and directions for further research.

6.2 Dealing with missing covariate values

The problem of missing data is encountered in many studies (Lessler and Kalsbeek, 1992), but can be a particular problem when data are combined from different sources. The most common method of dealing with missing data is to perform a *complete case analysis*. In this approach those subjects who have missing values for one or more of the covariates under consideration are *excluded* from the analysis *entirely*. The main advantage of this approach is its ease of implementation, with many statistical packages removing subjects with missing observations by default. The main disadvantage of the approach is the loss of information because some known data values for subjects, for whom some other value is missing, are excluded from the analysis, as well as the missing values themselves.

Another approach, sometimes used in epidemiology, is to use a *missing value indicator* for each covariate with missing data (Greenland and Finkle, 1995). This indicator is then included as an extra (dummy) covariate in the analysis. Although this method is simple, it has been shown that it can yield biased results even when the data is missing completely at random (Vach and Blettner, 1991; Greenland and Finkle, 1995). This is most severe when there is a large correlation between any covariates included in the model.

A third approach is *single value imputation* (Little and Rubin, 1989) where a value (the unconditional mean, a conditional mean based on other covariates, or a conditional mean based on other covariates and the response) is imputed for each missing data point. Imputing these values and performing a standard analysis will artificially deflate the standard errors of the coefficients estimated. This is because imputed values are more similar than the corresponding values would be if they were not missing.

Two other approaches that have been shown to be improvements on the above are *maximum likelihood* and *multiple imputation* methods (Little and Rubin 1989; Little, 1992). Maximum likelihood methods specify a joint model for the response and covariate distribution. For simple missing data structures the likelihood can be factorised into components that can be maximised separately, but for more complicated structures iterative procedures such as the EM algorithm (Dempster *et al.*, 1977) must be used. In multiple imputation, multiple copies of the original data set are generated, each with missing values replaced with values imputed from a model conditional on *both* the complete covariates and the response (Rubin, 1996). Each generated data set is then analysed as if it were complete data. The parameter estimates are taken as the means of the analyses with standard errors calculated taking into account both the variability within and between the imputed data sets. Multiple imputation is becoming more popular with both commercial (SOLAS, Statistical Solutions Inc.) and public domain software (Schafer and Olsen, 1998) available for data where variables are continuous, categorical or both.

In a recent paper, Van Buuren *et al.* (1999) used multiple imputation for survival analysis data where there was missing information for two covariates (systolic and diastolic blood pressure). In this approach log survival time was included as a predictor variable for generation of the multiple imputation data set. In addition the censoring indicator was included as a predictor variable. It is not clear how the choice of model will affect the results when compared to a full multivariate model. Multiple imputation has also been used in survival analysis for imputation of the missing (censored) survival times, so that a simple linear model could be used in the analysis of the data (Wei and Tanner, 1991; James, 1995).

Other techniques for missing covariate values with a censored response includes a Tree based method (Ahn and Loh, 1994), using log-linear models for missing categorical covariates (Schluchter and Jackson, 1989) and using pseudo-likelihood methods when the missing data can be considered to be missing completely at random (Robins *et al.*, 1994).

Use of Bayesian methods appears to be a sensible approach for dealing with missing data as the missing values can be considered in the same way as other unknowns in the model, namely as model parameters that need to be estimated. Bayesian methods that have been developed explicitly to deal with missing values are data augmentation (Tanner and Wong, 1987; Kong *et al.*, 1994) and the hierarchical sub-models of Arjas and Liu (1996). In both these methods the missing covariate is assumed to be a random variable with an associated stochastic mechanism. The data augmentation method repeatedly imputes missing data from their predictive distribution (based on the observed data) and then uses the (weighted) average posterior distribution, based on both observed and augmented data to derive an approximate posterior distribution for the underlying parameters. Data Augmentation is also used by Schafer in the generation of multiple imputation datasets (Schafer, 1997). In the hierarchical sub-model method missing covariate values are considered as unknowns in the same way as other model parameters are, so that the posterior distributions are obtained as well as the predictive distributions for the missing covariates. In this chapter I will be assuming that the missing data is missing at random, i.e. missingness does not depend on unobserved data. However, it is possible to extend Bayesian methods for missing data to situations where there is informative missing data (Best *et al.*, 1996).

6.3 A Simple Example

In order to demonstrate how the Bayesian framework can be used for missing data problems, I will use a simple example of multiple regression. The data comes from Armitage and Berry (1987) and consists of 53 subjects from one arm of a clinical trial comparing two drugs used to lower blood pressure during operations. The response variable is the time (in minutes) elapsing between the time at which the drug was discontinued and the time at which the systolic blood pressure returned to 100 mmHg. The question of interest is the extent to which the recovery time depends on the quantity of the drug used and the level to which blood pressure was lowered during hypotension. The two covariates are,

x_1 - log (quantity of drug used (mg)).

x_2 - mean level of systolic blood pressure during hypotension (mm Hg).

with response

x_1	x_2	y
2.26	66	7
1.81	52	10
1.78	72	18
1.54	67	4
⋮	⋮	⋮
⋮	⋮	⋮
2.10	51	25
1.80	61	44

Table 6.1 Data on the use of a hypotensive drug, x_1: log (quantity of drug used), x_2: mean level of systolic blood pressure during hypotension (mmHg), y: recovery time.

y - recovery time (mins).

In order to demonstrate how missing data can be dealt with within the Bayesian framework I have removed every second observation for x_1. Part of the complete data can be seen in Table 6.1.

6.3.1 Multiple Imputation using a hierarchical model

The multiple regression model to be fitted to the data is as follows

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i$$
$$e_i \sim N(0, \sigma^2) \qquad (6.1)$$

where subscript i refers to the i^{th} individual.

However, there is the problem that half of the observations for x_1 are missing and standard techniques of estimation cannot cope with this. In this situation maximum likelihood techniques could be used as the missing data has a relatively simple structure (Little and Rubin, 1989). However, in most practical cases (involving more covariates) there is no obvious pattern to the missing data and multiple imputation techniques should be used. The question, therefore, is how to generate the multiple imputation data sets.

The hypotension data set can be considered to have a multivariate structure. This can also be thought of as a hierarchical structure, with the 3 variables y, x_1 and x_2 being nested within individuals. One may expect there to be similarities between the variables measured

on the same individual, i.e. correlation between the variables. The level 2 units are the individuals and the level 1 units are the three variables. Let y_{ij} denote the j^{th} variable on the i^{th} individual. Let z_{0ij}, z_{1ij}, and z_{2ij} denote three dummy variables for y, x_1 and x_2 that take the value 1 when the variable is present and 0 otherwise. The multivariate model can be defined as follows

$$y_{ij} = \alpha_0 z_{0ij} + \alpha_1 z_{1ij} + \alpha_2 z_{2ij} + u_{0j} z_{0ij} + u_{1j} z_{1ij} + u_{2j} z_{2ij}$$

$$\text{where } \underline{u}_j \sim MVN\left[\begin{pmatrix}0\\0\\0\end{pmatrix}, \begin{pmatrix}\sigma_0^2 & \sigma_{01} & \sigma_{02}\\ \sigma_{01} & \sigma_1^2 & \sigma_{12}\\ \sigma_{02} & \sigma_{12} & \sigma_2^2\end{pmatrix}\right] \tag{6.2}$$

α_0, α_1 and α_2 represent the means of y, x_1 and x_2 respectively and σ_0^2, σ_1^2 and σ_2^2 their corresponding variances with the covariances also estimated in the model.

Half of the subjects will have all three variables measured and half of the subjects will have two variables measured as x_1 is missing. However, an estimate of the predicted value of x_1, x_1^*, conditional on the values of y and x_2, can be obtained by using the shrunken residual estimate u_{1j} as follows

$$x_{1j}^* = \alpha_1 + u_{1j} \tag{6.3}$$

When generating multiple imputation data sets one also needs to take into account the uncertainty in obtaining this prediction. This is obtained by adding a random value sampled from a Normal distribution with zero mean and variance equal to the variance of the residual u_{1j} (Goldstein, 1995). Thus, each imputed data set will have the missing values for x_1 imputed with values x_{1j}^{**} where

$$\begin{aligned} x_{1j}^{**} &= \alpha_1 + u_{1j} + v_{1j} \\ \text{where } v_{1j} &\sim N\left(0, Var(u_{1j})\right) \end{aligned} \tag{6.4}$$

The value of Var(u_{1j}) will be the same for each missing value of x_1.

Parameter	Estimate (standard error)
α_0	22.70 (2.216)
α_1	1.99 (0.058)
α_2	66.34 (1.052)
Variance matrix	$\begin{pmatrix} 260.2 & 1.09 & -13.29 \\ 1.09 & 0.12 & 1.48 \\ -13.29 & 1.58 & 58.6 \end{pmatrix}$

Table 6.2 Parameter estimates for multivariate model for hypotension data.

The parameter estimates from (6.2) can be seen in Table 6.2 which shows the means of y, x_1 and x_2 and the associated variance matrix. Converting the variance matrix to a correlation matrix gives,

$$\begin{pmatrix} 1 & & \\ 0.20 & 1 & \\ -0.11 & 0.60 & 1 \end{pmatrix}$$

It can be seen that y is positively associated with x_1 and negatively associated with x_2. There is a fairly high correlation between x_1 and x_2 indicating that when x_1 is missing the value of x_2 should provide some information to predict the value of x_1.

When the m imputed data sets have been generated, each data set can then be analysed using standard linear regression models. The estimates are combined as follows
Let $Q^{(t)}$ be the estimate of the parameter of interest for the t^{th} imputed data set (t=1,...,m) and $U^{(t)}$ be the associated variance estimate. The multiple imputation point estimate is the average of the estimates from each multiple imputation dataset,

$$\overline{Q} = \frac{1}{m}\sum_{t=1}^{m} Q^{(t)} \qquad (6.5)$$

The variance estimate associated with $\overline{Q}$ has two components, the *within-imputation variance* and the *between-imputation variance*. The within-imputation variance is the average of the estimates of the multiple imputation variance estimates,

$$\overline{U} = \frac{1}{m}\sum_{t=1}^{m} U^{(t)} \qquad (6.6)$$

Multiple Imputation Data set	β_0 (Variance)	β_1 (Variance)	β_2 (Variance)
1	23.583 (5.022)	15.898 (74.874)	-0.624 (0.129)
2	23.424 (4.356)	25.752 (68.013)	-0.750 (0.102)
3	22.725 (5.099)	6.693 (76.773)	-0.387 (0.130)
4	22.699 (4.554)	19.071 (54.953)	-0.657 (0.106)
5	23.406 (5.000)	16.151 (88.172)	-0.677 (0.151)
6	23.618 (4.623)	23.219 (77.176)	-0.885 (0.139)
7	22.561 (4.875)	15.276 (77.757)	-0.597 (0.128)
8	23.384 (4.402)	23.306 (59.753)	-0.752 (0.104)
9	22.650 (3.806)	27.849 (43.798)	-1.059 (0.104)
10	23.079 (4.427)	22.121 (58.630)	-0.885 (0.127)
	$\overline{Q}$ =23.113, $\overline{U}$ =4.616	$\overline{Q}$ =19.534, $\overline{U}$ =67.990	$\overline{Q}$ =-0.727, $\overline{U}$ =0.122
	B=0.175, T=4.809	B=38.912, T=110.793	B=0.035, T=0.160

Table 6.3 Parameter estimates of multiple imputation data sets for the hypotension data.

The between-imputation variance (B) is the variance of the multiple imputation data set estimates,

$$B = \frac{1}{m-1}\sum_{i=1}^{m}\left(Q^{(i)} - \overline{Q}\right)^2 \tag{6.7}$$

The total variance (T) is obtained by,

$$T = \overline{U} + \left(\frac{m}{1+m}\right)B \tag{6.8}$$

Ten multiple imputation data sets were generated and multiple regression performed on each one using (6.1). The parameter estimates for the ten multiple imputation data sets can be seen in Table 6.3. The results in this table will be compared with those from the Bayesian analysis in the following section.

6.3.2 A Bayesian model

If only complete cases are used then the following model can be fitted,

$$\begin{aligned} y_i &\sim N(\mu_i, \sigma^2) \\ \mu_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} \end{aligned} \tag{6.9}$$

This is a standard multiple regression model and can be easily fitted using maximum likelihood methods in most statistical software packages. From a Bayesian perspective

prior distributions need to be specified for β_0, β_1, β_2 and σ^2. It is usual to use the standard normal distribution for the β's with suitably large variances and an inverse Gamma distribution for σ^2. The problem with using a complete case analysis is that only information on 27 of the 53 subjects is used even though the remaining 26 subjects have complete data for y and x_2.

To include all the data, i.e. including those cases where y and x_2 were recorded, but not x_1, the Bayesian model needs to be extended. The unknown value for x_1 will depend on both y and x_2 so the joint distribution needs to be obtained. A sensible option for multiple regression is to assume multivariate normality for the 2 covariates and 1 response variable. In a similar way to the use of the product normal formulation in Chapter 3, the model can be extended as follows

$$\begin{aligned} y_i &\sim N(\mu_i, \sigma^2) \\ \mu_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} \\ x_{1i} &\sim N(\mu_{x_1 i}, \sigma^2_{x_1}) \\ \mu_{x_1 i} &= \theta_1 + \theta_2 x_{2i} \\ x_{2i} &\sim N(\mu_{x_2 i}, \sigma^2_{x_2}) \\ \mu_{x_2 i} &= \gamma \end{aligned} \qquad (6.10)$$

Prior distributions now need to be specified for β_0, β_1, β_2, σ^2, θ_0, θ_1, γ, $\sigma^2_{x_1}$, and $\sigma^2_{x_2}$. These can be defined as non-informative distributions as follows

$$\begin{aligned} &\beta's, \theta's, \gamma \sim N(0,1000000) \\ &\frac{1}{\sigma^2}, \frac{1}{\sigma^2_{x_1}}, \frac{1}{\sigma^2_{x_2}} \sim Gamma(0.001, 0.001) \end{aligned} \qquad (6.11)$$

Note that the final two lines of model (6.10) do not need to be specified in this case as there is no missing data for x_2. However, I have included it to demonstrate how the model can be extended to situations when more than one covariate has missing values using the product normal formulation.

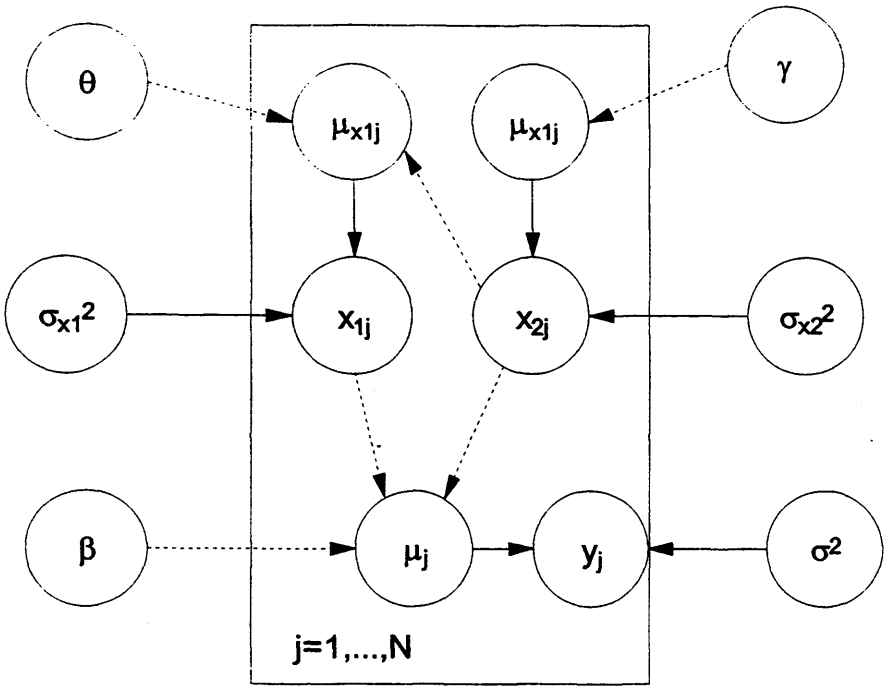

Figure 6.1 DAG for multiple regression model with missing data for hypotension data.

The DAG for model (6.10) can be seen in Figure 6.1. The DAG clearly shows the inter-dependencies between y, x_1 and x_2.

The advantages of formulating the model in this manner are;

1. The Wishart distribution does not need to be used as a prior distribution for the multivariate normal distribution. I find the Wishart distribution less intuitive than the product normal form for multivariate normality and in any case would not be possible using the current version of WinBUGS (1.3), as it cannot cope with partial missing data for multivariate normal distributions.
2. The desired regression model can be written down as part of the product normal formulation, so there is no need to transform any of the parameters.
3. It would be easier to include prior information if this were available. For example, if there is external information about the relationship between the variables then this could be incorporated through the use of informative prior distributions or possibly a sub-model.

4. There is not the problem of generating multiple data sets, analysing each data set and then combining the results as when using multiple imputation.

The complete case Bayesian model and the model incorporating missing data were fitted in WinBUGS with a burn-in of 1000 iterations and a further 5000 iterations for the generation of the sample. Fitting each of these models took about 5 seconds on a Pentium II 400 Mhz PC. The complete case model (6.9) was also fitted using standard maximum likelihood methods. The results of the two Bayesian analyses, the multiple imputation model and the maximum likelihood model on the complete data can be seen in Table 6.4

It appears from the results that a unit increase in log dose lengthens the recovery time by about 19 minutes and a unit increase in systolic blood pressure leads to a reduction in recovery time of about 0.8 minutes. It can also be seen that the parameter estimates for the complete case analysis for both the ML and Bayesian models give broadly similar answers with the Bayesian model having slightly larger standard errors.

When the model is extended to include *all* the data the standard errors are smaller. This is true for both the multiple imputation analyses and the Bayesian model, which again give

Parameter	Maximum Likelihood Analysis on Complete Cases	Bayesian Analysis on Complete Cases	Multiple Imputation	Bayesian Analysis including missing data
β_0	24.69 (3.05)	24.70 (3.20)	23.11 (2.19)	23.30 (2.34)
β_1	19.23 (11.25)	19.17 (11.71)	19.53 (10.53)	18.68 (10.59)
β_2	-0.78 (0.52)	-0.78 (0.54)	-0.73 (0.40)	-0.73 (0.41)
σ^2	248.20	255.70 (95.54)	241.31	245.10 (57.45)
θ_0	-	-	-	-0.029 (0.054)
θ_1	-	-	-	-0.027 (0.0076)
γ	-	-	-	0.021 (1.09)
σ^2_{x1}	-	-	-	0.084 (0.025)
σ^2_{x2}	-	-	-	62.23 (12.78)

Table 6.4 Parameter estimates (standard errors) for multiple regression on the hypotension data using a maximum likelihood and Bayesian estimation for the complete case data, and multiple imputation and Bayesian estimation for the complete data.

broadly similar estimates, but with the Bayesian model giving slightly larger standard errors. The reduction in standard error is most dramatic for β_2 which is sensible as the number of extra data values has doubled for x_2. However, the standard error has also decreased for β_1 even though there are no new data points for x_1. This is because for each missing value of x_1, both y and x_2 are known and due to the association between x_1 and x_2 and x_1 and y, a predictive distribution for each unknown is obtained. The stronger the relationship between the two covariates, the greater the precision of the predictive distribution.

6.4 Simulation to investigate bias in indirect models

Later in this chapter I extend the methods described above to deal with the situation where there are missing values for both continuous and dichotomous covariates with a censored response variable. The work is an extension of work by Arjas and Liu (1996) who investigated missing covariate data in a Cox proportional hazards model. Arjas and Liu discuss the use of *indirect* and *direct* models. A direct model models the relationship between the covariates and an indirect model assigns an appropriate distribution to each covariate with missing values, but does not model the relationship between covariates explicitly. However, Arjas and Liu argue that some of the inter-relationships between the covariates will be picked up through the linear predictor. It is more difficult and more time consuming to use direct models, especially when there are numerous covariates with missing data and this will lead to complex models. However, an important question is whether it matters if an indirect model is used rather than a direct model, and how the choice of model may affect the parameter estimates?

In this section I use simulation techniques to show that using indirect models can lead to biased results, especially when the correlation between the covariates is high. I do this by using a very simple simulated data set with missing data and argue that if there is bias in the simple case, then it is very likely that there is bias in more complicated scenarios. The results I find are important as they indicate that the full joint distribution of the data needs to be modelled in order to make valid statistical inferencesThe model considered is as follows:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i$$
$$e_i \sim N(0, \sigma^2) \tag{6.12}$$

The model is thus a multiple regression model with response y_i and two continuous covariates x_{1i} and x_{2i} (i=1,...,N). Each data set is generated as follows.

1) x_{1i} and x_{2i} are generated from a bivariate normal distribution with 0 mean, variances 1 and covariance/correlation ρ.
2) y is generated assuming β_0=0, β_1=1, β_2=1 and σ^2=1.

I investigate values of ρ, the covariance/correlation between x_{1i} and x_{2i}, of 0.0, 0.2, 0.4, 0.6 and 0.8. For each value of ρ, 100 data sets are generated each of size N=100. The last 50 observations of x_1 are removed from each data set. Each data set was fitted in WinBUGS using the following 2 models,

Indirect Model

$$y_i \sim N(\mu_i, \sigma^2)$$
$$\mu_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} \tag{6.13}$$
$$x_{1i} \sim N(\theta, \sigma_{x_1}^2)$$

Direct Model

$$y_i \sim N(\mu_i, \sigma^2)$$
$$\mu_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$$
$$x_{1i} \sim N(\mu_{x_1}, \sigma_{x_1}^2) \tag{6.14}$$
$$\mu_{x_1} = \theta_0 + \theta_1 x_{2i}$$

Non-informative prior distributions were used as follows

$$\beta's, \theta's \sim N(0, 1000000)$$
$$\frac{1}{\sigma^2}, \frac{1}{\sigma_{x_1}^2} \sim Gamma(0.001, 0.001) \tag{6.15}$$

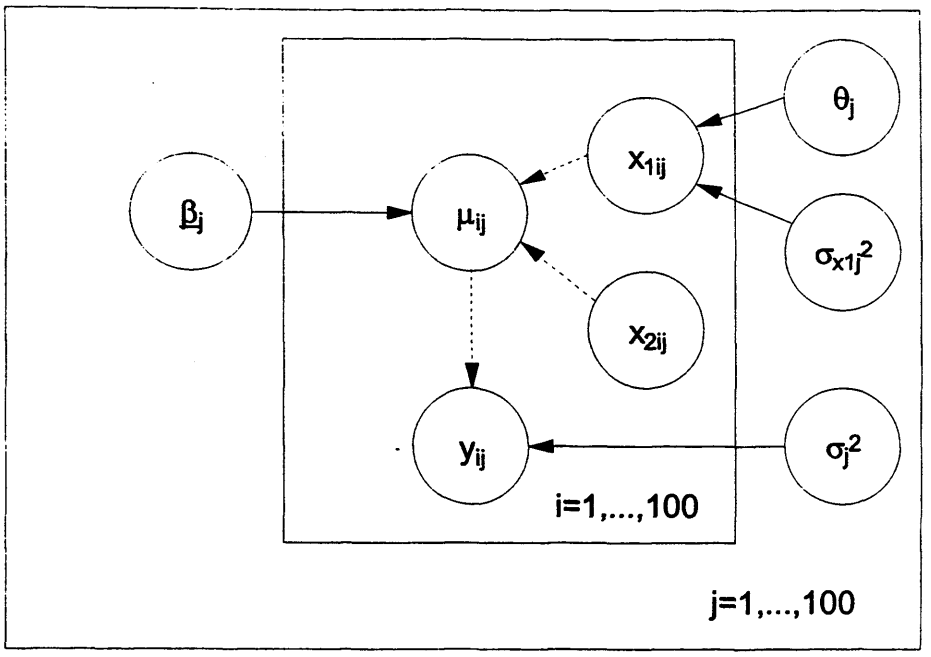

Figure 6.2 DAG for indirect model looping over each simulated data set.

At present it is not possible to use batch runs in WinBUGS so for each value of ρ, the 100 data sets are stacked on top of each other and each model fitted separately in a loop. This can be seen in the DAG for the indirect model in Figure 6.2, where subscript j represents the j^{th} simulated data set.

The results of the simulations can be seen in Table 6.5. It can be seen that when the two covariates are not related the choice of the direct or indirect model is not important. However, as the correlation between the two covariates increases β_1 is underestimated, whilst β_2 is overestimated for the indirect model. However when using the direct model, the mean of the posterior density means is approximately equal to one and hence not biased. In addition the mean of the residual variation, σ^2, appears to be overestimated for the indirect model whilst it is closer to one for the direct model (it is slightly greater than one as I am taking the mean of a positively skewed distribution).

ρ		β_0		β_1		β_2		σ^2	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.0	Indirect	0.01	0.12	0.99	0.13	1.00	0.12	1.03	0.20
	Direct	0.01	0.13	0.99	0.13	0.99	0.13	1.03	0.20
0.2	Indirect	0.01	0.12	0.98	0.13	1.07	0.12	1.04	0.20
	Direct	0.02	0.13	0.99	0.13	0.99	0.13	1.03	0.20
0.4	Indirect	0.01	0.12	0.93	0.14	1.16	0.13	1.05	0.20
	Direct	0.02	0.12	1.00	0.14	0.99	0.14	1.03	0.20
0.6	Indirect	0.01	0.12	0.81	0.16	1.29	0.14	1.08	0.21
	Direct	0.02	0.12	1.00	0.17	0.99	0.16	1.02	0.19
0.8	Indirect	0.01	0.12	0.55	0.19	1.54	0.15	1.12	0.20
	Direct	0.01	0.12	1.00	0.23	0.99	0.21	1.02	0.18

Table 6.5 Mean of posterior density means and standard deviations from 100 simulated datasets for 5 values of ρ for direct and indirect models, with β_0=0, β_1=1, β_2=1 and σ^2=1.

Thus, it can be seen that even in this very simple example, if the covariates are even moderately correlated, bias can arise in the parameter estimates. It seems sensible therefore to always attempt to model using a direct model.

6.5 Neuroblastoma Data

Neuroblastoma is the most common extracranial solid tumour of childhood accounting for approximately 10% of childhood malignancies. The incidence has increased over the last 20-25 years (Stiller, 1993) and currently occurs in about 10 per million per year in children under 15 years of age in both the United States and Europe (Parkin *et al.*,s 1988; Stiller and Parkin, 1992; Powell *et al.*, 1998). The clinical behaviour of the tumour varies widely, ranging from spontaneous regression in a small percentage of patients through to rapid disease progression with poor prognosis in others. Identifying prognostic factors is thus important not only for predicting long term outcome, but more importantly as a guide to the choice of the most appropriate therapy. Most of the improvements in outcome for children with Neuroblastoma have stemmed from a greater ability to distinguish between cases of varying risk, with intensive therapy being given to those most at risk, whilst those with a better prognosis can avoid morbidity associated with such treatment. Optimal identification of such factors is thus clinically very important. A number of prognostic factors have been clearly recognised or reported to be of importance in neuroblastoma, but the relative

importance of each of these factors, their inter-relations, and in some cases the biological mechanisms responsible have yet to be fully elucidated.

Currently the most predictive factors are the classical stage of the disease and age of the child. However, several paediatric oncology groups are now examining the value of a variety of other immuno-biological and histopathological markers.

In 1986 the first of a series of conferences to standardise definitions for diagnosis, staging and treatment response (Brodeur *et al.*, 1988) led to the International Neuroblastoma Staging System (INSS). This should allow survival data from different oncology centres to be compared according to a uniform set of definitions of stage, as prior to this a number of different staging systems were used. In an initial analysis to explore the feasibility of combining datasets from various oncology centres, partitioning analysis (Ciampi *et al.* 1986) was used on a database from five centres. This analysis only investigated the effects of age and stage since only these two variables were routinely recorded at all centres. Only limited analyses could be performed when looking at biological prognostic variables. These have mainly been explored factor by factor with relatively little work on investigating the joint influence of several or all of the potential prognostic factors. Where this has been attempted, the analyses have generally been based on selected or centre-specific patient databases. These almost inevitably (in a relatively rare disease) include a limited number of patients (Berthold *et al.*, 1992; Berthold *et al.*, 1994; Combaret *et al.*, 1996; Rubie *et al.*, 1997; Matthay *et al.*, 1998; Ladenstein *et al.*, 1998). However, the need to combine data from different sources has been recognised (Castleberry *et al.*, 1997) in order to explore the joint effects of a number of potential disease markers. To quote Favrot *et al.* (1996) '.. multifactorial analyses are fundamental if the respective impacts of biological abnormalities and other factors on neuroblastoma progression are to be explored and determined.'.

At the second INSS meeting in 1991 it was recommended to construct International Neuroblastoma Risk Groups (INRG) (Castleberry *et al.*, 1997) from a composite of age, INSS stage and the most predictive and widely available laboratory based variables. Data from 1991 consists of 2832 children with Neuroblastoma from five oncology groups.

Centre	Total	N-myc only	Ferritin only	Both measured	Neither measured	Included in Analysis
1	365 (100%)	126 (35%)	0 (0%)	0 (0%)	239 (65%)	126 (35%)
2	96 (100%)	22 (23%)	6 (6%)	5 (5%)	63 (66%)	33 (34%)
3	131 (100%)	6 (5%)	64 (49%)	44 (34%)	17 (13%)	114 (87%)
4	268 (100%)	29 (11%)	105 (39%)	96 (36%)	38 (14%)	230 (86%)
5	275 (100%)	2 (1%)	32 (12%)	1 (0%)	240 (87%)	35 (13%)
Total	1135(100%)	185 (16%)	207 (18%)	146 (13%)	597 (53%)	538 (47%)

Table 6.6 Number analysed at each centre.

However, it was soon recognised that there was a problem with missing data in that not all disease markers were recorded on each subject by each group. This may be because the disease marker was not recorded at all at a certain centre, or that it was only recorded after a specific date. Missing data represents a major problem for statistical analysis and interpretation, as standard techniques require all markers to be recorded on all subjects. Therefore, analysis of the data using standard techniques would need to be based on a reduced set of patients for whom *all* disease markers of interest were recorded.

In order to demonstrate how a Bayesian approach could be adopted in the presence of such missing data I concentrate on two laboratory based disease markers, N-myc and Ferritin. N-myc is a dichotomous marker (being either amplified or non-amplified) and Ferritin is a continuous marker. I also just concentrate on children who were over one year and had Evans stage IV at diagnosis. This is a subset of children with poor prognosis.

Table 6.6 displays the breakdown of missing observations at each centre for this data. It can be seen that at centres 4 and 5 at least one of N-myc or Ferritin was measured for most patients (87% and 86% respectively) in the subset. Centre 1 only has measurements for N-myc. Centres 3 and 6 contribute only a small number of subjects in the subset. If a complete case analysis was performed then the analysis would only be on 146 subjects, whereas, if one could analyse the data where at least one of N-myc or Ferritin were present then the analysis would be on 538 subjects. Thus, the information loss would be high if a complete case analysis were used.

6.6 Complete Case Analysis

With the neuroblastoma data discussed in the previous section the outcome of interest is time to death. Analysis of time to death data is complicated by the fact that some of the individuals have not yet been observed to die. This is known as *censoring* and is important because, although one does not know when these individuals will die, we do known that their time to death will be after the time they were last observed alive. Analysis of data with censoring makes use of *survival analysis* techniques. In survival analysis interest often lies in estimating the *hazard* at any time after the start of the study. The hazard is the instantaneous probability of death *given* that a subject has already survived to time *t*. As a consequence the *hazard function* is often modelled and the effect of a covariate described using the *hazard ratio* (Collett, 1994).

There are a number of techniques available for the analysis of censored data. The most common of these is the log-rank test (Mantel, 1966) which is used to compare two or more groups of individuals. However, some form of modelling is often required and when this is the case the proportional hazards model (Cox, 1972) is the most common method of analysis. The proportional hazards model is known as a semi-parametric method as no form of probability distribution is assumed for the survival times, with the baseline hazard being modelled non-parametrically, but the method does have a regression function to determine the changes in hazard associated with one or more covariates.

Another possibility when modelling survival data is to assume that the survival times follow a specific probability distribution. If this assumption is valid then inferences will be more precise than the proportional hazards model, though the gain is not that great which is one of the reasons why proportional hazards models remain so popular. In addition one has to choose which probability distribution to use. There are a number of distributions that are commonly used in the analysis of survival data including the exponential, Log-normal, log-logistic, gamma and Weibull distributions (Klein, 1997). The most commonly used of these distributions is the Weibull distribution and it has been previously used within the Bayesian framework using asymptotic approximations and Gaussian quadrature (Abrams *et*

al., 1996). Use of the Weibull distribution assumes that the baseline hazard, $\lambda_0(t)$, is constant, increasing or decreasing over time.

The Weibull distribution for survival time t_i for i^{th} individual with covariate vector x_i can be defined as follows.

$$f(t \mid x_i) = re^{\beta x_i} t_i^{r-1} \exp(-e^{\beta x_i} t_i^r) \tag{6.16}$$

The baseline hazard is the risk of death when each of the covariates take the value zero and can thus be obtained as,

$$\lambda_0(t_i) = re^{\beta_0} t_i^{r-1} \tag{6.17}$$

When $r>1$ then the baseline hazard function will increase over time, when $r<1$ it will decrease over time and when $r=1$ the hazard will be constant over time. In the latter case the Weibull distribution reduces to the simpler exponential distribution. As the value of r determines the shape of the hazard function, it is sometimes referred to as the *shape* parameter.

For the Neuroblastoma data described in the previous section, a complete case analysis would involve the following model.

$$\begin{aligned} t_i &\sim Weibull(r, \mu_i) \\ \log(\mu_i) &= \beta_0 + \beta_1 Logferr_i + \beta_2 Nmyc_i \end{aligned} \tag{6.18}$$

where subscript i refers to the i^{th} individual.

For censored observations survival is assumed to follow a truncated Weibull distribution with lower band corresponding to the censoring time. Formulating the model in this way leads to the censored survival times being treated as missing values and a predictive distribution being obtained for each missing (censored) survival time. This is similar to the multiple imputation work developed by Wei and Tanner (1991) and extended by James (1995). Within a Bayesian framework prior distributions need to be specified for the unknown parameters, r and the β's. In this case I have chosen these to be as follows

$$r \sim Gamma(1,0.001) \qquad \beta_0, \beta_1, \beta_2 \sim N(0,100000) \tag{6.19}$$

Parameter	ML Analysis		Bayesian Analysis	
	Estimate (SE)	95% CI	Estimate (SE)	95% CrI
β_0 (intercept)	-	-	-3.88 (0.322)	-4.56 to –3.29
β_1 (*Logferr*)	0.213 (0.107)	0.003 to 0.423	0.215 (0.105)	0.025 to 0.412
β_2 (*Nmyc*)	0.495 (0.206)	0.091 to 0.899	0.490 (0.205)	0.086 to 0.882
r (shape)	1.120	-	1.116 (0.086)	0.967 to 1.281

Table 6.7 Comparing Maximum Likelihood (ML) and Bayesian complete case analysis

These prior distributions are non-informative. The prior distribution for r is slowly decreasing from zero. The value of r will tend to be low leading to essentially a non-informative prior. The corresponding DAG for model (6.19) can be seen in Figure 6.3. The node $cens_i$ will take the value 0 for individuals that die and the censoring time for censored individuals. The stochastic node t_i representing the survival time will have missing values for censored observations and will be sampled from a truncated Weibull distribution with lower bound equal to $cens_i$.

Model (6.19) was fitted using WinBugs with a 'burn-in' of 1000 iterations and a further 5000 iterations to sample from. This took about 4 minutes on a Pentium II 400 Mhz PC. The results of fitting model (6.19) to the neuroblastoma data can be seen in Table 6.7. Also shown are maximum likelihood estimates obtained from SAS PROC LIFEREG. SAS uses

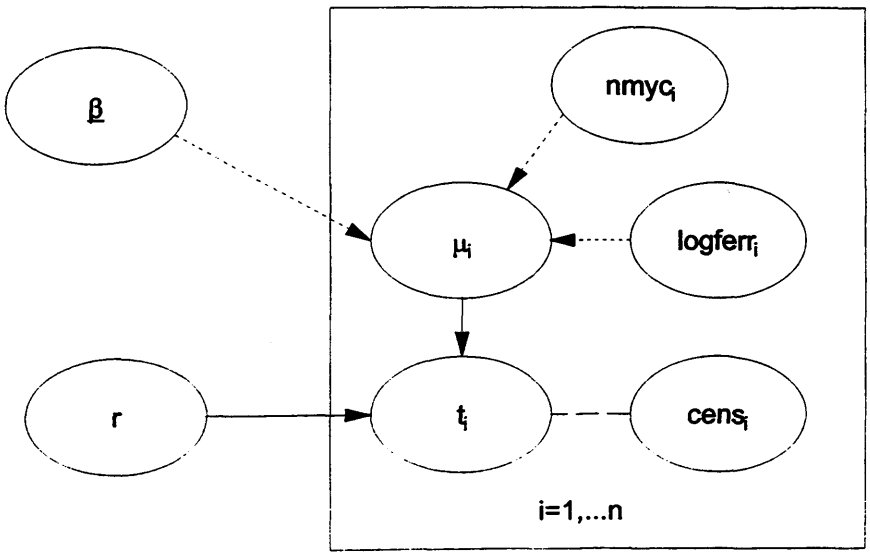

Figure 6.3 DAG for complete case analysis of neuroblastoma data.

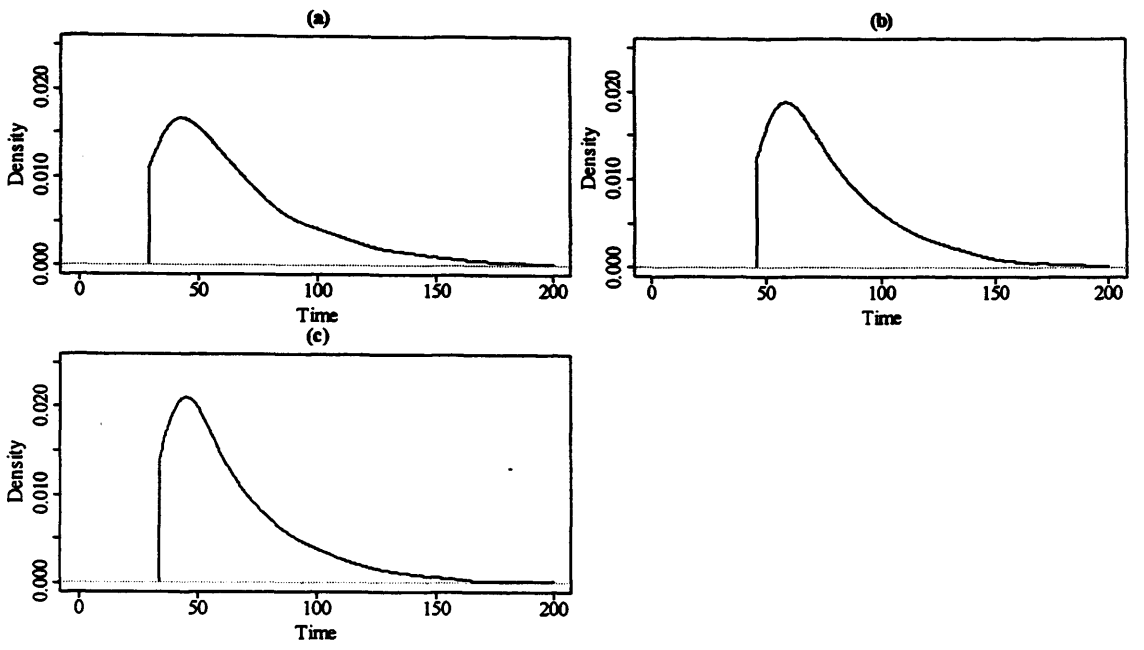

Figure 6.4 Example of predictive distributions for censored survival times

the accelerated failure time parameterisation of the Weibull model, so the parameters need to be transformed using the methods described in Collett (1994). Collett also described how to obtain the standard errors of the parameters of interest (the log hazard ratios). It can be seen that the estimates are broadly similar under the two methods of estimation, with slightly larger standard errors being obtained using the Bayesian model. Amplified N-myc is associated with an increased hazard (hazard ratio=1.63, 95% credible interval (1.09,2.42)) as is Log Ferritin (hazard ratio= 1.24, (1.03,1.51)). The positive value of r indicates that the hazard is increasing over time, but the 95% credible interval does include one, so there is some evidence that the use of the exponential model may be adequate.

Figure 6.4 shows the predictive distributions for three censored survival times. These subjects had censored survival times of (a) 29.2, (b) 45.8 and (c) 33.6 months. It can be seen that at each iteration of the Gibbs Sampler the survival times of subjects with censored observations are sampled from a distribution truncated at the censoring time.

6.7 Indirect Model for Missing Data

As discussed in section 6.4, Arjas and Liu (1996) introduced indirect and direct models to model the missing data. I have shown that when the covariates are correlated, the indirect model is biased. However, for comparison purposes I will first analyse the data using an indirect model. In the case presented here the choice of direct or indirect model should make little difference as the relationship between *Logferr* and *Nmyc* is relatively weak, with the mean (standard deviation) of *Logferr* being 5.34 (1.04) for non amplified *Nmyc* and 5.53 (0.98) for amplified *Nmyc*. The 95% confidence interval for the difference in means is –0.54 to 0.18.

The model in (6.3) can be extended as follows

$$\begin{aligned} t_i &\sim Weibull(r, \mu_i) \\ \log(\mu_i) &= \beta_0 + \beta_1 Logferr_i + \beta_2 Nmyc_i \\ Logferr_i &\sim N(\alpha, \sigma^2_{Lf}) \\ Nmyc_i &\sim Bernoulli(\delta_i) \\ \log it(\delta_i) &= \theta \end{aligned} \qquad (6.20)$$

In model (6.20) *Logferr* has been assigned a Normal distribution with mean α and variance σ^2_{Lf} and *Nmyc* has been assigned a Bernoulli distribution with the logit of the probability of *Nmyc* being amplified estimated by θ. The DAG for model (6.20) can be seen in Figure 6.5

The prior distributions for $\beta_0, \ldots, \beta_2$ and r remain the same as in (6.19). The prior distributions for α, θ and σ^2_{Lf} are assumed to be;

$$\begin{aligned} \alpha, \theta &\sim N(0, 1000000) \\ \frac{1}{\sigma^2_{Lf}} &\sim Gamma(0.001, 0.001) \end{aligned} \qquad (6.21)$$

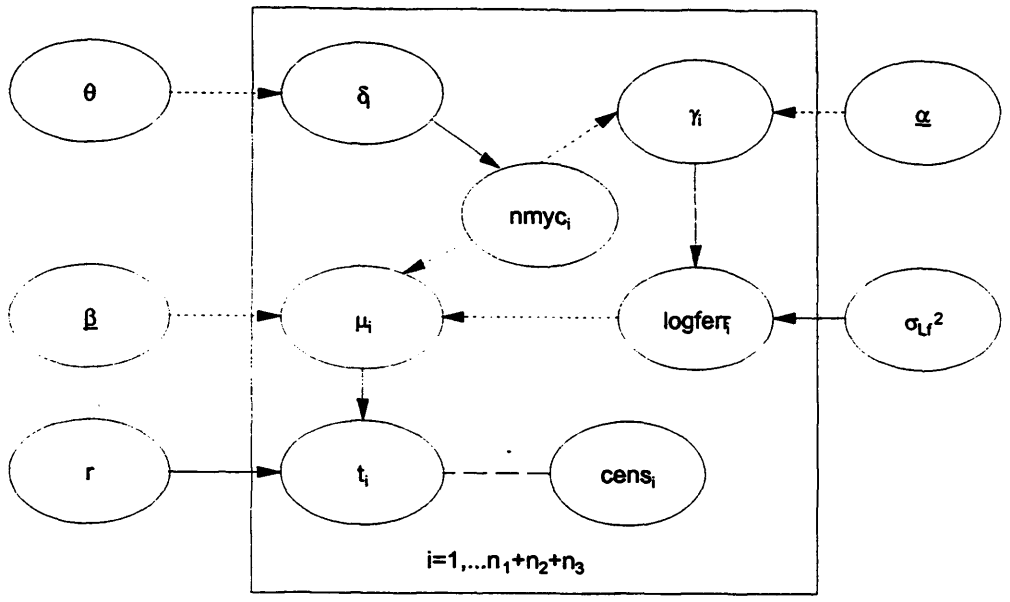

Figure 6.5 DAG for Indirect Model.

In order to clarify how the linear predictor affects the imputed values of *Logferr* and *Nmyc* the full conditional distributions can be obtained. Using the methods for obtaining full conditional distributions described in section 2.5.2, the full conditional distributions for missing values of *Logferr* and *Nmyc* will be given by;

$$p(Logferr_i \mid \cdot) \propto \prod_{i=n_1+1}^{n_1+n_2} Weib(t_i \mid r, \exp(\beta_0 + \beta_1 Logferr_i + \beta_2 Nmyc_i)) \times \prod_{i=n_1+1}^{n_{11}+n_2} N(Logferr_i \mid \alpha, \sigma_{Lf}^2) \tag{6.22}$$

and

$$p(Nmyc_i \mid \cdot) \propto \prod_{i=n_1+n_2+1}^{n_1+n_2+n_3} Weib(t_i \mid r, \exp(\beta_0 + \beta_1 Logferr_i + \beta_2 Nmyc_i)) \times \prod_{i=n_1+n_2+1}^{n_1+n_2+n_3} Bernoulli(Nmyc_i \mid \theta) \tag{6.23}$$

The parameters in model (6.20) were estimated using WinBUGS with a 5000 iteration 'burn-in' and 20,000 further samples. The results of fitting this model can be seen in Table 6.8 together with the estimates from the complete case Bayesian model. It can be seen that the standard errors are smaller for the indirect model, which is intuitive as there are more

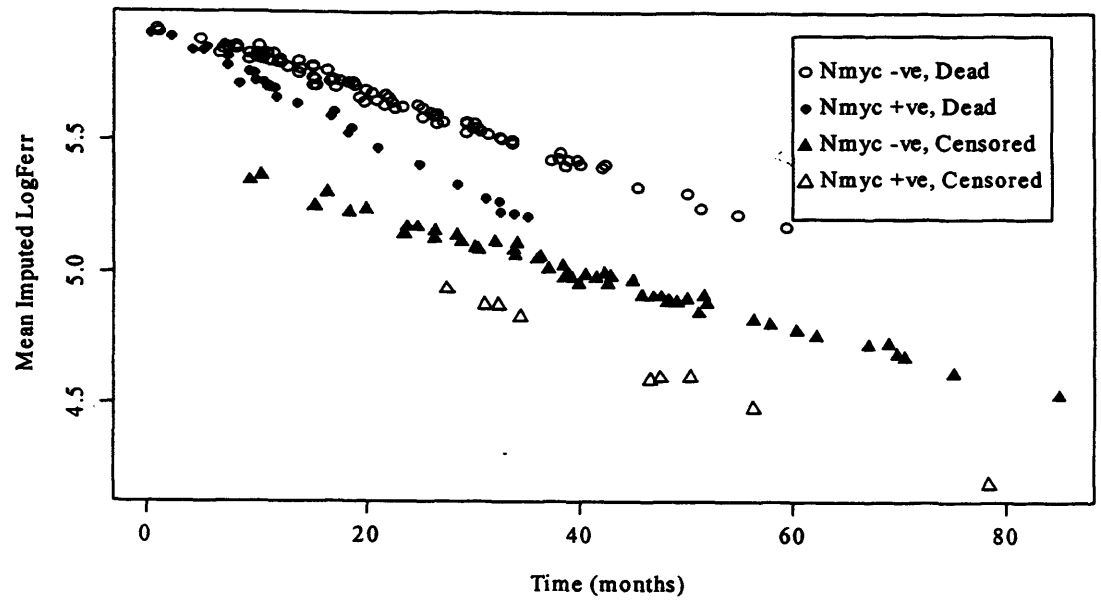

Figure 6.6 Relationship between survival time, censoring status and *Nmyc* with mean imputed values of *Logferr* using an indirect model for the neuroblastoma data.

data in this model than in the complete case model. The parameter estimates for β_1 and β_2 have both increased.

Figure 6.6 shows for the missing values of *Logferr*, a plot of the mean values of the predictive distribution of each missing value obtained when using the indirect model. It can be seen that as survival time increases the mean of the predictive distribution decreases, which is sensible, as lower values of *Logferr* are associated with longer survival. Censored observations appear to show a different pattern from non-censored values and tend to have lower posterior means. This is also sensible, censored individuals will survive longer than their censoring time and that lower values of *Logferr* are associated with longer survival. Thus, one would expect the imputed values of *Logferr* to be lower for a censored time than

Parameter	Indirect Model		Complete Case Model	
	Estimate	95% CrI	Estimate	95% CrI
β_0 (Intercept)	-4.01 (0.185)	-4.38 to -3.65	-3.86 (0.322)	-4.48 to -3.29
β_1 (*Logferr*)	0.37 (0.060)	0.25 to 0.49	0.215 (0.105)	0.02 to 0.41
β_2 (*Nmyc*)	0.57 (0.147)	0.27 to 0.56	0.490 (0.205)	0.09 to 0.88
r (Shape)	1.16 (0.051)	1.06 to 1.16	1.116 (0.086)	0.97 to 1.28

Table 6.8 Comparing indirect and complete case models for the neuroblastoma data.

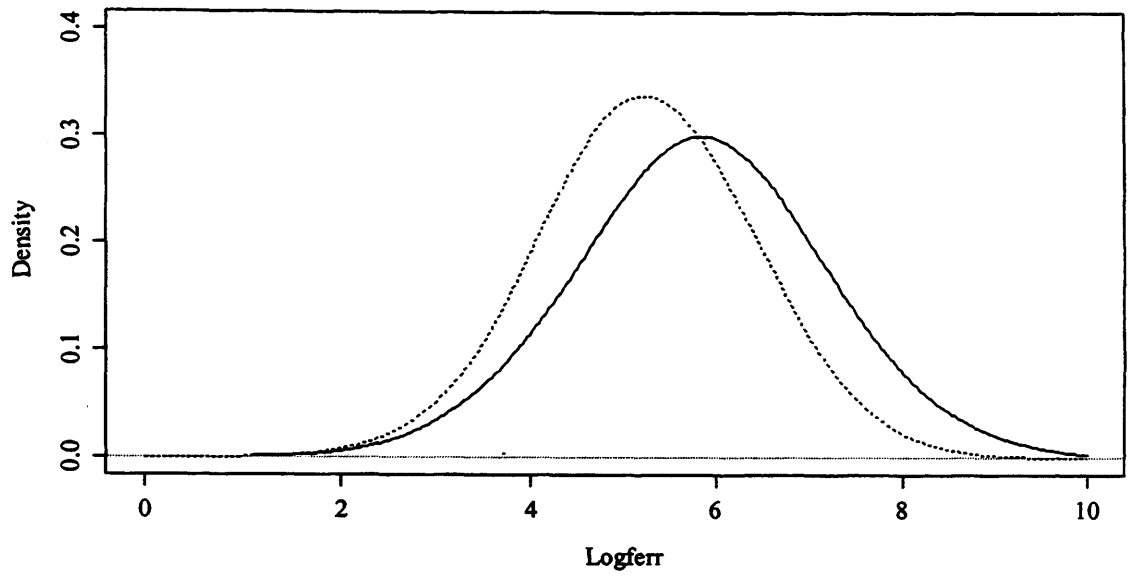

Figure 6.7 Predictive densities for two imputed values of *Logferr* for indirect model.

compared to an equivalent observed survival time where an individual died. The effect of *Nmyc* is small, with the means tending to be slightly lower for amplified values. The effect of *Nmyc* is small, probably because the relationship between *Nmyc* and *Logferr* is relatively weak.

It is important to realise that Figure 6.6 only shows the *mean* value for each missing value of *Logferr* and that each missing value has a predictive distribution. Figure 6.7 shows the predictive distribution for two missing values of *Logferr*. The solid line corresponds to the posterior density for a child who died at 6.6 months and the dotted line corresponds to the posterior density for a child who died at 55 months. Both had non-amplified *Nmyc*. It can be seen that the child who died earlier had a higher mean value for *Logferr*, but there is considerable overlap between the densities, demonstrating that the level of uncertainty associated with the prediction of the missing values is considerable. This uncertainty is appropriately taken into account when obtaining the estimates of the parameters of interest, i.e. the log hazard ratios. One would expect the uncertainty to decrease if the relationship between the two covariates was stronger.

6.8 Direct Model

In section 6.4 I demonstrated that bias can arise when using an indirect model even when the covariates are only moderately correlated. It therefore appears sensible to extend the indirect model for the neuroblastoma data to explicitly allow for the potential relationship between *Nmyc* and *Logferr*. However, it is unlikely to have a dramatic effect in this instance as the relationship between *Logferr* and *Nmyc* is weak. Thus model (6.20) can be extended as follows;

$$\begin{aligned} t_i &\sim Weibull(r, \mu_i) \\ \log(\mu_i) &= \beta_0 + \beta_1 Logferr_i + \beta_2 Nmyc_i \\ Logferr_i &\sim N(\gamma_i, \sigma^2_{Lf}) \\ \gamma_i &= \alpha_0 + \alpha_1 Nmyc_i \\ Nmyc_i &\sim Bernoulli(\delta_i) \\ \log it(\delta_i) &= \theta \end{aligned} \tag{6.24}$$

The important addition to the above model when compared to the indirect model is that it now includes the possibility of a relationship between *Logferr* and *Nmyc*. This can be seen in the DAG shown in Figure 6.8. Prior distributions need to be specified for the additional parameters α_0 and α_1.These are,

$$\alpha_0, \alpha_1 \sim N(0,1000000) \tag{6.25}$$

It is important to note that there is no need to include in the model a term relating *Logferr* to the outcome Nmyc. This can be seen from the functional form of the full conditional distributions,

$$\begin{aligned} p(Logferr_i \mid \cdot) \propto & \prod_{i=n_1+1}^{n_1+n_2} Weib(t_i \mid r, \exp(\beta_0 + \beta_1 Logferr_i + \beta_2 Nmyc_i)) \times \\ & \prod_{i=n_1+1}^{n_1+n_2} N(Logferr_i \mid \alpha_0, \alpha_1, Nmyc_i, \sigma^2_{Lf}) \end{aligned} \tag{6.26}$$

and

Parameter	Direct Model		Indirect Model	
	Estimate	95% CrI	Estimate	95% CrI
β_0 (Intercept)	-4.02 (0.180)	-4.37 to -3.67	-4.01 (0.185)	-4.38 to -3.65
β_1 (*Logferr*)	0.36 (0.061)	0.24 to 0.48	0.37 (0.060)	0.25 to 0.49
β_2 (*Nmyc*)	0.55 (0.151)	0.24 to 0.83	0.57 (0.147)	0.27 to 0.56
R (Shape)	1.16 (0.050)	1.07 to 1.26	1.16 (0.051)	1.06 to 1.16

Table 6.9 Comparison of indirect and direct models for the neuroblastoma data.

$$p(Nmyc_i \mid \cdot) \propto \prod_{i=n_1+n_2+1}^{n_1+n_2+n_3} Weib(t_i \mid r, \exp(\beta_0 + \beta_1 Logferr_i + \beta_2 Nmyc_i)) \times \prod_{i=n_1+n_2+1}^{n_1+n_2+n_3} Bernoulli(Nmyc_i \mid \theta) \prod_{i=n_1+n_2+1}^{n_1+n_2+n_3} N(Logferr_i \mid \alpha_0, \alpha_1, Nmyc_i, \sigma_{Lf}^2) \quad (6.27)$$

It can be seen that for both full conditional distributions there are terms for both covariates, even though *Logferr* is not included in the regression equation for *Nmyc*.

The results of fitting model (6.24) in WinBUGS, with a burn in of 5000 samples and then a further 20000 samples to obtain the estimates, can be seen in Table 6.9. It can be seen that in this case there is very little difference in the estimates or the standard errors between the direct and indirect models. This is because there is very little association between the two

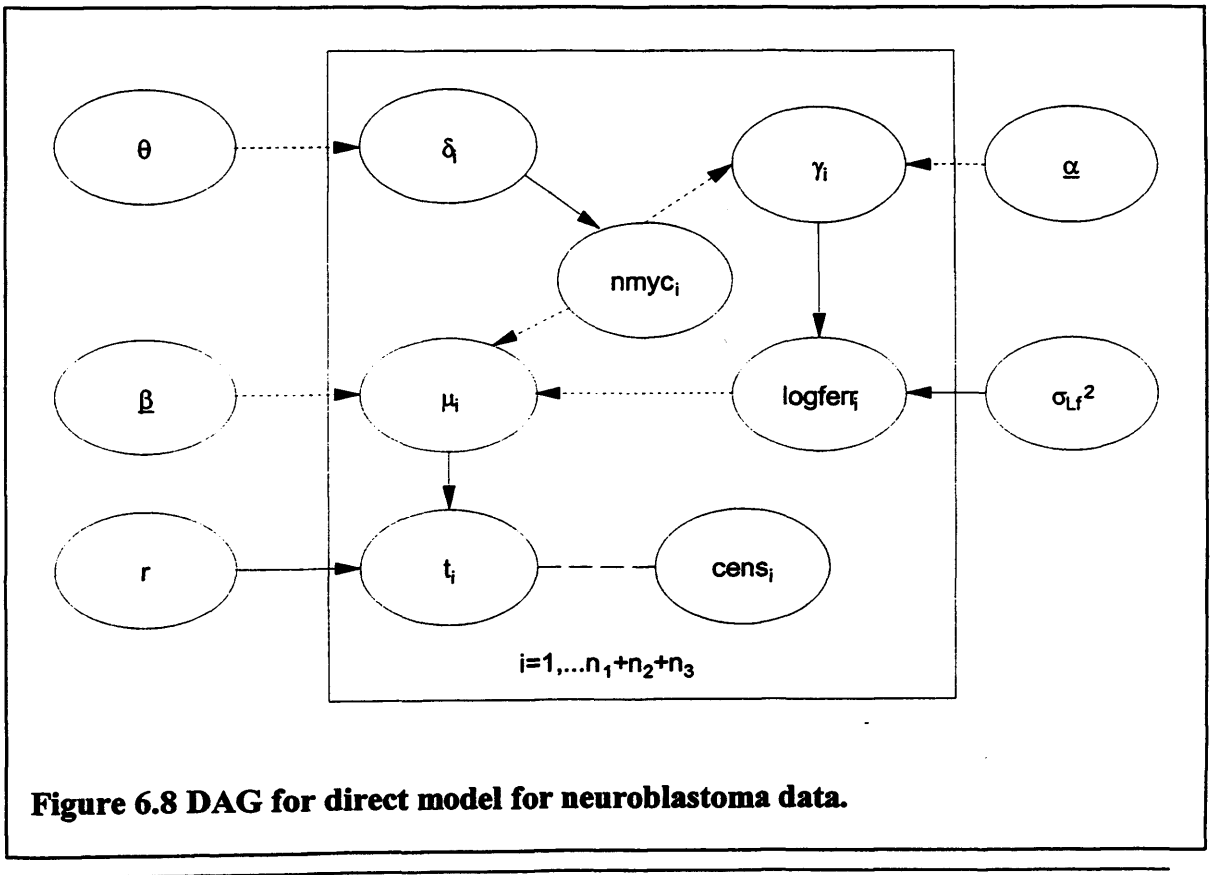

Figure 6.8 DAG for direct model for neuroblastoma data.

covariates. However, as I demonstrated in the simulation study in section 6.4, when the covariates are associated it is likely that the results from an indirect model will be biased. It is probably safer to model the relationship between the covariates even if this is weak. This also gives a way of assessing the strength of the relationship between the covariates.

Figure 6.9, shows the mean of the predictive distributions for each missing value of *Logferr* plotted against time. It is similar to Figure Figure 6.6 for the indirect model and shows a similar relationship. However, using the direct model the difference between negative and positive values of *Nmyc* is not quite so distinct.

6.8.1 Reversing Normal and Binomial Distributions

In the previous section I stated that it should not matter which way the relationship between *Nmyc* and *Logferr* is specified in the model, with either *Nmyc* or *Logferr* being the dependent variable. Figure Figure 6.10 shows the DAG for the direct model where now *Nmyc* is the dependent variable. The model is thus written

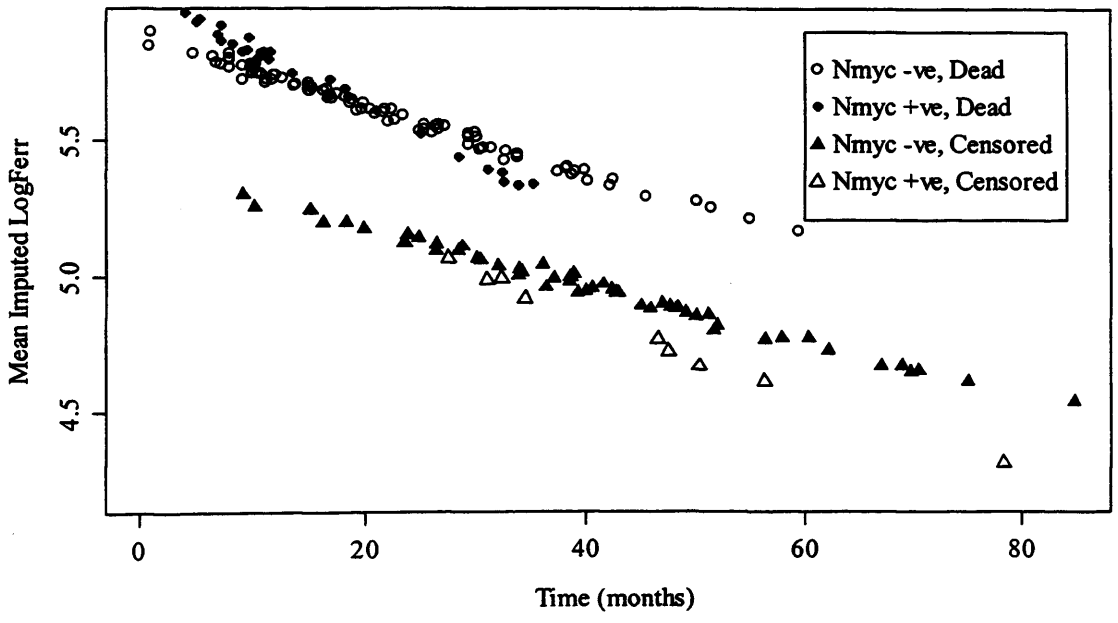

Figure 6.9 Relationship between survival time, censoring status and *Nmyc* with mean imputed values of *Logferr* for direct model.

$$
\begin{aligned}
t_i &\sim Weibull(r, \mu_i) \\
\log(\mu_i) &= \beta_0 + \beta_1 Logferr_i + \beta_2 Nmyc_i \\
Logferr_i &\sim N(\alpha, \sigma_{Lf}^2) \\
Nmyc_i &\sim Bernoulli(\delta_i) \\
\log it(\delta_i) &= \theta_0 + \theta_1 Logferr_i
\end{aligned}
\tag{6.28}
$$

The full conditional distributions for *Logferr* and *Nmyc* for model (6.28) are

$$
\begin{aligned}
p(Logferr_i) \propto & \prod_{i=n_1+1}^{n_1+n_2} Weib(t_i \mid r, \exp(\beta_0 + \beta_1 Logferr_i + \beta_2 Nmyc_i)) \times \\
& \prod_{i=n_1+1}^{n+n_2} N(Logferr_i \mid \alpha, \tau_{\log ferr}) \prod_{i=n_1+1}^{n+n_2} Bernoulli(Nmyc_i \mid \theta_0, \theta_1, Logferr_i)
\end{aligned}
\tag{6.29}
$$

and

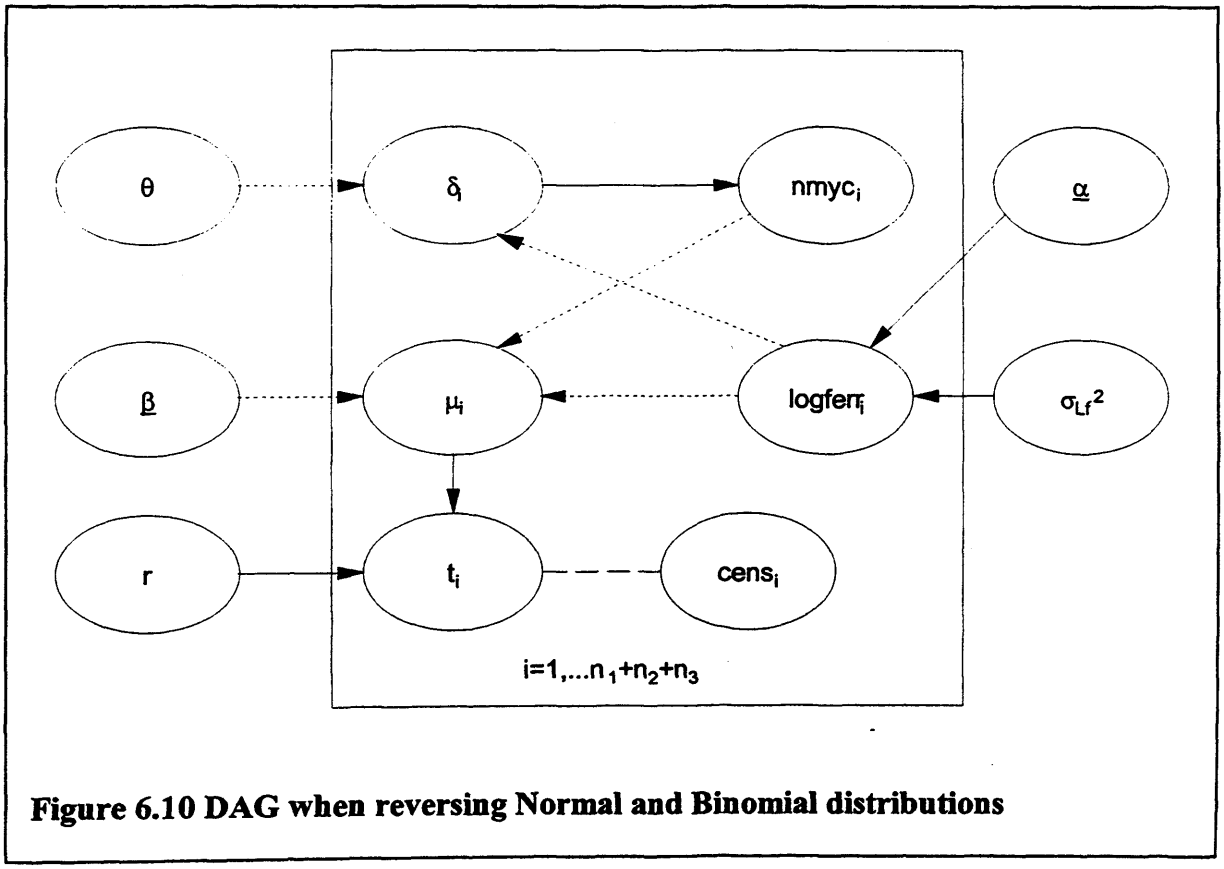

Figure 6.10 DAG when reversing Normal and Binomial distributions

$$p(Nmyc_i) \propto \prod_{i=n_1+n_2+1}^{n_1+n_2+n_3} Weib(t_i \mid r, \exp(\beta_0 + \beta_1 Logferr_i + \beta_2 Nmyc_i)) \times \prod_{i=n_1+n_2+1}^{n+n_2+n_3} Bernoulli(Nmyc_i \mid \theta_0, \theta_1, Logferr_i) \quad (6.30)$$

The results of fitting model (6.28) can be seen in Table 6.10. It can be seen that the estimates are almost identical, which one would expect.

6.9 Inclusion of Extra Covariates

The models I have described so far have been relatively simple in that there are only two covariates and the effect of centre has been ignored. When developing a prognostic model there are likely to be other covariates of interest. It is also possible that the effect of centre is important. One extra covariate that is likely to be related to survival is age. It is important therefore to include this in any clinically useful model. However, age may also be related to *Logferr* and *Nmyc* and it should also be included in any sub-models, i.e. the models that define the inter-relationships between the covariates. If age was strongly related to one of the covariates then it would decrease the uncertainty associated with any missing values associated with that covariate leading to reduced standard errors. Even if age is not related to survival, it could be related to *Logferr* and *Nmyc*, thus leading to improved prediction for the missing values. In order to allow for potential differences between the five centres, I include centre as a random effect in the Weibull regression model. Thus, model (6.24) can be extended to

Parameter	Direct Model		Reversed Direct Model	
	Estimate	95% CrI	Estimate	95% CrI
β_0 (Intercept)	-4.02 (0.180)	-4.37 to -3.67	-4.00 (0.180)	-4.36 to –3.65
β_1 (*Logferr*)	0.36 (0.061)	0.24 to 0.48	0.36 (0.061)	0.24 to 0.48
β_2 (*Nmyc*)	0.55 (0.151)	0.24 to 0.83	0.54 (0.151)	0.24 to 0.83
R (Shape)	1.16 (0.050)	1.07 to 1.26	1.16 (0.050)	1.06 to 1.26

Table 6.10 Comparing direct model with reversed normal / binomial assumption for neuroblastoma data.

$$
\begin{aligned}
& t_i \sim Weibull(r, \mu_i) \\
& \log(\mu_i) = \beta_0 + \beta_1 Logferr_i + \beta_2 Nmyc_i + \beta_3 Age_i + \varepsilon_j \\
& \varepsilon_j \sim N(0, \sigma_\varepsilon^2) \\
& Logferr_i \sim N(\gamma_i, \sigma_{LF}^2) \\
& \gamma_i = \alpha_0 + \alpha_1 Nmyc_i + \alpha_2 Age_{ij} \\
& Nmyc_i \sim Bernoulli(\delta_i) \\
& \log it(\delta_i) = \theta_0 + \theta_1 Age_i
\end{aligned}
\tag{6.31}
$$

where subscript j refers to the j^{th} centre so that the effect of centre is included as a random effect. This is sensible if the survival times in a particular centre are likely to be more similar than the survival times from different centres. The DAG can be seen in Figure 6.11. It can be seen as the number of covariates increases the DAG, and thus the model become more complicated. If the extra covariates having missing data then the model, will become even more complex and care will need to be taken so that all inter-relationships between covariates are being appropriately modelled. Prior distributions need to be given for all model parameters. These were again chosen to be relatively non-informative and are as follows

$$
\begin{aligned}
& \beta_0, \ldots, \beta_3, \alpha_0, \ldots, \alpha_2, \theta_0, \theta_1 \sim N(0,100000) \\
& \frac{1}{\sigma_{Lf}^2}, \frac{1}{\sigma_\varepsilon^2} \sim Gamma(0.001,0.001) \\
& r \sim Gamma(0.001,1)
\end{aligned}
\tag{6.32}
$$

The model was fitted using WinBUGS and again had a 5000 iteration 'burn-in' followed by 20000 samples. The parameter estimates, 95% credible intervals and Geweke Z scores can be seen in Table 6.11.

There is little change in the parameter estimates for β_1 and β_2 from the direct model (6.24), which ignored the effects of centre and age, with the log hazard ratio for *Logferr* increasing slightly and the log hazard ratio for *Nmyc* decreasing slightly. There is very little change in the standard deviations for these parameters. The value of α_1 is small relative to its standard deviation and the corresponding 95% credible interval clearly crosses zero

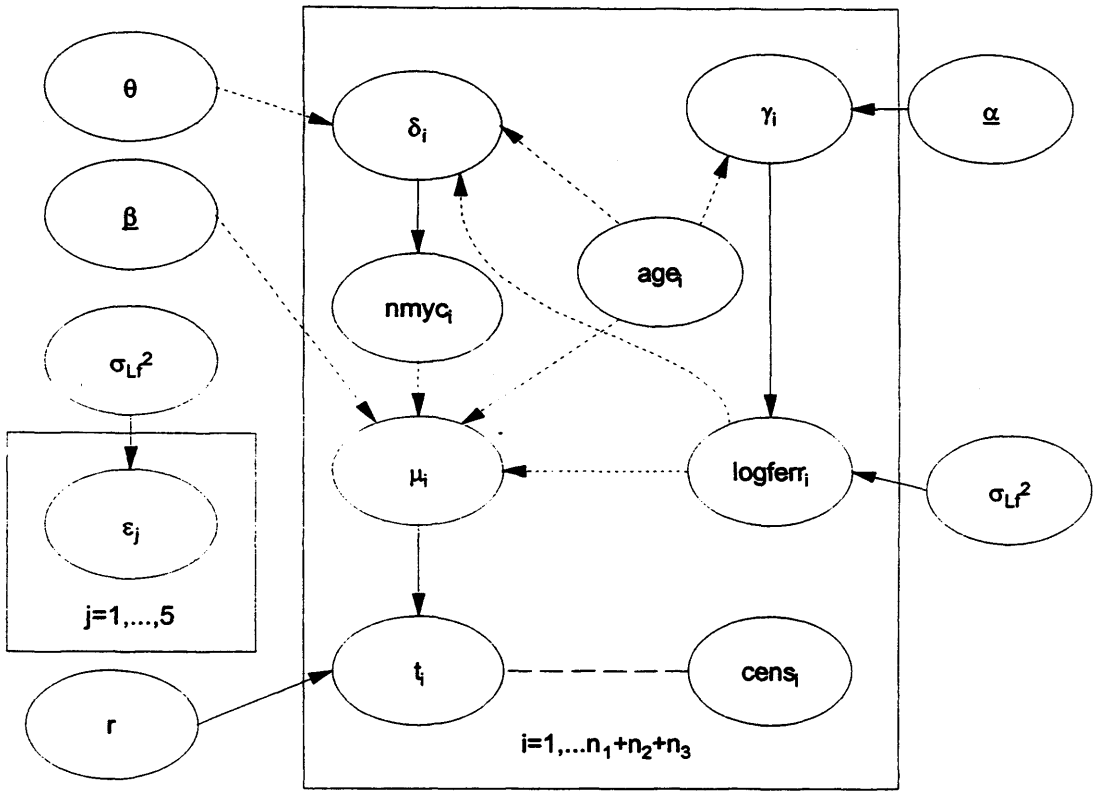

Figure 6.11 DAG for direct model with additional covariates, age and centre (as a random effect).

indicating little association between *Logferr* and *Nmyc*. This explains why the direct and indirect models give broadly similar parameter estimates. The negative coefficient for β_3 indicates that there is a reduction in the hazard as age increases. In addition, age is positively associated with *Logferr* (α_2) and negatively associated with *Nmyc* (θ_1) indicating the *Logferr* tends to increase with age and that the probability of amplified *Nmyc* decreases with age. The between centre variance (σ_ε^2) is small indicating very little variation between centres in terms of survival.

The Geweke scores are also shown in the table. All of the Z scores are clearly less than 2, indicating that there is relatively little evidence of non-convergence. As in Chapter 3, it can be more meaningful to plot the densities of the first 10% of each chain and the last 50% of each chain rather than just to obtain a Z score. These density plots can be seen in Figure 6.12 and show that there is very little difference between the two densities for all the parameters. There are, however, very slight differences for β_0, α_1 ,α_2, θ_0 and r. However,

Parameter		Estimate (Standard deviation)	95% Credible Interval)	Geweke Z score
β_0)	(Intercept	-4.06 (0.210)	(-4.47, -3.64)	0.77
β_1	(*Logferr*)	0.40 (0.063)	(0.28, 0.53)	0.17
β_2	(*Nmyc*)	0.51 (0.156)	(0.19, 0.80)	-1.14
β_3	(*Age*)	-0.05 (0.023)	(-0.10, -0.01)	-0.64
σ_ε^2		0.038 (0.136)	(0.00,0.21)	0.26
R	(Shape)	1.17 (0.052)	(1.07, 1.28)	-1.33
α_0		5.50 (0.071)	(5.36, 5.64)	0.86
α_1	*Inter-relationships between covariates*	0.22 (0.191)	(-0.17, 0.58)	0.14
α_2		0.06 (0.022)	(0.01, 0.10)	-0.21
σ_{Lf}^2		-1.18 (0.089)	(1.02, 1.37)	-0.77
θ_0		-1.18 (0.143)	(-1.47,-0.91)	-0.48
θ_1		-0.26 (0.075)	(-0.41,-0.12)	0.79

Table 6.11 Estimates from direct model with extra covariates

these differences are very small and I would expect them to disappear if a larger sample size was used. This is not to say that the present sample size is too small, but that the sample size is obviously reduced when one only looks at the first 10% of the chain.

6.10 Discussion

In this chapter I have demonstrated methods for dealing with missing data with an example of a meta-analysis including individual patient survival data. In section 6.3 I used a simple example, namely the hypotension data set, to demonstrate how multivariate data can be considered to have a hierarchical structure, and that one can take advantage of this structure when analysing data with missing values. I have shown how this can be done using classical hierarchical models using multiple imputation by generation of multiple imputation data sets using the estimated residuals for the missing values. When using a Bayesian model very similar parameter estimates were obtained. When using MCMC methods the Bayesian model can be considered to be a multiple imputation analysis with 1000's of multiple imputation data sets, as at each iteration of the Gibbs Sampler a value for the missing covariate is sampled from a predictive distribution. However, it must be acknowledged that the hypotension data is a simple example, and in practice one does not

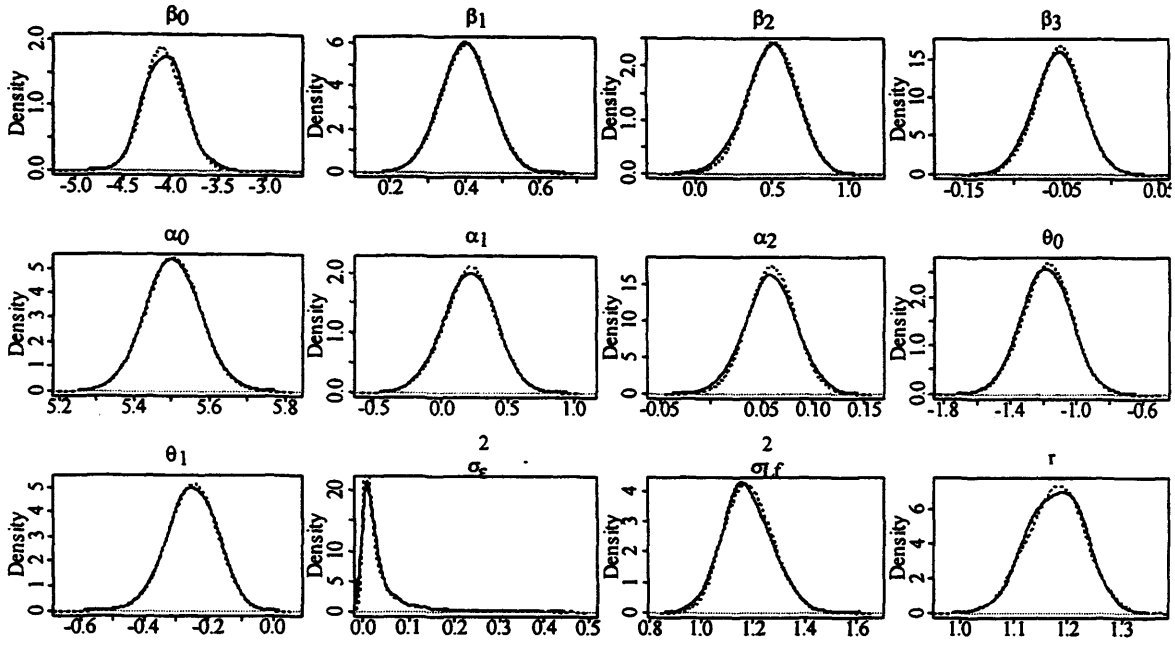

Figure 6.12 Density plots for first 10% (solid line) and last 50% (dotted line) of chains for each parameter.

always deal with continuous data where multivariate normality can be assumed. When dealing with categorical data, censored data, random effects or an unusual distribution the use of a full Bayesian model becomes more appealing. I have demonstrated this with the use of the neuroblastoma data set where the outcome was survival time with some censoring, and missing values for a continuous and a dichotomous covariate. In the simulation in section 6.4 I showed that the use of the indirect model discussed by Arjas and Liu (1996) can lead to serious bias in the parameter estimates when the covariates are correleted. Thus it would seem sensible to recommend that the inter-relationships between covariates should always be modelled unless one has very good reason to believe that the covariates are totally unrelated.

One potential problem with the fully Bayesian analysis is the time taken to fit the models. The final model with the centre random effect and age included in the model with a 5000 iteration 'burn-in' and a further 25000 samples took about 90 minutes on a Pentium II 400 Mhz. Although this is clearly acceptable, when dealing with much larger samples with many more missing covariates, the time taken could increase greatly. However, since values are sampled for each missing value at each iteration, one could generate multiple

imputation data sets, analyse each data set classically and combine the parameter estimates as in section 6.3.1. Care would need to be taken that sufficient iterations had passed between generation of the multiple imputation data sets due to the problem of autocorrelation.

An important question is what covariates to include in the sub-models. A sensible approach is to follow the guidelines for multiple imputation (Rubin, 1996; Schafer, 1997). Following these guidelines, all covariates included the main model should be included in the sub-models. In addition covariates not of interest in the main model could be included in the sub-models. For example, if gender was not related survival then this would not be included in the main model. However, if gender was predictive of *Nmyc* and/or *Logferr* the it would be sensible to include gender as a covariate in the sub-models as this would lead to improved prediction of the missing values. If there is uncertainty of what to include in the sub-models then it may be appropriate to perform a number of sensitivity analyses to see how the choice actually affects the estimates of the parameters of interest.

With a larger number of covariates the missing data part of the model will become more complicated. However, one can use simple parameterisation rules. If $x_1,\cdots,x_p$ are p covariates which have some missing data values, then the joint distribution of the p covariates is the product of a series of conditional distributions, i.e.

$$\begin{aligned} p(x_1,x_2,\ldots,x_p) &= p(x_1 \mid x_2,\ldots,x_p)p(x_2,\ldots,x_p) \\ &= p(x_1 \mid x_2,\ldots,x_p)p(x_2 \mid x_3,\ldots,x_p)p(x_3,\ldots,x_p) \\ &= p(x_1 \mid x_2,\ldots,x_p)p(x_2 \mid x_3,\ldots,x_p)\ldots p(x_{p-1} \mid x_p)p(x_p) \end{aligned} \qquad (6.33)$$

Thus, complex multivariate relationships can be expressed as the product of simple univariate models. For example, with a combination of categorical and continuous variables the *general location mode* (Olkin and Tate, 1961) can be fitted where the categorical variables are assumed to have a multinomial distribution and the continuous variables are assumed to be multivariate normal with separate means for each combination of the categorical variables, but with constant covariance matrix. The missing data model

presented in this chapter is the simplest form of the general location model having just one dichotomous covariate and one continuous covariate.

I have assumed that the survival times have a parametric form, namely they follow a Weibull distribution. This may be slightly restrictive in that it imposes a monotonic shape on the hazard function. Other possibilities include other parametric distributions, for example Gamma or Log-normal, or the use of a Cox proportional hazards model. Although it is possible to use Bayesian methods for fitting proportional hazards models (Kalbfleisch, 1978; Clayton, 1991), fitting can be very slow and in my experience the BUGS program often fails. One possibility is to define a number of time intervals, rather than estimate the hazard at each event time, as in the proportional hazards model, and to assume the hazard is constant within each of these intervals. This can be achieved through the use of a piecewise exponential model (Aitkin *et al.*, 1989). The more time intervals that are chosen, the closer the piecewise exponential model gets to a proportional hazards model.

7 DISCUSSION

7.1 Summary

In this thesis I have demonstrated the use and versatility of hierarchical models through a number of examples, showing that they cover a wide range of areas and applications in medical research. In fact hierarchical models are becoming a common tool for applied statisticians to use, due to many data sets having a clustered form. As with many other statistical techniques, one reason for the increase in the use of hierarchical models is the advancement of computer software. For all examples, I have contrasted the classical approach to estimation and model fitting with the Bayesian approach, with the aim of demonstrating the further versatility that the Bayesian approach can add, especially in complex situations.

7.2 Hierarchical Data and Hierarchical Models

It is becoming increasingly apparent that many forms of data have a clustered form. It is clearly necessary to take this into account. I have shown in section 2.6 that ignoring the hierarchical structure of such data can lead to inappropriate inferences being drawn due to errors in the calculation of the standard errors. There are many types of data that may exhibit a clustered form and I have demonstrated and developed the concept for repeated measures data, meta-analysis data and multivariate data.

It has long being acknowledged that repeated measures data induces a correlation structure. The type of models fitted until about the mid 1980's, namely repeated measures ANOVA and MANOVA, were restrictive in that they assumed that repeated measurements were taken at the same time points with no missing data. They also did not offer the opportunity to model the between unit variation. I have demonstrated how the use of hierarchical models, incorporating suitable random effects, can lead to more realistic repeated measures models that are more concerned with estimation rather than with hypothesis testing.

It is not always realised that meta-analysis data exhibits a hierarchical structure. However, it is clear that if between-study heterogeneity exists then this should be taken into account. Researchers who advocate the use of fixed effects meta-analysis models are ignoring this heterogeneity and thus will end up with pooled effect estimates that are too precise.

Multivariate data also exhibits a hierarchical structure and again it is not always acknowledged that this is the case. The relationship between hierarchical models and multivariate structured data is important to observe as many of the multivariate techniques such as repeated measures MANOVA generally require there to be complete data on all individuals. However, fitting these models within the hierarchical models framework allows there to be a differential number of measures on each subject leading to less wastage in removing subjects with incomplete data. However, as always with missing data, it is important to consider why the data is missing.

7.3 Interpretation of Hierarchical Models

It is clear that hierarchical models are more complex to fit and interpret when compared with standard linear models. There are a number of issues to be decided including definition of the clustering unit(s), how many levels of information there are and what variables should considered to vary between these. In some cases the definition of the clustering unit will be fairly obvious, for example in simple repeated measure or standard meta-analysis problems. In other situation it may be less clear, for example in chapter 4 I used a three level model for the peak flow data as this led to sensible interpretation of the variance components (between-subject, between-day within-subject and within-day within-subject). However, if one was just interested in the fixed effects then just a standard two-level model could be fitted, probably with some auto-regressive term incorporated into the within-subject variance to allow for the fact that measurements taken on the same day tend to be more similar than measurements on different days. For the attributable risk meta-analysis presented in chapter 5 a third level, the type of study, could be incorporated. This may be important if there is variation in the quality of the type of studies and is probably more appropriate when combining randomised and non-randomised evidence (Abrams and Jones, 1995).

Often in hierarchical models the variance components are considered to be nuisance factors rather than of direct interest. For example, in chapter 3 the main interest was in the mean profiles of the two blood pressure profiles. However, it is important to realise that the variance components do have a sensible interpretation that may be clinically meaningful, which is shown in chapter 4 for the peak flow analysis. Thus, when variability is itself of interest then the use of hierarchical models is a sensible approach. However, although the fixed effects tend to be robust to misspecification of the random effects it is unclear whether the reverse is true. In addition the standard errors associated with the variance components may be subject to bias in classical models. Therefore a Bayesian approach may be advisable in this situation.

For the ABPM analysis in chapter 3, I have introduced the possibility of using restricted cubic splines to model the mean profile. These are very powerful in that they are simple to use and can accommodate a wide variety of curved profiles. I have shown how they can be incorporated in to either the fixed or random component of the model. A potential criticism of using restricted cubic splines is the subjective choice of the number and location of the knots. I investigated this through the use of sensitivity analysis by varying the number and location of the knots and found that the models were robust to these changes as long as there were a sufficient number of knots. It may be worthwhile investigating the possibility of using models that treat the number and location of the knots as unknowns, but I feel that these models would probably add little apart from unreasonable amounts of computing time.

7.4 Classical and Bayesian Hierarchical Models

In this thesis I have contrasted model fitting and interpretation from both a classical and Bayesian perspective. When identical models are fitted using both the frameworks, and the number of level 2 units is large, there will be very little difference in the inference being drawn. This was generally the case in models fitted in this thesis as they all consisted of relatively large datasets with vague prior distributions being used, so virtually all information was contained in the likelihood. However, the classical analysis does not take account of the uncertainty associated with the estimates of the variance components, which

becomes important when the number of level 2 units is small. The Bayesian approach incorporates all relevant uncertainty automatically and thus one would expect the standard errors of the fixed effects to be larger with smaller datasets.

Despite leading to similar inferences for equivalent models, I feel that the Bayesian approach has a number of advantages when compared to the classical analyses. The main two of these are incorporating extra complexity and easier interpretation that is clinically meaningful.

When I refer to extra complexity I mean 'realistic complexity' in the context of Best *et al.* (1996). For the ABPM data in chapter 3 it was clear that the within-subject variability varied from subject to subject and that this violated one of the assumptions of a standard hierarchical random effects model. The use of Bayesian models enabled modelling of the within-subject variance. Although this is possible using classical models, it is not straightforward to model any unexplained heterogeneity in the within-subject variance using a between-subject random effect. Interestingly, although there was considerable heterogeneity in the within-subject variance between subjects, modelling this heterogeneity appeared to have very little impact on the mean profiles or more importantly their standard errors. An interesting area of further research would be to investigate if there are any situations when differing within-subject variances would have an impact on the standard errors of the fixed effects. The PEF models in chapter 4 could also be extended to allow for heterogeneity of the within-subject variance. The impact here could be important because interest lies in the quantification of the within-subject variance. In chapter 4 I showed how further complexity could be incorporated into the attributable risk meta-regression model by modelling the observed counts and frequencies using a Poisson or Binomial distribution. This is generally preferable to calculating summary estimates for each study prior to conducting the meta-analysis. In addition, for the infertility example this enabled the case-control study to contribute to both the prevalence estimate and the relative risk estimate. Finally, in chapter 6 I showed how complex models with missing data can be fitted from a Bayesian perspective. The importance here is that standard methods for missing data do not easily allow for censoring and/or the use of random effects. The Bayesian models could be extended to investigate the effect of informative missing data. At present the models

assume that the missing data are missing at random, but in some situations there may be informative missing data. The effect of this on parameter estimates could be investigated through simulation studies. The Bayesian approach has some distinct advantages in the study of informative missing data mechanisms, as they could be modelled in a similar way to Best *et al.* (1996). The main problem with informative missing data is that, by definition, the covariate related to 'missingness' has not been measured and thus assumptions need to be made regarding the missing data mechanism. A sensible approach would be to fit a number of models assuming different missing data mechanisms in order to investigate how robust the results are to different missing data assumptions.

In general I feel that the interpretation of the Bayesian models is simpler. Firstly, although interpretation of the parameter estimates is similar, when fitting classical models one should interpret any uncertainties in these estimates, either through p-values or confidence intervals, in the usual non-intuitive way as discussed in section 2.5.2. The Bayesian quantification of uncertainty, either through probability statements or the use of credible intervals, which I have used throughout this thesis, is much more intuitive. In chapter 4 I show how the use of density plots can be useful for quantification of uncertainty for variance components. As these models may appear complex to non-statisticians, it is important to present results in a way they can be clearly understood. I feel that this is easier to do within the Bayesian framework.

In this thesis I have used vague prior distributions for the Bayesian analyses, with the aim that virtually all the information regarding parameter estimates is contained in the likelihood. Although informative prior distributions could be used in the analyses I have presented, I feel that they would be of little benefit as the models have generally included large sets of data. Where there may be need for further research is when variances are being estimated with sparse amounts of data, which is often the case in meta-analyses. The use of a Gamma(0.001,0.001) prior distribution may actually be informative when data is sparse and/or the variance estimate is very small. The sensitivity of results to the choice of different vague priors is an area for future research, but the exploration of the sensitivity of results to various model assumptions is a crucial element to any Bayesian analysis.

I have shown how multivariate normality can be expressed using a series of conditional normal distributions known as the product normal formulation in chapters 3 and 6. In chapter 4 I took a different approach using the standard multivariate normal distribution available in WINBUGS. The disadvantage of the latter is that it requires a Wishart distribution to be used as a prior distribution. I find the Wishart distribution non-intuitive and would find it difficult to use informative priors using this formulation. Using the product normal formulation enables standard univariate prior distributions to be used. For the missing data models presented in chapter 6 it would not be possible to use the built in multivariate distribution in WINBUGS as it can not currently cope with missing data. In addition the use of the product normal formulation could aid the use of informative priors when modelling the relationship between covariates as the information may be available from external studies.

7.5 Model Comparison

An important issue I have not covered in this thesis is comparison of models from a Bayesian perspective. In classical models comparison of nested models is fairly straightforward using the likelihood ratio test. Non-nested models can be compared using the Bayesian Information Criterion (BIC) and the Akaike Information Criterion (AIC) as seen in Chapter 3. One possibility is the use of *Bayes Factors* (Kass, 1993; Kass and Raftery, 1995), which compares ratios of model probabilities, so for two models (M_1 and M_2) and complete data Y, the Bayes Factor is $p(Y \mid M_1)/p(Y \mid M_2)$. However, there are complications in the calculation of Bayes Factors, particularly in complex models such as some of those presented here. One possible solution is to compare the conditional predictive ordinates (CPO's) for each observation in the model (Pettit, 1990). For observation y_i, complete data Y, the CPO for the i^{th} observation can be defined as $p(y_i \mid Y_{(i)})$ where $Y_{(i)}$ is the complete data excluding y_i. Comparison of the CPO's from different models gives the preference for each model for each observation. The ratios of the sum of the CPO's can then be combined to obtain *Pseudo Bayes Factors* (Kass and Raftery, 1995). An advantage of the use of CPO's and pseudo Bayes Factors is that they can easily be incorporated into WinBUGS by the addition of two lines of code

(Spiegelhalter *et al.*, 1996). However, further research is required into their use and interpretation.

An area of my thesis where model comparison may be important is in reducing the complexity of the missing data models, particularly in problems with a large number of covariates. With a large number of covariates there will be many inter-relationships between the covariates that need to be modelled, but some of these inter-relationships may be negligible and a reduced model could exclude these. However, there are problems with the definition of 'negligible' and also the problem of comparing different models.

7.6 Conclusions

In this thesis I have demonstrated and developed a wide range of uses for hierarchical models and shown that they can be used to tackle a number of major issues in medical research. I have compared classical and Bayesian approaches and generally believe that the Bayesian approach offers a number of clear advantages over the classical approach. With the advances in computer technology and computationally intensive techniques, such as MCMC methods, there will be the potential to fit more and more complex, but realistic and appropriate, hierarchical models. This is likely to bring with it many challenges and research opportunities for the applied statistician.

BIBLIOGRAPHY

Abrams, K.R., Ashby, D. and Errington, R.D. (1996) Bayesian Weibull survival models - an application to a cancer clinical trial. *Journal of Lifetime Data Analysis* **2**, 159-174.

Abrams, K.R. and Jones, D.R. (1995) Meta-analysis and the synthesis of evidence. *IMA J Math Appl Med Biol* **12**, 297-313.

Ahn, H.S. and Loh, W.Y. (1994) Tree structured proportional hazards regression modelling. *Biometrics* **50**, 471-485.

Aitkin, M.; Anderson, D.; Francis, B., and Hinde, J (1989). *Statistical modelling in GLIM.* Oxford Science Publications: Oxford.

Aitkin, M. and Longford, N. (1986) Statistical modeling issues in school effectiveness studies (with discussion). *Journal of The Royal Statistical Society Series A-General* **149**, 1-43.

Akaike, H. (1980) Likelihood and the Bayes procedure. In: Bernardo, J.M., DeGroot, M.H., Lindley, D.V. and Smith, A.J.M., (Eds.) *Bayesian Statistics*, pp. 144-166. Valencia: University Press

Arjas, E. and Liu, L.P. (1996) Nonparametric Bayesian-approach to hazard regression - a case- study with a large number of missing covariate values. *Statistics in Medicine* **15**, 1757-1770.

Armitage, P. and Berry, G. (1987) *Statistical Methods in Medical Research*, 2nd edn. Blackwell Scientific Publications: Oxford.

Ayala, D.E., Hermida, R.C., Mojon, A., Fernandez, J.R. and Iglesias, M. (1997) Circadian blood pressure variability in healthy and complicated pregnancies. *Hypertension* **30**, 603-610.

Ayres, J.G. and Turpin, P.J. (1997) *Peak flow measurement. An illustrated guide*, Chapman and Hall: London.

Bailey, R.H. and Bauer, J.H. (1993) A review of common errors in the indirect measurement of blood pressure. *Archives of Internal Medicine* **153**, 2741-2748.

Beacon, H.J. and Thompson, S.G. (1996) Multi-level models for repeated measurement data: application to quality of life data in clinical trials. *Statistics in Medicine* **15**, 2717-2732.

Beacon, H.J., Thompson, S.G. and England, P.D. (1998) The analysis of complex patterns of longitudinal binary response: an example of transient dysphagia following radiotherapy. *Statistics in Medicine* **17**, 2551-2561.

Berry, D.A. and Stangl, D.K. (1996) *Bayesian Biostatistics*, Dekker: New York.

Berthold, F., Kassenbohmer, R. and Zieschang, J. (1994) Multivariate evaluation of prognostic factors in localized neuroblastoma. *American Journal Of Pediatric Hematology Oncology* **16**, 107-115.

Berthold, F., Trechow, R., Utsch, S. and Zieschang, J. (1992) Prognostic factors in metastatic neuroblastoma - a multivariate- analysis of 182 cases. *American Journal Of Pediatric Hematology Oncology* **14**, 207-215.

Best, N.G., Cowles, M.K. and Vines, S.K. (1995) *CODA manual version 0.39*, MRC Biostatistics Unit: Cambridge, UK.

Best, N.G., Spiegelhalter, D.J., Thomas, A. and Brayne, C.E.G. (1996) Bayesian-analysis of realistically complex-models. *Journal of The Royal Statistical Society Series A-Statistics In Society* **159**, 323-342.

Bethesda, M.D. (1995) *Global Initiative for Asthma. Global Strategy for asthma management and prevention*, National Institute of Health: NIH publication No96-3659A.

Biggerstaff, B.J. and Tweedie, R.L. (1997) Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Statistics in Medicine* **16**, 753-768.

Bingham, C., Arbogast, B., Cornelissen, G., Lee, J. and Hallberg, F. (1982) Inferential statistical methods for estimating and comparing cosinor parameters. *Chronobiologia* **9**, 397-439.

Breslow, N.E. and Clayton, D.G. (1993) Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9-25.

Breslow, N.E. and Day, N.E. (1980) *Statistical methods in cancer research. Volume 1 - the analysis of case-control studies*, International Agency for Research on Cancer: Lyon.

British Thoratic Society (1996) The British guidelines on asthma management. 1995 review and position statement. *Thorax* **52(suppl)**, S1-S21

Brodeur, G.M., Seeger, R.C., Barrett, A., Berthold, F., Castleberry, R.P., D'Angio, G., De Bernardi, B., Evans, A.E., Favrot, M., Freeman, A. *et al.* (1988) International criteria for diagnosis, staging, and response to treatment in patients with neuroblastoma. *Journal of Clinical Oncology* **6**, 1874-1881.

Brooke, A.M., Lambert, P.C., Burton, P.R., Clarke, C., Luyt, D.K. and Simpson, H. (1995) The natural-history of respiratory symptoms in preschool-children. *American Journal of Respiratory And Critical Care Medicine* **152**, 1872-1878.

Brooke, A.M., Lambert, P.C., Burton, P.R., Clarke, C., Luyt, D.K. and Simpson, H. (1996) Night cough in a population-based sample of children - characteristics, relation to symptoms and associations with measures of asthma severity. *European Respiratory Journal* **9**, 65-71.

Brooks, S.P. (1998) Markov Chain Monte Carlo method and its application. *The Statistician* **47**, 69-100.

Brooks, S.P. and Gelman, A. (1998) General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* **7**, 434-455.

Browne, W.J. and Draper, D. (2000) A comparison of Bayesian and likelihood methods for fitting multilevel models. *Journal Of The Royal Statistical Society Series B-Methodological* to appear

Bryk, A.D. and Raudenbush, S.W. (1992) *Hierarchical Linear Models*, Sage Publications: London.

Bryk, A.D., Raudenbush, S.W., Seltzer, M. and Congdon, R. (1988) *An introduction to HLM: Computer program and user's guide*, University of Chicago Department of Education: Chicago.

Burke, M.J., Towers, H.M., O'Malley, K. and et, a. (1982) Sphygmomanometers in hospital and family practice: problems and recommendations. *British Medical Journal* **285**, 469-471.

Burton, P., Gurrin, L. and Sly, P. (1998) Extending the simple linear regression model to account for correlated responses: An introduction to generalized estimating equations and multi-level mixed modelling. *Statistics in Medicine* **17**, 1261-1291.

Burton, P.R., Tiller, K.J., Gurrin, L.C., Cookson Wocm, Musk, A.W. and Palmer, L.J. (1999) Genetic Variance Components Analysis for Binary Phenotypes Using Generalized Linear Mixed Models (GLMMs) and Gibbs Sampling. *Genetic Epidemiology* **17**, 118-140.

Carlin, B.P. (1996) Hierarchical longitudinal modelling. In: Gilks, W.R., Richardson, S. and Spiegelhalter, D.J., (Eds.) *Markov Chain Monte Carlo in Practice*, pp. 303-319. London: Chapman and Hall

Carlin, B.P. and Louis, T.A. (1996) *Bayes and Empirical Bayes Methods for Data Analysis*, Chapman and Hall: London.

Castleberry, R.P., Pritchard, J., Ambros, P., Berthold, F., Brodeur, G.M., Castel, V., Cohn, S.L., De Bernardi, B., Dicks-Mireaux, C., Frappaz, D., Haase, G.M., Haber, M., Jones, D.R., Joshi, V.V., Kaneko, M., Kemshead, J.T., Kogner, P., Lee, R.E., Matthay, K.K., Michon, J.M., Monclair, R., Roald, B.R., Seeger, R., Shaw, P.J., Shuster, J.J. *et al.* (1997) The International Neuroblastoma Risk Groups (INRG): a preliminary report. *Eurpean Journal of Cancer* **33**, 2113-2116.

Churchill, D., Perry, I.J. and Beevers, D.G. (1997) Ambulatory blood pressure in pregnancy and fetal growth. *Lancet* **349**, 7-10.

Ciampi, A., Hogg, S.A. and Kates, L. (1986) Stratification by stepwise regression correspondence analysis and recursive partitioning: a comparison of three methods of survival analysis with covariates. *Computational Statistics and Data Analysis*

Clayton, D.G. (1991) A monte-carlo method for Bayesian-inference in frailty models. *Biometrics* **47**, 467-485.

Clough, J.B., Williams, J.D. and Holgate, S.T. (1991) Effect of atopy on the natural-history of symptoms, peak expiratory flow, and bronchial responsiveness in 7-year-old and 8- year-old children with cough and wheeze - a 12-month longitudinal- study. *American Review Of Respiratory Disease* **143**, 755-760.

Coats, A.J.S., Radaelli, A., Clark, S.J., Conway, J. and Sleight, P. (1992) The influence of ambulatory blood-pressure monitoring on the design an interpretation of trials in hypertension. *Journal of Hypertension* **10**, 385-391.

Cochran, W.G. (1954) The combination of estimates from different experiments. *Biometrics* **10**, 101-129.

Collett, D. (1994) *Modelling survival data in medical research*, Chapman and Hall: London.

Combaret, V., Gross, N., Lasset, C., Frappaz, D., Peruisseau, G., Philip, T., Beck, D. and Favrot, M.C. (1996) Clinical relevance of CD44 cell-surface expression and N-myc gene amplification in a multicentric analysis of 121 pediatric neuroblastomas. *Journal of Clinical Oncology* **14**, 25-34.

Conway, J. and Coats, A.J.S. (1991) Ambulatory blood-pressure monitoring in the design of antihypertensive drug trials. *Journal of Hypertension* **9**, S57-S58

Corrao, S., Scaglione, R., Arnone, S., Amico, G., Amato, V., Licata, A., Bova, A. and Licata, G. (1996) Analysis of 24-h noninvasive ambulatory blood-pressure profiles by a 3rd-degree polynomial approach. *Journal Of Cardiovascular Diagnosis And Procedures* **13**, 237-242.

Cowles, M.K. and Carlin, B.P. (1996) Markov-chain monte-carlo convergence diagnostics - a comparative review. *Journal of the American Statistical Association* **91**, 883-904.

Cox, D.R. (1972) Regression models and life tables (with discussion). *Journal of the Royal Statistical Society (B)* **74**, 187-220.

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

Crowder, M.J. and Hand, D.J. (1993) *Analysis of Repeated Measures*, London: Chapman and Hall.

Davey-Smith, G., Song, F. and Sheldon, T.A. (1993) Cholesterol lowering and mortality: the importance of considering initial level of risk. *British Medical Journal* **306**, 1367-1373.

De Boor, C. (1978) *A parctical guide to splines*, Springer-Verlag: New York.

de Finetti, B. (1972) *Probability, induction and statistics*, Wiley: London.

De Groot, M.E. (1970) *Optimal Statistical Decisions*, McGraw-Hill: New York.

de Leuuw, J. and Kreft, I.G.G. (1995) Questioning multilevel models. *Journal of Educational and Behavioral Statistics* **20**, 171-190.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society (B)* **39**, 1-8.

Dempster, A.P., Rubin, D.B. and Tsutakawa, R.K. (1981) Estimation in covariance components models. *Journal of the American Statistical Association* **76**, 341-353.

Denison, D.G.T., Mallick, B.K. and Smith, A.F.M. (1998) Automatic Bayesian curve fitting. *Journal Of The Royal Statistical Society Series B-Statistical Methodology* **60**, 333-350.

Dersimonian, R. and Laird, N. (1986) Meta-analysis in clinical trials. *Controlled Clinical Trials* **7**, 177-188.

Dickersin, K., Chan, S., Chalmers, T.C., Sacks, H.S. and Smith, H.J. (1992) Meta-analysis: state-of-the-science (review). *Epidemiol Rev* **14**, 154-176.

Dickson, D. and Hasford, J. (1992) 24-hour blood pressure measurement in antihypertensive drug trials: data requirements and methods of analysis. *Statistics in Medicine* **11**, 2147-2158.

Diggle, P.J. (1988) An approach to the analysis of repeated measurements. *Biometrics* **44**, 959-971.

Diggle, P.J., Liang, K.Y. and Zeger, S.L. (1994) *Analysis of Longitudinal Data*, Oxford: Oxford University Press.

Donner, A. (1998) Some aspects in the design and analysis of cluster randomized trials. *Applied Statistics* **47**, 95-113.

Draper, D. (1998) *Bayesian Hierarchical Modelling (Draft of Chapters 1 and 2)*, Department of Mathematical Sciences: University of Bath.

Draper, E.S., Kurinczuk, J.J., Abrams, K.R. and Clarke, M. (1999) Assessment of separate contributions to perinatal mortality of infertility history and treatment: A case-control analysis. *Lancet* **353**, 1746-1749.

Droitcour, J., Silberman, G. and Chelimsky, E. (1993) Cross-design synthesis: a new form of meta-analysis for combining results from randomized clinical trials and medical practice databases. *Int J Technol Assess Health Care* **9**, 440-449.

DuMouchel, W.H. and Harris, J.E. (1983) Bayes methods for combining the results of cancer studies in humans and other species (with comment). *Journal of the American Statistical Association* **78**, 293-308.

Durrleman, S. and Simon, R. (1989) Flexible regression models with cubic splines. *Statistics in Medicine* **8**, 551-561.

Enright, P.L., Lebowitz, M.D. and Cockroft, D.W. (1994) Physiological measures - pulmonary-function tests - asthma outcome. *American Journal Of Respiratory And Critical Care Medicine* **149**, S9-S18

Favrot, M.C., Ambros, P., Schilling, F., Frappaz, D., Combaret, V., Berthold, F., Dominici, C., Erttmann, R., Esteve, J., Jenkner, A., Kerbl, R., Mann, J., Mathieu, P., Parker, L., Powell, J. and Philip, T. (1996) Comparison of the diagnostic and prognostic value of biological markers in neuroblastoma. Proposal for a common methodology of analysis. SENSE group. *Annals of Oncology* **7**, 607-611.

Fleiss, J.L. (1993) The statistical basis of meta-analysis. *Statistical Methods in Medical Research* **2**, 121-145.

Freedman, J.H. and Silverman, B.W. (1989) Flexible parsimonious smoothing and additive modelling (with discussion). *Technometrics* **31**, 3-39.

Galland, R. and Fuller, W. (1973) Fitting segmented polynomial regression models whose join points have to be estimated. *Journal of the Americam Statistical Association* **68**, 144-147.

Gatzka, C.D. and Schmieder, R.E. (1995) Improved classification of dippers by individualized analysis of ambulatory blood pressure profiles. *American Journal of Hypertension* **8**, 666-671.

Gelfand, A.E., Sahu, S.K. and Carlin, B.P. (1995a) Efficient parametrizations for generalized linear mixed models. In: Bernardo, J.M., Berger, J., Dawid, A.P. and Smith, A.F.M., (Eds.) *Bayesian Statistics 5*, Oxford University Press: Oxford.

Gelfand, A.E., Sahu, S.K. and Carlin, B.P. (1995b) Efficient parametrizations for normal linear mixed models. *Biometrika* **82**, 479-488.

Gelfand, A.E. and Smith, A.F.M. (1990) Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398-409.

Gelman, A. (1996) Inference and monitoring convergence. In: Gilks, W.R., Richardson, S. and Spiegelhalter, D.J., (Eds.) *Markov Chain Monte Carlo in Practice*, Chapman and Hall: London.

Gelman, A. and Rubin, D.B. (1992) Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science* **7**, 457-511.

Germano, G., Damiani, S., Germano, U., Pecchioli, V., Pica, B. and Antonini, P. (1990) Evaluation of the effect-duration of once-daily enalapril compared with once-daily captopril. *Nephron* **55**, 65-69.

Geweke, J. (1992) Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In: Bernardo, J.M., Berger, J., Dawid, A.P. and Smith, A.J.M., (Eds.) *Bayesian Statistics 4*, pp. 169-193. Oxford: Oxford University Press.

Gilks, W.R. (1996) Full conditional distributions. In: Gilks, W.R., Richardson, S. and Spiegelhalter, D.J., (Eds.) *Markov Chain Monte Carlo in Practice*, Chapman and Hall: London.

Gilks, W.R., Clayton, D.G., Spiegelhalter, D.J., Best, N.G., Mcneil, A.J., Sharples, L.D. and Kirby, A.J. (1993) Modeling complexity - applications of gibbs sampling in medicine. *Journal of the Royal Statistical Society Series B-Methodological* **55**, 39-52.

Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1996) *Markov Chain Monte Carlo in Practice*, Chapman and Hall: London.

Gilks, W.R. and Roberts, G.O. (1996) Strategies for improving MCMC. In: Gilks, W.R., Richardson, S. and Spiegelhalter, D.J., (Eds.) *Markov Chain Monte Carlo in Practice*, Chapman and Hall: London.

Gilks, W.R., Wang, C.C., Yvonnet, B. and Coursaget, P. (1993) Random-effects models for longitudinal data using Gibbs sampling. *Biometrics* **49**, 441-453.

Gilks, W.R. and Wild, P. (1992) Adaptive rejection sampling for Gibbs sampling. *Applied Statistics* **41**, 337-348.

Glass, G.V. (1976) Primary, secondary and meta-analysis of research. *Educ Res* **5**, 3-8.

Goldstein, H. (1986) Multilevel mixed linear-model analysis using iterative generalized least-squares. *Biometrika* **73**, 43-56.

Goldstein, H. (1989) Restricted unbiased iterative generalized least-squares estimation. *Biometrika* **76**, 622-623.

Goldstein, H. (1995) *Multilevel Statistical Models*, London: Edward Arnold.

Goldstein, H., Healy, M.J.R. and Rasbash, J. (1994) Multilevel time-series models with applications to repeated- measures data. *Statistics in Medicine* **13**, 1643-1655.

Goldstein, H. and McDonald, R.P. (1988) A general model for the analysis of multilevel data. *Psychometrika* **53**, 455-467.

Goldstein, H. and Rasbash, J. (1992) Efficient computational procedures for the estimation of parameters in multilevel models based on iterative generalized least-squares. *Computational Statistics & Data Analysis* **13**, 63-71.

Greenland, S. and Finkle, W.D. (1995) A critical-look at methods for handling missing covariates in epidemiologic regression-analyses. *American Journal of Epidemiology* **142**, 1255-1264.

Guttman, I. (1982) *Linear Models: An Introduction*, John Wiley & Sons: New York.

Hallberg, F. (1969) Chronobiology. *Annual Review of Physiology* **31** , 675-725.

Hamid, S., Corden, Z.M., Ryan, D.P., Burnett, I. and Cochrane, G.M. (1998) Evaluation of an electronic hand-held spirometer in patients with asthma. *Respiratory Medicine* **82**, 1177-1180.

Hammond, C.B. (1994) Infertility. In: Scott, J.R., Disaia, P.J., Hammond, C.B. and Spellacy, W.N., (Eds.) *Danforth's Obstetrics and Gynaecology, 7th Edition*, Philadelphia: JB Lippencott Co.

Hand, D.J. and Crowder, M.J. (1996) *Practical Longitudinal Data Analyses*, London: Chapman and Hall.

Hardy, R.J. and Thompson, S.G. (1996) A likelihood approach to meta-analysis with random effects. *Statistics in Medicine* **15**, 619-629.

Harrell, F.E., Lee, K.L. and Mark, D.B. (1996) Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* **15**, 361-387.

Heinzl, H. and Kaider, A. (1997) Gaining more flexibility in Cox proportional hazards regression models with cubic spline functions. *Computer Methods And Programs In Biomedicine* **54**, 201-208.

Herndon, J.E. and Harrell, F.E. (1995) The restricted cubic spline as base-line hazard in the proportional hazards model with step-function time-dependent covariables. *Statistics in Medicine* **14**, 2119-2129.

Hetzel, M.R. and Clark, T.J.H. (1980) Comparison of normal and asthmatic circadian rhythms in peak expiratory flow rate. *Thorax* **35**, 732-738.

Higgins, B.G., Britton, J.R., Chinn, S., Cooper, S., Burney, P.G.J. and Tattersfield, A.E. (1992) Comparison of bronchial reactivity and peak expiratory flow variability measurements for epidemiologic studies. *American Review of Respiratory Disease* **145**, 588-593.

Hunink, M.G.M. and Wong, J.B. (1994) Meta-analysis of failure time data with adjustment for covariates. *Medical Decision Making* **14**, 59-70.

Idema, R.N., Gelsema, E.S., Wenting, G.J., Grashuis, J.L., Van Den Meiracker, A.H., Brouwer, R.M.L. and Veld, A.J. (1992) A new model for diurnal blood pressure profiling: square wave fit compared with conventional methods. *Hypertension* **19**, 595-605.

James, I.R. (1995) A note on the analysis of censored regression data by multiple imputation. *Biometrics* **51**, 358-362.

Jeng, G.T., Scott, J.R. and Burmeister, L.F. (1995) A comparison of metaanalytic results using literature vs individual patient data - paternal cell immunization for recurrent miscarriage. *Journal of the American Medical Association* **274**, 830-836.

Kadane, J.B. (1995) Prime time for Bayes. *Controlled Clinical Trials* **16**, 313-318.

Kalbfleisch, J.D. (1978) Non-parametric Bayesian analysis of survival time data. *Journal of the Royal Statistical Society (B)* **40**, 214-221.

Kass, R.E. (1993) Bayes factors in practice. *The Statistician* **42**, 551-560.

Kass, R.E. and Raftery, A.E. (1995) Bayes factors. *Journal of the American Statistical Association* **90**, 773-795.

Klein, J.P. (1997) *Survival Analysis: techniques for censored and truncated data*, Springer: London.

Kong, A., Liu, J.S. and Wong, W.H. (1994) Sequential imputations and Bayesian missing data problems. *Journal of the American Statistical Association* **89**, 278-288.

Kotses, H., Haver, A. and Creer, T.L. (1984) An intraindividual comparison of standard and mini-Wright scores. *Annals of Allergy* **52**, 419-422.

Kreft, I.G.G., Deleeuw, J. and Vanderleeden, R. (1994) Review of 5 multilevel analysis programs - BMDP-5V, GENMOD, HLM, ML3, VARCL. *American Statistician* **48**, 324-335.

L'Abbe, K.A., Detsky, A.S. and O'Rourke, K. (1987) Meta-analysis in clinical research. *Ann Intern Med* **107**, 224-233.

Ladenstein, R., Philip, T., Lasset, C., Hartmann, O., Garaventa, A., Pinkerton, R., Michon, J., Pritchard, J., Klingebiel, T., Kremens, B., Pearson, A., Coze, C., Paolucci, P., Frappaz, D., Gadner, H. and Chauvin, F. (1998) Multivariate analysis of risk factors in stage 4 neuroblastoma patients over the age of one year treated with megatherapy and stem-cell transplantation: a report from the European Bone Marrow Transplantation Solid Tumor Registry. *Journal of Clinical Oncology* **16**, 953-965.

Laird, N.M. and Ware, H. (1982) Random-effects models for longitudinal data. *Biometrics* **38**, 963-974.

Lambert, P.C. and Abrams, K.R. (1995) Meta-analysis using multilevel models. *Multilevel Modelling Newsletter* **7**, 17-19.

Langford, I.H. and Lewis, T. (1998) Outliers in multilevel data. *Journal Of The Royal Statistical Society Series A-Statistics In Society* **161**, 121-153.

Lebowitz, M.D., Holberg, C.J., Knudson, R.J. and Burrows, B. (1987) Longitudinal-study of pulmonary-function development in childhood, adolescence, and early adulthood - development of pulmonary-function. *American Review of Respiratory Disease* **136**, 69-75.

Lessler, J.T. and Kalsbeek, W.D. (1992) *Nonsampling error in surveys*, Wiley: New York.

Levin, M.L. (1958) The occurrence of lung cancer in man. *Acta Un Int Cancer* **9**, 531-541.

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

Lilford, R.J. and Braunholtz, D. (1996) The statistical basis of public policy - a paradigm shift is overdue. *British Medical Journal* **313**, 603-607.

Lin, X.H., Raz, J. and Harlow, S.D. (1997) Linear mixed models with heterogeneous within-cluster variances. *Biometrics* **53**, 910-923.

Lindley, D.V. (1985) *Making decisions*, Wiley: London.

Lindley, D.V. and Smith, A.F.M. (1972) Bayes estimates for the linear model. *Journal of the Royal Statistical Society (B)* **34**, 1-41.

Lindsey, J.K. (1993) *Models for Repeated Measurements*, Oxford: Oxford University Press.

Littel, R.C., Milliken, G.A., Stroup, W.W. and Wolfinger, R.D. (1996) *SAS system for mixed models*, SAS Institute Inc: Cary, NC.

Little, R.J.A. (1992) Regression with missing X's - a review. *Journal of the American Statistical Association* **87**, 1227-1237.

Little, R.J.A. and Rubin, D.B. (1989) *Statistical Analysis with missing data*, Wiley: New York.

Longford, N.T. (1987) A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika* **74**, 817-827.

Longford, N.T. (1993) *Random Coefficient Models*, Oxford: Oxford University Press.

Luyt, D.K., Burton, P.R. and Simpson, H. (1993) Epdiemiological study of wheeze, doctor diagnosed asthma, and cough in preschool children in Leicestershire. *British Medical Journal* **306**, 1386-1390.

Mancia, G., Bertinieri, G., Grassi, G., Parati, G., Pomidossi, G., Ferrari, A., Gregorini, L. and Zanchetti, A. (1983) Effects of blood pressure measurement by the doctor on patient's blood pressure and heart rate. *Lancet* **ii**, 695-698.

Mantel, N. (1966) Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports* **50**, 163-170.

Mathsoft (1996) *Splus Version 3.4 for Unix Supplement*, Mathsoft, Inc: Seattle.

Matthay, K.K., Perez, C., Seeger, R.C., Brodeur, G.M., Shimada, H., Atkinson, J.B., Black, C.T., Gerbing, R., Haase, G.M., Stram, D.O., Swift, P. and Lukens, J. (1998) Successful treatment of stage III neuroblastoma based on prospective biologic staging: a Children's Cancer Group study. *Journal of Clinical Oncology* **16**, 1256-1264.

Matthews, J.N.S. (1993) A refinement to the analysis of serial data using summary measures. *Statistics in Medicine* **12**, 27-37.

Matthews, J.N.S., Altman, D.G., Campbell, M.J. and Royston, P. (1990) Analysis of serial measurements in medical-research. *British Medical Journal* **300**, 230-235.

McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models*, Chapman and Hall: London.

Mitchell, E.A., Taylor, D.J., Ford, R.P.K., Stewart, A.W., Becroft, D.M.O., Thompson, J.M.D., Scragg, R., Hassell, I.B., Barry, D.M.J., Allen, E.M. and Roberts, A.P. (1992) Four modifiable and other major risk factors for cot death: The New Zealand Study. *Journal of Paediatric Child Health* **28 (suppl 1)**, S3-S8

Morrison, D.F. (1990) *Multivariate Statistical Methods*, McGraw-Hill: London.

O'Brien, E. (1996) Ave atque vale: the centenary of clinical sphygmomanometry. *Lancet* **348**, 1569-1570.

Olkin, I. and Tate, R.F. (1961) Multivariate correlation models with mixed discrete and continuous variables. *Annals of Mathematical Statistics* **32**, 448-465.

Oxman, A.D., Clarke, M.J. and Stewart, L.A. (1995) From science to practice. Meta-analysis using individual patient data are needed. *Journal of the American Medical Association* **274**, 845-846.

Pan, H. and Goldstein, H. (1998) Multi-level repeated measures growth modelling using extended spline functions. *Statistics in Medicine* **17**, 2755-2770.

Parkin, D.M., Stiller, C.A., Draper, G.J., Bieber, C.A., Terracini, B. and Young, J.L. (1988) *International incidence of childhood cancer*, IARC Scientific Publications 87: IARC.

Patterson, H.R. (1984) Sources of error in recording the blood pressure of patients with hypertension in general practice. *British Medical Journal* **289**, 1661-1664.

Peixoto Filho, A.J., Mansoor, G.A. and White, W.B. (1995) Effects of actual versus arbitrary awake and sleep times on analyses of 24-h blood pressure. *American Journal of Hypertension* **8**, 676-680.

Penny, J.A., Halligan, A.W.F., Shennan, A.H., Lambert, P.C., Jones, D.R., Deswiet, M. and Taylor, D.J. (1998) Automated, ambulatory, or conventional blood pressure measurement in pregnancy: Which is the better predictor of severe hypertension? *American Journal of Obstetrics and Gynaecology* **178**, 521-526.

Pepys, J. (19.75) Skin tests in diagnosis. In: Gell, P.H., Coombs, R.R.A. and Lach, P.J., (Eds.) *Clinical aspects of Immunology*, pp. 55-80. Oxford: Blackwell Scientific Publications.

Peto, R. (1987) Why do we need systematic overviews of randomised trials. *Statistics in Medicine* **6**, 233-240.

Pettit, L. I. (1990) Measuring the effect of observations on Bayes factors. *Biometrika* **77**, 455-466.

Pickering, T.G. (1990) *Ambulatory Blood Pressure Monitoring and blood pressure variability Part 1*, London: Science Press.

Pignon, J.P. and Arriagada, R. (1993) Meta-analysis. *Lancet* **341** , 418-422.

Powell, J.E., Esteve, J., Mann, J.R., Parker, L., Frappaz, D., Michaelis, J., Kerbl, R., Mutz, I.D. and Stiller, C.A. (1998) Neuroblastoma in Europe: differences in the pattern of disease in the UK. SENSE. Study group for the Evaluation of Neuroblastoma Screening in Europe. *Lancet* **352**, 682-687.

Prasad, N. and Isles, C. (1996) Ambulatory blood pressure monitoring: a guide for general practitioners. *British Medical Journal* **313**, 1535-1541.

Punzi, H.A. (1998) Why ambulatory blood pressure monitoring?. *American Journal of Health-System Pharmacy* **55 (Suppl 3)**, S12-S16

Rasbash, J., Browne, W.J., Goldstein, H., Yang, M., Plewis, I.F., Healy, M.J.R., Woodhouse, G., Draper, D., Langford, I.H. and Lewis, T. (1999) *A user's guide to MLwiN*, Institute of Education: University of London.

Rasbash, J. and Woodhouse, G. (1995) *MLn Command Reference*, edn. London: Institute of Education, University of London.

Raudenbush, S.W. (1994) Random effects models. In: Cooper, H. and Hedges, L.V., (Eds.) *The handbook of research synthesis*, pp. 301-322. Russell Sage Foundation: New York

Reddel, H., Jenkins, C. and Woolcock, A. (1999) Diurnal variability - time to change asthma guidelines? *British Medical Journal* **319**, 45-47.

Rice, J.R. (1969) *The approximation of functions*, Addison-Wesley: Reading, Massachusetts.

Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1994) Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89**, 846-866.

Rodriguez, G. and Goldman, N. (1995) An assessment of estimation procedures for multilevel models with binary responses. *Journal Of The Royal Statistical Society Series A-Statistics In Society* **158**, 73-89.

Royston, P. (1983) Some techniques for assessing multivariate normality based on the Shapiro-Wilk W. *Applied Statistics* **32**, 121-133.

Royston, P. and Altman, D.G. (1997) Approximating statistical functions using fractional polynomial regression. *The Statistician* **46**, 411-422.

Rubie, H., Hartmann, O., Michon, J., Frappaz, D., Coze, C., Chastagner, P., Baranzelli, M.C., Plantaz, D., Avet-Loiseau, H., Benard, J., Delattre, O., Favrot, M., Peyroulet, M.C., Thyss, A., Perel, Y., Bergeron, C., Courbon-Collet, B., Vannier, J.P., Lemerle, J. and Sommelet, D. (1997) N-Myc gene amplification is a major prognostic factor in localized neuroblastoma: results of the French NBL 90 study. Neuroblastoma Study Group of the Societe Francaise d'Oncologie Pediatrique. *Journal of Clinical Oncology* **15**, 1171-1182.

Rubin, D.B. (1996) Multiple imputation after 18+ years. *Journal of the American Statistical Association* **91**, 473-489.

Rutter, C.M. and Elashoff, R.M. (1994) Analysis of longitudinal data - random coefficient regression modeling. *Statistics in Medicine* **13**, 1211-1231.

Sackett, D.L. (1996) *Evidence Based Medicine*, edn. Edinburgh : Churchill Livingstone.

Schafer, J.L. (1997) *Analysis of Incomplete Multivariate Data*, Chapman and Hall: London.

Schafer, J.L. and Olsen, M.K. (1998) Multiple imputation for multivariate missing-data problems: a data analyst's perspective. *Multivariate Behavioral Research* **33**, 545-571.

Schluchter, M.D. and Jackson, K.L. (1989) Log-linear analysis of censored survival-data with partially observed covariates. *Journal of the American Statistical Association* **84**, 42-52.

Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics* **6**, 461-464.

Seligman, S.A. (1971) Diurnal blood pressure variation in pregnancy. *The Journal of Obstetrics and Gynaecology of the British Commonwealth* **78**, 417-422.

Selwyn, M.R. and Difranco, D.M. (1993) The application of large gaussian mixed models to the analysis of 24 hour ambulatory blood pressure monitoring data in clinical trials. *Statistics in Medicine* **12**, 1665-1682.

Shi, M.G., Weiss, R.E. and Taylor, J.M.G. (1996) An analysis of pediatric CD4 counts for acquired-immune- deficiency- syndrome using flexible random curves. *Applied Statistics-Journal Of The Royal Statistical Society Series C* **45**, 151-163.

Siersted, H.C., Hansen, H.S., Hansen, N.C.G., Hyldebrandt, N., Mostgaard, G. and Oxhoj, H. (1994) Evaluation of peak expiratory flow variability in an adolescent population-sample - the Odense schoolchild study. *American Journal of Respiratory and Critical Care Medicine* **149**, 598-603.

Silverman, B.W. (1985) Some aspects of the spine smoothing approach to nonparametric regression curve fitting (with discussion). *Journal Of The Royal Statistical Society Series B-Methodological* **50**, 413-436.

Simon, R. and Altman, D.G. (1994) Statistical aspects of prognostic factor studies in oncology. *British Journal Of Cancer* **69**, 979-985.

Skene, A.M. and Wakefield, J.C. (1990) Hierarchical models for multicentre binary response studies. *Statistics in Medicine* **9**, 919-929.

Sly, P.D., Cahill, P., Willet, K. and Burton, P. (1994) Accuracy of mini peak flow metres in indicating changes in lung function in children with asthma. *British Medical Journal* **308**, 572-574.

Smith, B.J. (2000) *Bayesian Output Analysis Program (BOA). Version 0.50. User Manual*, College of Public Health: University of Iowa.

Smith, P.L. (1979) Splines as a Useful and Convenient Statistical Tool. *The American Statistician* **33**, 57-62.

Smith, T.C., Spiegelhalter, D.J. and Thomas, A. (1995) Bayesian approaches to random effects meta-analysis: a comparitive study. *Statistics in Medicine* **14**, 2685-2699.

Somes, G.W., Harshfield, G.A., Arheart, K.L. and Miller, S.T. (1994) A fourier series approach for comparing groups of subjects on ambulatory blood pressure patterns. *Statistics in Medicine* **13**, 1201-1210.

Spiegelhalter, D.J. (1998) Bayesian graphical modelling: a case-study in monitoring health outcomes. *Applied Statistics-Journal Of The Royal Statistical Society Series C* **47**, 115-133.

Spiegelhalter, D.J., Dawid, A.P., Lauritzen, S.L. and Cowell, R.G. (1993) Bayesian analysis in expert systems (with discussion). *Statistical Science* **8**, 219-283.

Spiegelhalter, D.J., Myles, J.P., Jones, D.R. and Abrams, K.R. (1999) An introduction to Bayesian methods in health technology assessment. *British Medical Journal* **319**, 508-512.

Spiegelhalter, D.J., Thomas, A. and Best, N.G. (1999) *WinBUGS Version 1.2 User Manual*, MRC Biostatistics Unit: Cambridge.

Spiegelhalter, D.J., Thomas, A., Best, N.G. and Gilks, W.R. (1996) *BUGS: Bayesian Inference Using Gibbs Sampling, Version 0.50.*, MRC Biostatistics Unit: Cambridge.

Spiegelhalter, D.J., Thomas, A., Best, N.G. and Gilks, W.R. (1997) *BUGS version 6 manual addendum*, MRC Biostatistics Unit: Cambridge.

Stanton, A., Cox, J., Atkins, N., O'Malley, K. and O'Brien, E. (1992) Cumulative sums in quantifying circadian blood pressure patterns. *Hypertension* **19**, 93-101.

Stewart, L.A. and Clarke, M.J. (1995) Practical methodology of meta-analyses (overviews) using updated individual patient data. Cochrane Working Group. *Statistics in Medicine* **14**, 2057-2079.

Stewart, L.A. and Parmar, M.K. (1993) Meta-analysis of the literature or of individual patient data: is there a difference? *Lancet* **341**, 418-422.

Stiller, C.A. (1993) Trends in neuroblastoma in Great Britain: incidence and mortality, 1971-1990. *Eurpean Journal of Cancer* **29A**, 1008-1012.

Stiller, C.A. and Parkin, D.M. (1992) International variations in the incidence of neuroblastoma. *International Journal of Cancer* **52**, 538-543.

Stone, C.J. and Koo, C. (1986) Additive splines in statistics. In: *Statistical Computing Section Proc. Amer. Statist. Assoc.*, pp. 45-48. Washington D.C.: American Statistical Association

Streitberg, B., MeyerSabellek, W. and Baumgart, P. (1989) Statistical analysis of circadian blood pressure recordings in controlled clinical trials. *Journal of Hypertension* **7**, S11-S17

Sutton, A. J.; Abrams, K. R.; Jones, D. R.; Sheldon, T. A., and Song, F. (1998) Systematic reviews of trials and other studies. *Health Technol Assess.* **2**(19).

Tanner, M.A. and Wong, H.W. (1987) The calculation of posterior distributions by data augmentation - rejoinder. *Journal of the American Statistical Association* **82**, 548-550.

Taylor, J.M.G. and Law, N. (1998) Does the covariance structure matter in longitudinal modelling for the prediction of future CD4 counts? *Statistics in Medicine* **17**, 2381-2394.

Thisted, R.A. (1988) *Elements of statistical computing - numerical computation*, Chapman and Hall: New York.

Thompson, S.G. (1993) Controversies in meta-analysis: the case of the trials of serum cholesterol reduction (review). *Statistical methods in medical research* **2**, 173-192.

Thompson, S.G. (1994) Why sources of heterogeneity in meta-analysis should be investigated (review). *British Medical Journal* **309**, 1351-1355.

Thompson, S.G. and Pocock, S.J. (1991) Can meta-analysis be trusted? *Lancet* **338**, 1127-1130.

Thompson, S.G., Smith, T.C. and Sharp, S.J. (1997) Investigating underlying risk as a source of heterogeneity in meta-analysis. *Statistics in Medicine* **16**, 2741-2758.

Tian, J.J., Shukla, R. and Buncher, C.R. (1994) On prediction of future observations in a growth curve model. *Statistics in Medicine* **13**, 2205-2217.

Toelle, B.G., Peat, J.K., Salome, C.M., Mellis, C.M. and Woolcock, A.J. (1992) Toward a definition of asthma for epidemiology. *American Review of Respiratory Disease* **146**, 633-637.

Toogood, J.H., Andreou, P. and Baskerville, J. (1996) A methodological assessment of diurnal variability of peak flow as a basis for comparing different inhaled steroid formulations. *Journal of Allergy and Clinical Immunology* **98**, 555-562.

Turney, E.A., Amara, I.A., Koch, G.G. and Stewart, W.H. (1992) Evaluation of alternative statistical methods for linear model analysis to compare two treatments for 24-hour blood pressure response. *Statistics in Medicine* **11**, 1843-1860.

Vach, W. and Blettner, M. (1991) Biased-estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values for confounding variables. *American Journal of Epidemiology* **134**, 895-907.

Van Buuren, S., Boshuizen, H.C. and Knook, D.L. (1999) Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine* **18**, 681-694.

Wade, A.M. and Ades, A.E. (1998) Incorporating correlations between measurements into the estimation of age-related reference ranges. *Statistics in Medicine* **17**, 1989-2002.

Walter, S.D. (1976) The estimation and interpretation of attributable risk in health research. *Biometrics* **32**, 829-849.

Walter, S.D. (1997) Variation in baseline risk as an explanation of heterogeneity in meta-analysis. *Statistics in Medicine* **16**, 2883-2900.

Wang, Y.X. and Taylor, J.M.G. (1995a) Flexible methods for analyzing longitudinal data using piecewise cubic polynomials. *Journal of Statistical Computation And Simulation* **52**, 133-150.

Wang, Y.X. and Taylor, J.M.G. (1995b) Inference for smooth curves in longitudinal data with application to an AIDS clinical-trial. *Statistics in Medicine* **14**, 1205-1218.

Wegman, E.J. and Wright, I.W. (1983) Splines in statistics. *Journal of the American Statistical Association* **78**, 351-365.

Wei, G.C.G. and Tanner, M.A. (1991) Applications of multiple imputation to the analysis of censored regression data. *Biometrics* **47**, 1297-1309.

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

Whittemore, A.S. (1983) Estimating attributable risk from case-control studies. *American Journal Of Epidemiology* **117**, 76-85.

Wille, S. and Svensson, K. (1989) Peak Flow in Children Aged 4-16 Years: Normal Values for Vitalograph Peak Flow Monitorm Wright and Mini Wright Peakflow Meters. *Acta Paediatr Scand* **78**, 544-548.

Wold, S. (1974) Spline functions in data analysis. *Technometrics* **16**, 1-11.

Wright, B.M. (1978) A miniature Wright peak flow meter. *British Medical Journal* **2**, 732-738.

Wypij, D., Pugh, M. and Ware, J.H. (1993) Modelling pulmonary-function growth with regression splines. *Statistica Sinica* **3**, 329-350.

Zeger, S.L. and Diggle, P.J. (1994) Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics* **50**, 689-699.

Zeger, S.L. and Karim, M.R. (1991) Generalized linear-models with random effects - a Gibbs sampling approach. *Journal of the American Statistical Association* **86**, 79-86.

Zeger, S.L. and Liang, K.Y. (1992) An overview of methods for the analysis of longitudinal data. *Statistics in Medicine* **11**, 1825-1839.