Probabilistic Risk Assessment of Dietary Data

Ayona Chatterjee

Doctor of Philosophy
University of Edinburgh
2005

Abstract

Food consumption data are recorded for monitoring the dietary habits of a population, for judging the nutritional adequacy of diets and for food risk assessment. Food risk assessment involves looking at risk from high consumption levels of potentially harmful products such as alcohol, low or high intakes of nutrients such as retinol and also consumption of pesticides through foods. Interest lies in finding out not only the average consumption of certain food products but also the probabilities of exceeding the safe levels of consumption for these products.

Until recently, deterministic methods have been used to estimate food risk. These methods give one risk estimate for the whole population, mostly ignore the variability in the consumption data and produce unrealistic results. Recently there has been an increase in the use of probabilistic models for food risk assessment. Some of these probabilistic models are based purely on empirical distributions and others on probability models in a frequentist or Bayesian framework.

In this thesis we improve on existing Bayesian models for food consumption data collected on successive days, and propose possible models to assess various types of food risk. Bayesian hierarchical modelling provides a natural framework for risk assessment, and allows us to account for the various sources of variability present in the data. We discuss general problems associated with dietary data such as large proportions of zeros and extreme intakes, and suggest models to account for these. We look at ways to model the intake of several food products which may all contain the same pesticide, and combine this with pesticide residue data for exposure assessment for that pesticide. We also discuss a non-Bayesian approach to study extreme intakes using extreme-value theory. We use our models to produce predicted probabilities of exceeding recommended and safe levels of consumption for individual days and longer periods.

Acknowledgements

I would first like to acknowledge the Late Rob Kempton, whose ideas motivated me to work in the area of food risk assessment. I would like to thank my supervisor Dr. Chris Theobald for his guidance during my PhD. His critical thinking and insight have helped me learn immensely. I also would like to acknowledge Dr. Graham Horgan for his support in the completion of my PhD. I express my gratitude to all the staff and students at BioSS for their valuable inputs towards my work and also creating a warm and friendly environment. I am grateful to BioSS for partially funding my studies.

A big thank you to Alex for all his positive suggestions towards my work and for being such a wonderful officemate. I also would like to thank Yash for his love and faith in me. For being patient with me and for always being there for me. Finally I would like to thank my parents for their constant love and support. Their encouragement and belief in me has helped in the completion of this thesis.

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text.

(*Ayona Chatterjee*)

Table of Contents

Chapter 1

Food Risk Assessment

1.1 Introduction

Dietary data are collected to study dietary habits of a population and for food safety assessment. Also dietary studies are conducted to provide decision-makers with guidelines in formulating programmes to educate people on their eating habits and also to evaluate effectiveness of such programmes.

Risk from food products can be of various kinds, and is generally associated with very high or low levels of consumption. For example, consumption of alcohol in large amounts can have detrimental effect on human health. Another source of food risk is nutrient intakes. Low levels of nutrient intakes such as vitamins and calcium can lead to deficiencies in the body, whereas large intakes of a nutrient such as vitamin A can actually lead to negative health effects. An important area of food risk assessment concerns intakes of toxic substances through food. Most food products contain naturally occurring, intentionally or unintentionally a wide range of substances: some are desired and some undesired such as natural toxins, pesticide residues and mycotoxins. When discussing food risk, we mostly focus on large intakes, values above certain safe levels of consumption. These safe levels of consumptions have been defined by medical experts or nutritionists or toxicologists. With pesticides, one would be interested in seeing if an individual's intake of a pesticide is less than the safe level or with a nutrient if the intake is less than a certain upper limit.

Food risk can be either acute or chronic. Acute risk is associated with large intakes on a single day or even due to one meal. Chronic risk is risk due to high intakes over many days. Certain products such as alcohol, if consumed in large amounts over a long period of time, pose a risk to human health. In such cases it might be of interest to study chronic risk. Assessing risk from intake of harm-

ful substances requires information on the consumption patterns of individuals. The period of time over which the intakes are recorded for will depend on the substance that we are studying, and can be anything from one day to one week (Voltier et al. (2002)).

From a statistical point of view we are interested in modelling the intakes of certain substances and also predicting the probability of intakes exceeding the certain safe and high levels of consumption. We do not comment on the nature of the risk from exceeding the safe levels or on formulating these safe limits.

According to Verger et al. (2002), dietary studies are conducted for the national governments of several countries for many different reasons. The methods used in conducting these studies vary between countries and there is a constant effort to standardise the steps in conducting a survey and also to combine various data sets available all over Europe. Verger et al. (2002) emphasise the need for comparable data at European level and identifies solutions to make existing food consumption data from nationally representative databases more comparable. Verger et al. (2002) give a list of all data sources available in the UK relating to food habits of individuals. Studies for collecting dietary data are designed to suit the questions that need to be answered and necessary information regarding the questions is collected. The sample of individuals is assumed to represent the population. Before conducting a dietary study one has to decide the population under study, for example if the study concerns only pregnant women or infants less than two- years, the sample size required, number of days on which data will be collected and also the type of survey method to be used.

Work has been done to identify the minimum sample size required for a dietary survey and also the minimum number of days for which each individual's intakes should be observed. Voltier et al. (2002) suggests that a minimum sample size of 2000 adults in each European country will be needed in order to identify trends in the mean intake of most foods and nutrients in Europe, and the sample size should be larger if trends have to be identified for socio-demographic subgroups. The number of repeated observations required for each individual depends on the objective of the study and also the product whose intake is being monitored. A study by Karkeck (1987) gives the number of days required to estimate energy intake of a population as 3-7 days whereas for Iron the required number of observed days can be as high as 12-19. According to Voltier et al. (2002) for estimating habitual intake well, a single 24 hour recall does not give sufficient data.

Information about food consumption patterns is collected using a food frequency questionnaire, a 24-hour recall method or a food diary method. For the food frequency questionnaire, respondents have to fill in information about what they normally eat, how often and in what amounts. In the 24 hour recall method, the individual gives information about exact amounts of foods consumed in the previous day. This method is commonly used for determining the long-term average consumption of a product by an individual. For the food diary the individual has to weigh each food consumed and record it in a diary for a fixed period of time. Such data collected over consecutive days help in looking at consumption patterns over several days.

Once information has been obtained about what product has been consumed by an individual and how much of it, the products are broken in to primary agricultural units to find the approximately the exact contribution of each product to a particular meal or food product. Also this is essential in finding out the exact amounts of nutrients, protein and fat intake through that product. The food composition databases that are used for conversion of food intakes to estimate nutrient intakes also vary between countries. The data for these conversions are obtained by analysis of foods in laboratories or from published or unpublished sources.

Other forms of data that are used in food risk assessment are levels of concentrations of impurities in food products. Concentration is the measure of certain pesticide on products of interest. Most agricultural products are sprayed with pesticides, and according to Ferrier et al. (2002) the main source of exposure to pesticide is by ingestion of products with pesticide residues on them. However consumption of pesticides are more variable than those of e.g. fat and protein. Hence to model pesticide intake we need to model the consumption of the products on which these pesticides are sprayed along with modelling concentration levels.

1.2 Common Problems Associated with Dietary Data

One of the main problems with using information from dietary data about consumption of products is that there is a chance that the values have been misreported. For products such as alcohol and confectionery there is a possibility

that people might over-report or under-report the amounts consumed. Also when individuals are asked to participate in a study to record dietary habits, the individuals might change their eating habits and thus eat less fatty products or eat more products which are socially perceived to be healthy such as fruits. Kim et al. (1984) reported a mean decrease in energy intake of 13% for 29 subjects recording food intakes for one week during four different seasons. Correcting such inaccuracies using statistical models is not considered in this thesis.

Many products, including certain dairy products, nutrients, alcoholic beverages and fruits, are not consumed daily. Some individuals may have them daily, for example milk, some occasionally, for example alcohol, and some never, for example the total alcohol consumption for a teetotaller over his or her lifetime is zero. For a study conducted over seven days for alcohol, we might have individuals whose total alcohol intake over the seven days is zero; these individuals might be teetotallers or occasional drinkers. Thus individuals appear to have different propensity of consumption for products. Irregular eating patterns may lead to large number of zero observations. We can have products which have a large range of possible intake values. For certain nutrients we may have few very large daily intakes giving rise to a highly (right) skewed data set in contrast to total energy which is less skewed. This implies that intakes are not Normally distributed, and models have to be developed to take in to account the possibility of a large number of non-consumption days and a distribution of intakes with a long tail for a certain product.

We are interested in capturing the daily variability in consumptions of products. For dietary data collected over several days it has been observed that there is a large within-individual variance among intakes. Often the within-individual variance is larger than the between-individual variance. For example, in the case of vitamins such as A, E or C, the day-to-day variability in intakes can be four or even six times as large as the between-individual variability in the means (Sempos et al. (1985)). Any dietary assessment based on the distribution of the mean of a few days of intakes may be biased. According to Carriquiry (2003), the bias can be greatly reduced by increasing the number of daily intakes that are collected on each individual in the sample. This might not be realistic due to the time and costs involved. Several methods to account for the within-individual variability among intakes have been developed and will be discussed later in this chapter.

Consumption of products can be affected by day of the week, by week of the

year or by personal characteristics such as body weight, sex and age. Intakes on consecutive days could be correlated with each other. For each data set, such issues should be examined and incorporated in the model when necessary.

Concentrations of pesticide residues on food products are mostly measured in laboratories (Kroes et al. (2002)). Techniques used to assess these levels vary between laboratories depending on the expertise of the analysts involved and differences in equipment. Data collection on successive days or years may become more precise due to improvements of methods. Also the actual composition of a pesticide may change with time. Pesticide levels can also be affected by week of the year and geographical location of fields. Commercial processing, washing and cooking have different effects on different pesticides. It is difficult to find data sets which give information about variation in the pesticide levels at different time of the year or due to processing. Small residues are often reported as zeros due to limitations of the measuring instruments. According to Kroke et al. (1999), recently improved sensitivity has led to an increased incidence of reported positive findings of low levels of contaminants in food. However, as Kroke et al. (1999) points out, it is also essential to make sure that these instruments are capable of accurately measuring high levels of contaminants in food products.

While modelling dietary data these are some problems we need to keep in mind. We try to develop models that are able to account for these issues in dietary data sets.

1.3 Deterministic and Probabilistic Modelling Approaches

Exposure assessment, which is a part of risk assessment, is defined by World Health Organization (1997) as

> The qualitative and/or quantitative evaluation of likely intake of biological, chemical or physical agents via food as well as exposure from other sources if relevant.

In studying intakes of food products we are interested in finding not only the average intake for the population but also the probabilities of exceeding certain levels of consumption. When studying consumption of nutrients we only need consumption data for exposure assessment. However, while assessing exposure to

a pesticide through a certain product, we need data on consumption of that food along with the concentration level of the pesticide on that product. There are two possible approaches for exposure assessment, one using deterministic methods and the other is using probabilistic methods. Here we look at the pros and cons of the two approaches.

1.3.1 Deterministic Methods

Deterministic methods involve just obtaining summary statistics from the data. We do not fit any distribution to the data. Lunchick (2001) provides a list of advantages and disadvantages of using the deterministic and probabilistic approach. Deterministic methods have been widely used for exposure assessment to pesticides. Deterministic methods use point estimates to generate a single estimate of exposure based on various assumptions about the exposure scenario. For nutrient intake the point estimate can be just the mean or the median intake or an upper percentile for the data set. For pesticides, the point estimate is often the worst-case estimate: for example, this method might assume that every food sample contains residues at the highest level found in concentration data and then the consumption of that product is always the 97.5^{th} percentile of all the consumption values. The point estimate is then the product of the maximum residue concentration and the 97.5 percentile for consumption. Since with the deterministic approach we just get a single estimate for exposure assessment, it is easy to compute and comprehend. The pesticide industry and pesticide regulators in the EU are familiar with these estimates. However, there are some disadvantages of using this deterministic approach. Since with the deterministic approach one might just look at one scenario, the variability and uncertainty in the data are not accounted for. This approach sometimes over-estimates or underestimates the probability of exceeding the safe level of consumption depending upon the percentile of the consumption values used, which is undesirable. Also if one looks at just the worst case scenario, one might get estimates which are biologically improbable and unrealistic. These estimates might mislead regulators and consumers. When considering more than one product at a time, these unrealistic estimates are compounded exponentially according to Petersen (2000).

1.3.2 Probabilistic Methods

Probabilistic methods are based on either empirical distributions or probability models. Such probabilistic methods take in to account the variability present in the whole data. For modelling nutrients we have a single distribution for describing the consumption patterns of that nutrient. The pesticide exposure estimate is represented as a distribution of all possible combinations of product intake and concentration of a pesticide on that product. Probability estimates required for risk assessment can be obtained for the exposure distribution. However, a probabilistic approach is labour and resource intensive and not yet very popular among pesticide regulators.

The empirical distribution method for exposure assessment of pesticides involves sampling from the consumption and concentration data set, multiplying these sampled values and using this combined sample to estimate the distribution of that pesticide intake through that food product. Such a method does not account for repeated observations on the same individual and also correlations between consumption of food products. There are no probability distributions involved here. Such approaches are used in risk-assessment software such as @RISK (Palisade (2005)).

A commonly used method for probabilistic assessment is Monte Carlo (MC). Petersen (2000) provides an introduction to Monte Carlo analysis in exposure assessment. Suppose we want to infer the intake of pesticide through the consumption of a certain food product. MC analysis involves sampling from the empirical or probability distribution for the consumption data of the food product and the empirical or probability distribution for the pesticide concentration on that product, and multiplying the values of consumption and concentration to give an intake for that pesticide. The process of sampling from the empirical or probability distributions for the consumption and concentration data sets are repeated a number of times, and an estimate of the distribution of the pesticide intake though that food is obtained. There are two ways to perform MC assessment. One is where we just use the data for sampling purpose, that is we use the empirical distributions for the consumption and concentration data sets. The other approach is where we assume that there exists a probability distribution that describes the consumption and concentration data sets. We then sample from these probability distributions to obtain the distribution of pesticide intakes. In some cases the product of two distribution functions can be a known distribution function, for example if we have Lognormal distributions describing

both the consumption and concentration data sets then the product of these two distributions is also Lognormal and hence we can sample from the Lognormal distribution which describes the pesticide intakes. According to Petersen (2000),

> The results from a MC assessment are only as valid as the input parameters, data and assumptions.

Making parametric assumptions about the distribution of consumption and concentration data helps us to obtain more precise estimates of exposure assessment.

United States Environmental Protection Agency (US EPA) already uses probabilistic methods for exposure assessment (Ferrier et al. (2002)). However in the UK the process is being reviewed by pesticide regulators, and there is a growing acceptance of the need to use probabilistic techniques for exposure assessment. The US EPA uses a tiered approach which includes both the deterministic and probabilistic methods for assessing acute dietary exposure to pesticide residues. Suhre (2000) gives an example of the four tiers used by the EPA using data on the intake of a hypothetical chemical OrganoPHOS. For tiers one and two, point estimate for consumption of the pesticide for a worst-case scenario are determined, using a high percentile for the residue estimate and a distribution for the consumption data. Tiers three and four use a probabilistic approach and have a probability distribution for the residue and the consumptions to provide a more realistic acute exposure than the point estimate.

Ferrier et al. (2002) give a UK perspective on probabilistic methods used for exposure assessment. The authors give details of available software for exposure assessment such as Calendex, DEEM, SHEDS and EASE. Most of the software requires choosing input distributions, and have in-built data bases which can be a limitation in certain studies. However certain software do allow data sets to be imported for use also and it may not be possible to handle repeated observations for an individual. Ferrier et al. (2002) also lists all the organisations and research activities related to consumer exposure to pesticides. Few fully operational probabilistic models are available for assessors to work with, and hence there is a need for more work in this area.

1.4 Existing Probabilistic Models for Dietary Data

In this section we review the existing probabilistic models for studying dietary data. There is a wide range of models available for dietary data, some of which

we will discuss in the following sections. Some use Monte Carlo analyses with probability distributions, some focus on estimating the long-term average intake for a product, and recently few papers have discussed Bayesian models for dietary data. We divide this section into two parts; we first discuss models to study nutrient intakes and then look at existing models to study exposure to pesticides or toxins. The models used to study consumption data for pesticide exposure can also be used to study nutrient intakes.

1.4.1 Consumption Models

The 'usual intake' of a nutrient, food or chemical can be defined as the long-term average intake of that nutrient, food or chemical by an individual (Nusser et al. (1996), Hoffmann et al. (2002)). Estimation of usual intake distributions includes estimation of percentiles, and allows subsequent calculation of statistical summaries such as mean and standard deviation.

A method of estimating usual daily intake distribution is given by Nusser et al. (1996). Since the distribution of most consumption data may be skewed and non-Normal, the authors give a method to normalise the data using a combination of power and grafted polynomial transformations. The model involves five main steps, and has been illustrated using 24-hour recall data collected on four non-consecutive days on 737 women aged 25-50 years. Nusser et al. (1996) give results for usual intakes of calcium, energy, iron, vitamin A and vitamin C. Steps I and II are intended to remove effects of day of the week and interview sequence, and create an equal weight sample from the original sample to remove any selection of day bias. Nusser et al. (1996) asserts that intakes recorded on the first sample day are most accurate and hence the intakes on the other days are adjusted to have a mean and variance equal to that on the first sample day. Step III is to transform the data to normality. Their measurement error model is defined as

$$Y_{ij} = y_i + u_{ij}, \tag{1.1}$$

where Y_{ij} is the adjusted intake of the i^{th} individual on the j^{th} day. The usual intake that is to be estimated is given by y_i, and it is assumed that y_i has a Normal distribution with mean μ_y and variance σ_y^2, and that u_{ij} is also Normal with mean zero and variance σ_{ui}^2. Also σ_{ui}^2 has mean μ_A and variance σ_A^2. The paper does not specify the distribution of σ_{ui}^2, and justifies this by adding that to fit the model only σ_A^2 is required. Step IV is to estimate the parameters for the usual intake distribution, and the last step is to back-transform the data to the

original scale. The transformation to normality is complicated, and for skewed data such in the case of Vitamin A the number of join points needed for a grafted cubic polynomial is large. Also converting the intakes back to the original scale using a inverse non-linear transformation results in biased intakes which need to be corrected. One can use C-SIDE, Iowa State University (2002) which has been developed at the Iowa State University for fitting this model given by Nusser et al. (1996).

Hoffmann et al. (2002) compare exposure assessment methods by Slob (1993), Wallace et al. (1994), Buck et al. (1995) and Nusser et al. (1996) and also suggest a simplified version of the latter. The first three models are discussed in the next subsection. Hoffmann et al. (2002) applied their approach to data from three European food consumption surveys, a French survey conducted in 1998/99, a Belgian survey conducted in 1997/1998 and a Swedish survey conducted in 1997/98. Hoffmann et al. (2002) compares the estimated total fat and vegetable intake, by the four methods for a German validation study conducted in 1995/96, see Kroke et al. (1999) for details. Hoffmann et al. (2002) also conclude that two non-consecutive days of 24 hour diet recalls are adequate to describe usual intake, instead of using more days to study usual intakes. They believe that the effect of intra-individual variability on the usual intake vanishes if the number of repeated observations per individual is made large. The variance of the usual dietary intake in the framework of food consumption surveys can be estimated by

$$\hat{\sigma}^2_{usual} = \hat{\sigma}^2_{\overline{y}} - \frac{1}{k}\hat{\sigma}^2_{\delta}. \tag{1.2}$$

Here k is the number of repetitions, $\hat{\sigma}^2_{\overline{y}}$ is the observed variance of individual mean intake and $\hat{\sigma}^2_{\delta}$ is the estimated average within-individual variance.

Hoffmann et al. (2002) use a two-parameter Box-Cox function to transform data to normality, and they call their method the S-Nusser as it is a simplified version of that of Nusser et al. (1996). They conclude that the latter method is most flexible, though it is computationally intensive one surely achieve Normality for the data set provided the data is not highly skewed and does not have a large proportion of zeros. However, the methods of Slob (1993), Wallace et al. (1994), Buck et al. (1995) and Nusser et al. (1996) do not perform well for skewed data and also when large number of zeros are present in the data. Nusser et al. (1997) recommends a three-step procedure for such data.

1. First to estimate the distribution of zeros and non-zeros.

2. Second to model the usual intake distribution for non zero consumptions using the method by Nusser et al. (1996).

3. Estimate the usual intake distribution from the joint distribution of usual intakes and individual consumption probabilities.

Guenther et al. (1997) provide a review of estimating usual nutrient intake distributions at the population level. According to Guenther et al. (1997) the methods developed cannot be used for individual usual intake estimation. This model discussed by the authors is similar to the approach adopted by Nusser et al. (1996). The measurement error model treats the intake observed for any individual on a given day as the sum of the individual's true usual intake and a 'measurement error' for that individual on that day. Guenther et al. (1997) study intakes for fat, folate and vitamin A, and compare results obtained using one-day data and three-day means.

The method developed by Guenther et al. (1997) assumes that a reported one-day nutrient intake found in a food intake survey data set can be conceptually represented as a generalisation of Equation 1.1

$$Y_{itj} = y_i + c_t + b_j + e_{it}. \tag{1.3}$$

The nutrient intake for individual i on date t and for a member j of a day sequence is Y_{itj}. Here t is the date on day j, which is the sequence number of the day for which the individual has provided intake information for the survey. The usual intake for individual i is denoted by y_i. Here c_t is the temporal effect on nutrient intake due to the particular day of the week and time of the year, and b_j is the bias associated with intakes on a particular reporting day of the survey. The method is implemented using the software C-SIDE (Iowa State University (2002)). This method also does not work well when the data has large number of zeros.

Carriquiry (2003) combines information about vitamin supplement intakes with food consumption data to study nutrient intakes. The author discusses the problems of zeros in consumption data sets, and advocates the use of a propensity questionnaire which determines how often people consume these nutrients. The author uses the approach of Nusser et al. (1996) to obtain estimates for their model parameters. The only difference between the methods of Carriquiry (2003) and Nusser et al. (1996) is in the last stage where the estimated usual intakes are transformed to the original scale. Carriquiry (2003) uses a mean

back-transformation that greatly reduces the bias. For data sets with zeros, the author extends Nusser's model as follows. Let y_i denote the usual intake for individual i so that

$$y_i = E\{Y_{ij}|i, Y_{ij} > 0\}. \tag{1.4}$$

Let p_i denote the probability to consume the food by individual i. The model assumes that the propensity to consume and the amount of consumption are independent and the usual intake can be given as $y_i^* = y_i \times p_i$.

Gay (2000) also gives an approach to transforming dietary data to Normality and then uses analysis of variance (ANOVA) to characterise day-to-day variation in nutrient intake. The author suggests that usual intake distribution can be accurately constructed using as little as two day weighed dietary data. The method corrects for any bias that might be introduced due to uneven coverage of days of the week. The author uses a transformation $x = y^\lambda$ where $\lambda = 0$ corresponds to a logarithmic transformation, y is the original intake data and x the transformed intakes. Appropriate λ values for seven nutrients under study are determined to achieve normality for the intakes. The author does not mention how the zeros in the data are handled. Using ANOVA the overall variation in nutrient intakes is partitioned into components due to differences between individuals, systematic differences between days and random within-individual variation. The model assumes that the intakes on consecutive days are as variable as intakes on non-consecutive days and the within-individual variance is the same for all individuals.

1.4.2 Models for Assessment of Exposure to Pesticides

While presenting models for exposure assessment, two types of data are generally used. The first kind is where one directly works with the exposure data. That is using MC techniques, pesticide or toxin intakes are generated from the consumption and concentration data sets. The other is where one works with the consumption and concentration data sets and then combine them to obtain the distribution for the pesticide or toxin intake. The two methods by Hamey (2000) and Slob (1993) discussed in this section use data of the first kind, where directly using some undisclosed method the data on toxin or pesticide intakes are available. The Bayesian method by Paulo et al. (2004*a*) in Section 1.5 uses data on both consumption and concentration to model pesticide intakes.

An example of exposure assessment using an empirical distribution approach can be seen in Hamey (2000). Hamey (2000) uses MC methods for exposure

assessment of the pesticide *carbarly* on toddlers aged $1\frac{1}{2}$ to $4\frac{1}{2}$ years, through daily intake of apples, pears, peaches and nectarines. Since the numbers of samples representing some combinations of fruits were low, consumption values were simulated using observed marginal frequencies of consumption and amount eaten from an existing database. Exposure estimates were obtained using @RISK, which is an add-in in Excel. Hamey also looks at possible association between the amounts of fruits consumed and body weight of individuals using Spearman's rank correlation coefficient.

One of the first papers discussing a probability model for estimating long-term intakes of toxins was by Slob (1993). He proposed a simple model to estimate usual exposure of a population and also predict long-term exposure to pesticide or toxins in food. Usual exposure can be defined in a similar way to usual intake as the long-term average intake of a pesticide or a toxin through a certain food by an individual. Slob illustrates his methods using data on dioxin and cadmium exposure of the Dutch population recorded on two days for 5898 individuals. A logarithmic transformation is assumed to normalise this data set. The model for exposure is given as

$$\log[Y_{ij}(t)] = f(t) + \epsilon_i + \delta_{ij}. \tag{1.5}$$

Here $Y_{ij}(t)$ is the intake of cadmium or dioxin by individual i on day j at age t, $f(t)$ is the usual log intake of an individual at age t which is to be determined, ϵ_i and δ_{ij} correspond to the between-individual and within-individual variances respectively. Both these variances are assumed to be homogeneous between and within individuals, even though testing for validation of these assumptions showed that this was not the case. The author validates his assumptions by using simulated examples to show the negligible effects the assumptions have on the model estimates. No day-of-the-week or week-of-the-year effect is considered. A regression analysis of the log-intakes was performed using a fourth order polynomial in time. The between-individual and within-individual variances were estimated from the residuals.

To estimate percentiles for the intakes, the usual log intakes were assumed to be Normal and the $(1-\alpha)^{th}$ percentile and according to Slob (1993) can be given as $Q_{1-\alpha}(t) = \exp[f(t) + q_{1-\alpha}\sigma_\epsilon]$. Here $q_{1-\alpha}$ is the $(1-\alpha)^{th}$ percentile of the standard Normal distribution and σ_ϵ is the estimated standard deviation of the between-individual variability. The quantiles along with the estimate of the usual intake, $\exp[f(t)]$, represent a description of the populations' intake.

Slob (1993) also suggested that the model can be extended to study lifelong exposures by assuming that the between-individual variability in Equation (1.5) to be constant with age. Thus for each individual we have a usual intake which is a function of age. Using the estimated long-term usual intake for each individual the distribution of long term intakes can be obtained. This assumes that concentration and consumption patterns are constant throughout the lifetime of the individual.

Wallace et al. (1994) propose a method to estimate long-tern distributions based on few repeated short-term measurements. The authors use data on human exposure to volatile organic chemicals (VOC) such as carbon tetracholirde, chloroform, trichloroethylene, tetrachloroethylene and paradichlorobenzene. The multiplicative model by Wallace et al. (1994) assumes the between-individual and within-individual variances are independent and log-normally distributed. The observations are collected throughout the year to take in to account any seasonal variation. Using the geometric mean and geometric standard deviation of the variances a distribution for the exposures averaged over time is obtained. From the distribution, percentiles such as the 50^{th} and 97.5^{th} percentiles are determined for each VOC for assessing individual exposure levels. However there are cases when the hypothesis of log-normality is rejected and results obtained for such cases may not be accurate.

Another approach to estimate long-term exposure from short-term measurements is given by Buck et al. (1995). The daily exposure Y_{ij} is given by an additive model $Y_{ij} = \mu_i + \tau_{ij}$. Here μ_i is the true daily average exposure for the i^{th} individual and τ_{ij} is the deviation from μ_i on day j. The model assumes that mu_i has a distribution G with mean μ_p and variance σ_p^2. Also τ_{ij} has a distribution F with mean 0 and variance σ_T^2. Further μ_i and τ_{ij} are assumed to be independent and the variance σ_T^2 is the same for all individuals. Buck et al. (1995) demonstrate their model using simulated data and estimate percentiles of population exposure. The authors assume G to be a Lognormal distribution. Buck et al. discuss effects on the long-term distribution of exposures when the model assumptions are relaxed. They also suggest having a different distribution for each individual and having varying within-individual variance. This approach has been used by Myles et al. (2003) and will be discussed in the next section. According to Buck et al. (1995), two-ways in which such exposure models may be incorrect are when there is correlation between day-to-day exposure levels or if there are long-term trends in daily exposure. Buck et al. also discuss issues such

as the best sampling plan and the sample size required for estimating long-term exposure from short-term measurements.

1.5 Bayesian Models for Dietary Data

We now discuss two Bayesian models that have been used to study consumption data. The first model is that of Myles et al. (2003) to study retinol intakes and the second is by Paulo et al. (2004*a*) for modelling pesticide intakes.

Myles et al. (2003) discuss two models, the first one is in a non-Bayesian framework and the latter one uses a Bayesian framework for parameter estimation and exceedance probability predictions. The paper uses data on daily retinol intakes on 2197 individuals in the age-group 16-64 years over seven successive days. This data set is also used in this thesis, and is described in Chapter 4 and is obtained from the UK Data Archive (1987).

The model by Myles et al. (2003) may be expressed as

$$y_{ij} = \alpha + \beta_{k(i)} + \gamma_{m(i)} + \delta_i + \epsilon_{ij}. \tag{1.6}$$

Here y_{ij} represents the i^{th} individual's log transformed retinol intake on the j^{th} day, $k(i) = 1, 2$ represents the sex of the i^{th} individual, $m(i) = 1, 2, 3, 4$ represents the age category to which the i^{th} individual belongs. The authors do not mention how the zeros are handled. The between-individual variation is represented by δ_i and is Normally distributed with mean 0 and variance σ_b^2, and ϵ_{ij} represents the random variation within individual i between days and is assumed to be Normally distributed with mean 0 and variance σ_w^2, which is constant across different individuals.

The other model defined by Myles et al. (2003) is the same as the model given in Equation (1.6) except that the within-individual variation in intakes from day to day is assumed to be normally distributed with mean zero and variance $\sigma_{w(i)}^2$, that is the within-individual variability may be different for different individuals. The authors assume that $\log(\sigma_{w(i)})$ is from a Normal distribution with some unknown mean and standard deviation. The model is fitted using the WinBUGS software which will be discussed in more detail in Section 1.5. Myles et al. (2003) do not completely specify prior distributions they use for their Bayesian model. They give predicted percentages of excessive retinol intakes for males and females. We compare results from the model of Myles et al. (2003) with a mixture model

proposed by us in Chapter 4.

Paulo et al. (2004*a*) outline some advantages of using Bayesian modelling over other approaches such as the deterministic and empirical approaches used in exposure assessments. The authors use a Lognormal distribution to describe the non-zero consumptions of food products and pesticide concentrations on a single day within a Bayesian perspective. Paulo et al. (2004*a*) use data on the consumption of endive, lettuce, grape and kiwi fruit along with the residue levels of the pesticide Iprodione on them for their analysis and these data sets are described in Chapter 3 of this thesis.

The model by Paulo et al. (2004*a*) for the consumption of a food product is defined as

$$\begin{aligned} B_f &\sim \text{Bernoulli}(\pi_f) \\ P(y_f|B_f) &= \begin{cases} \delta(0) & \text{if } B_f = 0 \\ LN(\mu_f, \sigma_f^2) & \text{if } B_f = 1. \end{cases} \end{aligned} \tag{1.7}$$

Here B_f is an indicator function for consumption and is from a Bernoulli distribution with probability π_f, y_f is the amount of product f consumed and $\delta(0)$ represents a spike at $y_f = 0$ and LN stands for the Lognormal distribution. The authors define a hierarchical Bayesian model for the non-zero consumptions, and extend the model to a multivariate case for modelling intake of p products simultaneously. Then $\ln(\mathbf{y})$ is given a multivariate Normal distribution $N_p(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$; $\boldsymbol{\mu}_y$ is also given a multivariate Normal distribution $N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and $\boldsymbol{\Sigma}_x$ is given a Wishart distribution. We will discuss hierarchical Bayesian models like this one in the next section. The model is fitted for simultaneous consumption of four commodities, endive, cabbage-lettuce, grape and kiwi fruit. Paulo et al. (2004*a*) present results of the posterior statistics for frequency of intakes and mean log amounts for the four products along with their corresponding standard deviations.

The concentration model to study pesticide residue by Paulo et al. (2004*a*) is defined as

$$Pr(C_j^*|I_j) = \begin{cases} \delta(0) & \text{if } I_j = 0 \\ f(\theta) & \text{if } I_j = 1, \end{cases} \tag{1.8}$$

where C_j^* is the concentration of the pesticide present in a food sample, I_j is 0 if the pesticide was not used on the population from which the sample j was collected and 1 otherwise. The measured concentration of the pesticide C_j on

sample j is then defined as

$$C_j = \begin{cases} \text{ND} & \text{if } C_j^* < \text{LOD} \\ C_j^* & \text{if } C_j^* \geq \text{LOD} \end{cases} \tag{1.9}$$

The level of detection is denoted by LOD and ND stands for non-detects.

A Lognormal distribution is used to describe the concentrations exceeding the LOD. This model is useful to study intake of a pesticide through consumption of more than one food product.

Paulo et al. (2004*a*) do not study any effects of age and sex of an individual on the intakes. Also for the consumption data they do not mention how to handle repeated observations, that is when we have successive intakes over several days for each individual. In Chapter 3 of this thesis, we combine information from the consumption and concentration data to assess exposure from Iprodione with repeated observations for each individual.

1.6 Thesis Structure

We have discussed existing models for dietary data in the previous sections. No model explicitly deals with the presence of large number of zeros in consumption or concentration data sets. Many existing models use a log transformation to the intakes to achieve Normality or use a Lognormal distribution to describe the consumptions. This assumption of Lognormally distributed consumptions is clearly false in the presence of large number of zeros. In Chapter 2 we look at a data set on daily alcohol intakes in which more than half the values are zeros. There we develop two models to study such data sets. The first model incorporates the idea similar to Carriquiry (2003) on propensity to consume. However for our model, the propensity to consume is not a probability. The model in Chapter 2 is also based on the model of Myles et al. (2003) with varying within-individual variance. The second model is a latent variable model. Details of a latent variable model are discussed in Chapter 2. We compare the models and give predicted exceedance probabilities for daily alcohol intake for various sex-by-age groups. We also look at maximum alcohol intake over a week. The models developed in Chapters 2, 3 and 4 are all in a Bayesian framework. We follow Paulo et al. (2004*a*) in choosing to work with Bayesian techniques.

Chapter 3 focuses on exposure assessment and we illustrate our model for intake of the pesticide Iprodione through endive, lettuce, grape, strawberry and

currant. Here we have information about consumption patterns for five food products and data on concentrations of Iprodione on the same products. Both these data sets have large proportions of zeros. We extend the univariate latent variable model developed in Chapter 2 to a multivariate model, and study the consumption of all the five products simultaneously. For the concentration data we use a latent-t model as well as a latent Gaussian model, and combine predictions from the concentration and consumption data for exposure assessment.

Some dietary models fail to work satisfactorily when the observed data are highly skewed. In Chapter 4 we work with a data set on retinol intakes which is highly skewed and has a few very large retinol intakes. We discuss a mixture model to predict intakes for such skewed dietary data. The chapter also includes an extension to deal with correlated intakes on consecutive days. The model is able to predict the proportion of people exceeding the safe level of retinol consumption and also the proportion of people exceeding their recommended daily allowances.

In Chapter 5 we look at a non-Bayesian approach to model extreme dietary intake. We continue to work with the retinol intakes that will be described in Chapter 4 and look at possible models using extreme-value-theory.

For most of our models we use power transformations on the intakes. Using such simple transformations allow us to back transform these responses to the original scale in straightforward manner.

1.7 Bayesian Model Fitting and Predictions

The data sets on consumption used in this thesis have repeated observations for each individual. The number of successive days on which data are available range from two to twelve for our data sets. We also have information about the sex and age of all the individuals and for some of our models we create sex-by-age groups to classify each individual. This kind of data possess a kind of hierarchical structure in which we have the whole sample from the population and then within the sample we have various sex-by-age groups. For each of these groups we have numerous individuals for whom we have intakes over several days.

Bayesian models can take into account the various sources of variation present in such data. Our choices of prior distributions for our Bayesian models reflect the

hierarchical or multi-level structure in the data sets, in which there are variations in the response within individuals and between individuals, variations within sex-by-age groups and between these groups and also variations between days. The intakes are modelled conditionally on certain parameters, known as hyperparameters, corresponding to a higher level of the hierarchy (Gelman et al. (2003)). We can incorporate prior information on parameters whenever it is available. Another attraction to Bayesian modelling is the scope of developing hierarchical models

For all our Bayesian models we follow a common practice of giving the reciprocal of variances a Gamma prior distribution. Here $Ga(\alpha, \lambda)$ denotes a Gamma distribution with expectation α/λ and variance α/λ^2. Also for convenience we take the lower-level effects to be Normal and independent.

In a Bayesian model, we use the posterior distributions for our model parameters to simulate consumption for a random individual not in the data set over any time period; these we call the predicted intakes. This is with the assumption that the same hyperparameters govern the intakes in the future also. For models in which we use information about the sex and age of an individual, we simulate intakes for a random individual in each sex-by-age group. To predict for the whole population together, we combine the predicted intakes in each sex-by-age group such that the number of simulated intakes in each group is proportional to the observed group frequencies. This is a convenient way of sampling in proportion to the group frequencies. The predicted intakes are then used to determine the probabilities of exceedance within a sex-by-age group or for the whole population .

Explicit evaluation of the posterior distributions of model parameters is often not possible in a Bayesian framework. This is mainly due to the need of integrating complex and high dimensional functions. MCMC method provides a solution to this problem by allowing us to sample from the posterior distribution directly and thereby performing the integration implicitly (Brooks (1998)). A clever way to achieve this is by constructing a Markov chain whose stationary distribution is the targeted posterior distribution of the parameter. When the Markov chain is run for a sufficiently long time the simulated values from the chain can be treated as a sample from the target distribution. Quantiles of the parameters of interest can be estimated from sampling the posterior distributions. Various algorithms such as the Metropolis-Hastings by Metropolis et al. (1953) and Hastings (1970) and the Gibbs sampler (Geman & Geman (1984) and Besag & York (1989)) are available to construct such a Markov chain.

A widely used program for implementing MCMC in a Bayesian framework is WinBUGS (Spiegelhalter et al. (2004)). We use this software for estimating the posterior distribution of our Bayesian model parameters and for prediction. WinBUGS stands for Windows based Bayesian inference using Gibbs sampling. The software is freely downloadable from the UK Medical Research Council Biostatistics Unit's website on the BUGS project. WinBUGS uses several samplers: it chooses an implements an appropriate one and samples from the joint posterior distribution of the model parameters. The package uses a S-PLUS type language for specifying statistical models, and graphical and statistical summaries can be obtained. History plots for parameter distributions can be examined to monitor convergence. These give the value of a parameter against the iteration number, and can be plotted for all the iterations performed for a model.

WinBUGS also allows graphical representation of models using Doodles. Relationships between variables in a model can be graphically represented by using nodes and edges. The nodes represent the model parameters and the edges link the parameters which are related. Hierarchical structures can be presented by allowing variables with the most influence on the data places close to the top or bottom of the doodle diagram and those of lesser influence places in decreasing order down or up the graph (Whittaker (1990)) respectively.

WinBUGS requires that the initial values of certain parameters to be specified, but an option of allowing WinBUGS to generate the initial values is also available. Though inferences obtained using MCMC are independent of the specified initial values, they affect the speed and ease of detection of convergence. The number of iterations may depend on the rate of convergence. To reduce possibility of inferential bias caused by the effect of starting value, the initial iterations are normally discarded and treated as burn-ins.

The number of chains that should be run while implementing MCMC algorithms is a topic of debate (Gelman & Rubin (1992), Geyer (1992)). Many argue that running a single long chain will get the estimated posterior distribution closer to the target distribution than it would with any number of shorter runs. However many advocate the use of multiple chains with dispersed initial values. This method ensures that the sampler output covers the entire sample space. According to Cowles & Carlin (1996), running multiple chains is inefficient as compared to running one single long chain. We run single long chains for our parameters.

The WinBUGS manual (Spiegelhalter et al. (2003*c*)) suggest a rule of thumb that the simulations should be run until the Monte Carlo error (MC error) for each parameter of interest is less than about 5% of the sample standard deviation. The MC error is an estimate of the difference between the mean of the sampled values (which we are using as our estimate of the posterior mean for each parameter) and the true posterior mean. We use history plots and the MC error to check for convergence for our parameter distributions in our Bayesian models.

Chapter 2

Modelling Data Sets with Many Zeros

Dietary data obtained on consumption of certain food products may have a large number of zeros. This is observed while monitoring consumption of products consumed infrequently such as alcohol or specific fruits such as bananas or apples. Distribution of intakes for such products have a peak at zero. We need to take into account the occurrence of zero consumption while modelling such data. In this chapter we look at two models for such dietary data with large number of non-consumption days. We use data on daily alcohol intakes to illustrate our models.

2.1 Health Effects of Alcohol Consumption

Alcohol consumption studies are important due to the effect high alcohol consumption has on health and the social system. It is well established that high levels of alcohol consumption causes cancer of the mouth, larynx, oesophagus and the liver (Sieri et al. (2002)). Excess alcohol intake also increases the risk of cerebrovascular disease (haemorrhagic stroke) and increases blood pressure. High levels of alcohol intake lead to vitamin deficiencies, including vitamin B-1, vitamin B-2, niacin, vitamin B-6, folacin and vitamin C. There is a decrease in absorption levels of minerals such as zinc and magnesium in the body with high levels of alcohol consumption. These negative health effects are mostly associated with high consumption levels of alcohol over a long period of time. However, a single occasion of large alcohol consumption can also be detrimental to one's health (Frezza et al. (1990)). It is accepted that binge drinking is a growing trend among British teenagers. Several studies such as Sales et al. (1989) have shown the association of high alcohol consumption levels with mortality due to cancer of the trachea and lung, cirrhosis of the liver. Alcoholism is also associated with

anti-social behaviour, and in many cases contributes to homicide, suicide and traffic accidents. Alcohol misuse costs the National Health Service (NHS) of UK an estimated £3 billion in hospital services - 12% of the total NHS hospital costs according to an article published by The Guardian (Nov 2002).

However there are positive health effects of alcohol consumption also. Moderate levels of consumption lower fibrinogen and clotting factors and hence lowers the risk of cardio-vascular disease (Criqui et al. (1987)). Ethanol in alcohol can reduce the chance of a heart attack. Doll (1997) has shown that the relationship between drinking and mortality is a 'U' or 'J' shaped. This means the mortality risk is higher for teetotallers, dips for consumers of one or two drinks per day and rises sharply for those who drink excessively, putting themselves at risk from alcohol related diseases. For middle-aged humans, though moderate alcohol consumption can have a negative impact such as increased risk of breast cancer in women, it reduces mortality from heart diseases by about a third. According to Doll (1997) it appears that the beneficial health effects of moderate consumption in total outweigh the harmful ones.

One unit of alcohol contains eight grams of pure alcohol (ethanol). To calculate the number of units in a drink, one can use $PV/100$, where P is the strength of the drink expressed as % alcohol by volume (abv) and V is the volume in millilitres. A 175 ml glass of wine at 13% abv is worth 2.3 alcoholic units, and one pint of 5% lager contain 3 units of alcohol.

The recommended maximum levels of alcohol intake according to the British government are 2-3 units of alcohol per day for women and 3-4 units of alcohol per day for men. These are considered to be the 'safe levels' of alcohol consumption. Looking at intervals of several days, for an average male 21 units of alcohol per week and for a female 14 units of alcohol per week are the upper levels for consumption so as not to have any harmful health effects (Webb et al. (1996)). Binge drinking is said to be an occasion when one consumes more than twice the safe level of alcohol in a day.

Compared to wine-producing European countries, the UK's alcohol consumption levels used to be moderate. However in the recent past while levels in most wine-producing countries have fallen or stabilised, levels in the UK are still rising. If the present trend continues, the UK will overtake France's level of alcohol consumption in the next ten years (Institute of Alcohol Studies (May 2004)). France

has one of the highest per capita alcohol consumptions in the world (Arves & Choquet (1999)). The report by the Institute of Alcohol Studies (May 2004), UK states that alcohol consumption per head in the UK rose sharply between 1950 and 1975 and thereafter appeared to have reached a plateau. However the report also adds that since 1995 total alcohol consumption seems to be rising again, and in 2000 it was the highest since 1910. An alcohol consumption fact-sheet published by Alcohol Concern Factsheets (2003) showed that the proportion of women drinking over the safe levels rose steadily from 1984 to 1996, whereas men's drinking has remained stable during this period. The UK Medical Council on Alcohol (Alcoholis (2002)) claims that between 1984 and 2000 there has been more than a 15% increase in the number of women in the age group 18-24 consuming more than 14 units of alcohol per week. The report also found the proportion of men and women drinking more than the safe levels decreases with age, and the proportion of abstainers increases with age. In the UK about 9% males and 14% females do not drink alcohol but 27% males and 14% females consume more than 21 and 14 units of alcohol per week respectively (Alcohol Concern Factsheets (2003)). Males aged 16 and over drank on average 16 units of alcohol per week and females drank 6.3 units per week.

From a statistical viewpoint we are interested in developing a model for daily alcohol intakes so as to predict the probability of an individual's daily and longer term intakes being greater than the safe levels.

2.2 The Alcohol Data Set

The data that have been used in this chapter are a part of a larger data set obtained to detect and model misreporting of food intake at the Rowett Research Institute (RRI), Aberdeen, for a study as reported by Stubbs et al. (Aug 2001). The whole data set gives information about the intakes of various macro and micro nutrients for 59 individuals. Here we use only information about alcohol intakes. Subjects were recruited from Aberdeenshire by press release, newspaper advertisements and posters in community halls, business centres and universities. Thirty men and 29 women in the age group 18-64 years volunteered to be a part of the study. It was held at the Human Nutrition Unit (HNU) at RRI. All subject were healthy and smokers were excluded. Each subject was studied for 14 days, starting on the same day of the week. During days one and two the subjects were fed a fixed mandatory diet designed to meet their energy requirements.

During the next twelve days the food intakes of these individuals were covertly quantified by trained staff at the HNU. Subjects were given access to a variety of familiar foods that they would normally eat during the study. The 12-day study period was split into two 3-day recording periods where the subjects recorded the type and weight of the food consumed alternating with two 3-day periods where they did not. The order of these 3-day periods was randomised across subjects. Investigators secretly weighed all the foods disappearing from the individuals' larders every 24 hours for days 3-14. We re-number these observed days as 1-12 for our study. Each subject's feeding behaviour was continuously observed on video surveillance. More details can be found in Stubbs et al. (Aug 2001). Exact information about the amount of alcohol consumed by each individual per day is available. We use these alcohol intake values. The data for alcohol intake were recorded in megajoules (MJ). One gram of alcohol has about seven calories of energy and so one unit is equal to 56 calories. One unit of alcohol is about 0.23 MJ. The safe levels of daily alcohol intakes in MJ are 0.69 MJ for females and 0.92 MJ for males.

The subjects in this study may be considered atypical and may not represent the population. They are self selected and do not include any smokers. The results obtained using this data set may not be applicable to the whole population, however here we are interested in illustrating our methods for such data sets.

2.3 Summary Statistics

We have 59 individuals' daily alcohol intakes observed over 12 consecutive days. The age, sex and weight of each individual were noted before the start of the experiment and are available. A day on which an individual consumed no alcohol will be called 'zero consumption day' for that individual and the other days as 'consumption days'. We have in all 708 alcohol intake values, by combining data for all persons over the 12 days. The histogram of daily alcohol consumption in Figure 2.1 shows a positively skewed distribution with a large number of zero alcohol consumption days. Out of the 708 alcohol intake values, 55% are zero. We also observe that there are very few non zero intakes less than 0.2 MJ.

We define three age categories, the same for males and females, and thus have six sex-by-age groups. Table 2.1 shows the mean and median alcohol consumptions along with the upper quartiles and the maximum intakes for the six sex-by-age groups. We see that the average alcohol intake is around 0.36 MJ per

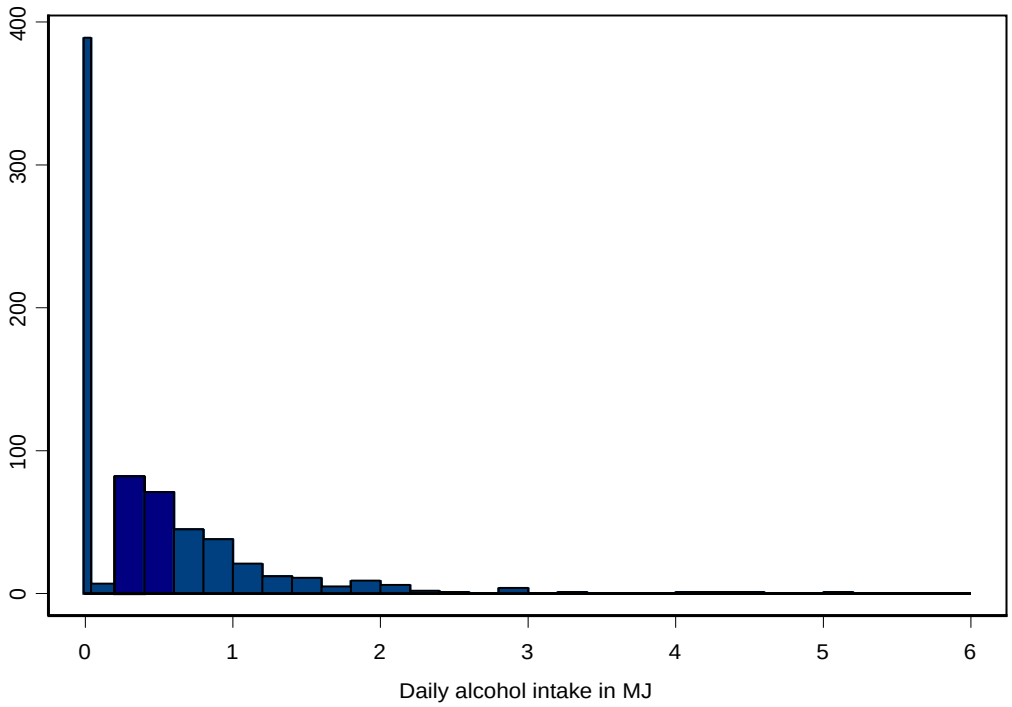

Figure 2.1: Histogram of daily alcohol consumption (MJ). The bar corresponding to the zero intakes is shown separately.

person, but if we consider only the non-zero alcohol consumption days this figure increases to 0.81 MJ of alcohol per day. Males have a higher average alcohol consumption than females if we ignore the sex-by-age groups. The maximum observed intake is 5.7 MJ which is more than 5 times the safe level on a given day for a female. For females it appears that average alcohol consumption decreases with age.

Figure 2.2 shows the box plot for daily alcohol intakes for the six sex-by-age groups. There is substantial variation within and between the sex-by-age groups. This could be due to the small sample size. The medians for most of the sex-by-age groups are zero. Men in the age group 50-64 years and women in the age group 18-34 years appear to have high intakes.

Since the data were collected in an artificial environment we prefer to ignore any possible day of the week effect. We thus assume that alcohol consumption on a given day is independent of the day of the week, that is, the distribution

Sex	Age	Number of individuals	Mean	Median	Upper quartile	Maximum intake
Male	18-34	8	0.30	0.00	0.42	1.87
Male	35-49	11	0.27	0.00	0.40	4.57
Male	50-64	11	0.65	0.39	1.01	4.31
Female	18-34	9	0.41	0.31	0.68	5.70
Female	35-49	10	0.33	0.00	0.58	3.24
Female	50-64	10	0.20	0.00	0.34	2.91
All individuals		59	0.36	0.00	0.50	5.70

Table 2.1: Summary statistics for daily alcohol intake in MJ according to the sex-by-age groups.

of alcohol intake is the same for all the 12 days. Though if we look at the data on the 10^{th} day the average alcohol consumption seems to be higher than on the other days. The reason for this is not known. However on testing for day effect formally using a Kruskal-Wallis test, we get a significance probability of 0.9 and we conclude that the mean intake for all the twelve days are equal.

A bar chart of the number of alcohol consumption days in Figure 2.3 shows a peak at zero, since we have 14 people who did not consume any alcohol on all of the 12 days.

It is not clear if the individuals in the data set who have zero alcohol consumption on all the 12 days are teetotallers or occasional drinkers who drink only on special days like birthdays and festivals. The data set does not give us any information about such long term consumption patterns.

2.4 Existing Methods for Modelling Zeros in Data

Several methods have been developed to account for high proportions of zeros in data. Mixture models developed to take into account high proportions of zeros in a data set are often called zero-modified distributions or distributions with added zeros (Johnson et al. (1992)). These are normally a combination of a discrete distribution together with the degenerate distribution with all probability concentrated at the origin. However a similar process can be applied to continuous distributions and can be used for the distribution of alcohol intakes. If F is a distribution function for a non-negative random variable X and $0 < \omega < 1$ then $\omega + (1-\omega)F(x)$ for $(x > 0)$ defines a modified distribution with extra probability

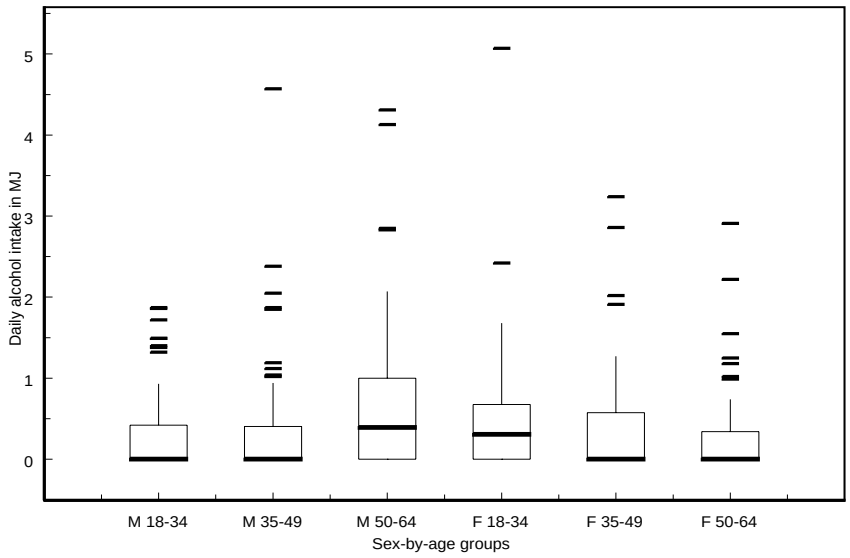

Figure 2.2: Box plot of daily alcohol consumption according to the sex-by-age groups.

ω at 0 (Johnson et al. (1992), Berk & Lachenbruch (2002)).

Another approach commonly used in econometrics is the Tobit model. Tobit models were introduced by Tobin (1958) for studying household expenditure on durable goods. He used a regression model in which expenditure was the dependent variable and could not take negative values. In his case the lower limit was zero, but any limiting value can be used in a Tobit model. Tobit models are also known as censored or truncated regression models. The regression model is truncated if the observations outside a specified range are totally lost and censored if one can at least observe the explanatory variable (Amemiya (1984)).

The standard Tobit model defined by Amemiya (1984) is as below

$$z_i = \mathbf{x}_i^T \boldsymbol{\beta} + u_i, \qquad i = 1, 2, \ldots, n, \tag{2.1}$$

$$y_i = \begin{cases} z_i & \text{if } z_i > 0 \\ 0 & \text{if } z_i \leq 0, \end{cases} \tag{2.2}$$

where u_i are assumed to be independent and identically distributed drawings from

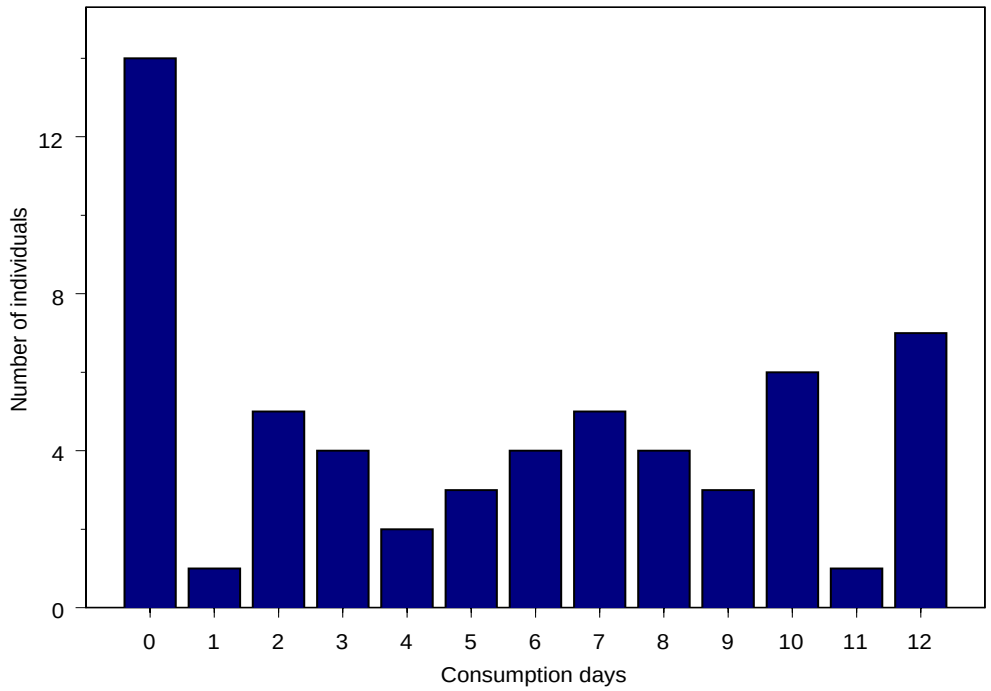

Figure 2.3: Bar chart for number of consumption days per individual.

$N(0, \sigma^2)$. It is assumed that y_i and x_i are observed for $i = 1, \ldots, n$, but z_i are unobserved if $z_i \leq 0$.

More recently Allcroft & Glasbey (2003) gave a method using Tobit analysis which they call the latent Gaussian model to analyse crop lodging (flattening of the crop) data. The latent Gaussian model is constructed such that zero observations correspond to the part of the distribution below some threshold and non-zero observations are transformed to fit the Gaussian distribution above the threshold. The threshold might be zero as in this example or positive: for example it can be the level of detection when recording pesticide residue on crops. Allcroft & Glasbey (2003) use data on per cent crop lodging from three seasons for 32 varieties of crops and seven trials, sixty-six per cent of the values are zero. A square-root transformation of the non-zero values fits the upper tail of a Normal distribution. The model is described by Allcroft & Glasbey (2003) as follows. Let y_{ij} be the square root of the observed lodging for variety i in trial j, and z_{ij} is the corresponding latent variable. Thus

$$y_{ij} = \begin{cases} z_{ij} & \text{if } z_{ij} \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.3}$$

Further the authors consider a fixed effects model for the complete block design with varieties V and trials T:

$$z_{ij} = \mu + \nu_i + \tau_j + \varepsilon_{ij}.$$

Here μ is the overall mean, ν_i the variety effect, τ_j is the the trial effect and the errors ε_{ij} are independent and identically distributed as $N(0, \sigma^2)$. Allcroft & Glasbey (2003) use maximum likelihood to obtain parameter estimates.

In the next sections we develop two possible models to study data sets with high proportions of zeros, as in the case of our alcohol intakes, and compare the two approaches. Model I is based on each individual's propensity to drink whereas Model II is a latent Gaussian model.

The three main features we need to account for in modelling the data are:

1. Skewness in the data
2. The large number of zero values
3. Varying within-individual variance between individuals.

2.5 Model I: Propensity Model

To model the zeros and the non-zero intakes in our data set, we first model if the individual drinks alcohol or not on a particular day; if yes then we define a distribution to determine how much the individual drinks. From Figure 2.1 we observe that the non-zero intakes are positively skewed. We investigate if there exists a transformation such that the non-zero intakes are from a Normal distribution. We want a transformation that will not affect the zero intakes, i.e. the zeros remain as zeros whereas the non-zero intakes are transformed to fit a Normal distribution. This makes back transformation to the original scale simple. A log transformation of the amount of alcohol consumed is not feasible because of the zero consumptions on certain days. Deleting these zero intakes would lead to an over-estimation of the mean consumption, as would adding a small quantity to the zero intakes. We try various simple power transformations to normalize the data. For the non-zero intakes, the quantile-quantile (QQ) plot in Figure 2.4 gives a poor fit to Normality, and a square root transformation improves this only slightly.

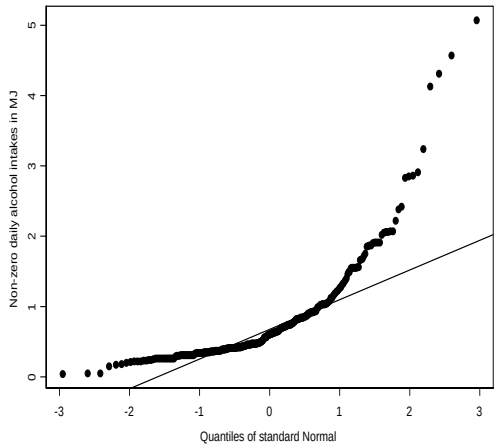

Figure 2.4: Normal QQ plot for untransformed non-zero intakes.

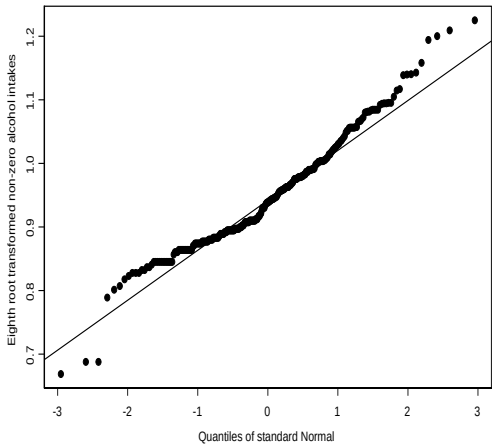

Figure 2.5: Normal QQ plot for eighth root transformed non zero intakes.

For the eighth root transformation, the QQ plot is given in Figure 2.5, but further transformation to the 16^{th} root gives no improvement in the fit of the data. Hence we decide to work with the eighth root of the data which transforms the data to Normal the most satisfactorily among all the power transformations considered. This is close to taking logarithms of the non-zero intakes.

We also calculate the Anderson-Darling (AD) statistic in Minitab (1972-2005), which is a measure of how far the plotted points fall from the fitted line in a probability plot. The statistic is a weighted squared distance from the points to the fitted line with larger weights in the tails of the distribution. For the eighth root transformation, the AD statistic is 3.2, whereas for the untransformed data it is 21.2. Smaller the AD statistic, better is the fit to normality. So the transformation has successfully reduced the AD statistic.

Figure 2.3 shows how the number of days on which alcohol is consumed varies between individuals. The simplest model is one which everyone has the same probability of drinking on any day, leading to a Binomial distribution, but days on which alcohol is consumed are more variable than this. Under the propensity model, on a given day j, an individual i has a real-valued propensity to drink, denoted by π_{ij}. The distribution of π_{ij} depends on i, so some individuals are more likely to drink than others. An individual not drinking on a given day has a zero or negative π_{ij}, while for someone drinking it will be positive. More generally, we may allow the distribution of π_{ij} to depend upon the day of the week, so that

on a weekend we would expect an individual i to have a higher π_{ij} than on a weekday. For convenience, we let the range of π_{ij} be from $-\infty$ to ∞ as we give π_{ij} a Normal distribution. This propensity π_{ij} thus takes into account the fact that on a given day a person might not consume alcohol but consume alcohol on any other day. The concept of propensity here is similar to the propensity factor used by Carriquiry (2003), except Carriquiry (2003) use it as a probability of consumption and hence is between $(0, 1)$.

In Figure 2.6, we have six randomly selected individuals from our data set who have non-zero alcohol consumption over the 12 days. Their daily alcohol intake is plotted, and the horizontal line represents the individual's average alcohol consumption in MJ. We observe that not only there is variation in intakes between individuals but also among intakes within an individual. The variation in intakes within an individual is not the same for all individuals. We take this varying within-individual variance into account in our model by assuming the within-individual variance to differ between individuals.

To model the alcohol intake on days with positive propensity to drink we develop a model similar to the one given by Myles et al. (2003) for dietary data. Henceforth in this section we refer to the eighth root transformation of the daily alcohol intakes as the response. The notional response for an individual i on a given day j is denoted by α_{ij}, where α_{ij} is given by Equation (2.4). The additive effect of the six sex-by-age groups is represented by the effect $\gamma_{k(i)}$.

$$\alpha_{ij} = \gamma_{k(i)} + \xi_i + \varepsilon_{ij}, \tag{2.4}$$

where $\gamma_{k(i)}$ is the effect of the group $k(i)$ including individual i. We drop the i from $\gamma_{k(i)}$ to simplify the notation and hence k becomes an index rather than a function. Also ξ_i and ε_{ij} represent between-individual and within-individual effects. We take $\xi_i \sim N(0, \sigma_b^2)$ and $\varepsilon_{ij} \sim N(0, \sigma_{wi}^2)$, where σ_b^2 is the between-individual variance and σ_{wi}^2 is the within-individual variance for individual i.

We define actual response a_{ij} for individual i on day j by

$$a_{ij} = \begin{cases} \alpha_{ij} & \text{if } \pi_{ij} > 0 \\ 0 & \text{if } \pi_{ij} \leq 0. \end{cases} \tag{2.5}$$

The total variance for an individual i is the sum of the within-individual variance σ_{wi}^2 and between-individual variance σ_b^2.

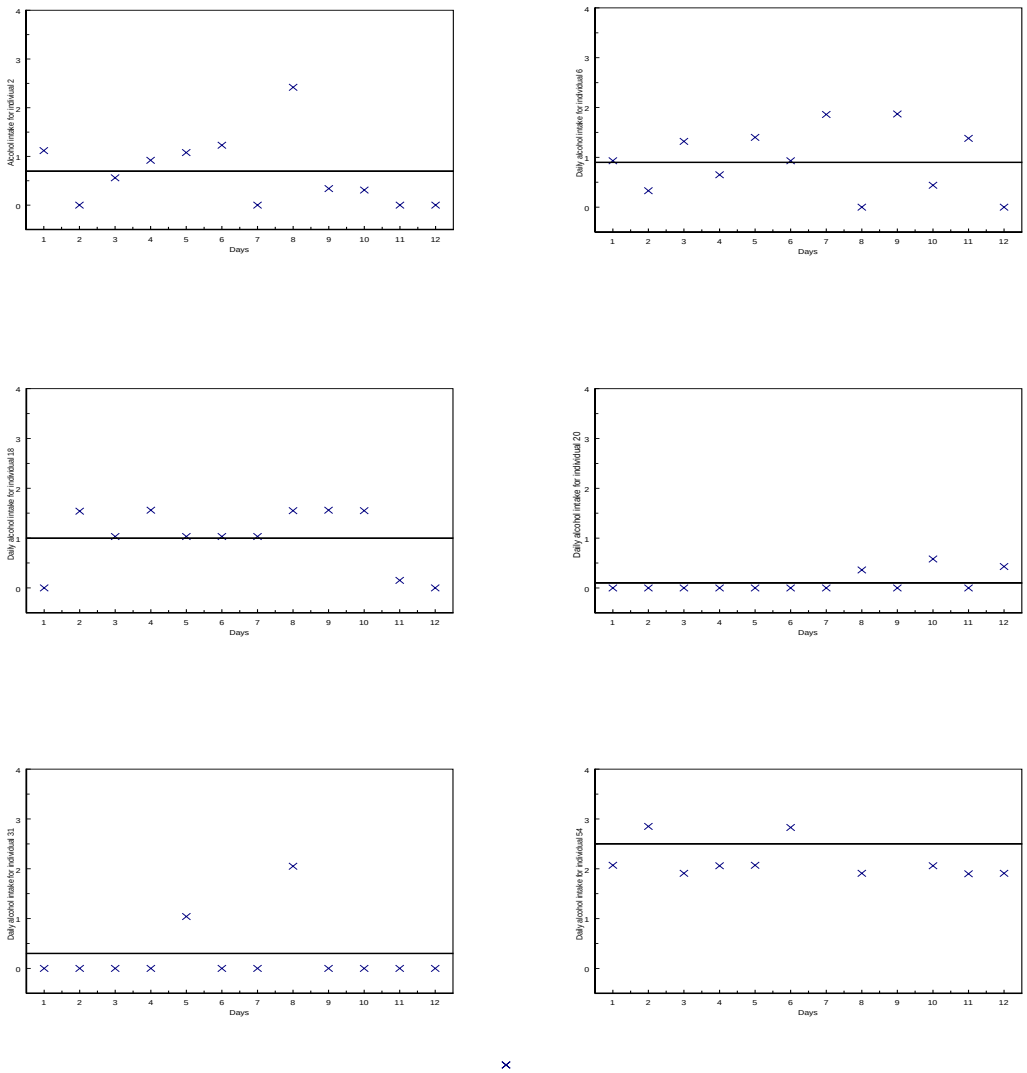

Figure 2.6: Alcohol intake for 6 randomly selected individuals

2.5.1 Prior Distributions for Propensity Model

We develop a hierarchical Bayesian model for our alcohol intakes. We account for the variation in the response over the twelve days within each individual, variation between individuals and within sex-by-age groups.

The expectation of the notional daily response for an individual i belonging to group k as given in equation (2.4) is assumed to be from a Normal distribution with mean γ_k and variance σ_b^2. The parameters for the sex-by-age effects (γ_k) are assigned a Normal prior distributions with mean ω and a common variance of four. We give ω a Normal prior distribution with mean 1 and variance 4. These values are chosen with the assumption that a person with a positive propensity to

drink on a given day will drink about three units of alcohol which is equivalent to almost one $MJ^{1/8}$ response on the eighth root scale. Though alcohol consumption cannot be negative, we still prefer to work with the Normal distribution for the convenience of modelling random effects at different levels.The prior probability of a negative consumption is negligible under this model.

Unfortunately we do not have any prior information about the parameters of the between-individual and within-individual variances as we could not find any publications relevant to this study. The precision (σ_b^{-2}) for the between-individual variance is given a Gamma prior distribution which we denote as $Ga(0.1, 0.1)$. Here $Ga(\alpha, \lambda)$ denotes a Gamma distribution with mean α/λ and variance α/λ^2. This gives σ_b a prior distribution with 5% upper and lower quantiles 1.3×10^6 and 0.41.

We assign the within-individual precisions σ_{wi}^{-2} a common Lognormal distribution. This is similar to Myles et al. (2003) model for dietary data. We define the $\ln(\sigma_{wi}^{-2})$ to have a Normal distribution with mean μ_w and precision δ_w. The prior for μ_w is chosen as $N(2, 1.5)$. We give δ_w a vague prior distribution of $Ga(0.0005, 0.0005)$.

The expected propensity for an individual i is also assumed to depend upon the sex-by-age group to which the individual belongs. The expected propensity for an individual i is given by β_i. We assume that β_i is from a Normal distribution with mean ρ_k and variance 1. The priors for ρ_k are chosen to be $N(0, 4)$ for all six sex-by-age groups. Thus within a sex-by-age group, the expected propensity has a common distribution. The prior distributions for the individual propensity give every individual an equal chance of drinking or not drinking on a given day.

From the above model and the data we can calculate the posterior distributions for the sex-by-age effects. Using the posterior distributions for the sex-by-age effects and the group propensities ρ_k along with the posterior distributions of the between and within-individual variances, we can simulate the daily alcohol intake for a new individual in a given sex-by-age group. We can also predict the total alcohol intake over a longer period of time, for example a month.

The alcohol intakes for these new individuals, each in a different group, will generate the predicted actual alcohol consumption; any realization can be expected to have some zero alcohol consumption days. We can predict the prob-

ability of a zero consumption day for each sex-by-age group from the predicted propensities. To look at longer-term alcohol intakes we can extend the model over several days and observe the overall consumption by summing the actual predicted alcohol consumption over the required number of days. We can then look at the predicted intakes and determine the proportion of individuals in each sex-by-age group exceeding the safe levels of alcohol consumption over a week. Estimates of these probabilities are given in Section 2.7 of this chapter.

To generate the alcohol intake for a new individual belonging to the sex-by-age group 'k' we introduce a set of new variables (these are represented with an ending of $.n$ in the WinBUGS code in Appendix A). The doodle diagram corresponding to the propensity model is given in Figure 2.7. The nodes and edges in blue present the extension to the model to predict intakes in the future for a new individual. The expectation for the new individual's alcohol consumption is given by γ_k. The notional intake is from a Normal distribution whose mean is the posterior distribution of γ_k and the variance is given by the posterior distribution of σ_b. The within-individual variance for this new individual is estimated using the posterior distributions of μ_w and σ_w. This new model is then simulated over number of days. The actual alcohol consumption on a future day d for this new individual is given as

$$a_{kd} = \begin{cases} \alpha_{kd} & \text{if } \pi_{kd} > 0 \\ 0 & \text{if } \pi_{kd} \leq 0. \end{cases} \tag{2.6}$$

The π_{kd} are from $N(\beta_k, 1)$ and β_k is from a Normal distribution with mean as the posterior distribution of ρ_k and variance 1. Also α_{kd} is the notional intake.

2.5.2 Results for the Propensity Model

We use 100,000 iterations to estimate our model parameters in WinBUGS. The first 5000 samples were discarded as they were used as burn-ins. Refer to Appendix A for the code of the model as used in WinBUGS. The results presented in this section are for the eighth root transformed data unless mentioned otherwise.

The posterior expected notional response for each sex-by age group (γ_k) is given in Table 2.2. We also have the standard error (SE) for these parameter estimates and the upper 2.5 percentile. It appears the mean for the posterior expectation for the group effects are higher in the age group 35 to 49 years for each sex. For males however the 97.5% value for ages 18 to 34 years is the same as that for 35 to 49 years.

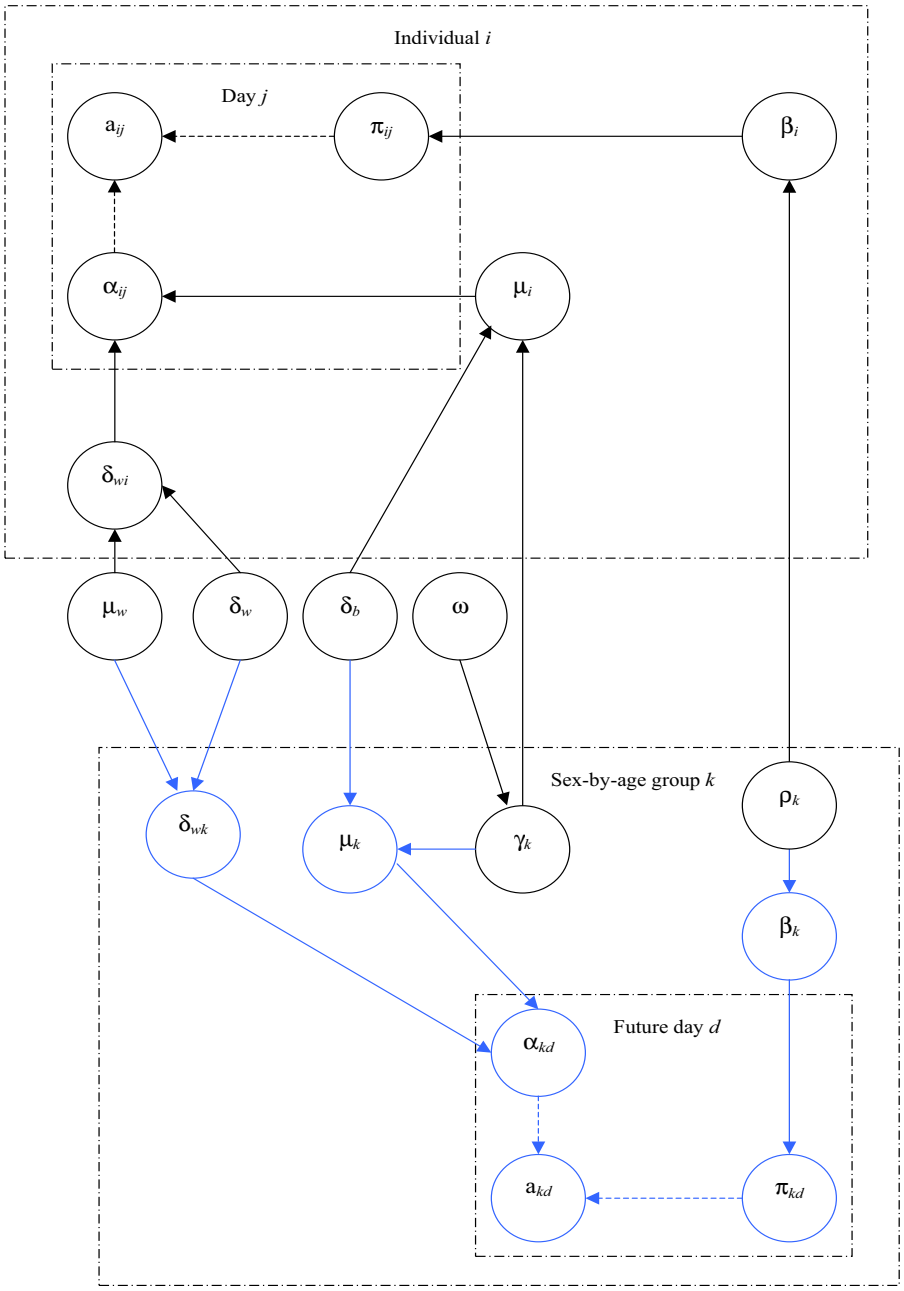

Figure 2.7: Doodle diagram for Propensity model

Sex	Age	Mean	SE	MC error	97.5 percentile
Male	18-34	0.9450	0.0453	0.0015	1.032
Male	35-49	0.9624	0.0354	0.0008	1.032
Male	50-64	0.9429	0.0319	0.0005	1.006
Female	18-34	0.9487	0.0383	0.0009	1.025
Female	35-49	0.9517	0.0378	0.0009	1.029
Female	50-64	0.9326	0.0417	0.0016	1.015

Table 2.2: Posterior moments for sex-by-age effects for Model I.

The posterior expectation of β_i gives the posterior expected propensity to drink for individual i. The smaller the number of consumption days for a person, the lower will be his or her posterior expected propensity to drink, and hence he or she should have a smaller value of the estimated β_i. From Figure 2.8 we can see that the posterior expectations for β_i increase with the number of consumption days. For people with zero alcohol consumption belonging to the same group, one would expect the posterior expectations of the β_i values to be similar. After 100,000 simulations these values matched to one or two decimal places. Increasing the number of simulations resulted in the values becoming closer to each other.

The posterior expectation for the between-individual variance, σ_b^2 is 0.0096 and it has a standard deviation of 0.0024.

In the model given by Equation (2.4), individual i has his or her own variance for alcohol intake, σ_{wi}^2. For individuals with zero alcohol consumption over the twelve days, the data provides no information for σ_{wi}^2 values. For these 14 individuals, their posterior distribution for σ_{wi}^2 is similar to the prior distribution. Among the remaining 45 posterior expected values for within-individual variance σ_{wi}^2, the maximum and the minimum are 0.001 and 0.015.

2.5.3 Sensitivity to Choice of Prior Distributions

We have made some arbitrary decisions about the prior distributions for the between-individual and within-individual variances. We examine the effects on the posterior expectations of changing the parameter values of the prior distributions.

When the prior expectations for the sex-by-age effects (γ_k) and the prior expectation for the between-individual variance (σ_b^2) are doubled simultaneously,

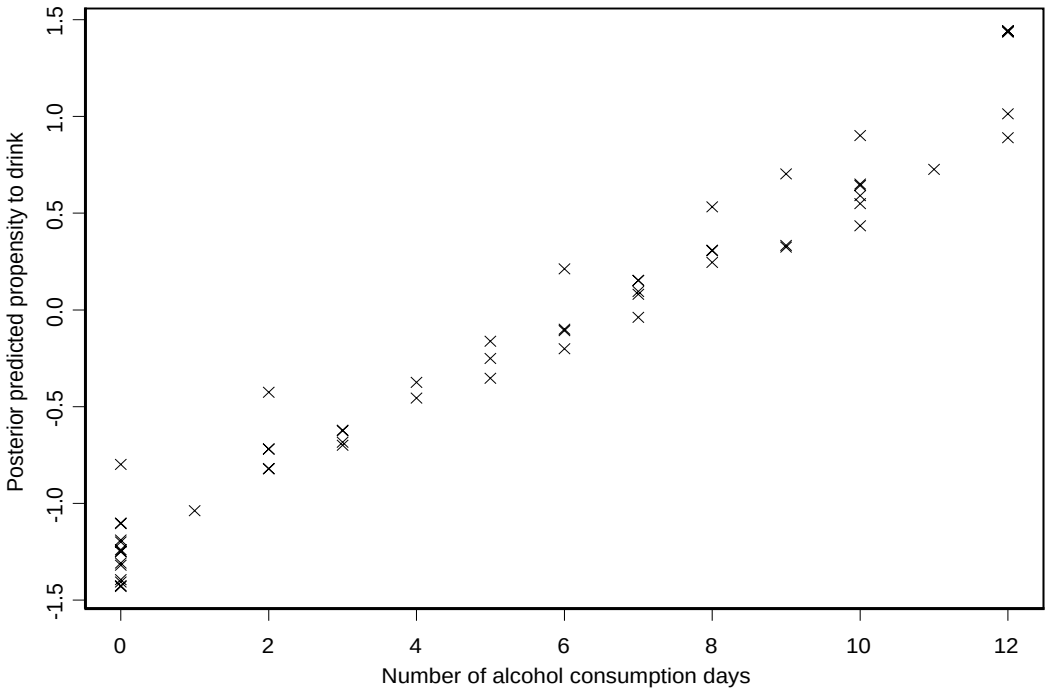

Figure 2.8: Scatter plot of the posterior expected propensity to drink along with the number of alcohol consumption days per individual.

the changes in the posterior expectations for γ_k are less than 5% and there is less than 1% change in the expectation for σ_b^2. We also decrease the prior expectations for the sex-by-age effects from one to zero and the changes in the posterior expectations for these parameters are less than 2%. A 50% decrease in the prior expectation for the between individual variance causes the posterior expectation for this parameter to decrease by 5%, but the overall effects on the predicted intakes are negligible.

Similarly we change the in the prior expectation for the group propensity parameters $\rho_{k(i)}$ from 0 to 1 and -1, these result in less than 1% changes in their posterior expectations. However, changing the variance has an effect on the posterior expectations of the group propensities and in turn affects the predicted proportions of zeros in each sex-by-age group. The larger the prior variance for $\rho_{k(i)}$ the poorer is the match between the observed and predicted proportions of zeros. Decreasing the variance from one to 0.5 does not help in improving the predicted proportions of zeros any further. The effect on the predicted exceedance

probabilities is however negligible.

We also change the prior expectation for the log of within-individual precision, μ_w. An increase and decrease by 100% in the prior expected value μ_w causes the prior expectation of the log of the within-individual variance to increase and decrease by 100% respectively. The posterior expected values of the within-individual variances showed less than 1% change in both cases. Changing the expected variance for the log of the within-individual precision also causes less than 1% change to the posterior expected within-individual variances.

2.5.4 Model Adequacy

In Figure 2.9 we compare the cumulative distribution function (cdf) of the data with the predictive cdf from the Propensity model in the original scale for the whole population. The predicted probability of a zero intake matches well with the observed probability but the two cdf curves do not lie very close to each other. The model predicts more intakes less than 0.4 MJ and fewer intakes less between 0.4 and 4 MJ as compared to the data. An alternate and better transformation to Normality may improve the fit of the model to the data. We propose an alternative model for studying the alcohol intakes in Section 2. 6.

2.5.5 Convergence of Parameter Estimates

The history plots for the sex-by-age effects, between and within-individual variances were examined. The parameter distributions seem to have reached convergence. As an example the history plots for sex-by-age effects (γ_k) defined in equation (2.4) are given in Figure 2.10 for iterations between 5000 and 80000.

The Monte Carlo errors are less than 5% of the sample standard deviation for all the results obtained: the MC error for the sex-by-age effects can be seen in Table 2.2.

More robust methods for convergence testing can be used. There is debate as to whether it is better to run more than one chain with dispersed initial values and monitor convergence instead of running a single long chain (Cowles & Carlin (1996)). According to Cowles & Carlin (1996), running multiple chains is inefficient as compared to running one single long chain. We have run single long chains for our parameters.

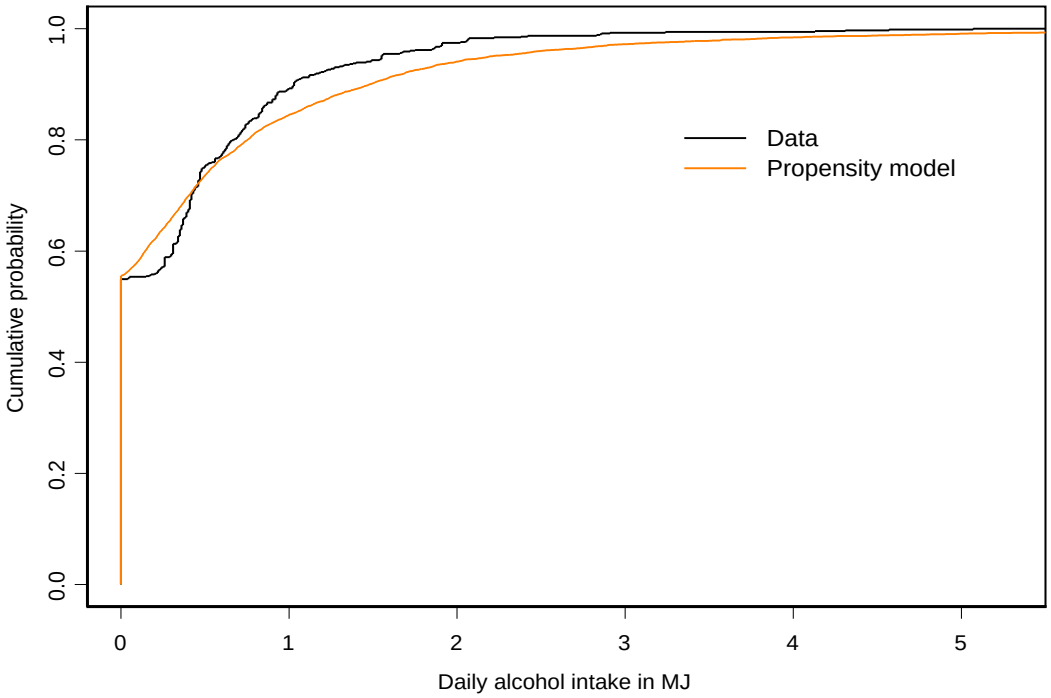

Figure 2.9: Empirical cdf of the daily alcohol intakes with the predictive cdf of daily alcohol intakes from the Propensity model all the sex-by-age groups combined.

2.6 Model II: Latent Gaussian Model

In contrast with the Propensity model, the latent Gaussian model allows us to model both the occurrence and the amount of the measured quantity to be described by a single random variable. The latent Gaussian model as described by Allcroft & Glasbey (2003) assumes that zero observations are actually censored observations, smaller than a known threshold. These zero observations can be false zeros; for example data obtained from applying blood-alcohol tests to motorists might have false zeros, which arise when the level of alcohol in the blood is less than the level of detection. Since for our data set we know the exact amount of alcohol intake by each individual, we treat the zeros as true zeros and set our threshold to be zero.

The model assumes that there exists a transformation such that the non-zero part of the data fits the tail of a Normal distribution above the threshold. This is also different from Model I where we assume there exists a transformation such

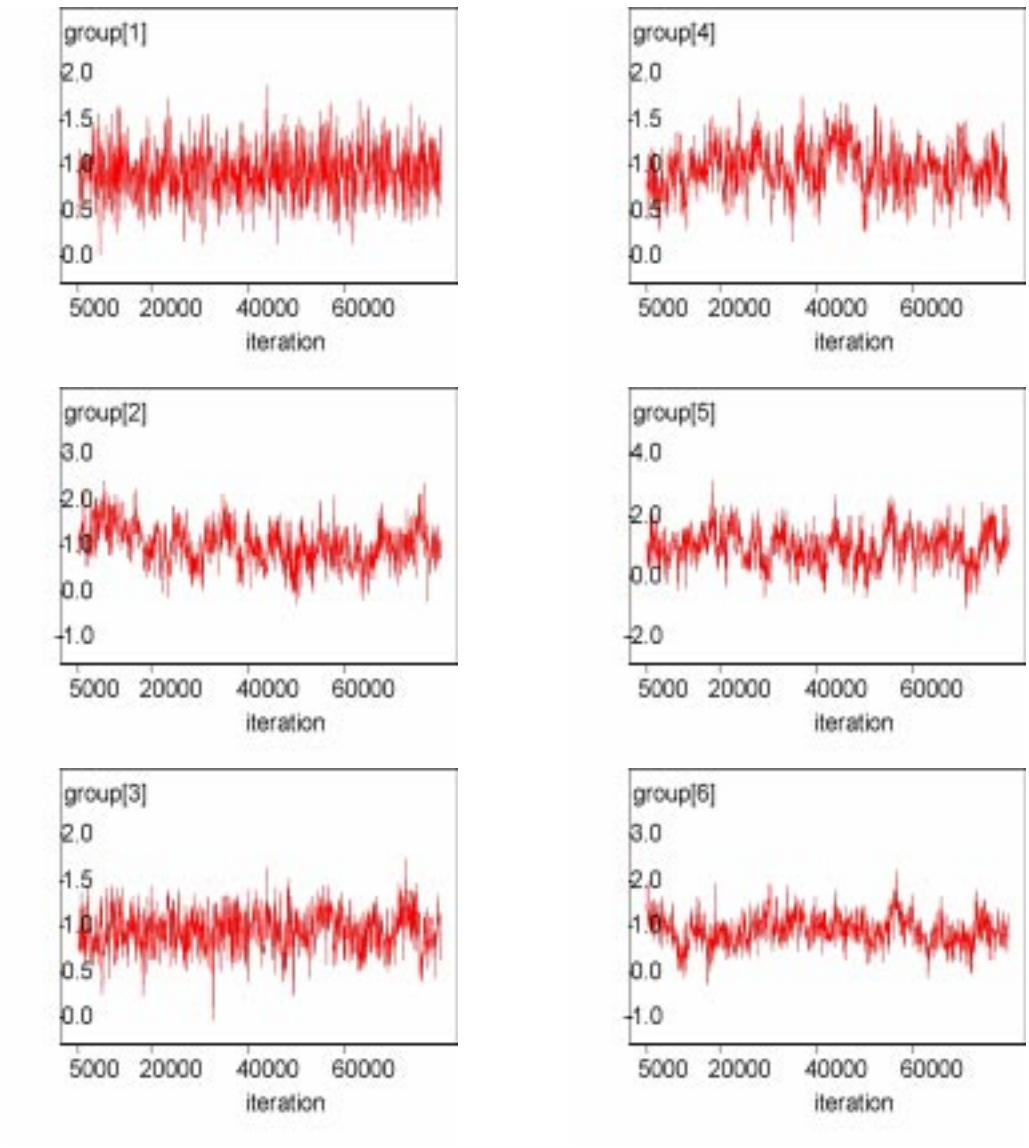

Figure 2.10: History plots for the six sex-by-age effect parameters in the Propensity Model for iterations between 5000 and 80,000.

that the notional responses are from a Normal distribution and not just the tail of one.

In Figure 2.11 we have a QQ plot for the daily alcohol intakes. The group of points forming a horizontal line on the plot correspond to the zero intakes. We want the remaining points to fit the right tail of a Normal distribution. The diagonal straight line is through the 66^{th} and 87^{th} quantile of a standard Normal distribution. These quantiles correspond to the quartiles of the positive intakes. These values are chosen since for our data we want the 45% non-zero values to fit the right tail of a Normal distribution. So dividing this 45% in to approximately four equal parts, the lower and upper quartiles correspond to the 66^{th} and 89^{th} quantile of a standard Normal. If this is true the non-zero intakes will lie close to the diagonal straight line in Figure 2.11. This is clearly not the case. We see that for large intakes, the fit to the right tail of a Normal distribution is particularly poor.

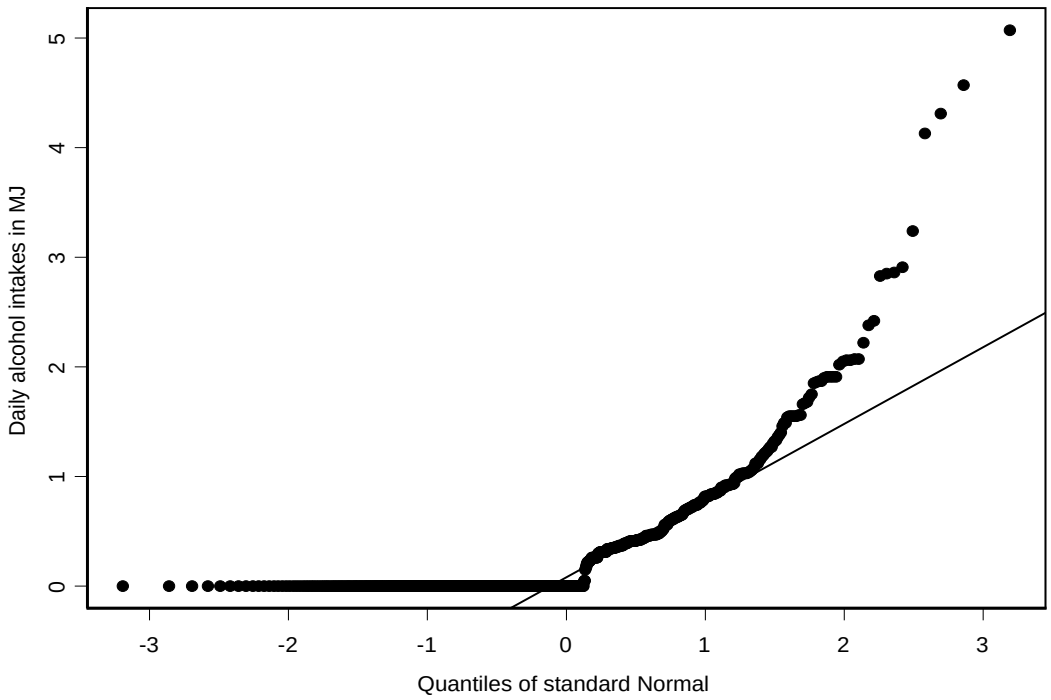

Figure 2.11: QQ plot for daily alcohol intakes.

To improve the fit of the positive intakes to a tail of a Normal distribution we

try simple power transformations such as taking the square root of the intakes. In Figure 2.12 we have the QQ plot for the square root intakes. We see that the transformed non-zero points do not all lie close to the diagonal straight line and the deviation from the line is more for smaller intakes. However there is an improvement for the points corresponding to large intakes as they are now closer to the diagonal straight line as compared to the untransformed data. Since the fit is good for most of the points we decide to work with the square root of intakes.

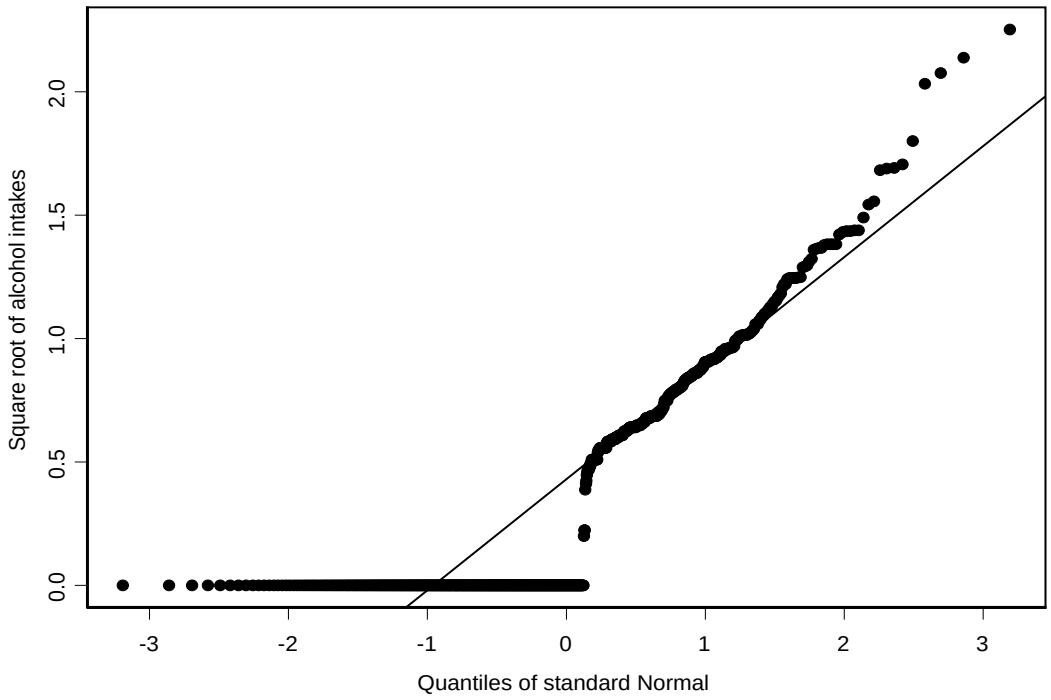

Figure 2.12: QQ plot for square roots of daily alcohol intakes.

The data here have a hierarchical structure. We define a Normal distribution to describe each individual's notional alcohol intakes. However we have a single transformation to all the non-zero intakes as opposed to having a transformation for each individual's non-zero intakes to fit the right tail of a Normal distribution. This is only an approximate method.

Each individual's intakes are assumed to be from a Normal distribution which is left-censored at zero. The within-individual variances for the intakes over the 12 days differ between individuals. Here we work with the square root of the

intakes and refer to these transformed intakes as responses henceforth.

As before α_{ij} denotes the i^{th} individual's notional response on the j^{th} day. We assume that α_{ij} are from a Normal distribution with mean μ_i and within-individual variance σ^2_{wi}. Here we refer to μ_i as the expected notional response for individual i. The actual response for the i^{th} individual on the j^{th} day is

$$a_{ij} = \begin{cases} \alpha_{ij} & \text{if } \alpha_{ij} > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{2.7}$$

The likelihood for this model can be written as

$$L(a_{ij}|\mu_i, \sigma_{wi}) = \left\{ \prod_{a_{ij}>0} \sigma_{wi}^{-1} \phi\left(\frac{a_{ij} - \mu_i}{\sigma_{wi}}\right) \right\} \left\{ \prod_{a_{ij}=0} \Phi\left(\frac{-\mu_i}{\sigma_{wi}}\right) \right\}, \tag{2.8}$$

where ϕ and Φ denote the probability density function and the cumulative distribution function of the standard Normal distribution respectively.

The notional alcohol response α_{ij} is from a Normal distribution with mean μ_i and variance σ^2_{wi}. We let the expectation of the notional response for each individual depend on the sex and age of that individual, so that

$$\mu_i = \gamma_{k(i)} + \xi_i. \tag{2.9}$$

Here $\gamma_{k(i)}$ is the sex-by-age effect for the i^{th} individual and ξ_i are independently identically distributed $N(0, \sigma^2_b)$. As before we drop the subscript i from $\gamma_{k(i)}$ for convenience.

2.6.1 Prior Distributions for the Latent Gaussian Model

We develop a hierarchical Bayesian latent Gaussian model for the responses.

The sex-by-age effects γ_k are given Normal prior distributions, around ω and have a variance of 200. Since most daily intakes are zero, we might expect μ_i to be negative. Here ω is given a Normal prior distribution with mean 0 and variance 200.

The between-individual precision σ_b^{-2} is given a Gamma distribution $Ga(0.01, 0.01)$. Thus the 5% upper and lower quantiles for σ_b are 1.9×10^{68} and 2.5. The logarithms of the within-individual precision $\log(\sigma_{wi}^{-2})$ are given a Normal distribution

with mean μ_w and precision δ_w. These hyper parameters are then given a Normal and a Gamma prior respectively, $N(0, 0.1)$ and $Ga(1, 10)$. For the individuals who have zero alcohol consumption over the 12 days, the posterior distributions for the within-individual precisions σ_{wi}^{-2} should be close to their prior distribution.

We specify the censored observations in our model using the $I(lower, upper)$ function in WinBUGS. Since for our latent Gaussian model we assume the zeros to be censored we replace the 0's by NA in our data file. We also specify the lower and upper values between which the censored and uncensored observations lie. When censoring is specified the censoring node contributes a term to the full conditional distribution of its parents. Thus for censored observations the interval is $I(-\infty, 0)$ and for uncensored observations it is $I(-\infty, 10000)$. WinBUGS allows only one limit in the interval to vary between individuals. Here we fix the lower limit to $-\infty$ and the upper limit is 0 or 10000 for censored and uncensored observations respectively. WinBUGS does not allow the varying limit to be infinity and hence we fix the upper limit for non-zero observations to be 10,000.

We used WinBUGS to obtain posterior parameter distributions of our model and also predicted daily and longer term intakes. The model was run for 100,000 simulations and among these the first 5000 simulations were discarded as burn-ins.

2.6.2 Results for Latent Gaussian Model with Square-root Transformation

The results are for the square root transformed data. The posterior expectations for the group effects in equation (2.9) are in Table 2.3. The smaller the posterior mean value of the sex-by-age effect, the larger is the posterior probability of getting an intake less than zero.

Sex	Age	Mean	SE	MC error
Male	18-34	−0.2602	0.3262	0.007
Male	35-49	−0.4699	0.2905	0.007
Male	50-64	0.4535	0.2486	0.003
Female	18-34	0.0619	0.2889	0.005
Female	35-49	−0.1716	0.2822	0.006
Female	50-64	−0.5953	0.3122	0.009

Table 2.3: Posterior moment for sex-by-age group effects for Latent Gaussian Model.

The posterior expectation for the between individual variance is 0.63 with a standard error of 0.21. Considering only the individuals who drank on at least one day, the maximum and minimum posterior expectations for the within-individual variances are 2.34 and 0.02 respectively. Using the posterior distributions for μ_i and the between-individual and within-individual variances, we can determine the probability of a censored intake for individual i, $Pr(a_{i*} = 0)$, where $*$ represents a future day. We have for each individual

$$Pr(a_{i*} = 0|\mu_i, \sigma_{wi}) = \Phi(-\mu_i/\sigma_{wi})$$

and we find the probabilities using the posterior expectations of $\Phi(-\mu_i/\sigma_{wi})$. In Figure 2.13 we compare the number of consumption days with the posterior predicted probability of non-zero intakes for each individual. Figure 2.13 suggests that the two values are close to each other, the diagonal line represents equality between the observed number of days and posterior predicted probability to drink. The points at the top right hand corner of the plot represent individuals who had non-zero alcohol consumption on all the 12 days and the points on the bottom left hand corner represent daily drinkers.

2.6.3 Sensitivity to Choice of Prior Distributions

We examine the effects on the posterior expectations of changing the parameter values of the prior distributions as we did for Model I.

The prior expectation ω for the sex-by-age effects γ_k is changed from 0 to 2 and the expectation for the between-individual variance is doubled from 1 to 2. Simultaneously the expectation of the logarithms of the within-individual variance, given by μ_w is given a prior expectation 10. The largest change in the posterior expectation for the sex-by-age effects is observed for males in the age group 18-34 years, where the posterior expectation decreases from - 0.26 to -0.34. The smallest change observed is for females in the same age group, where there is an increase of about 4% in the posterior expectation. The change in the prior for the between-individual variance causes its posterior expectation to increase by about 25% of the previously stated value. The average of the posterior expectations for the within-individual variance increases by about 14%.

We also study the effect of decreasing the prior expectation for the sex-by-age parameter from 0 to -2 and halving the prior expectation for the between individual variance. We reduce the prior expectation of the expectation of the

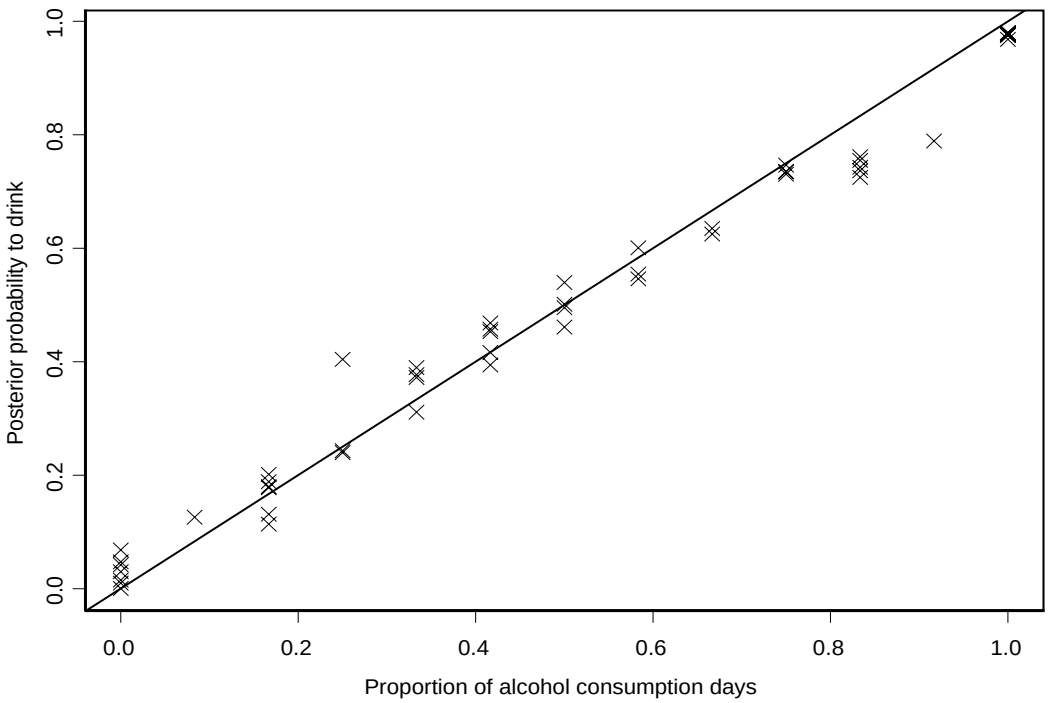

Figure 2.13: Scatter plot of predicted probability of having non-zero alcohol intake on a particular day and observed number of consumption days for each individual.

logs of within-individual variance from 0 to -10. The posterior expectations for the sex-by-age effects show the same changes as mentioned in the above paragraph but in the opposite direction. The posterior expectation for the between individual variance increases by less than 15%, whereas the change in the average of the posterior expectations of the within-individual variance is less than 1%.

History plots are used to judge the convergence of the parameter distributions of the model. The MC errors corresponding to the parameters were less than 5% of the sample standard deviation. All parameter distributions appear to converge.

2.6.4 Model Adequacy

A predicted negative response from our model corresponds to a censored observation and is set to zero. As for Model I, we consider a random individual from the population with probabilities of group membership proportional to the observed numbers in the group. The predicted daily alcohol intakes for such an individual are determined by the posterior distributions of the model parameters. We com-

bine the simulated intakes for all the groups, and in Figure 2.14 we compare the cdfs of the data and the predicted alcohol intakes on the original scale. The two cdf lines do not appear to be very close to each other. As in Figure 2.9 the main difference is the lack of observed positive intakes less than 0.4 MJ.

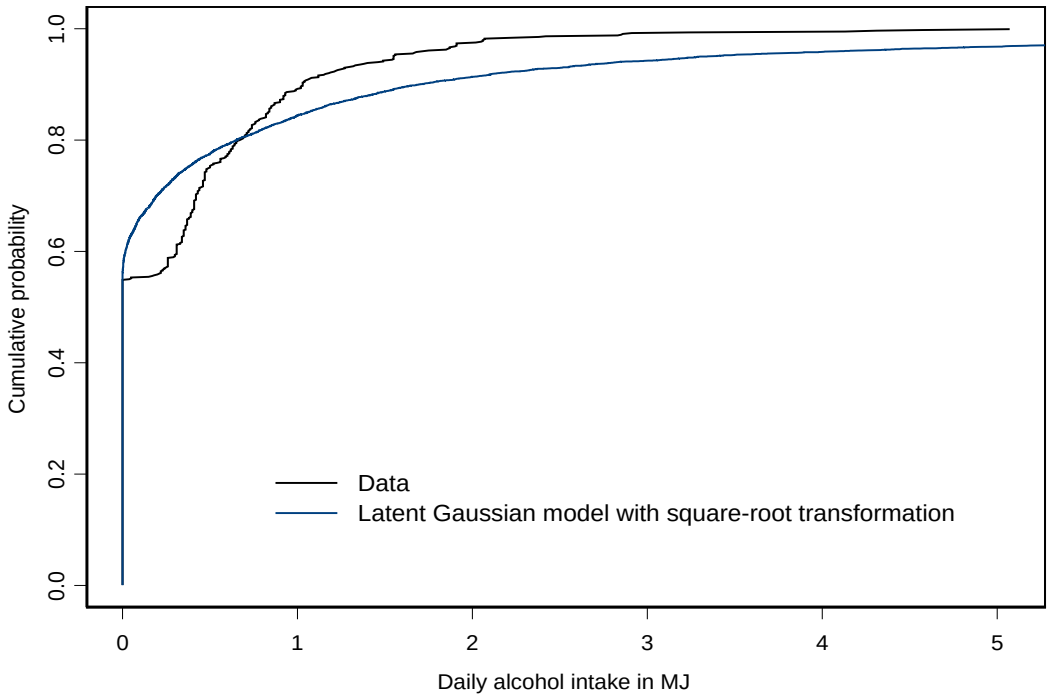

Figure 2.14: Empirical cdf of daily alcohol intakes with predicted cdf from latent Gaussian model fitted to square root transformation for all the sex-by-age groups.

To improve the fit of the model it is essential to find a better transformation so that the non-zero values fit a right tail of a Normal distribution well. From Figure 2.14 we infer that we require a transformation such that the smaller values will be close together.

2.6.5 An Alternative Transformation for the Latent Gaussian Model

Since we want the small intakes to be closer to each other without separating out the large intakes, the transformation we work with is

$$a_{ij} = \begin{cases} (y_{ij}/0.69)^2 & \text{if } a_{ij} \leq 0.69 \\ (y_{ij}/0.69)^{0.7} & \text{if } a_{ij} > 0.69. \end{cases} \quad (2.10)$$

Here y_{ij} are our observed alcohol intakes. We call this the two-part transformation. The change point of 0.69 MJ is chosen as this corresponds to 3 units of alcohol which we assume to be the approximately the average amount of alcohol a person consumes, provided he or she consumes alcohol. The above function is monotonic and continuous.

The QQ plot for the two-part transformation is in Figure 2.15. We see that the non-zero points all lie close to the diagonal straight line. Using this transformation we fit the same latent Gaussian model with the same prior distributions as in Section 2.6.1. We again simulate predicted daily intakes for an individual in each sex-by-age group.

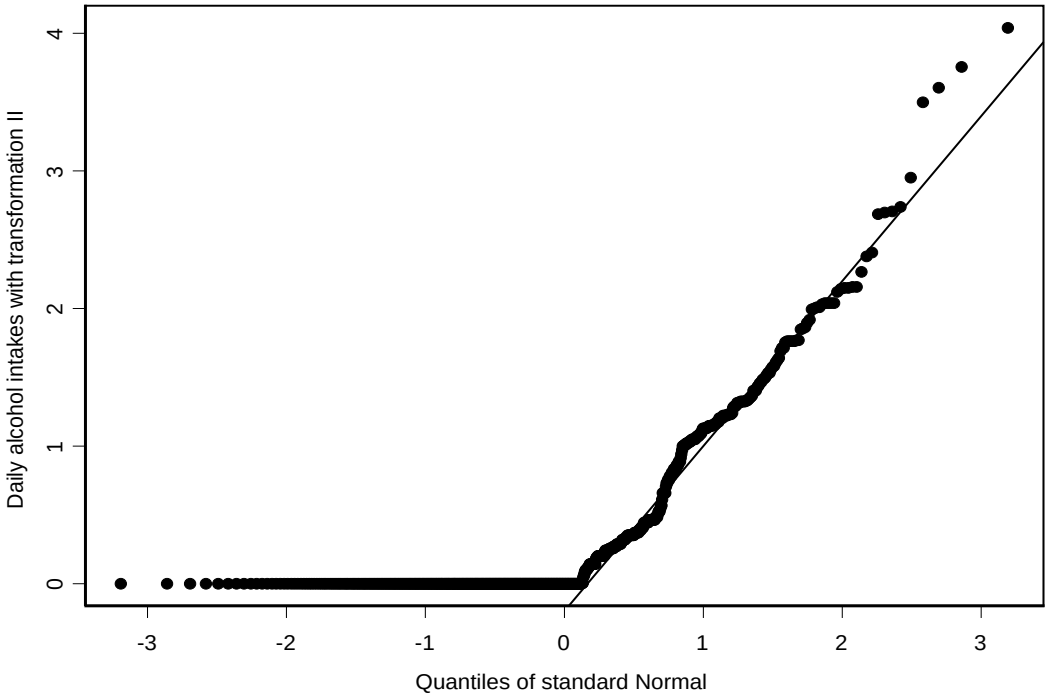

Figure 2.15: QQ plot for daily alcohol intakes with two-part transformation

Figure 2.16 compares of the cdfs of the observed data and the predicted intakes using this transformation. In this case the fit of the model to the data is

much better, and there is a large improvement in predicting daily intakes less than 0.4 MJ.

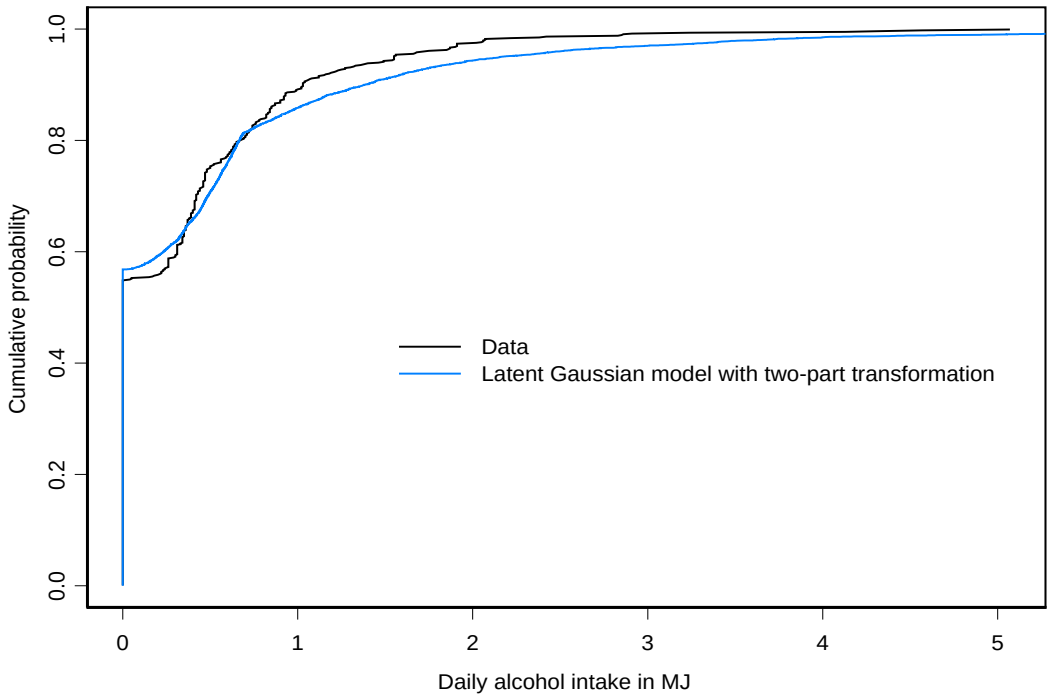

Figure 2.16: Empirical cdf with the predicted cdf of daily alcohol intakes from latent Gaussian Model with two-part transformation for all the sex-by-age groups.

2.7 Model Comparison

We have discussed two models to study daily alcohol intakes. In this section we compare the results from the two models, the Propensity model and the latent Gaussian model. Using the predicted intakes from the Propensity model and the latent Gaussian model with the two-part transformation we compare the cdf of the predicted intakes with the observed data for each sex-by-age group in Figure 2.17. We see that the Propensity model and the latent Gaussian with two-part transformation predict the proportion of zeros well in all cases, expect for males 50-64 years, where the latent Gaussian over-estimates this proportion. However, the latent Gaussian model fits the empirical cdf of the combined data better. For intakes greater than 1 MJ, both models behave similarly.

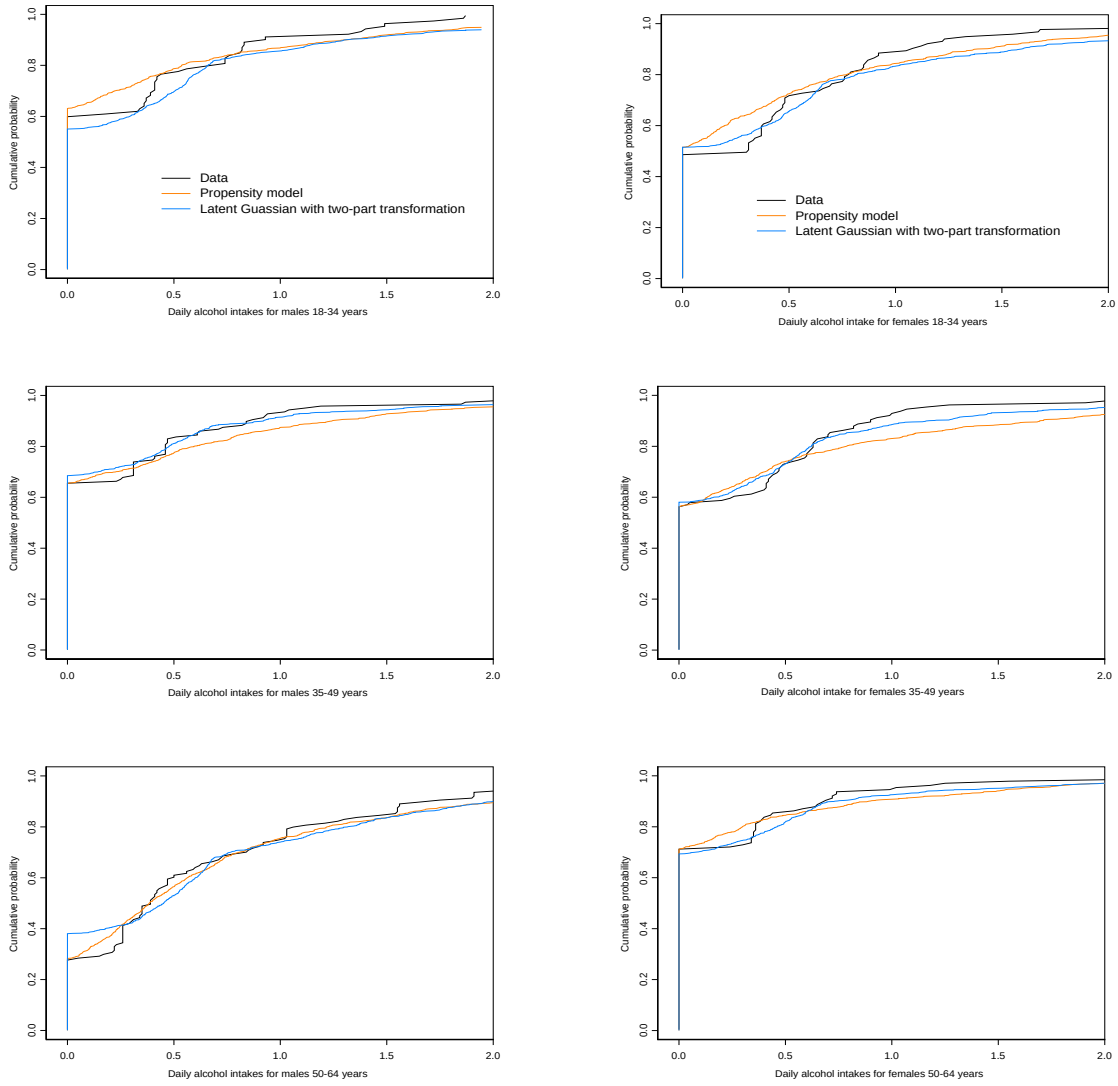

Figure 2.17: Comparison of empirical cdf of observed daily alcohol intakes with cdf of predicted intakes from the Propensity model and the latent Gaussian model with the two-part transformation for each sex-by-age group.

From the predicted daily intakes we can find the predictive probability of zero intakes in each group. Table 2.4 has these probabilities from the Propensity model and the latent Gaussian model with two-part transformation along with the observed proportions. Both models perform well, except for one case mentioned before, where the latent Gaussian model with the two-part transformation over-estimates the proportion of zeros for males 50-64 years by almost 35% of the observed proportion.

We also determine the predicted exceedance probabilities from the model. The exceedance probabilities considered are the probabilities of exceeding the recommended alcohol intake levels, which are 0.69 MJ for females and 0.92 MJ for males. Table 2.4 has the exceedance probabilities from the Propensity model and the latent Gaussian Model along with the observed proportions from the data. Again there is not much to choose between the two models. However the Propensity Model over estimates the exceedance probability by about 1.7 times the observed values for males aged 35-49 years. The latent Gaussian model underpredicts the exceedance percentages for all the sex-by-age groups.

Sex	Age	Percentage of zero intakes			Exceedance Percentage		
		Data	Propensity model	Latent Gaussian	Data	Propensity model	Latent Gaussian
Male	18-34	60.4	62.8	57.5	10.4	13.2	10.2
Male	35-49	65.9	66.1	68.2	8.3	14.4	6.9
Male	50-64	28.0	27.3	35.8	25.1	27.1	21.1
Female	18-34	49.1	52.9	46.5	24.1	23.1	19.6
Female	35-49	56.7	57.0	57.3	16.7	19.7	15.0
Female	50-64	71.7	73.2	71.7	9.1	10.8	7.9

Table 2.4: Observed and predictive percentages of zero intakes and daily alcohol intakes exceeding the safe levels of 0.69 MJ for females and 0.92 MJ for males. The predicted probabilities for the latent Gaussian are with the two-part transformation.

According to the British government guidelines, a male drinking more than 4.8 MJ of alcohol in a week and a female drinking more than 3.2 MJ exposes themselves to health risks. From the posterior predictive distribution of total alcohol consumption over a week we can find the probability of an individual exceeding these safe limits in each group. The exceedance probabilities are given in Table 2.5 for both models along with the corresponding proportions from the data. To find the total alcohol consumption for an individual over a week, we use

the average intake over twelve days and multiply by seven. This is done to reduce the bias that might be introduced by choosing the seven days over which we find the total intake. For example if we choose the first seven days from our data set and look at the total alcohol consumption over these days, females aged 50-64 years have an observed exceedance probability of zero. The predicted percentages are not very close to the observed percentages. We also compare the cdf of the predicted distributions from the Propensity model and the latent Gaussian model with two-part transformation for the total weekly alcohol consumption with the cdf from the data in Figure 2.18. The Propensity model gives a very low predicted probability of total weekly intake being zero as compared to the data. The latent Gaussian model appears to over estimate the probability of the total weekly intake being less than 2 MJ.

Sex	Age	Percentage of exceedance		
		Data	Propensity Model	Latent Gaussian
Male	18-34	12.5	17.4	17.4
Male	35-49	9.1	18.1	12.1
Male	50-64	45.4	38.0	33.1
Female	18-34	55.5	38.1	38.5
Female	35-49	30.0	43.3	30.1
Female	50-64	20.0	25.0	18.8

Table 2.5: Observed and predictive percentages of total weekly alcohol intakes exceeding the safe levels of 3.22 MJ for females and 4.82 MJ for males. The predictive percentages for the latent Gaussian model are with the two-part transformation.

2.8 Discussion

One common problem with monitoring alcohol intakes is with mis-reporting of actual consumption. This was not a problem here as accurate alcohol consumption values were available.

The data in this study have been collected from an artificial environment where the individuals were in an enclosed area. The individuals taking part in the study may not be representative of the whole population. It justifiable to not assume any day of the week effect for this data set. In a more natural environment one might expect higher alcohol intakes over weekends. One could also consider any seasonal effects on alcohol intakes. For this one would then need data on alcohol consumption during various times of the year. Again alcohol con-

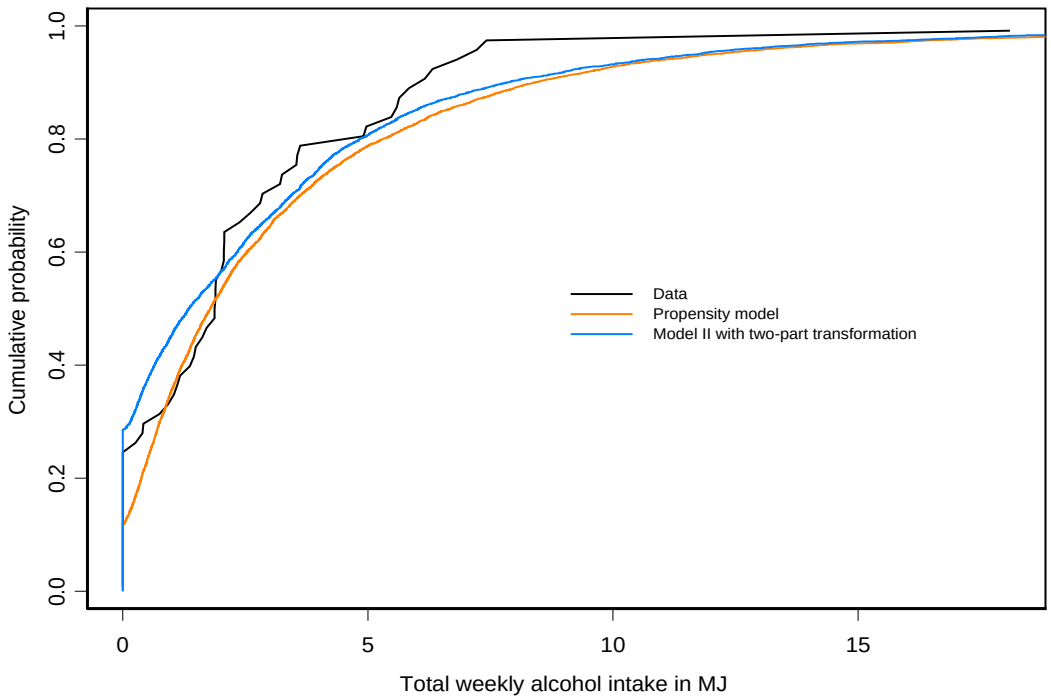

Figure 2.18: Empirical cdf with the predictive cdfs of total weekly alcohol intakes from the Propensity model and latent Gaussian model with two-part transformation for all the sex-by-age groups.

sumption might be higher during festive seasons like in the month of December. Alcohol is not a typical food product, as from the data set we can see that there are very few non-zero values less than 0.4 MJ. This is due to the nature of alcohol consumption, where if you consume alcohol you are likely have at least a glass of some alcoholic beverage which would give you an alcoholic intake of around 0.46 MJ (2 alcoholic units). Thus for our Propensity model, though the model gives good predictions, the match between the data and the model is poor for non-zero intakes less than 0.4 MJ. May be an alternate transformation to Normality will improve the fit. Results obtained using log transformation to the non-zero intakes are very similar to those obtained using the eighth root.

For the latent Gaussian model, it can be quite difficult to obtain a transformation which makes the non-zero values fit the right tail of a Normal distribution above the threshold well. The choice of the transformation can be as in our case through trial and error method and arbitrary. The fit of the latent Gaussian model to the data very much depends on finding a suitable transformation. One

parameter transformations such as the square root may not work, and one may have to look at two or three parameter transformations with unknown parameters. It is also difficult to choose a prior distribution for such a transformation.

In developing both our models we ignore any possible within-individual correlation in intakes between consecutive days. The possibility of dependence between consumption of food products on consecutive days is discussed in Chapter 4 with retinol intakes as an example.

The predicted cdf from the latent Gaussian model with the two-part transformation seems to fit the cdf for the combined data better than the Propensity model I as seen from Figure 2.17. Looking at the combined predicted probability of getting a zero, the Propensity model gives a predicted probability for a zero intake of 0.57 whereas the latent Gaussian with the two-part transformation gives 0.56. The observed proportion of zero intakes is 0.55. The observed proportion of intakes exceeding the safe limits for males and females combined is 0.16. The models predict the probability of exceedance for males and females combined as 0.18 and 0.14. For males in age group 50-64 years, the latent Gaussian model with the two-part data gives inaccurate prediction for the probability of a zero intake and also for the exceedance probabilities. The data suggest that males in the age group 50-64 have the highest percentage of drinking more than their safe limit, both models predict so correctly. Both models approximately take the same time to run in WinBUGS. We must remind ourselves that the samples within each sex-by-age group are quite small.

The data set that has been used here is quite small. Some of the prior distributions used for the models were rather vague. Obtaining more information about the variability in drinking patterns might help us introduce informative priors in our model. Also a larger data set with more individuals in each group should improve the performance of both the models.

Chapter 3

Exposure Assessment for Pesticide Intake

An important area of food safety risk assessment involves monitoring intake of pesticides through food. While studying pesticide intakes it is worth considering ingestion of the pesticide through various foods. Consumption of certain products may be correlated and we can have groups of food products which are sprayed with the same pesticide. For such cases it is important to study consumption of correlated food products simultaneously to obtain estimates of pesticide intakes.

In this chapter we present a method for exposure assessment for the intake of a pesticide through multiple food products simultaneously. We illustrate our model for daily intake of the fungicide Iprodione through five food products by combining data on consumption of these five products with Iprodione concentration data on them. We develop a multivariate latent Gaussian model for consumption of five food products which may contain Iprodione residues. We also suggests a latent Gaussian and a latent t distribution-model for the concentration data. Our model allows us to predict the probability of an individual's Iprodione intake to exceed the safe level through the consumption of five products simultaneously and individually.

3.1 Iprodione

Iprodione is a white odourless crystal, and its chemical name is *3-(3,5- dichlorophenyl)-N-(1-methyethy)-2,4,-dioxo-1-imidazolidinecarboxamide*. It is a dicarboximide fungicide used against a range of fungal diseases such as *Botrytis, Fusarium* and *Rhizoctonia* in vines, black and red currant, raspberries and vegetables such as lettuce, cabbage, cauliflower, fennel and potatoes. The compound is used as a foliar

spray on several crops, and as a post-harvest dip for fruits. Iprodione inhibits the germination of spores and the growth of fungal mat (mycelium). The product is authorised for use in France, Germany, the Netherlands and the UK, and also in Japan, USA and Canada (IPCS (1977)).

Toxicological studies on rats have shown that intake of Iprodione is associated with reductions in fertility, body weight gain and food consumption. Special studies to examine the effect of Iprodione intake on mutagenicity in rats showed no negative effects (Extenion Toxicology Network, USA (1992)). Iprodione is slightly toxic to wild-fowl and moderately toxic to fish species. It does not appear toxic to plants. It has potential to contaminate ground water. In humans, a daily intake of up to 60 micrograms of Iprodione is considered safe; this is called the acute reference dose (ARD) (United States Environmental Protection Agency (1998)).

In this chapter we look at Iprodione intake in certain fruits and salad leaves. We are interested in modelling the intake of Iprodione by an individual, and estimating his or her probability of exceeding the ARD on a given day.

Acute dietary exposure to pesticides is calculated mostly using point estimates (Boon et al. (2004)). In these estimates a single high residue concentration, for example the maximum observed concentration, is multiplied by a single high consumption level such as the 97.5^{th} percentile for each product and divided by the average consumer body weight. This provides a single value for the intake of pesticide. However these point estimates do not account for the variation in consumption patterns. Also using point estimates we can address only one product at a time.

In this chapter we provide a probabilistic approach for modelling pesticide intake from consumption and concentration data sets. We compare results using our probabilistic approach with an empirical approach in section 8.

3.2 The Consumption and Concentration Data Sets

There are two data sets used in this study. One is for the daily consumption of endive, cabbage lettuce, grape, strawberry and currant, and the other is for the concentrations of Iprodione in these five products. These products were chosen as

they have been found to have the highest concentrations of Iprodione, and hence we assume that the most ingestion of Iprodione is through these products. We henceforth refer to cabbage lettuce as just lettuce.

The consumption data set is derived from the Dutch National Food Consumption Survey (DNFCS), (Anonymous & Nederlan (1998), Kistemaker et al. (1998)). In this survey, 6250 individuals were randomly selected in the range 1-97 years and their food intakes over two consecutive days were recorded. Out of all the individuals in the study, 5756 individuals completed the study and kept a record of food consumed on both the days. The individuals had to weigh the food and record the type and amount of food consumed. With the use of the conversion model for primary agricultural products, developed at the RIKILT - Institute of Food and Safety in Netherlands, the consumptions of the products were then converted to approximately accurate amounts of raw agricultural commodities: see van Dooren et al. (1995) for details. In this study we use information on amounts of the five products consumed on two days by the 5756 individuals.

Table 3.1 is based on the consumption of the five products on 11512 days, (5756 individuals' intake on 2 days). From Table 3.1, we observe that the average consumption of grape is much higher than of the other four products, though the median is very similar for all the five products. The maximum observed intake is for grape.

Product	Mean	Median	Upper quartile	Maximum	Percentage of zero
Endive	6.37	0.00	0.00	530.02	91.73
Lettuce	3.55	0.00	0.00	300.70	89.72
Grape	14.85	0.00	4.09	1154.66	59.70
Strawberry	4.84	0.21	1.65	505.17	47.92
Currant	1.59	0.00	0.68	561.00	68.72

Table 3.1: Summary statistics for daily consumption in grams for the five products

Days on which consumption is zero for endive and lettuce are very large at about 90% of the total. On the other hand, for strawberry about 50% of the values are zero. Figure 3.1 shows histograms of the non-zero intakes over the two days for the five foods. It shows that the distributions of intake for all the five products are highly skewed. For all the products, most of the intakes are less than 100 g but endive appears to have relatively larger number of intakes greater

than 100 g.

We arbitrarily define four age groups as in Table 3.2. Table 3.2 shows the number of individuals and the average consumption for each product in each sex-by-age group. There does not appear to be any pattern among the average consumptions between the age groups. We formally test the effect of sex and age effects on the consumptions of the five products separately using Mann-Whitney and Kruskal-Wallis tests respectively. From Table 3.3 we see that the significance probabilities for sex and age effects for all the five products are less than 0.5. Consumption of none of the products appears to depend upon the sex or age of the individual, since for all the products the significance probability is quite large. We therefore do not include any sex or age effects in our model.

Sex	Age	Number of individuals	Mean				
			Endive	Lettuce	Grape	Strawberry	Currant
Male	1-20	770	5.04	3.60	14.32	5.06	1.28
Male	21-40	998	6.71	3.84	14.58	4.15	1.59
Male	41-60	831	5.28	3.52	15.43	5.33	1.97
Male	61-97	480	7.60	3.84	13.18	4.94	1.92
Female	1-20	774	7.70	3.13	15.32	4.81	1.61
Female	21-40	816	6.60	3.39	15.90	4.86	1.74
Female	41-60	749	4.81	3.61	14.37	5.52	1.38
Female	61-97	328	8.86	3.76	14.28	3.68	1.27

Table 3.2: Mean daily intakes in grams for endive, lettuce, grape, strawberry and currant according to the sex-by-age groups.

Product	Significance probability	
	Sex effect	Age effect
Endive	0.6	0.8
Lettuce	0.6	0.2
Grape	0.8	0.3
Strawberry	0.8	0.7
Currant	0.8	0.4

Table 3.3: Significance probabilities obtained using Mann-Whitney and Kruskal-Wallis tests for sex and age effect respectively.

Figure 3.2 shows box plots of daily intakes for the five food products on each of the two days. The plot has been truncated to intakes less than 100 g. It appears the distribution of intakes on the two days is similar. The small horizontal

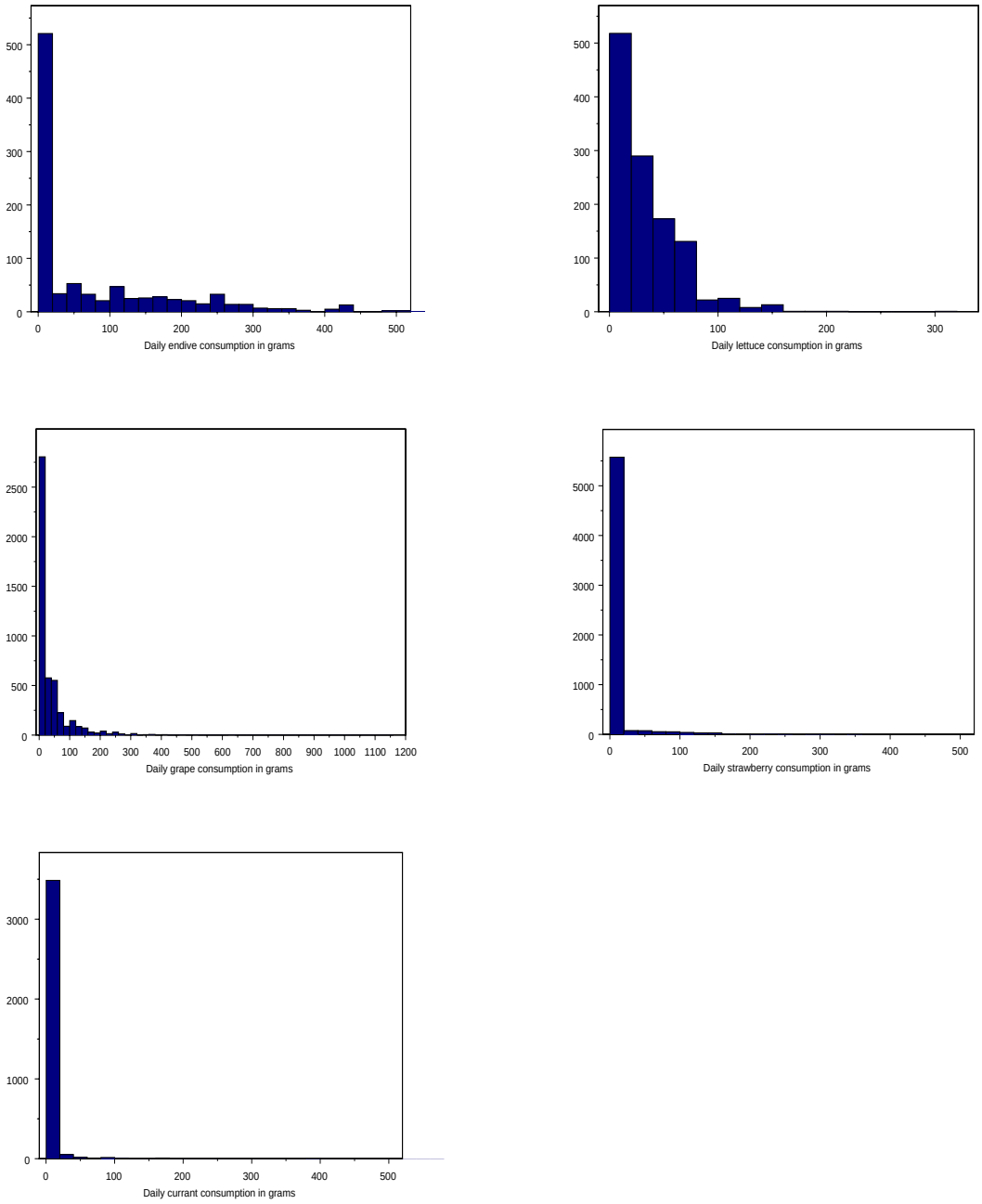

Figure 3.1: Histogram of non-zero intakes for the five products in grams

lines in the plot represent intakes which are outliers, points more than 1.5 times the inter-quantile range of the intakes.

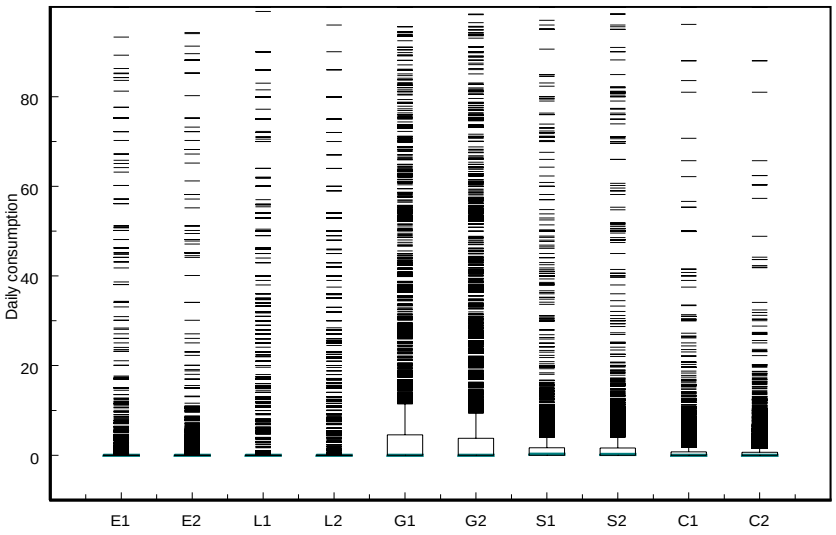

Figure 3.2: Box plot of daily intakes in grams for endive (E), lettuce (L), grape (G), strawberry (S) and currant (C) on the two observed days. E1 indicates consumption of endive on day 1, E2 consumption of endive on day 2 and so on.

Data on concentration of Iprodione in the five products are also available from a separate study. These concentrations are expressed in micrograms of pesticide per gram of commodity. The concentrations of Iprodione in the five food products were collected by the Dutch Ministry of Agriculture, Nature Management and Fisheries (LNV) through the programme for the Quality of Agricultural Products (KAP) over a period of five years, (van Klaveren (1999)). Samples of the five products were bought randomly from various shops and analysed in the laboratory to determine levels of Iprodione in them. The concentrations of Iprodione were measured and recorded for each sample. The data are stored in the KAP database and were obtained from Gerda van Donkergod of the RIKILT Institute of Food Safety. Table 3.4 shows the number of samples for each product along with the proportion of zero concentration and the average concentration of Iprodione on these products.

Product	Number of samples	Mean	Median	Maximum	Percentage with zero concentration
Endive	700	0.266	0.00	17.50	78.43
Lettuce	975	0.521	0.00	26.00	69.95
Grape	712	0.120	0.00	3.50	76.28
Strawberry	1367	0.121	0.00	6.42	80.18
Currant	131	0.610	0.00	18.00	72.52

Table 3.4: Summary statistics for concentration levels of Iprodione in micrograms per gram of the five products

The zeros in the concentration data set can be true zeros, where the pesticide is actually absent in the product, or can be a false zero where the level of pesticide residue on the product is less than the level of detection (LOD). The LOD was 0.02 $\mu g/kg$ for all the products. The maximum number of zeros (non-detects) or highest proportion is observed for strawberry, while lettuce has the least number of zeros. The average concentration level is the largest for currant. We model the concentrations of Iprodione on these products in section 5 of this chapter.

Figure 3.3 shows histograms of the non-zero concentrations of Iprodione on the five products. We see that all these distributions are also skewed. The maximum concentration is the smallest for grape and largest for lettuce.

We see that the mean Iprodione concentration is largest for currant; however until we combine this information about how often and how much currant individuals consume we are unable to comment on the intake of Iprodione. Thus for exposure assessment there is more variability and randomness associated with the data as compared to just modelling intakes of certain products such as nutrients.

Interest lies in predicting the Iprodione intake by an individual through the five products. We need to combine information on individual consumption of each product with the residue level on that product to predict the amount of pesticide residue consumed though the product. To achieve this, we model the consumption of these five products simultaneously and the concentration of Iprodione in them separately and then combine outputs from the two models. Since we have large numbers of zero values in both the consumption and concentration data sets, we work with the latent variable model similar to the one developed in Chapter 2. The models allow us to predict the consumption of the five products

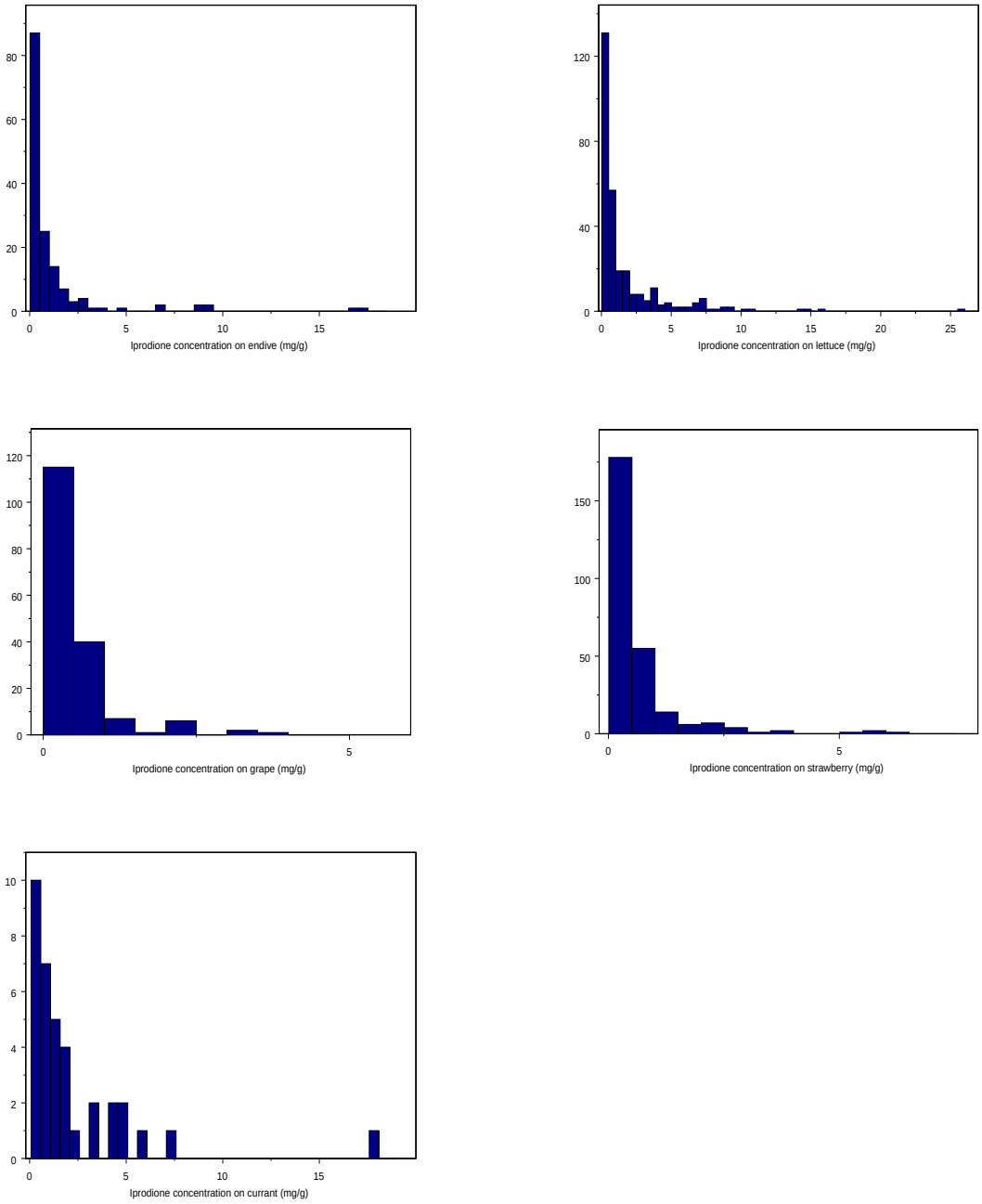

Figure 3.3: Histograms of non-zero concentrations of Iprodione on the five products in micrograms per gram of food product.

simultaneously for a random individual and also the concentration of Iprodione on each of the five products. In section 6 we discuss combining these predicted consumption and concentrations to simulate Iprodione intake for an individual.

Using our model we predict the probability of exceeding the ARD of Iprodione on a given day through the five products separately and together.

3.3 Multivariate Tobit and Latent Gaussian Models

This section gives an introduction to existing literature on multivariate Tobit and latent Gaussian models.

Multivariate Tobit models have been used in many econometric studies. This is an extension of the Tobit model described in Section 2.4. For the multivariate Tobit model the number of latent variables is more than one and the latent variables are assumed to be correlated. An example of multivariate Tobit model can be seen in Hamilton (1999), where the authors discuss a Bayesian MCMC estimation method for the model. The model has two latent variables, one for Medicare expenditure and one to model mortality and these variables are assumed to be correlated. Both variables are modelled by explanatory covariates and an error term.

Chavas & Kim (2004) develop a dynamic multivariate Tobit model to study the implications of a government price-support programme to the US dairy market. The price-support programme creates a censoring mechanism such that when the market price drops below the government-determined price floor, the market price is unobserved and is replaced by the government price. Chavas & Kim (2004) model multiple products simultaneously and use a standard maximum likelihood approach to obtain the correlation matrix between product prices.

Cornick et al. (1994) use a similar multivariate Tobit model to study household expenditure on fluid milk. The choice of this method is influenced by two main reasons: expenditures may be censored at zero and may be interdependent across milk types.

In the previous chapter we have defined an univariate latent Gaussian model to study alcohol intakes; we now extend this to a multivariate latent Gaussian

model. Instead of modelling the intake of one product, we consider the intake of several products simultaneously. We thus assume the intakes are from a multivariate Normal distribution.

The consumption data set for the intakes for all the five products contains large numbers of zeros as observed before. We could model the consumption and non-consumption of a product using a binary variable and then have some distribution for the non-zero intakes. However, since we want to model the possible correlation between the products we use the approach of Allcroft et al. (2005) of an underlying multivariate Gaussian distribution for the intakes. The idea is similar to the univariate case where intakes of each product are transformed to fit the part of the distribution above the threshold. Considering consumption of the five products together gives us a multivariate latent Gaussian model.

Allcroft et al. (2005) discuss fitting a multivariate latent Gaussian model to study daily consumption of 51 food types simultaneously. Here we assume $\mathbf{y_{if}}$ is assumed to be the i^{th} individual's intake for the f^{th} food, $i = 1, \ldots, n$ and $f = 1, \ldots, k$, and $\mathbf{z_{if}}$ the corresponding latent Gaussian variable. Then $\mathbf{z}_i$ has multivariate Normal distribution $N_f(\mathbf{0}, \mathbf{\Sigma})$ where $\mathbf{\Sigma}$ is the $k \times k$ symmetric covariance matrix for the individual mean $z_{i.}$. The authors assume that certain quadratic power functions transform consumptions for each individual food type to fit the Gaussian distribution above the thresholds. For each food type Allcroft et al. (2005) have a invertible function. They do not assume any age or sex dependence on the consumption values.

In contrast to the latent Gaussian model the multivariate Tobit model does not assume that the non-zero values fit the right tail of a Normal distribution above the chosen thresholds. For our consumption data we treat the zero intakes as censored observations. We develop a multivariate latent Gaussian model, as this approach allows us to model the consumption/non-consumption and the amount of consumption using a single variable and also take in to account the correlation between consumptions of the five products.

3.4 A Model for Consumption of Multiple Products

There appears to be an individual effect on the consumption of the five products. Also an individual having a salad may consume endive and lettuce simultane-

ously. We assume the intakes of the five products are correlated, but conditional on the individual effects the intakes are assumed to be uncorrelated between days. This also reduces the number of parameters in the model. We fit a multivariate latent Gaussian model for the consumption data set, we assume that a common transformation for all the five products ensures the positive consumptions for each product fit the right tail of a univariate Normal distribution. For simplicity we restrict ourselves to power transformations and here a square root transformation for the intakes for the five products is chosen for this purpose. We refer to these transformed intakes as responses.

We use p_{ijf} to denote the i^{th} individual's response on the j^{th} day for the f^{th} food, so that $i = 1, \ldots, n$, $j = 1, \ldots, d$ and $f = 1, \ldots, k$ for endive, lettuce, grape, strawberry and currant respectively. The corresponding latent variable is ρ_{ijf}. Here $n = 5756, d = 2$ and $k = 5$.

Since we want to model the responses for the five food products together, we have a multivariate Normal distribution for the latent variable $\boldsymbol{\rho}_{\mathbf{ij}}$ with mean vector $\boldsymbol{\mu}_i$ which is a 5-vector and variance matrix $\boldsymbol{\Sigma}$, which is 5×5 matrix. The univariate version of the latent Gaussian model is given in Section 2.6.

We have

$$p_{ijf} = \begin{cases} \rho_{ijf} & \text{if } \rho_{ijf} > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{3.1}$$

For the notional responses we have an additive model for the individual and product effects. Each element in the mean vector $\boldsymbol{\mu}_i$ is given by

$$\mu_{if} = \pi_f + \iota_i + \varepsilon_{if} \tag{3.2}$$

where π_f is the f^{th} food effect, ι_i is the i^{th} individual effect and ε_{if} is distributed independently as $N(0, \sigma_b^2)$. For simplicity we assume the between-individual variance represented by σ_b^2 in this model to be the same for all foods.

For simplicity here we have assumed that the individual and product effects on notional response are additive. We have a common variance matrix, $\boldsymbol{\Sigma}$, for the five products for all the individuals on both the days. Only the mean vector, $\boldsymbol{\mu}_i$ differs between individuals. Unlike the model for alcohol intakes in Chapter 2, here for simplicity we do not consider a varying within-individual variance term.

3.4.1 Prior Distributions for the Model

We choose our prior distributions to reflect the multi-level structure of the data set, in which there is variation in the response between individuals and between products.

The expectation of the individual effect ι_i is denoted by α and is given a Normal prior distribution with mean 0 and standard deviation of 100. Sensitivity to the prior distributions will be discussed in a later section.

The between-individual precision σ_b^{-2} is given a Gamma prior distribution $Ga(1, 50)$. Thus the lower and upper 5% quantiles for the between-individual standard deviation are 4 and 31 respectively. The food effects π_f are given the same Normal distribution, $N(0, 0.5)$.

The inverse of the variance matrix, $\mathbf{\Sigma}^{-1}$ for the five products is given a Wishart prior distribution $W_k(\mathbf{S}, t)$, where t denotes the degrees of freedom and $\mathbf{S}$ is the estimate of $t^{-1}\mathbf{\Sigma}^{-1}$. To represent our vague prior knowledge about the inverse of the variance matrix we choose the degrees of freedom to be as small as possible: here it is 5, the rank of $\mathbf{\Sigma}^{-1}$. The matrix $\mathbf{S}$ is set to be $200I_5$, where I is an identity matrix. It is worth noting here that except for cases with very few individuals, the choice of $\mathbf{S}$ has very little effect on the posterior distribution of $\mathbf{\Sigma}^{-1}$, according to Spiegelhalter et al. (Jan 2003*b*). Also increasing the values of the diagonal elements of $\mathbf{S}$ from 200 has little effects on the posterior estimates of the model parameters. However WinBUGS is unable to generate initial values for our model when the diagonal elements are less than 200.

3.4.2 Robustness to Changes in Prior Distributions

With so little information available from the data set, we have made some strong choices for the prior distribution parameters. In this section we examine the effects on the posterior distributions for our model parameters by changing the prior distribution parameters.

We increase the variance of both the food effects (π_f) and the individual effects (ι_i) to 2 and 5000 and decrease the expectation of the between individual variance from 50 to 1. These prior choices for the variance parameters for the product effects cause convergence problems. Even after 80,000 simulations the model parameters do not seem to converge as observed from the history plots.

The problem of convergence appears to be associated with increasing the variance for the food effect.

Keeping the initial prior distributions for π_f and α, we only change the prior distribution for the between-individual precision from $Ga(1, 50)$ to $Ga(1, 1)$ and $Ga(10, 1)$. The changes in the posterior expectation for the between-individual precision are less than 10%.

We next try decreasing the variance for product effect (π_f) and the individual effect from 0.5 and 10000 to 0.1 and 200 with the same prior distribution for the between-individual precision. The posterior expectations for the individual effects show a decrease by almost 0.3 times the previously observed values. The posterior expectations for the product effects all increase, for endive the posterior expectation for the product effect increases from -13.9 to -9.0 and for lettuce the increase is from -10.4 to -6.7. A larger change is observed in the posterior expectations for the remaining product effects. The posterior expectation for grape effect changes from -0.5 to 0.5, for strawberry the change is from 0.4 to 1.2 and the posterior expectation for currant effect changes from -9 to 0.9. These changes also affect the predicted percentages of zero consumption. Except for grape, these percentages slightly decrease.

The distributions for the model parameters seem to converge for variances in the range $(0.1, 1)$ for π_f. The changes in the prior distribution for α have no observable effect on the posterior distribution of the parameter. If the posterior expectations of the product effects increase then those of the individual effects decrease. The overall effect on the predicted consumption levels is small and the predicted percentages of zero are all within 5% of the observed percentages. The predicted upper percentiles are within 25% of the observed quantiles.

3.4.3 Model Adequacy

As discussed in Chapter 2, we use the posterior predictive distributions to predict consumption of each of the five products by some new individuals. These predicted daily intakes are used to estimate the probability of a zero intake for each of these five products. We use 25,000 simulated intakes for each product. In Table 3.5 we compare the observed and predicted percentages of zero intake for each product. The model appears to slightly under-predict these percentages for endive, lettuce and grape whereas it slights over-predicts the percentage of zeros

for strawberry and currant. Table 3.5 also has the observed and predicted upper 5% and 1% quantiles for the intakes. The agreement is not very good between the observed and predicted upper quantiles. For currant the model over-estimates both the quantiles, for the remaining food products if the models over-estimates one of the quantiles it under-estimates the other. The difference between the observed and predicted upper 5% quantile for strawberry is the largest.

Product	Percentage of zeros		Upper 5%		Upper 1%	
	Data	Model	Data	Model	Data	Model
Endive	92	91	5.4	10.5	225.2	180.9
Lettuce	90	89	25.0	18.1	80.0	120.7
Grape	59	57	82.6	90.6	213.1	196.5
Strawberry	48	51	10.1	29.3	122.2	89.7
Currant	69	72	5.6	13.4	20.1	37.0

Table 3.5: Comparison of observed and predicted percentages of zero consumption levels along with the upper 5 and 1 percentiles for the intake of the five products.

We also compare the cdfs of the daily intakes for the five products with the corresponding predictive cdfs in the original scale in Figure 3.4. For endive, grape and strawberry the fit is bad for intakes less than 50 grams. But since high intakes are more important for this study we conclude that the fit is good. The main reason for the poor fit may be that the common square root transformation to fit the non-zero consumptions to the right tail of a Normal distribution may not be suitable for all the products. Different transformations may be necessary for each product.

We are modelling the intakes of the five products simultaneously. Table 3.6 tells us how often each product is consumed along with the others. The first cell implies that 29% of the time all the five products were not consumed. The second value on the first line implies 1% of the time only currant is consumed. The values in the parentheses give the corresponding model predictions. The joint frequencies are expressed as percentages. A one in the left most column implies that the consumption of that product is non zero and a zero implies non-consumption.

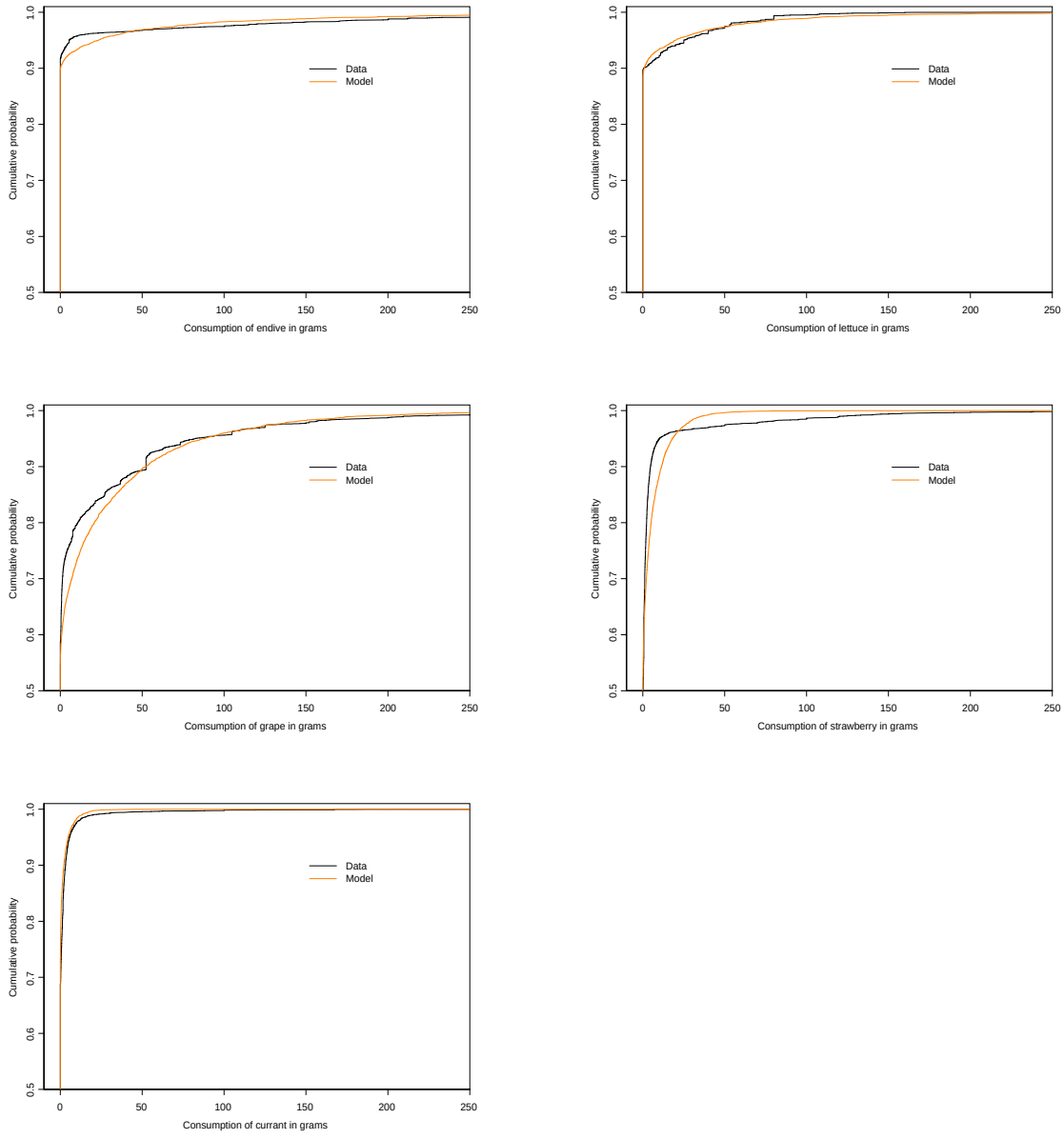

Figure 3.4: Empirical cdfs of the data and the predicted daily intakes for the five products.

(endive, lettuce, grape, strawberry, currant)	currant	
	0	1
(0,0,0,0,)	29 (21)	1 (3)
(0,1,0,0,)	2 (2)	0 (0)
(1,0,0,1,)	3 (3)	0 (1)
(1,1,1,1,)	1 (1)	0 (0)
(0,0,1,0,)	9 (11)	1 (3)
(0,1,1,0,)	1 (1)	0 (0)
(1,0,1,0,)	1 (2)	0 (0)
(1,1,1,0,)	0 (1)	0 (0)
(0,0,0,1,)	11 (15)	8 (8)
(0,1,0,1,)	1 (2)	1 (0)
(1,0,0,1,)	1 (2)	1 (1)
(1,1,0,1,)	0 (0)	0 (0)
(0,0,1,1,)	9 (10)	16 (10)
(0,1,1,1,)	1 (1)	1 (0)
(1,0,1,1,)	1 (0)	1 (1)
(1,1,1,1,)	1 (3)	0 (0)

Table 3.6: Comparison of observed and predicted (in parentheses) percentages of joint occurrences for the five products.

3.5 Model for Iprodione Concentration: Latent Variable Model

From the summary statistics in Table 3.4, we see that for strawberry, 20% of the samples have any detectable Iprodione residue whereas this value is the highest for lettuce where 30% of the samples have Iprodione residue. The histograms in Figure 3.3 show that the distributions of Iprodione concentrations for the five products are positively skewed and have long tails. The five products are not grown together and hence the residues of Iprodione on them will not be related to each other. Hence we assume the concentrations of Iprodione on the five products are uncorrelated among each other. We thus develop separate univariate models to study the concentrations. Since we have large numbers of zero concentrations, we again use the latent variable approach as used in Section 2.6. We show results obtained using an underlying Normal distribution and t-distribution. The choice of a t-distribution may be better suited for heavy tailed distributions to handle the skewness in the data.

3.5.1 Latent Gaussian Model

Initially we work with a univariate Latent Gaussian model similar to the one developed to study alcohol in Chapter 2.6. Here we do not have repeated observations per individual and hence the model for the five concentration data sets are simpler. We assume that the data in its raw form without any transformation fits the right tail of a Normal distribution above the LOD for all the five concentration data sets.

Let the observed concentrations of Iprodione be denoted by c_{if}, where $i = 1, \ldots, n$ and n is the number of samples collected for each product and $f = 1, \ldots, k$ for the k products respectively. Let ς_{if} be the associated latent variable. We have

$$c_{if} = \begin{cases} \varsigma_{if} & \text{if } \varsigma_{if} > LOD \\ 0 & \text{otherwise.} \end{cases} \tag{3.3}$$

Here ς_{if} are assumed to be from a Normal distribution with mean μ_f and variance σ_f^2.

The expectation for the Normal distribution, μ_f, is given a Normal prior with mean 0 and variance 100. The precision σ_f^{-2} is given a Gamma prior, $Ga(1, 1)$ for all the five products. Thus the lower and upper 5% quantiles for σ_f are 0.58 and 4.4. The posterior expectations for μ_f and σ_f^2 exhibit less than 1% change in the parameter values for the prior distributions. We have the same set of priors for the concentrations of Iprodione on all the five products.

Using the posterior distributions for our model parameters we simulate the predicted concentrations on the five products. In Figure 3.5 we have the empirical cdfs for the concentrations along with the predicted cdfs. The model does not perform well. Here we use the data in its raw form, a transformation to the data might be required so that the non-zero values fit the right tail of a Normal distribution better. However modelling the square root of the concentrations does not seem to improve the fit either.

3.5.2 Latent t Model

We propose an alternative model similar to the one in Section 3.5.1 but instead of assuming that they are from a Normal distribution we assume the concentrations are from a t-distribution. As before we assume the non-zero data fit the tail of a t-distribution above the LOD. We do not use any transformation on the data.

Here ς_{if} is assumed to be from a t-distribution $t(\mu_f, \tau, m)$ with parameters μ_f, τ and m degrees of freedom. Thus ς_{if} has a pdf

$$f(\varsigma_{if}) = \frac{\Gamma(\frac{m+1}{2})}{\Gamma(\frac{m}{2})}\sqrt{\frac{\tau}{m\pi}}\Big[1 + \frac{\tau}{m}(\varsigma_{if} - \mu_f)^2\Big]^{\frac{-(m+1)}{2}} \tag{3.4}$$

for $-\infty < \varsigma_{if} < \infty$. Like the latent Gaussian model, here also we have the same parameter values for the prior distributions for each concentration data set. The mean of the t distribution μ_f is given a Normal distribution prior for all the five products with the mean 0 and variance 100.

For all the five data sets we give τ a Gamma distribution prior, $Ga(1, 1)$, the distribution Gamma has been chosen after referring to an example on regression to stack-loss data in WinBUGS manual which uses a t-distribution (Spiegelhalter et al. (Jan 2003a)). Also since the Normal distribution is the limiting case for the t distribution, τ is analogous to σ_f^{-2} in the Latent Gaussian Model in Section 3.5.1. Thus both $sigma_f^{-2}$ and τ are given a prior expectation of 1. The degrees of freedom are 2 in all the models.

For all the five data sets on concentrations of Iprodione we investigate the robustness for the posterior expected values for our model parameters. For all five concentration models we increase and decrease the prior expectation of μ_f by 2. The change in the posterior expectation for μ_f for all five is less than 1% for both changes.

We also change the prior variance for μ_f and observe the change in its posterior expectation. We summarise the observations in Table 3.7. The right-most column gives the percentage change in the posterior expectation of μ_f. Less than 1% change was observed in the posterior expectation.

Increasing the degrees of freedom for the t-distribution causes the posterior expectation of μ_f to decrease and the predicted proportions of zeros to go further from the observed proportions.

Using the posterior distributions for the means of the t-distributions along with the posterior distribution of τ, we simulate the concentrations of Iprodione on the five products. Figure 3.5 compares the cdfs of the data and the predicted concentration levels. Compared to the latent Gaussian model, the latent t model appears to give a better fit to the data for all the five concentrations.

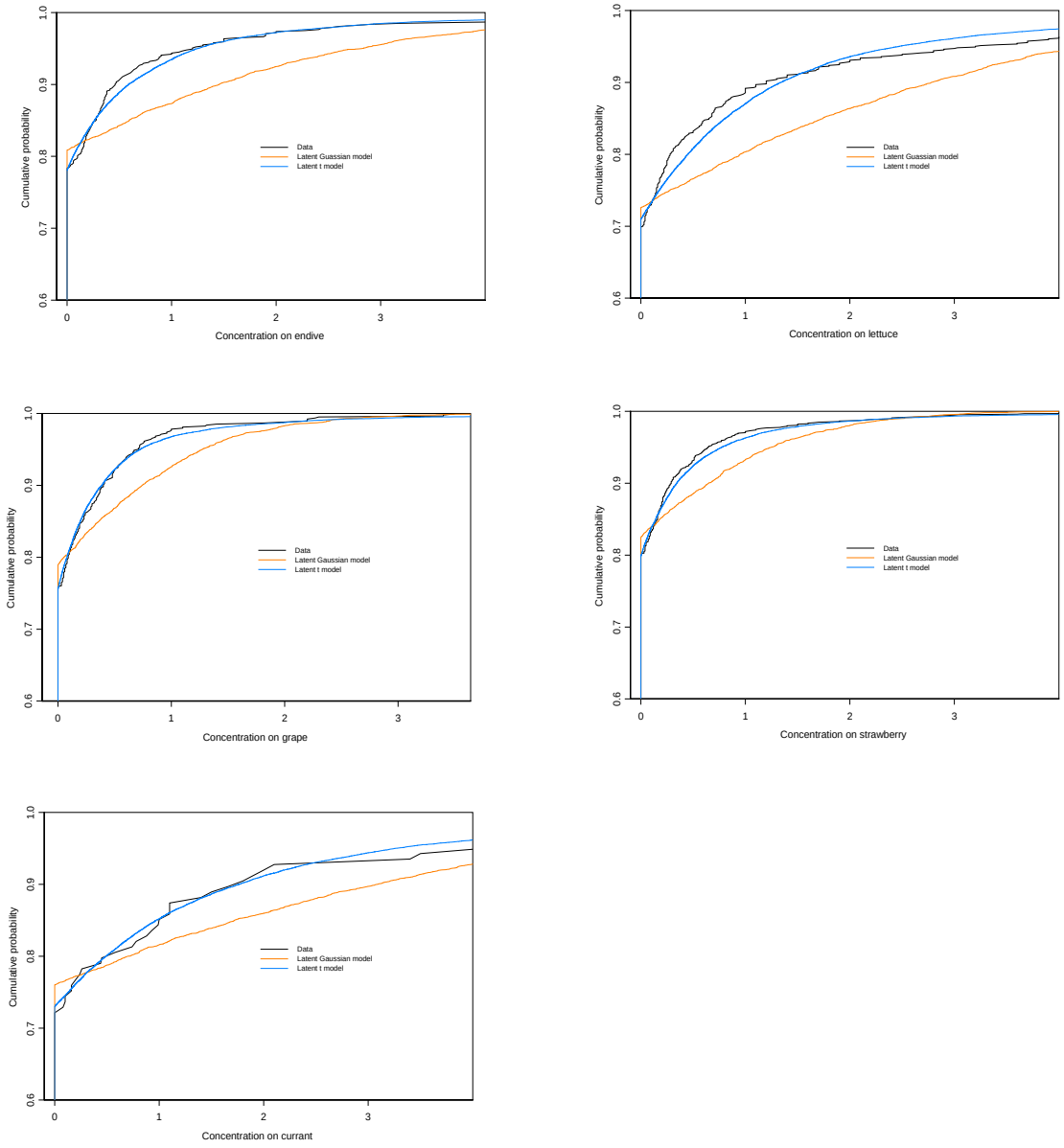

Figure 3.5: Empirical cdfs of the data and the predicted concentrations of Iprodione on the five products using the latent Gaussian and latent t models.

Product	Prior variance for μ_f	Effect on the posterior expectation (%)
Endive	1	−2
	100	+1
Lettuce	1	−1
	100	+1
Grape	1	0
	100	+1
Strawberry	1	−1
	100	0
Currant	1	0
	100	+2

Table 3.7: Changes in the posterior expectation for μ_f for various values of the prior variance of μ_f.

Table 3.8 compares the predicted percentages of zero concentrations and the upper 5% and 1% quantiles obtained from the latent t model with those observed from the data. The latent t model performs very well in predicting the proportions of zeros accurately. The model however under-predicts the upper 1% quantiles for endive and lettuce and over-predicts for currant.

Product	Percentage of zeros		Upper 5%		Upper 1%	
	Data	Latent t	Data	Latent t	Data	Latent t
Endive	78	78	1.2	1.3	6.8	4.0
Lettuce	70	71	3.2	2.5	8.4	7.0
Grape	76	76	0.7	0.7	2.2	2.2
Strawberry	80	80	0.6	0.8	2.4	2.4
Currant	73	73	3.8	3.3	6.9	9.4

Table 3.8: Comparison of observed and predicted percentages of zero concentrations along with the observed and predicted upper quantiles.

3.6 Predicting Iprodione Intake for the Five Products

Using the multivariate latent Gaussian model in Section 3.4 we have predicted intakes of the five products. The latent-t model allows us to predict the concentration of Iprodione on the same products. To predict the intake of Iprodione through a product f, we draw a random sample from the predicted daily intakes for that product and another sample of the same size from the predicted concentrations on that product and multiply these intakes and concentrations together.

Here we draw random samples of size 25,000 to predict Iprodione intakes.

We compare results obtained using our model with those from an empirical approach. In an empirical approach, the daily intakes of each products for an individual on a given day is multiplied by a randomly selected concentration for that product. By repeating this procedure many times, an empirical distribution for the Iprodione intakes is obtained. Table 3.9 shows the upper 5% and 1% quantiles of the predicted daily Iprodione intakes for the five products using our probabilistic model and from the empirical approach.

Product	Upper 25%		Upper 5%		Percentage exceeding the ARD	
	Emp	Prob	Emp	Prob	Emp	Prob
Endive	0.4	0.3	90.0	93.4	6.8	5.7
Lettuce	5.6	2.9	109.9	106.7	8.0	7.7
Grape	3.4	6.9	47.3	93.3	3.9	4.0
Strawberry	0.7	1.9	8.7	9.7	1.0	0.7
Currant	1.8	2.2	21.4	37.2	1.9	2.9

Table 3.9: Upper quantiles for the predicted Iprodione intakes in micrograms along with percentage exceedance for each product using our probability models (Prob) and an empirical approach (Emp).

As mentioned in Section 3.1, a daily intake of Iprodione of more than 60 micrograms is not considered safe. Table 3.9 gives the predicted percentage for an individual to consume more than 60 micrograms (ARD) of Iprodione through each of the products separately. These are compared with the percentages obtained from the empirical approach defined above. For our predicted intakes the samples of the predicted consumptions of all the five food products are taken at a time, thus taking into account the possible correlation between the intakes. For the empirical approach too we do the same. Figure 3.6 compares the distributions of Iprodione intakes for the five products using the empirical approach and the probabilistic approach. Both approaches seem to give similar cdf curves.

We are also interested in looking at the total Iprodione intake through these five products simultaneously. For this we sum the predicted intakes of Iprodione by each of the five products and then look at levels greater than 60 which is our ARD. The predicted percentage exceeding the safe level by our probabilistic approach is 23.3, whereas the empirical approach gives us 23.9.

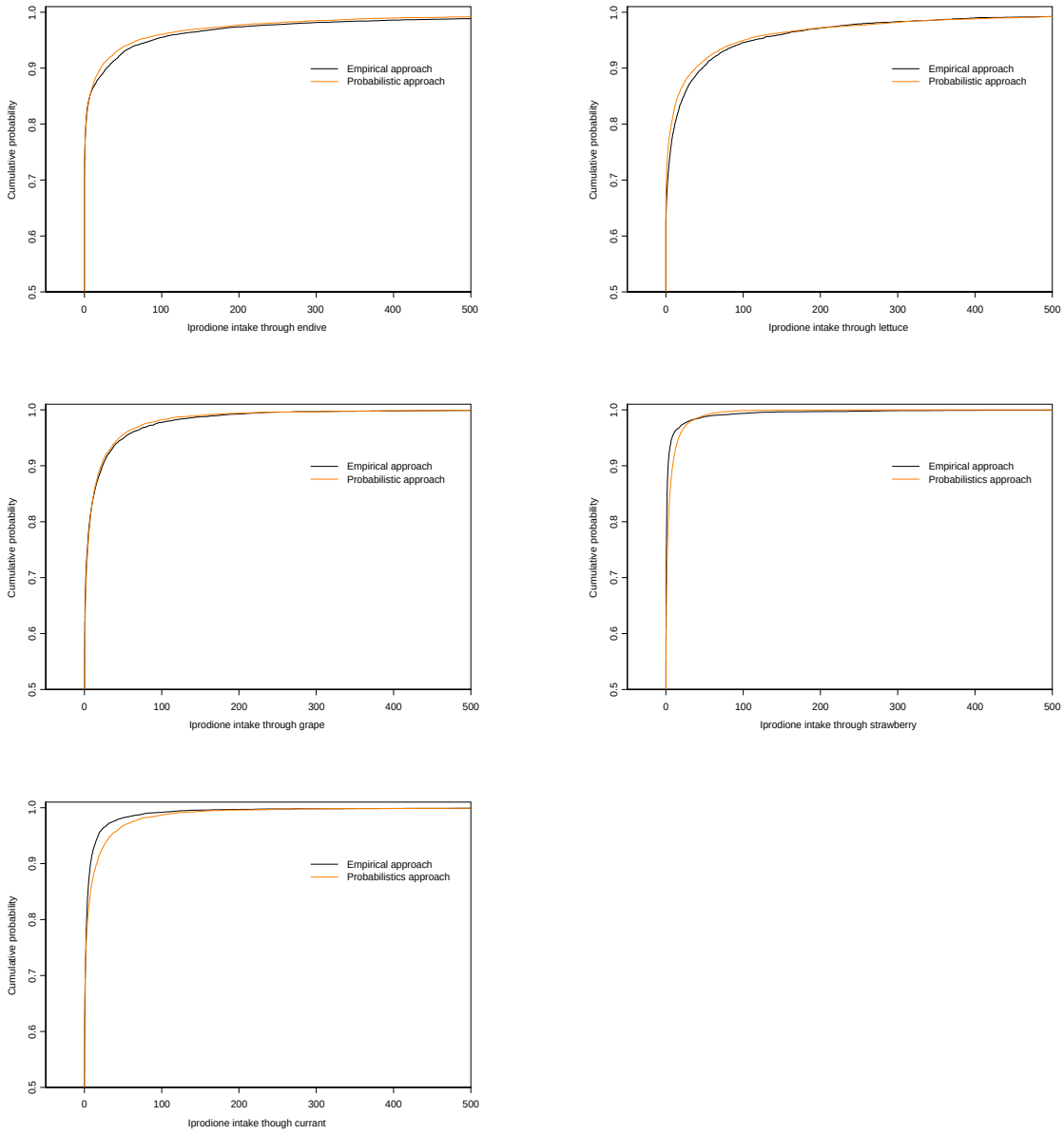

Figure 3.6: Cdfs for predicted Iprodione intake through the five products using an empirical approach and our probabilistic approach.

3.7 Discussion

In this chapter we have combined information from consumption and concentration data sets to predict daily intakes of the fungicide Iprodione from eating endive, lettuce, grape, strawberry and currant. This method of modelling the intakes simultaneously and modelling the concentrations, and then combining them allows us to take into account the various sources of variation in the data. We do not restrict ourselves to looking at only maximum possible consumption of the pesticide but we have a realistic approach for prediction of exceedance probabilities for Iprodione intakes. Probabilistic modelling provides us with a distribution over the possible pesticide intakes as opposed to a single exposure level from point estimate approach. With the probability modelling approach we are able to model the intakes of the five products simultaneously, thus taking into account any possible correlation between the consumption of the five products. Sometimes the data available are very sparse and using an empirical approach can give few values to work with.

According to the Committee on Toxicity of Chemicals in Food, Consumer Products and the Environment (2002) using probability models to describe the consumption and concentration data sets yields more robust estimates of high percentiles of the exposure distribution as compared to using empirical distribution approach.

In this chapter we are able to study Iprodione intake through multiple products. However we model intake of only one pesticide. For a cumulative risk assessment, we should study the intake of multiple pesticides on food products. The concentrations of different pesticides may be correlated and in such a case we should sample from the joint distribution of concentrations to obtain the exposure distribution. To obtain ARD for studying multiple intake of pesticides will requite a lot of chemical and toxicological studies to understand how these pesticides react among each other. Also the effect of washing and cooking on pesticide levels should be analysed. For certain products such as lettuce or endive in our data set, the level of pesticide on the inner and outer leaves will vary. Our data set does not provide information on whether individuals discard outer leaves and how this effects the level of pesticide intake.

Chapter 4

Modelling Dietary Data With Extreme Intakes

In this chapter we look at a model to account for dietary data sets with a large range of intakes which are positively skewed. The large skewness makes it difficult to obtain a suitable transformation to Normality for applying previously discussed methods for studying daily intakes. We illustrate our method on daily retinol intakes. We are interested in developing a model which can incorporate skewness in the daily intakes and predict daily and longer term intakes well. We want to find the distribution of daily retinol intakes and we also want to estimate the probability of exceeding the recommended daily intake (RDA) and the safe level of consumption of retinol in a given period. We use one week as our longer time period, but the period of study can be chosen to suit the purpose of the study.

Myles et al. (2003) have proposed a Bayesian model for dietary data recorded on successive days on the same individual, which allows each individual to have his or her own within-individual variance. This model has been described in detail in Chapter 1. The authors give results for log-transformed and untransformed data. The data set used in this chapter is the same one as used in their study and is described in Section 4.2. Our model is also Bayesian, but is more flexible, and allow for realistic extreme values to occur, improving the performance of the models. We compare results from our approach with that of Myles et al. (2003) in Section 4.8.

4.1 Retinol

Vitamin A occurs in two forms, retinol and beta carotene. Retinol is the more common form in which Vitamin A occurs in food products. Retinol is a fat-soluble vitamin and it is carried through the body and stored in the form of fat.

The chief storage tissue for retinol is the liver where almost 80% of the retinol in our bodies is stored. Vitamin A is beneficial to our bodies, but excess retinol can lead to hypervitaminosis. Hypervitaminosis is the excess accumulation of Vitamin A. Symptoms for it are bone pain and swelling of body parts, hair loss, skin irritation, vomiting, drowsiness and decreased appetite. Our study here focuses on retinol intakes, and we next look at the good and bad effects of its consumption.

Retinol is important for good vision and plays an important role in bone growth, reproduction, cell division and cell differentiation. Retinol maintains the lining of the eyes, digestive tracts, and the skin, and prevents them from breaking down and being susceptible to bacteria and viruses. Thus retinol regulates the immune system.

Retinol toxicity can occur in both the acute and the chronic form. In the acute case the symptoms are usually associated with nausea, blurry vision and headaches. In the chronic case, high levels of retinol intake cause increased levels of cholesterol in the body: this is because retinol is stored in our body in the form of fat. It also causes liver damage and hair loss. Long-term consumption of high levels of retinol can promote osteoporosis and weakening of the bones. In a BBC (2002) news bulletin there were confirmed reports indicating high levels of retinol causing hip fractures. Melhus et al. (1998) also showed the damaging effects of high levels of retinol consumption.

The recommended daily allowance (RDA) for Vitamin A is expressed in terms of 'retinol activity equivalent' (RAE), where 1 microgram (μg) RAE equals 1 μg retinol. The RDA of Vitamin A for an adult male is 900 μg RAE per day and that for an adult female is 700 μg RAE per day. Pregnant women have a RDA of 770 μg and infants less than six months have a RDA of 300 μg RAE per day. These guidelines have been provided by the US National Institutes of Health (2001) and UK Food Standards Agency (2001). The US Office of Dietary Supplements (2003) defines a Tolerable Upper Intake (TUI) level of 3000 μg RAE per day for adults of both sexes. Retinol intakes more than the TUI expose an individual to the negative effects of retinol consumption in both the acute and chronic case.

The chief sources of retinol are liver, eggs and fatty fish. Retinol can also be found in many fortified foods such as breakfast cereals. Eighty five g of beef liver has about 9000 micrograms (μg) of retinol whereas 85 g of chicken liver has about 4000 μg of retinol. A medium sized egg has only about 84 μg of retinol.

4.2 The Data Set

A weighed dietary survey among adults aged 16 to 64 living in private households in Great Britain was carried out by the Social Survey Division of the Office of Population Censuses and Surveys for the Ministry of Agriculture, Fisheries and Food (MAFF) and Department of Health and Social Security (DHSS). The data set has been obtained from the UK Data Archive (1987).

The survey was to provide information about adult's dietary habits. The data was also intended to help DHSS in assessing people who are at risk from cardiovascular diseases. In particular DHSS needed information on the working age population which could help relate dietary behaviour and intake with problems such as obesity, high blood pressure, anaemia and cholesterol levels. MAFF wanted to ensure that people ate wholesome food.

A simple random sample of addresses was selected from the Electoral Register for all U.K. constituencies. The exact sample size is not known. From the sampled addresses one individual between the ages of 16 and 64 years was selected from each household. If there were no individuals aged between 16 and 64 years then that household was not used. A letter informing the individual about the purpose of the study and the instructions on how the dietary record was to be kept was sent in advance. There were four study periods starting in October 1986, January, April and July of 1987. The total study period was seven days for each individual, during which the informant had to keep a detailed diary of what food was consumed at each meal, time of meal, description of meal and weight of the meal. The informant had also to keep a record of food consumed away from home. The starting date was not the same for all individuals and the same individuals were not used for the four study periods. In all we have information from food diaries for 2197 individuals.

Each constituency had an interviewer, and he or she had to explain in person to the informants the nature of the survey in detail. During the study period, the interviewer had to make five calls to the informant's house. The interviewer had to explain the weighing techniques and encourage the informant to complete the diary and not to miss any detail. Electronic weighing scales were provided to the informants. They were given a comprehensive list of food codes for various

types of foods. The interviewer had to make sure the food codes were entered correctly, and in case the food code was unknown the details of the food consumed had to be noted. The informant had to fill up a questionnaire and also give anthropometric measurements and blood pressure readings. He or she had to visit an appointed doctor to give a blood and urine sample. Not all informants agreed to do all the above, and such cases were then treated with slight changes which are not mentioned in the report. Where food codes were not known, nutritionists were involved to develop food codes for such food products. A pilot study was conducted in 1985 to judge the feasibility of such a study, and for the actual study an incentive payment of £10 was made on successful completion of the 7-day dietary record diary. The informants were not told about the blood and urine tests initially so as to make sure that the diet of the person remained unchanged.

A sample of food codes is given in Table 4.1. Nutritionist then used these food codes and the weight of the intake to calculate the amounts of energy, protein, vitamins, minerals, carbohydrates and fat consumed daily by each informant.

Code	Food description
R 348	Almond macaroons
691	Brie cheese any
2277	Kitkat
R 1643	Prawn and vegetarian curry; no rice
2525	Vinegar any

Table 4.1: Sample of various foods along with their codes.

For food codes marked with the prefix 'R', recipe details had to be given if these items were homemade.

4.3 Summary Statistics

For analysis we have combined all four data sets of retinol intake to study them together. Retinol intakes for 2197 individuals over 7 days are available from this data set. A histogram for all the 51379 days of retinol intake can be seen in Figure 4.1. To get a better view of the data set, Figure 4.2 gives the histogram of retinol intake for values up to 5000 μg per day.

The data set has retinol intakes from 0 to around 80,000 μg. From Figure 4.1 we do not get much information since for most classes the frequency is too small

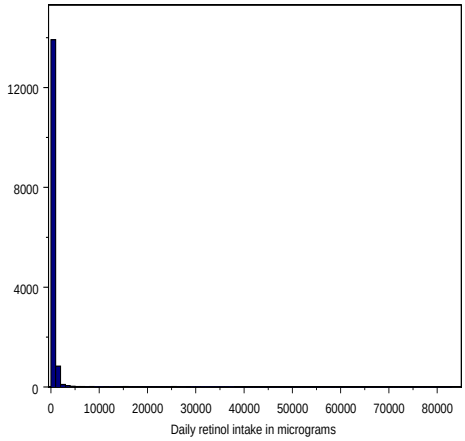

Figure 4.1: Histogram of daily retinol intakes.

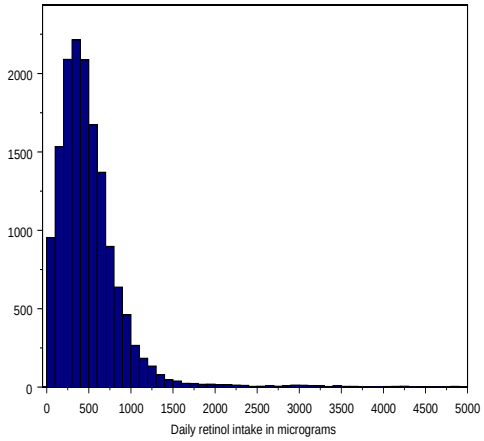

Figure 4.2: Histogram of daily retinol intakes less than 5000 μg per day.

and the histogram bars are not visible beyond 4000 μg. Figure 4.2 is restricted to values which are less than 5000, and we can see that most of the intakes are between 200 and 600 μg per day. Here the class widths are smaller as compared to the histogram in Figure 4.1, where class width is 5000 μg. Only 0.5% of the intakes are zeros. Table 4.2 gives summary statistics for the 2197 individuals over seven days for retinol intake. The mean retinol intake is around 1100 μg per day, whereas the median value is only about 440 μg.

Figure 4.3 is the histogram of the square root transformed retinol intakes. From Figure 4.3 we see that in most cases the square root transformed retinol intakes of an individual appear to be from a roughly Normal distribution on the left of the graph, and some intakes are from a second roughly Normal distribution on the right. This second distribution has a much larger variance than the first and takes into account the possible large retinol intakes. The second Normal distribution corresponds to a small minority of the intakes, has a large standard deviation and hence a flat peak and is barely visible in the histogram.

The age and sex of each individual are available. We have defined eight groups to identify the age and sex of an individual as can be seen in Table 4.2. The ages have been divided into four categories as used by Myles et al. (2003). Table 4.2 gives summary statistics for the retinol intakes based on the sex and age of each individual.

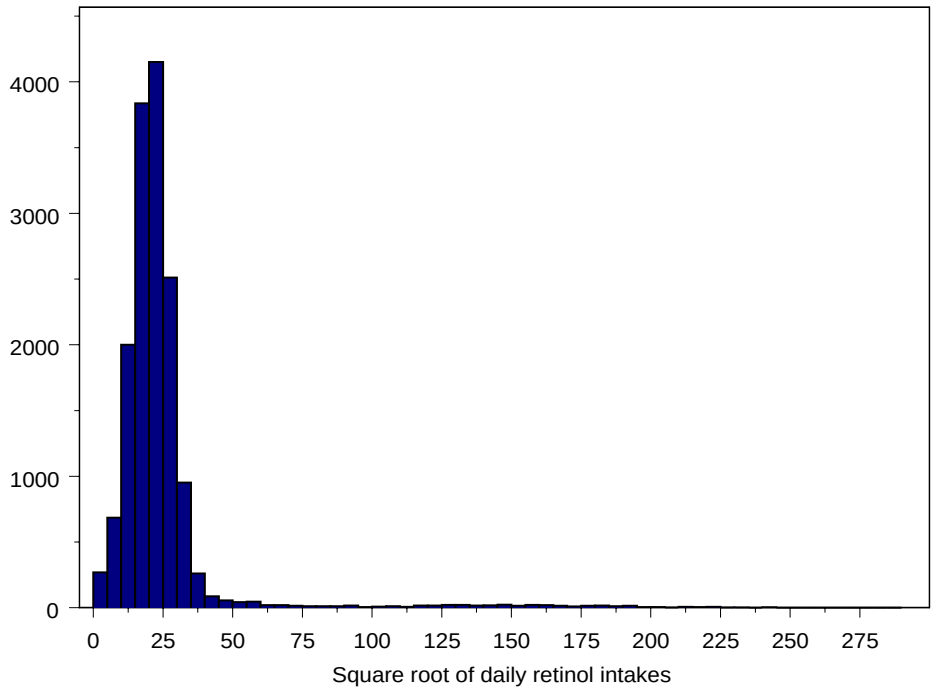

Figure 4.3: Histogram for the square roots of daily retinol intakes.

From the summary statistics it appears that males and females in the age group 50 to 64 years have the highest mean and median intakes for each sex. For males in the age group 35-49 years the upper 5% quantile is higher than that for males in the age group 50-64 years. Females aged 35-49 years have a slightly higher 1% quantile than that of females in the age group 50-64 years. This is different from the general trend of intakes increasing with age. This could mean that females in the age group 35-49 years have higher extreme intake values as compared to the other females.

4.4 Examination of the Effects of Age, Sex and Day of the Week on Retinol Intake

We want to know which effects should be included in the model and hence check if these variables have any significant effects on retinol intakes. From Figure 4.4 we see that males have a higher median retinol intake than females. This is natural as males consume more food and hence have a higher RDA for retinol. It also

Sex	Age	Number of individuals	Mean	Median	Upper quantile 5%	1%
Male	16-24	214	845	424	1275	17639
Male	25-34	254	1180	490	1591	25030
Male	35-49	346	1328	519	1894	25875
Male	50-64	273	1423	564	1839	29430
Female	16-24	188	771	326	955	16242
Female	25-34	256	916	371	1168	20969
Female	35-49	383	1141	406	1458	26365
Female	50-64	283	1263	412	2171	25696
All individuals		2197	1139	441	1457	25543

Table 4.2: Summary statistics for daily retinol consumption in micrograms according to sex-by-age groups.

appears that for both males and females the median retinol intake increases with age. The box plot has been truncated to means up to 5000 μg

To test these effects we use non-parametric tests, since t-tests require the assumption of an underlying Normal distribution, which is clearly false for our data set.

To check if there is any sex effect on retinol consumption, we use a Mann-Whitney test on the individual mean retinol intakes over the seven days. A significance probability of 0.001 indicates that the location parameter for the distribution of intakes for males is different from that for females.

We also check if the four age categories have any effect on retinol consumption. A Kruskal-Wallis test on the same means for the four age categories also gives a significance probability of 0.001 and hence we conclude that the mean retinol intake differs between the age groups.

Next we check if there is any day of the week effect. In the data set we have days numbered from one to seven, one being for Sunday. For some nutrients, we would expect the intakes to depend upon the day of the week, with perhaps consumption levels going up during the weekends. A box-plot of retinol intakes for the seven days in Figure 4.5 shows the distributions for all seven days appear to be very similar. The plot has been truncated to 5000 μg.

We formally test the day of the week effect using a Friedman test, where we

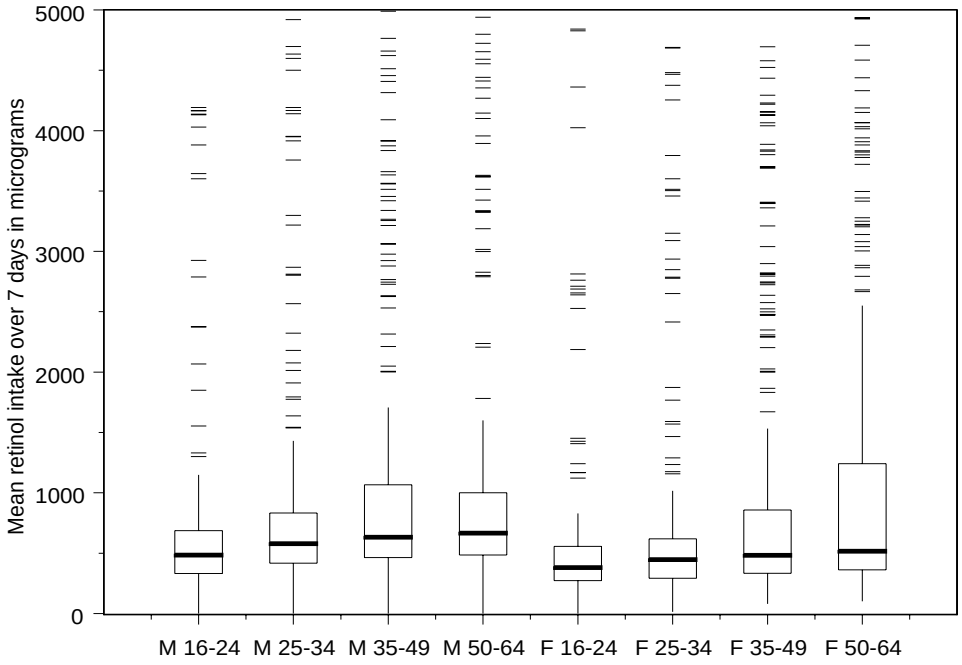

Figure 4.4: Box plot for mean retinol intakes over the seven days for the 2197 individuals grouped according to the eight sex-by-age categories. The plot has been truncated for mean intakes less than 5000 μg.

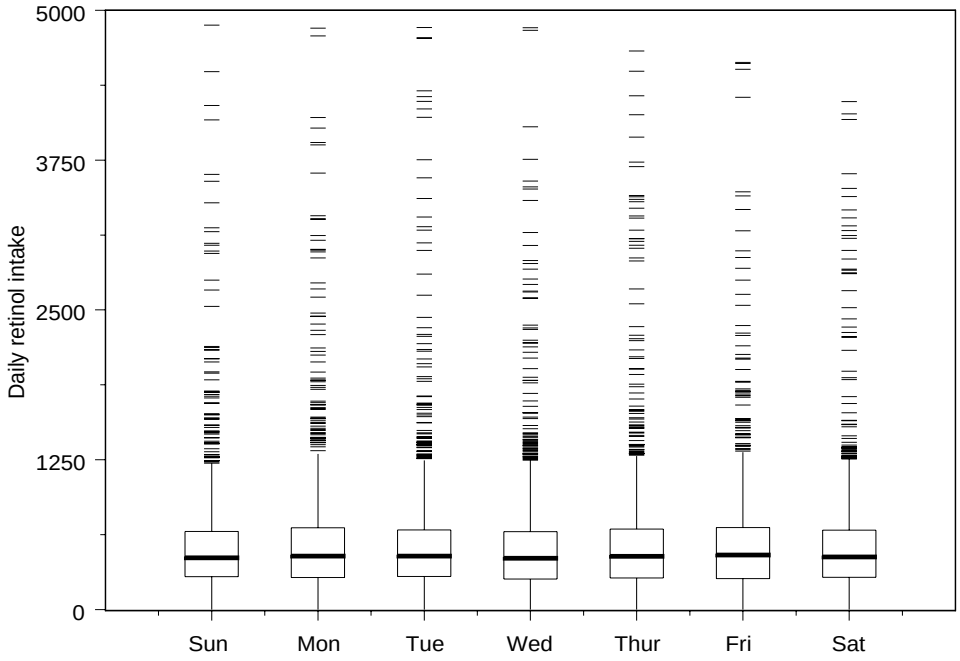

Figure 4.5: Box plot for daily retinol intake values less than 5000 for the 2197 individuals grouped according to the 7 days.

compare the daily retinol intakes for the seven days with each individual as a block and each day of the week as a treatment. We get a significance probability of 0.2. This provides little evidence that retinol intakes depend upon on the day of the week. Our model for the daily intakes therefore includes dependence on the age and sex of the individual but not on the day of the week.

We also check for any effects of study period on retinol intakes. We have data collected over four different times of the year. A Kruskal-Wallis test for the mean intake for each individual over the four different study periods gives a significance probability of 0.01. However we want a common model for all the four study periods and choose to ignore effect of study period while building our model.

For our retinol intake data set we develop a model with the following points in mind:

- Model the effect of age and sex on retinol intake. Also take into account the possible correlation among retinol intake values on consecutive days.
- Estimate probability of exceeding the RDA and also estimate the probability of exceeding the TUI of 3000 μg per day of retinol consumption on a particular day and over a week.
- Find the predictive distribution for the maximum retinol consumption for an individual over a week.

4.5 Motivation for a Mixture Model

We have investigated transformations to remove the skewness in the intake data, but it appears that a simple transformation such as a power transformation on its own is not adequate to normalize the data. It also appears that a simple distribution such as the Normal will not be able to provide a satisfactory model for variation in the observed data. It seems that for some individuals on most days the intakes are moderate, but on few days the intakes are very large and appear to be from a distribution different from the moderate intake distribution. Figure 4.3 suggests that a mixture of Normal distributions provides a plausible model for the square root transformed data. In such a model we regard the square roots of the intakes as coming from two distributions: the majority are from a 'moderate' distribution and the remainder from an 'extreme' distribution with a much higher mean and standard deviation. In this chapter we use the term 'response' for the square root transformed retinol intakes.

4.5.1 Mixture Models

Mixture distributions comprise a finite or infinite number of components, possibly of different distribution types, that can describe different features of the data. This provides a good description of even complex systems. One of the early uses of mixture modelling was by Bertillon (1863) to study heights of young men in France and then by Livi (1883) who commented on Bertillon's work. The height of men in Bertillon's study was explained by the mixing of two populations of military men, one from the plains and one from the mountains. Since there were two different distributions from which the heights arise, a mixture model was used. Pearson (1894) worked on mixtures of two Normal probability distributions and his work was one of the first major analyses involving mixture models. Use of mixture models has increased in the last 20 years due to the advance in computing facilities. Common problems in mixture models are presence of multiple maxima in the mixture likelihood function and the unboundedness of the likelihood function for the Normal components with unequal covariance matrix (McLachlan & Peel (2000)). However, better understanding of the method and the use of EM algorithm have helped in addressing these issues. These problems do not arise in Bayesian framework.

Marin et al. (2005) defines any convex combination,

$$\sum_{i=1}^{k} p_i f_i(x), \qquad \text{with} \qquad \sum_{i=1}^{k} p_i = 1 \qquad , p_i > 0 \quad (i = 1, \ldots, k) \qquad \text{and} \qquad k > 1 \tag{4.1}$$

of any probability density functions f_i as a mixture. The moments of this convex combination are the convex combinations of the moments, provided they exist of the f_i's. If X is a random variable from a mixture distribution then the m^{th} central moment can be given by

$$E(X^m) = \sum_{i=1}^{k} p_i E^{f_i}(X^m). \tag{4.2}$$

Here $E^{f_i}(X^m)$ represents the m^{th} central moment of the probability density function f_i in the mixture.

In a Bayesian framework one of the things we are interested in is the posterior expectations of our parameters. The use of MCMC methods provide a possible approach to the computation of posterior distributions and posterior predictive

distributions for mixtures of distributions.

For mixtures with components belonging to the same parametric family, the likelihood is invariant under permutations of the indices of the components. This implies that the component parameters (θ_i say) are not identifiable marginally: we cannot distinguish component 1 (or θ_1) from component 2 (or θ_2) from the likelihood because they are exchangeable. This problem is called the 'label switching problem' (Marin et al. (2005), McLachlan & Peel (2000)). By imposing an identifiability constraint on the components, the parameter identifiability problem is solved and in most cases the mixture model is then identifiable as described by Titterington et al. (1985). For a Normal mixture model with two components this can be done by creating the two-component mixture as

$$pN(\mu, \sigma^2) + (1-p)N(\mu + \sigma\varepsilon, \sigma^2\varpi^2), \tag{4.3}$$

where ε and ϖ can be given appropriate prior distributions.This ensures that the mean and the variance of one of the two Normal distributions are larger than the other one.

We develop a mixture model of Normal distributions for our retinol responses in the next section.

4.6 Model I: Bayesian Mixture Model

In the present study we have seven consecutive responses per individual. Rather than assume that each individual's seven responses are all from the moderate or extreme distribution, we allow these responses to come from one of the two populations on each day.

Thus the responses for each individual come from a mixture of two Normal distributions, one of two distributions is labelled 'moderate' and the other 'extreme'. The expectations of these two Normal distributions in the mixture depend upon the sex and age of that individual. There is a pair of Normal distributions for each sex-by-age group. Let ν_{ij} denote the response for the i^{th} individual on the j^{th} day with $i = 1, \ldots, 2197$ and $j = 1, \ldots, 7$. Let $g(i)$ denote the sex-by-age group to which the i^{th} individual belongs. For convenience we drop the subscript i from $g(i)$ and we denote these groups by just the subscript g with $g = 1, \ldots, 8$.

Since the sex and age of an individual appear to have significant effects on

retinol response, the means of the Normal distributions for each individual are allowed to depend on the individual's sex-by-age group. Thus for each of the eight sex-by-age groups we have different mean values μ_{1g} and μ_{2g} for the moderate and extreme Normal distributions respectively. The standard deviation for the moderate Normal distributions are taken to have a common value σ_1 for all sex-by-age groups and similarly for the standard deviation for the extreme Normal distribution (σ_2).

We assume that ν_{ij} comes from a mixture of two Normal distributions,

$$\pi_i N(\mu_{1g}, \sigma_1^2) + (1 - \pi_i) N(\mu_{2g}, \sigma_2^2) \tag{4.4}$$

The extra information available from having multiple observations per individual allows us to assume that the probability π_i of a moderate response for individual i differs between individuals.

On a particular day of the study, the individual i who is a member of the sex-by-age group g has a moderate response from $N(\mu_{1g}, \sigma_1^2)$ with probability π_i. Otherwise with probability $1 - \pi_i$ the individual has an extreme response, which we assume to be from $N(\mu_{2g}, \sigma_2^2)$. The choice between moderate and extreme response is assumed to be independent between days for each individual. Using Equation (4.2), the expected mean response for the i^{th} individual belonging to the g^{th} sex-age group is $\pi_i \mu_{1g} + (1 - \pi_i)\mu_{2g}$. Thus individuals have different expected intakes and different variances even within the same sex-by-age group.

The individual mixing probabilities π_i are regarded as random effects. Random effects are most conveniently taken as Normal and hence are defined on the interval $(-\infty,\infty)$, but the π_i must be in the interval (0,1). Hence we define individual effects κ_i which are Normal, and use a logit transformation of these effects to give random effects on $(0, 1)$. We have

$$\pi_i = \frac{e^{\kappa_i}}{1 + e^{\kappa_i}} \qquad (i = 1, \ldots, 2197). \tag{4.5}$$

4.6.1 Prior Distributions for Model Parameters

Our choice of prior distribution reflects the hierarchical or multi-level structure of the data set, in which there is variation in the response over the seven days within each individual, variation between individuals within sex-by-age groups and variation between these groups. Thus the daily intakes for individual i are

assumed to come from a common mixture distribution with parameters π_i, μ_{1g}, μ_{2g}, σ_1 and σ_2. Then the π_i are taken to arise from a common distribution defined by a small number of parameters, and the pairs of expectations (μ_{1g}, μ_{2g}) are also assumed to follow a common parametric distribution. In an alternative, more complex, hierarchical model we might allow the distribution of the π_i to depend on parameters at the sex-by-age level. Parameters, such as σ_1 and σ_2, whose distributions are not defined in terms of hyperparameters have to be given a prior distribution directly.

The choice of prior distributions for our mixture model has to reflect the assumption that the extreme Normal distribution for each sex-by-age group has larger expectation and variance than for the corresponding moderate distribution. We assume that the expected responses μ_{1g} for the moderate Normal distributions come independently from a Normal distribution $N(\omega, 50)$: the unknown value ω is itself given a Normal prior distribution $N(20, 100)$. To give the expectation μ_{2g} for the extreme distribution a larger value than that of the moderate distribution, we let $\mu_{2g} = \mu_{1g} + \varepsilon_g$, where the ε_g are assigned a common Normal distribution with a large positive expectation and variance: we take the common distribution for ε_g to be $N(200, 10000)$. Thus the lower and upper 5% quantiles for μ_{1g} and μ_{2g} are (0,40) and (55,386) respectively. The constraints on the means and variances of the Normal distributions helps solve the label-switching problem.

We assign σ_1^{-2}, the reciprocal of the variance for the moderate Normal distributions, the distribution $Ga(1, 100)$. This gives σ_1 lower and upper 5% quantiles of 5 and 45 respectively. To ensure that the variance (σ_2^2) of the extreme Normal distributions is larger than σ_1^2, we give the ratio σ_1^2/σ_2^2 a Beta prior distribution, $Beta(1.1, 10)$, so that the lower and upper 5% quantiles of σ_2 are 16 and 256.

For the individual probabilities π_i of a moderate response in equation (4.5), the κ_i are assumed to have a Normal distribution $N(\vartheta, \varphi^2)$. We give ϑ a Normal prior distribution $N(1, 50)$, and φ^{-2} a Gamma prior distribution, $G(1, 100)$. These choices give an expected prior probability of a response from the moderate distribution of about 0.9.

We fit this Bayesian mixture model using WinBUGS and obtain the posterior distributions of the parameters defining the components of the Normal mixtures and also the individual probabilities.

4.6.2 Results for Mixture Model I

The posterior expectations for the parameters of the moderate and extreme Normal distributions for each sex-by-age group along with their standard errors (SE) are given in Table 4.3. These results have been obtained using 50000 iterations. The values obtained are for the square root transformed retinol intakes. We note first that, because the moderate distributions predominate, their expectations are similar to the square roots of the medians in Table 4.2. They therefore follow the same pattern – higher for males, and increasing with age. For the extreme Normal distribution, however, females have higher posterior expectations than males except in the age group 50-64 years. A possible explanation for these higher expectations is that females are more likely to take Vitamin A supplements (Expert Group on Vitamins and Minerals (2002)), giving rise to high intake values. The posterior expected standard deviation for the moderate Normal distributions is 6.68 with a standard error of 0.05, whereas the extreme Normal distribution has a much larger value of 62.36 with a standard error of 1.57.

Sex	Age	Moderate		Extreme	
		Expectation	SE	Expectation	SE
Male	16-24	20.4	0.19	62.5	8.39
Male	25-34	21.7	0.17	85.1	7.02
Male	35-49	22.6	0.14	88.8	5.54
Male	50-64	23.1	0.16	109.1	7.16
Female	16-24	17.8	0.19	91.7	10.95
Female	25-34	18.9	0.17	94.4	8.77
Female	35-49	19.7	0.13	107.4	6.14
Female	50-64	20.0	0.16	108.5	6.37

Table 4.3: Posterior moments of the expectations of the moderate and extreme Normal distributions for square root of daily retinol intake according to sex-by-age groups.

For each individual we have a posterior expectation for the mixing probability π_i, which is the probability of the individual's intake being moderate on any day. Figure 4.6 is a histogram of these posterior expected probabilities. It shows that the values are mostly close to 0.97, giving an overall 97% probability of having a moderate intake. The minimum and maximum posterior expected probability of having a moderate intake are 0.34 and 0.98 respectively.

In Figure 4.7 we have the cdfs of the expected posterior probability for the eight sex-by-age groups. It appears that females aged 16-24 have the largest prob-

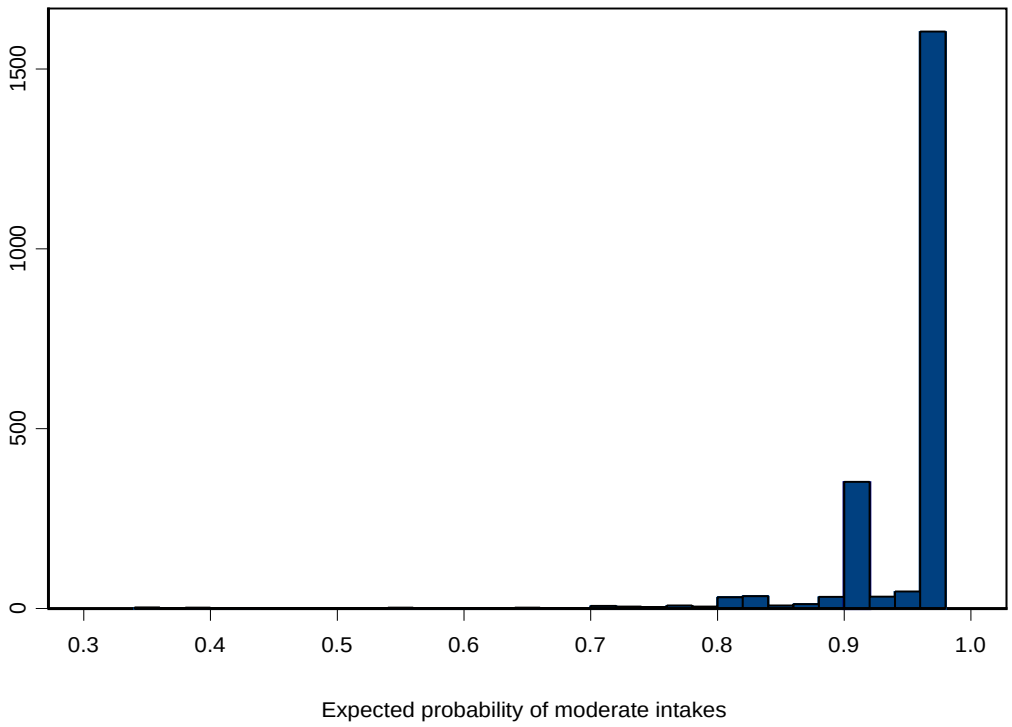

Figure 4.6: Histogram of posterior expected probability of moderate intake for 2197 individuals for Mixture model.

ability of a moderate intake.

4.6.3 Sensitivity to the Choice of Prior Distributions

We examine the effects on the posterior estimates of changing the parameter values of the prior distributions. The prior expectations for ω and ε_g are increased and decreased by 50% of the previously stated value. The prior expectation for σ_1 is also increases and decreases by 50% by changing the value of α and λ. The value of the ratio of the variances is also increased an decreased by changing the parameters of the Beta distribution. These changes in the prior distributions do not cause any substantial changes in the posterior expectations of the parameters. There is about 1% change in the posterior expected means for the moderate Normal distribution for both increase and decrease in ω and ε_g, whereas there is less than 5% change in the expected means for the extreme Normal distributions. For the variances, an increase in the prior expectation of σ_1 resulted in its posterior expectation to increase by 5% and for σ_2 a increase by

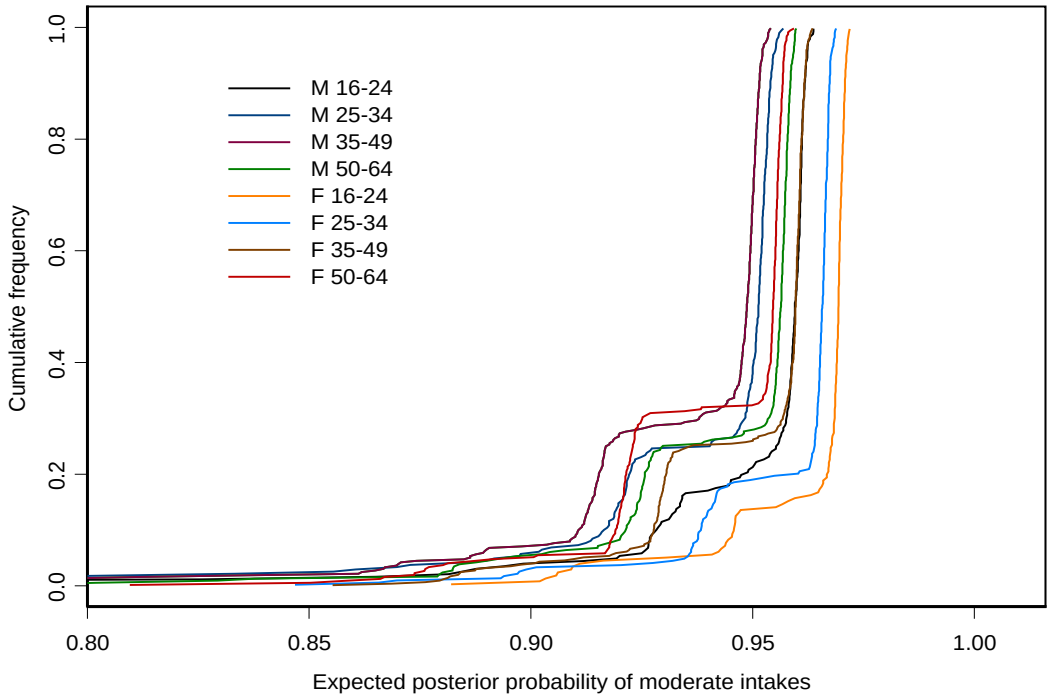

Figure 4.7: Cdfs of the expected posterior probability of a moderate intake according to sex-by-age groups.

less than 10% is observed. A decrease in the prior distribution parameters effects less than 5% change in the posterior expected values.

4.7 Model II: Extension to Mixture Model I with Markov Dependence Between Days

In the previous model we assumed, given π_i, that an individual i's response on a given day is independent of the response on the previous day. This might not be true for dietary data: there might be positive or negative dependence between consumption on successive days. This could occur because some individuals may want to vary their diets between successive days and others may want to use up unfinished dishes. The transition of intakes between the moderate and extreme Normal distributions might not be independent in such a case. This can be taken into consideration by allowing transitions between moderate and extreme components of a mixture to be governed by a first-order Markov chain.

We assume that for individual i, belonging to the g^{th} group, and for day j we have a latent variable s_{ij} which takes the value 1 if the response is moderate, i.e. it is from $N(\mu_{1g}, \sigma_1^2)$, and 2 if the response is extreme, i.e. from $N(\mu_{2g}, \sigma_2^2)$. We also assume, for each i, that s_{ij} follow a Markov chain over two states 1 and 2. The transition probabilities ρ_{iab} denote the probability of the i^{th} individual's response going from state a on day $j-1$ to state b on day j, $(j = 2, \ldots, 7;\ \ a, b = 1, 2)$. The transition matrix (P_i) for individual i is given below: see Kijima (1997) for further details on two-state Markov chains.

$$P_i = \begin{bmatrix} \rho_{i11} & \rho_{i12} \\ \rho_{i21} & \rho_{i22} \end{bmatrix}$$

Here $\rho_{iab} \neq 0$. Our two-state Markov chain is irreducible (every state is accessible from every other state), aperiodic (there exists at least one state for which transition from that state to itself is possible) and positive recurrent (expected return time is finite for every state) and hence has a unique stationary distribution. Further $\rho_{ia1} + \rho_{ia2} = 1$. The transition probabilities determine the stationary probabilities (π_{i1}, π_{i2}) as below,

$$(\pi_{i1}, \pi_{i2}) = \left(\frac{p_{i21}}{p_{i21} + p_{i12}}, \frac{p_{i12}}{p_{i21} + p_{i12}} \right) \tag{4.6}$$

$$\tag{4.7}$$

The stationary probabilities give the long term probability of an individual's response being either moderate or extreme, assuming that the consumption pattern for that individual remains constant within a sex-by-age group. Thus π_{i1}, which is the long term probability of having a moderate intake, is equivalent to π_i in model I.

Similar to Mixture model I, the means of the Normal distributions for an individual's response depend upon his or her sex-by-age group.

For the first day of observation, we allow the Normal distribution from which the response comes to be determined by the stationary probabilities of that individual. Thus

$$Pr(s_{i1} = a) = \pi_{ia} \qquad (a = 1, 2) \tag{4.8}$$

Thus given $s_{i1} = a, \quad \nu_{i1} \sim N(\mu_{ag}, \sigma_a^2)$.

From day two onwards we use the transition probabilities to determine the distribution of the response given its distribution on the previous day. The joint

distribution of $s_{i1}, \ldots, s_{i7}$ is defined by the marginal distribution of s_{i1} and the conditional distributions of s_{ij} given s_{ij-1} for $j = 2, \ldots, 7$ and we have

$$Pr(s_{ij} = b | s_{ij-1} = a) = \rho_{iab} \qquad (a, b = 1, 2). \tag{4.9}$$

Then the corresponding response on day j is given by $\nu_{ij} \sim N(\mu_{bg}, \sigma_b^2)$.

We have a logit transformation on the transition probabilities defined by

$$\rho_{ia1} = \frac{e^{\kappa_{ia}}}{1 + e^{\kappa_{ia}}}. \tag{4.10}$$

Though the transition or stationary probabilities do not depend upon the sex-by-age group of an individual, the different means for the groups and the varying stationary probabilities between individuals allow the individuals within the same sex-by-age group to have different expected intakes.

Thus using a hidden Markov model which assumes a Markov dependence between the latent variables, we allow the response on a given day to depend upon the response on the previous day.

4.7.1 Priors Distributions for Model Parameters

The prior distributions used in this extended Mixture model are the same ones as those used in Mixture model I except for κ_{ia}. We give κ_{i1} and κ_{i2} in equation (4.10) a common parametric distribution. The κ_{ia} are $N(\vartheta_a, \varphi_a)$ independently, and similar to model I ϑ_a is $N(1, 50)$ and φ_a^{-2} has a Gamma distribution, $G(1, 100)$.

4.7.2 Results for Mixture Model with Markov Dependence Between Days

We run our model for 55,000 simulations and discard the first 5000 as burn-ins.

In Table 4.4 we have the expected posterior means of the two Normal distributions for the eight groups along with the their standard errors (SE). As before, the results are for the square root transformed data set. The expectations for the moderate and extreme Normal distributions are very similar to the ones obtained for Mixture model I. The expectations of the extreme Normal distributions differ slightly more between the two models as compared to the moderate Normal distributions. Extreme expectations for males are slightly lower in Table 4.4, while those for females are slightly higher as compared to the ones in Table 4.3.

Sex	Age	Moderate		Extreme	
		Expectation	SE	Expectation	SE
Male	16-24	20.5	0.19	61.1	8.22
Male	25-34	21.7	0.17	83.6	6.61
Male	35-49	22.6	0.14	88.4	5.31
Male	50-64	23.1	0.16	108.9	7.23
Female	16-24	17.8	0.20	93.3	10.93
Female	25-34	18.9	0.16	95.6	8.94
Female	35-49	19.7	0.14	107.9	6.50
Female	50-64	20.0	0.16	109.1	6.18

Table 4.4: Posterior expectations of the moderate and extreme Normal distributions for square root of daily retinol intake for the sex-by-age groups from Mixture model II with Markov dependence between days.

We have also plotted the posterior expected stationary probabilities (π_{i1}) of having moderate intakes in Figure 4.8. The histogram shows that most values are close to 0.95, which is the average probability of having a moderate intake. The maximum and minimum posterior expectation for π_{i1} are 0.59 and 0.97 respectively.

Figure 4.9 is the scatter plot of the expected probability of having a extreme intake on day j given that the intake on day $j-1$ was moderate versus the expected probability of having extreme intakes on day j given that the intake on day $j-1$ was also extreme. We see that there are a small number of males with high probability of extreme intakes on consecutive days. Thus it appears that more males do not vary their diet as compared to females.No other difference in intake patterns between males and females are apparent from the plot. It also appears that most individuals have a smaller probability of having an extreme intake on a day given they had a moderate intake on the previous day as compared to having extreme intakes on consecutive days.

4.7.3 Sensitivity to the Choice of Prior Distributions

As for model I, we change the values of the parameters in the prior distributions and observe the change in the posterior estimates for those parameters.

The prior expectations for ω and ε_g are increased and decreased by 50% of the previously stated value. We change ω to N(30,100) and N(10,100) and simultaneously we change ε_g to N(150,1000) and N(50,1000) respectively. We change the

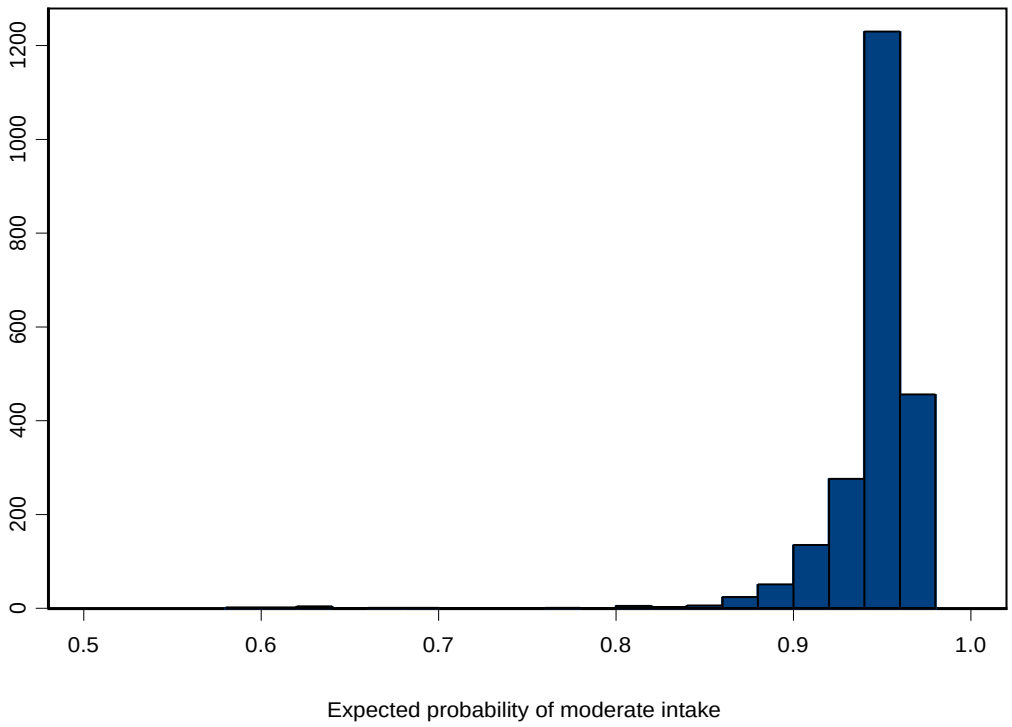

Figure 4.8: Histogram of posterior expected stationary probability of moderate intakes for 2197 individual from model II

prior expectation for the precision of σ_1 and the prior expectation for the precision for σ_2 by 50% in both directions. These changes in the prior distributions do not cause any substantial changes in the posterior expectations of these parameters. There is less than 1% change in the posterior expected means for the moderate Normal distributions for both increase and decrease situations. For the posterior expected means of the extreme Normal distributions, a change of about 2% is observed in these values. An increase (decrease) in the prior expectation for σ_1 and σ_2 causes the posterior expectations for these parameters increase (decrease) by about 5%.

Increasing the expectations of the prior distributions for the hyper parameters for the transition probabilities by 50% causes less than 2% change in their posterior expectations.

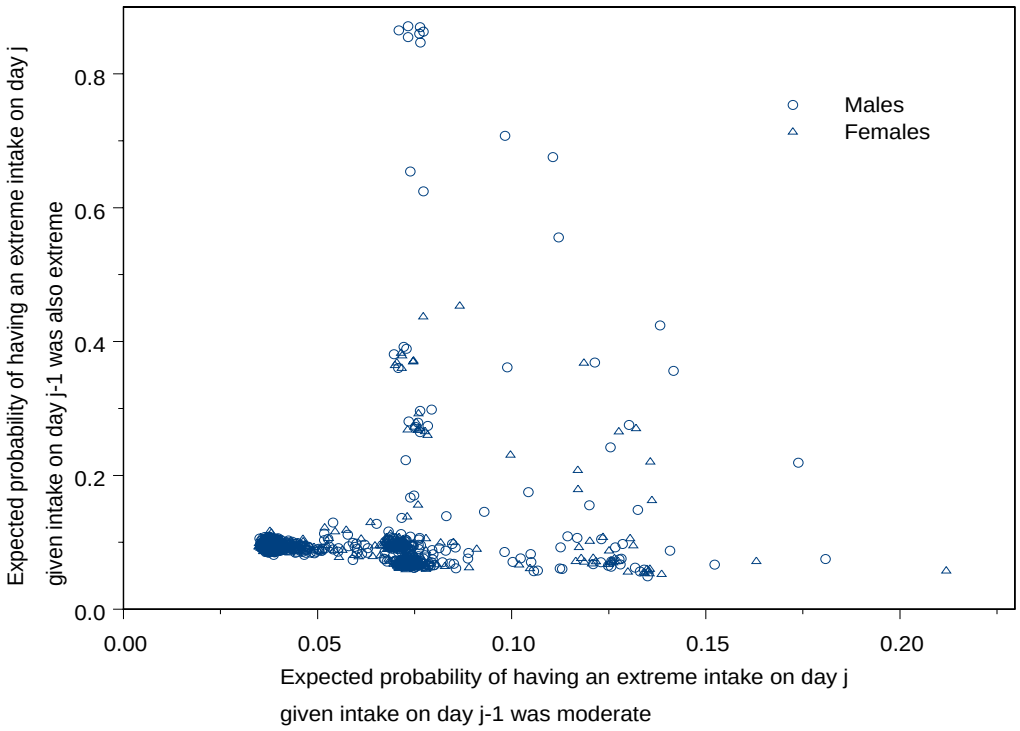

Figure 4.9: Scatter plot of expected probability for retinol intake going from moderate to extreme on consecutive days (ρ_{i12}) against expected probability of retinol intake going from extreme to extreme on consecutive days (ρ_{i22}) for males and females.

4.8 Predictions and Model Adequacy

In this section we compare performance of both the Mixture models proposed in this chapter with the observed data. We also give results from the model proposed by Myles et al. (2003) for untransformed and log transformed intakes.

Using the joint posterior distribution of the parameters defining the components of the Normal mixtures and the mixing or transition probabilities, we can simulate the daily intake of retinol for another individual in each sex-by-age group. We give predictive distributions of daily intake for both our Mixture models. In particular we can estimate for an individual in each sex-by-age group the probabilities of exceeding the RDA, the TUI level of 3000 μg of retinol per day and the high level of 10000 μg. The model can be extended to look at consumption patterns of retinol over several days. For illustration we simulate the maximum retinol intake over seven days and estimate the probability that the maximum

daily retinol intake in a week exceeds 3000 μg and 10000 μg.

The models ability to predict intakes over one or more days is more important for our study than inferences about model parameters. In Figure 4.10 we compare the cdf of the data with the predictive cdfs from both the Mixture models in the original scale. We combine the predictive intakes for each group in order to the cdf.

Figure 4.10 also shows the predictive cdf of the simulated intakes from applying the model with varying within-individual variance proposed by Myles et al. (2003) to both the intakes and their logarithms. Myles et al. (2003) uses a logarithmic transformation on the data to achieve normality. Negative simulated values from the model applied to the raw data are set to zero giving rise to the vertical part of the cdf curve in Figure 4.10. The cdf curves for the two Mixture models give a close fit to the data, and the lines from the predictions largely obscure that for the data. The cdf curves for the model of Myles et al. (2003) in both cases give fits substantially worse than those of the Mixture models. The plot has been truncated to 5000 μg for clarity.

We now look at the cdf comparison plots for each of the eight sex-by-age group. In Figure 4.11 we have eight plots comparing the cdf of the data with the predictive cdfs from the Mixture models and also those of Myles et al. (2003) model with untransformed and log transformed intakes. For all the plots, the maximum of the horizontal axis is set to 5000 μg and the vertical axis is between 0.7 and 1.0. This magnifies the top part of the curve and gives a clearer picture for comparison. For all the sex-by-age groups, Myles et al. (2003) model with the raw intakes gives the poorest fit. There is not much difference between the fit of the two mixture models; for most groups the two cdf curves from the Mixture models are close to each other and to the data, though for females aged 50-64 years Mixture model II gives a better fit. For females between 16 and 24 years, Myles et al. (2003) model with log transformed intakes appears to give a good fit. For all the groups, both versions of the models of Myles et al. (2003), the cdfs seem to go towards 1 on the y-axis in the plots sooner than the Mixture models and the data. This means that both the models of Myles et al. (2003) are estimating too few large intakes.

The maximum intake over the seven days of study could be of interest when considering acute retinol toxicity. In Figure 4.12 we compare the cdf of the ob-

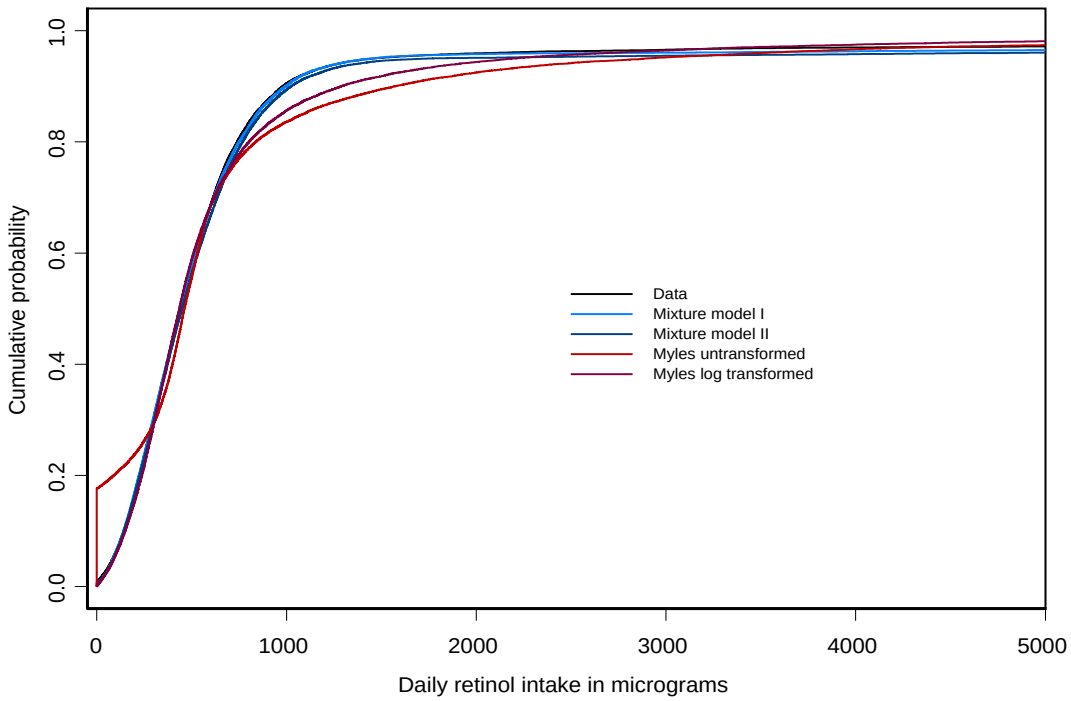

Figure 4.10: Comparison of empirical cdfs of observed daily retinol intakes with cdf of predicted intakes from both the Mixture models and Myles et al's model with raw and log transformed data for all sex-by-age groups combined.

served individual maxima with those of the predicted maximum retinol intake over seven days from the mixture models and that of Myles et al. (2003) applied to the raw data and logarithms. Our Mixture models again give better fits than those of Myles et al. (2003), although the Mixture models are less good in this case than for the daily intakes. Mixture model I appears closest to the data.

Using the predicted daily retinol intakes in each sex-by-age group we can calculate probabilities of an intake exceeding certain thresholds: we call these probabilities exceedance probabilities. Table 4.5 shows the proportions in each sex-by-age group exceeding the appropriate RDA along with the predicted exceedance probabilities from our mixture models and Myles et al. (2003) model applied to logarithms of the intakes. We remind ourselves that the RDA for males is 900 μg and for females it is 700 μg per day. For Myles et al. we give the probabilities only for the log intakes since these predictions appear to be better than those with the raw data. The exceedance probabilities calculated using the Mixture models are quite close to the observed proportions from the data. Myles

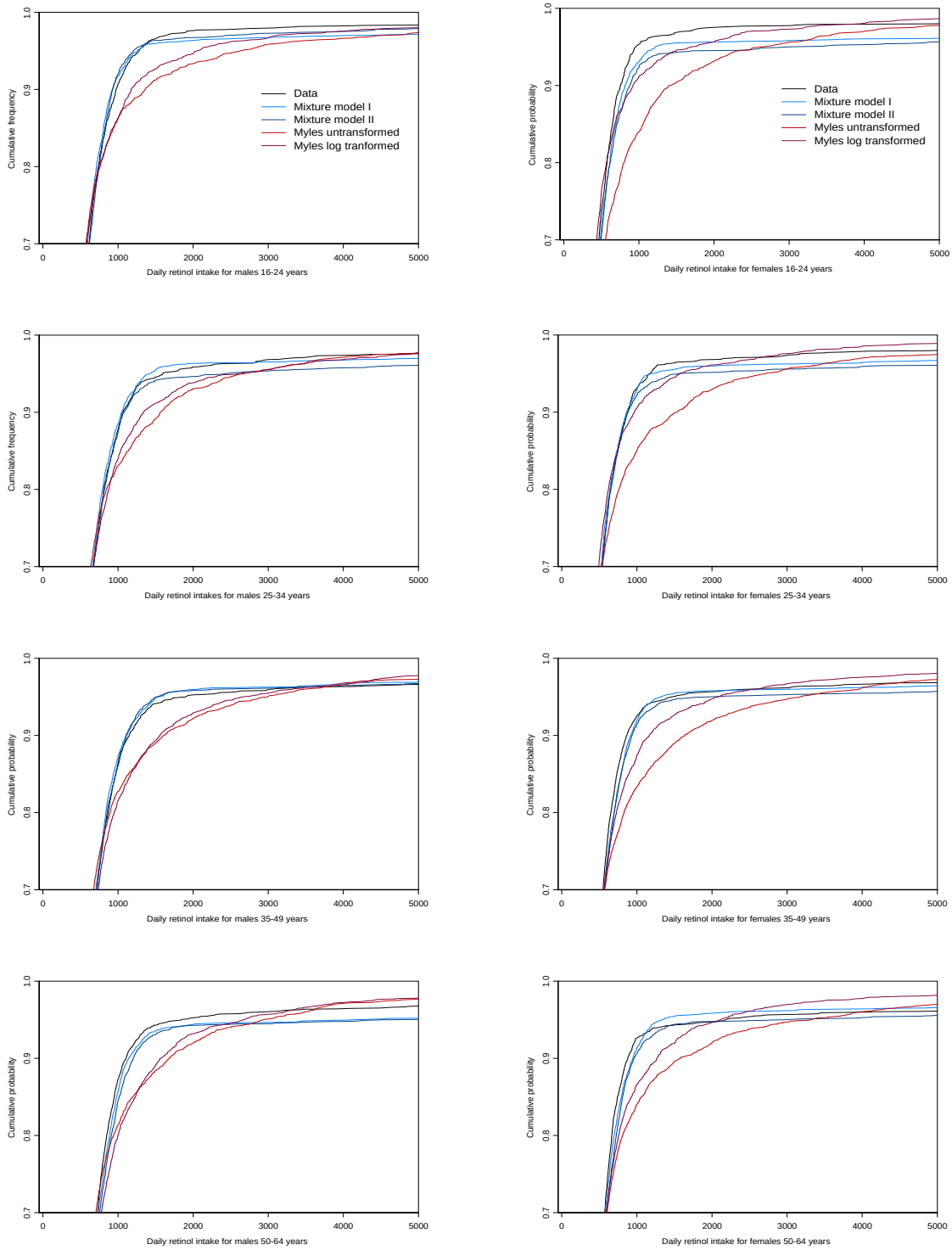

Figure 4.11: Comparison of empirical cdf of observed daily retinol intakes with cdf of predicted intakes from both the Mixture models and Myles et al's model with raw and log transformed data for each sex-by-age group.

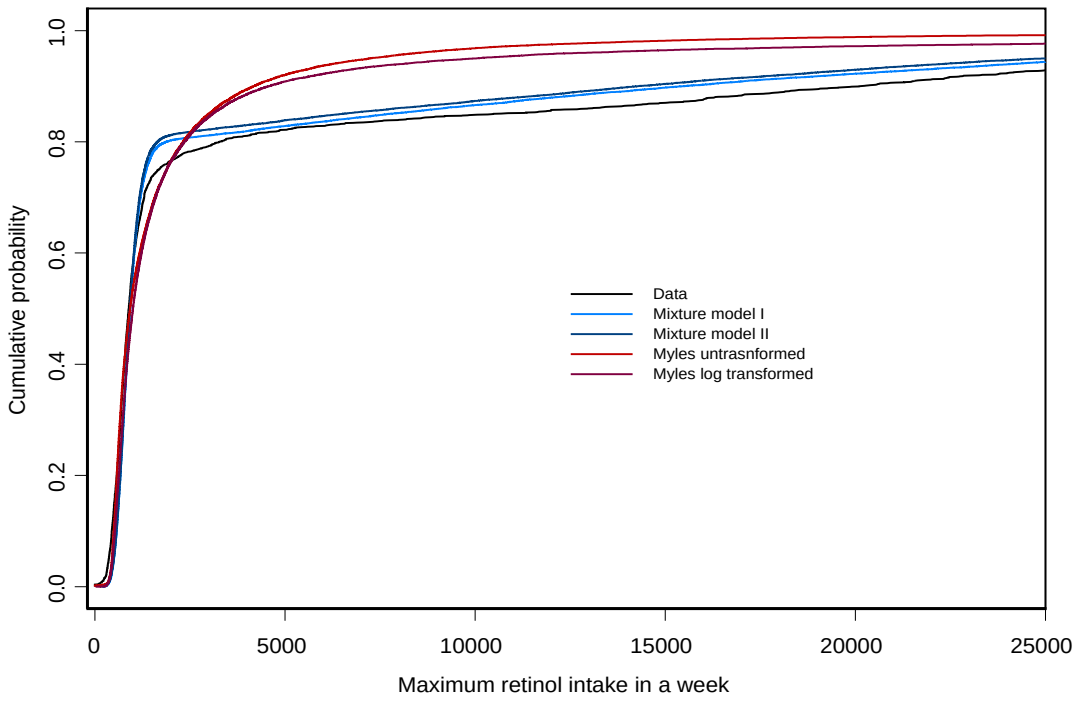

Figure 4.12: Comparison of empirical cdf of observed maximum daily retinol intakes over one week with cdf of predicted maximum intake over one week from both the mixture models and Myles et al's model with raw and log transformed data for all sex-by-age groups combined.

et al. (2003) model with log transformed intakes appears to over-estimate these probabilities for all the sex-by-age groups.

In Table 4.6 we have the predicted exceedance probabilities for 3000 and 10000 μg from the Mixture models, and that of Myles et al. (2003) applied to the log intakes. For predicted probabilities of daily retinol intake exceeding 3000 μg, both Mixture models do well and mostly better that Myles et al. model, although this gives gives the best estimate for females aged 16-24. For 10000 μg, Myles et al. (2003) under-predicts the exceedance probabilities for all the sex-by-age groups. The predicted exceedance probabilities calculated from the mixture models are mostly accurate.

Over all individuals, the proportions of maximum daily retinol intake over seven days which are greater than than the thresholds of 3000 μg and 10000 μg are 0.21 and 0.15 respectively. These can be compared with the predicted exceedance

Sex	Age	Intakes exceeding the RDA			
		Data	Model I	Model II	Myles
Male	16-24	12.5	12.0	11.0	14.9
Male	25-34	16.1	14.8	15.0	18.4
Male	35-49	18.2	16.7	18.2	18.7
Male	50-64	17.4	19.2	19.4	20.4
Female	16-24	11.5	13.7	14.7	14.5
Female	25-34	16.9	17.3	16.8	17.3
Female	35-49	17.1	18.7	19.6	18.7
Female	50-64	17.3	20.5	19.9	21.5

Table 4.5: Observed percentages of daily retinol intake greater than the RDA for the sex-by-age groups, with corresponding predictive exceedance percentages from the two Mixture models and from Myles et al's model applied to log intakes.

probabilities for the maximum intake from the Mixture models I and II, which are $(0.19, 0.12)$ and $(0.18, 0.11)$ respectively. The corresponding predicted exceedance probabilities from Myles' model with logarithms and the raw data are $(0.15, 0.05)$ and $(0.11, 0.03)$ respectively. The Mixture model under-estimate these exceedance probabilities, but to a much smaller extent that the model of Myles et al. (2003).

4.9 Discussion

Monitoring retinol intake is of importance for the positive and negative effects it has on human health. Much work has been done studying the effects of low Vitamin A, especially among children in developing countries (Ahmed et al. (2002)). Our main interest has been to develop a model which can predict high intake values realistically as these large intakes have a negative health impact. The large skewness present in the data made it challenging for modelling the intakes. We have modelled daily retinol intakes using only a simple transformation, i.e. the square root of intakes, in contrast to Nusser et al. (1996), who try to transform the another data set on retinol intakes to Normality as discussed in Chapter 1. In many cases power transformations fail to normalise dietary data. The presence of zeros prevents us from using the log transformation without adding some small quantity to the zeros. Our model does not rely on complicated transformations, and it predicts high retinol intakes well. This helps us to obtain good estimates of the probability of exceeding high levels.

If we had only one observation per individual then we could assume a mixture model in which individuals have moderate and extreme responses with common

Sex	Age	Intake exceeding (μg)							
		3000				10000			
		Data	Model I	Model II	Myles	Data	Model I	Model II	Myles
Male	16-24	2.1	2.9	2.9	3.0	1.4	1.4	1.5	0.7
Male	25-34	3.2	3.8	3.6	3.9	2.0	2.3	2.0	1.2
Male	35-49	4.1	3.9	5.0	4.4	2.4	2.4	3.0	1.1
Male	50-64	3.9	4.4	4.5	4.7	3.0	3.1	3.0	1.2
Female	16-24	2.3	3.8	4.2	2.4	1.4	2.4	2.8	0.7
Female	25-34	2.6	4.1	4.2	3.0	1.7	2.7	2.7	1.1
Female	35-49	3.8	4.6	4.4	3.5	2.7	3.1	3.1	1.1
Female	50-64	4.3	5.0	4.4	3.1	3.3	3.1	3.2	0.8
All individuals		3.4	4.0	4.1	3.5	2.3	2.5	2.6	1.0

Table 4.6: Observed percentages of daily retinol intake greater than 3000 μg and 10,000 μg for the sex-by-age groups, with corresponding predictive exceedance percentages from the two Mixture models and from Myles et al's model applied to log intakes

probabilities say π and $1-\pi$ respectively. For successive individual intakes, Mixture models I and II represent the distribution of intakes as a mixture, but allow the mixing proportions to differ between individuals. This is similar to Myles et al. (2003) in having a different marginal distribution for each individual, but Myles et al. assume the raw and the log transformed data to be Normal, which is clearly unreasonable.

Mixture model II gives us a possible approach for studying dietary data on successive days when intakes on consecutive days are correlated by allowing the transitions between moderate and extreme intakes on consecutive days to be governed by a Markov chain. For our retinol intakes, we do not seem to have improved the fit of the model to the data by adding the extra condition of dependence between intakes on consecutive days. For the predicted exceedance probabilities we can see that both the Mixture models give similar predictions. The model of Myles et al. (2003) applied to the logarithms of intakes results in the probabilities of daily retinol intake greater than 10000 μg which are much lower than the observed proportions. The model does not perform well in predicting high intake values.

Such mixture models can be applied to a wide variety of dietary data which exhibit long tails. Also if the data set has a large proportion of zeros both the mixture models can be extended by including a third component in the mixtures with all values equal to zero. One can also look at seasonal effects on consumption

patterns. For our study we wanted one model for all the retinol intakes so we did not consider seasonal effects. However if information about dietary habits for a set of individuals is available for different times of the year, one can allow the intakes to depend on the time of the year along with the sex and age of these individual.

Chapter 5

Extreme Value Theory for Modelling Dietary Data with Extreme Intakes

In the previous chapter we looked at modelling moderate and extreme retinol intakes using a mixture model. Our focus was on modelling high intakes of retinol. Another approach to study the extreme retinol intakes, ie the tail of the distribution is to use *extreme value theory*.

Extreme-value theory is used in areas such as risk assessment of financial markets, in telecommunications and environmental and reliability modelling. Recently Paulo et al. (2004*b*) have looked at applying multivariate extreme value theory to study intake of toxic chemicals and nutrients. According to Coles (2001), the distinguishing feature of an extreme value analysis is the objective to quantify the stochastic behaviour of a process at unusually large or small levels. Extreme value theory can be used to extrapolate beyond the observed sample maxima or minima, ie to predict something more extreme than what has already been observed.

A common problem in inferring about the tail of a distribution is that data in this region are scarce. According to Coles (2001) standard modelling approaches fit well where the data have greatest density, but can be biased in estimating tail probabilities.. If the interest is only in the tail of the distribution why should one model the body of the distribution? Extreme value theory provides procedures for tail estimation which are scientifically and stochastically rational. Extreme value models are developed using asymptotic arguments and with the assumption that the underlying process is non-changing to enable extrapolation. If these assumptions are not valid, the extrapolation might not be accurate.

For our retinol intake data set discussed in the previous chapter, we want to model the high retinol intakes; thus we look at the upper tail of the data. In this chapter we look at two distributions to study retinol intakes, one using block maxima and the other a threshold model.

5.1 Generalised Extreme Value Distribution

For 2197 individuals we have retinol intakes on seven consecutive days. Let $\{\nu_{i1}, \ldots, \nu_{i7}\}$ for $i = 1, \ldots, 2197$ be the observed retinol intakes assumed to have a common distribution function F. These vectors are assumed to be independent with the distribution function F. The assumption of independent and identically distributed random variables is necessary to develop the extreme value theory.

In extreme value theory while analysing time series data, we have a long sequence of random variable and the sequence is then split into disjoining blocks. This is common for example in studying rainfall data, where the daily rainfall over several years may be available and the data are then split into blocks of one year each. However though the retinol intakes are also time series, we only have seven observations per individual. Instead of having a long sequence of data points, we have successive intakes for many individuals. We treat the seven observations for each individual as a block. In most extreme value analysis choosing the block size is critical. The block size should be such that the maximum from one block is independent of the maximum from the next block. This is true with our data as the maximum retinol intake for one individual is independent of the maximum intake of another individual. Let M_i represent the maximum intake for individual i over the seven days which is our block size.

We define $M_i = \max\{\nu_{i1}, \ldots, \nu_{i7}\}$. We want to estimate G, the distribution function of the block maxima M_i. If there exists a sequence of constants $\{a_i > 0\}$ and $\{b_i\}$ such that

$$Pr\left\{\left(\frac{M_i - b_i}{a_i}\right) \leq z\right\} \to G(z) \quad \text{as} \quad i \to \infty \tag{5.1}$$

then according to Coles (2001),

$$G(z) = \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\} \tag{5.2}$$

defined on the set $\{z : 1+\xi(z-\mu)/\sigma > 0\}$, where the parameters satisfy $-\infty < \mu < \infty, -\infty < \xi < \infty$ and $\sigma > 0$. Here a_i and b_i are sequences of normalizing coefficients.

Equation (5.2) represents the generalised extreme value (GEV) family of distributions. The family has three parameters, μ the location, σ the scale and ξ the shape parameter.

The maximum likelihood estimates (mle) for the GEV distribution can be obtained using the FinMetrics module in S-PLUS, (Insightful Corporation (2002)). For values of $\xi > -0.5$, the mle have the usual asymptotic properties. For $\xi \leq -0.5$ the distribution has a very short bounded tail and rarely occurs in application of extreme values. For $-1 < \xi < -0.5$, mle are non-regular. When $\xi < -1$ mle are not obtainable.

We fit the GEV distribution to the maximum retinol intakes for the individuals over the seven days. The mle for the distribution parameters as obtained from S-PLUS are in Table 5.1, the hat on the parameters represent their estimates.

Parameter	Mle	95% CI
$\hat{\xi}$	0.8657	[0.8283,0.9031]
$\hat{\mu}$	799.5	[753.8, 845.1]
$\hat{\sigma}$	821.7	[786.5, 857.0]

Table 5.1: Mle of the GEV model for the maximum retinol intakes along with their 95 % confidence intervals (CI).

The estimated variance matrix for the parameter estimates is

$$\mathbf{\Sigma} = \begin{bmatrix} 0.0004 & 0.2798 & 0.0414 \\ & 542.6 & 339.2 \\ & & 323.9 \end{bmatrix}$$

Using matrix **Sigma**, we can find the corresponding standard errors (SE) for the parameter estimates and an approximate 95% confidence intervals (CI) for each of them are given in Table 5.1. The CI are constructed as estimate $\pm$ 1.96 $\times$ the corresponding SE for the parameter estimate.

The estimates of extreme quantiles of the GEV distribution are obtained inverting equation (5.2) to give

$$z_p = \mu - \frac{\sigma}{\xi}\left[1 - \{-\log(1-p)\}^{-\xi}\right] \quad (\xi \neq 0). \tag{5.3}$$

In extreme-value theory, z_p is called the return level associated with the return period $1/p$. The estimates of the parameters are used to obtain the return levels. However since in our case the maximum retinol intakes are not time series points this concept is not directly applicable. We can interpret the return levels for our case as the maximum retinol consumption which will be larger than z_p with probability p. The variance estimate for z_p as given by Coles (2001) is

$$\mathrm{Var}(z_p) \approx \nabla z_p^T \Sigma \nabla z_p, \tag{5.4}$$

where

$$\begin{aligned} \nabla z_p^T &= \left[\frac{\partial z_p}{\partial \mu}, \frac{\partial z_p}{\partial \sigma}, \frac{\partial z_p}{\partial \xi}\right] \\ &= [1, -\xi^{-1}(1 - y_p^{-\xi}), \sigma\xi^{-2}(1 - y_p^{-\xi}) - \sigma\xi^{-1}y_p^{-\xi}\log y_p] \end{aligned} \tag{5.5}$$

and $y_p = -\log(1-p)$.

By substituting the mle of the GEV distribution parameters in Equation (5.3), the mle of z_p for $0 < p < 1$ can be obtained. We can make the following statements from our parameter estimates. For $p = 1/100$, $\hat{z}_{0.01} = 51557$ with a standard deviation of 2803. Hence an approximate 95% confidence interval for $z_{0.01}$ is $[46063, 57051]$. We can estimate that for a given individual, the maximum intake exceeds 51557 μg with probability 0.01. For $p = 1/10$, $\hat{z}_{0.1} = 6568$ and the corresponding approximate 95% confidence interval is $[5987, 7149]$. Thus the maximum intake for an individual exceeds 6568 μg with probability 0.1.

Figure 5.1 gives the QQ plot of the residuals for the GEV fit. Data are converted to unit exponentially distributed residuals under null hypothesis that GEV fits in S-PLUS. The reference distribution is Exponential, and if the transformed residuals lie close to the diagonal line the data fits a GEV distribution. In our case the graph does not appear linear, and it is clear that the GEV distribution is not the appropriate distribution.

As mentioned earlier, a general consideration when fitting block maxima in a GEV distribution is the size of the block. In our case our block size is seven as we

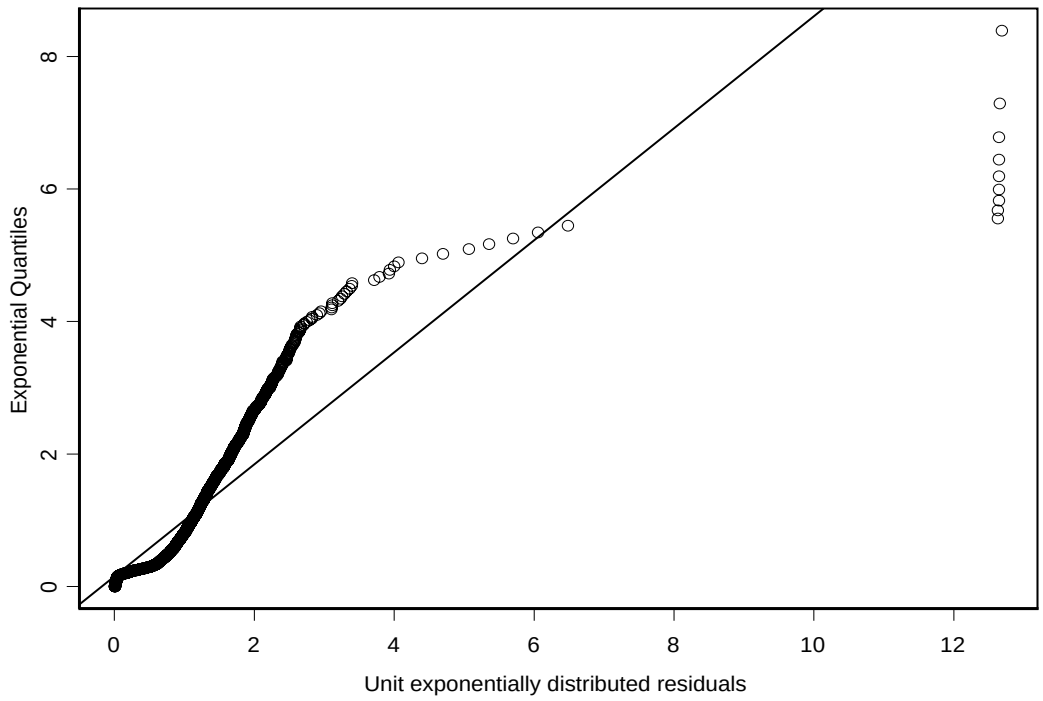

Figure 5.1: QQ plot of residuals for the GEV model fitted to the maximum retinol intakes with exponential reference distribution.

have only seven observations per individual. A small block size might lead to bias in the estimation of the parameters while too large a block size might increase the estimation variance. The choice of block size might be critical when there is correlation between values in consecutive blocks. For our data each individual's seven intakes are assumed to be independent of the other individual's intakes. However a block size of seven is quite small and can be the cause of the poor fit of the GEV distribution. Also here we use only one data point per individual and instead we could use r largest order statistics (Coles (2001)). We now look at a procedure where we can avoid blocks and use all the data available to us.

5.2 Generalised Pareto Distribution

In this section we consider the generalised Pareto distribution (GPD) as a conditional model for excesses of a high threshold. We model the intakes which are greater than a certain chosen threshold using the GPD.

As in the previous section, M_i is defined as the maximum intake over seven

days for individual i. As before M_i is assumed to have a distribution function given by Equation (5.2). For large enough u, which we call the threshold, the distribution function of $(\nu_{ij} - u)$ conditional on $\nu_{ij} > u$ is approximately,

$$H(y) = 1 - \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)^{-1/\xi} \tag{5.6}$$

defined on $\{y : y > 0 \text{ and } (1 + \xi y/\tilde{\sigma}) > 0\}$, where

$$\tilde{\sigma} = \sigma + \xi(u - \mu).$$

The parameter ξ in the GPD is equal to that of the corresponding GEV distribution. The family of distribution defined by Equation (5.6) is called the Generalised Pareto family. For more details see Coles (2001) and Davison & Smith (1990). In this case ξ and σ are both invariant to block size.

If we choose to use all the data available to us for fitting the GPD, we ignore the fact that we have seven observations per individual. The alternative is to work with the maximum per individual as for the GEV distribution. In the first case we use all the intakes and we are not wasting any data but losing the information of having repeated observations for each individual. In both cases we assume the observations for all individuals are independent and identically distributed. We present results using both the data sets, the whole data set and the maximum per individual.

Threshold selection is analogous to block size selection and is the most important part in fitting the Generalised Pareto distribution (GPD) (Coles (2001)). The threshold value should not be too low or else the asymptotic basis of the model will be violated leading to a bias. A very high value of threshold will leave very few values with which the model can be estimated, leading to a very high variance of the parameter estimates. Given that an intake is greater than the threshold, the difference between the intake and the threshold is called the mean excess. Exploratory techniques can be used to determine the threshold, such as plotting the mean excess versus threshold values, which is called the mean residual life plot. A straight line with positive gradient above some threshold is a sign of Pareto behaviour in the tail. For our whole data set and for the maximum intakes the mean residual life plot can be seen in Figures 5.2, 5.3, 5.4 and 5.5. We have truncated the horizontal axes on the plots 5.3 and 5.5 at 10,000 for the threshold. There is some evidence of linearity above $u = 2000$ for both the plots.

We choose to work with a lower value of threshold with $u = 1500$.

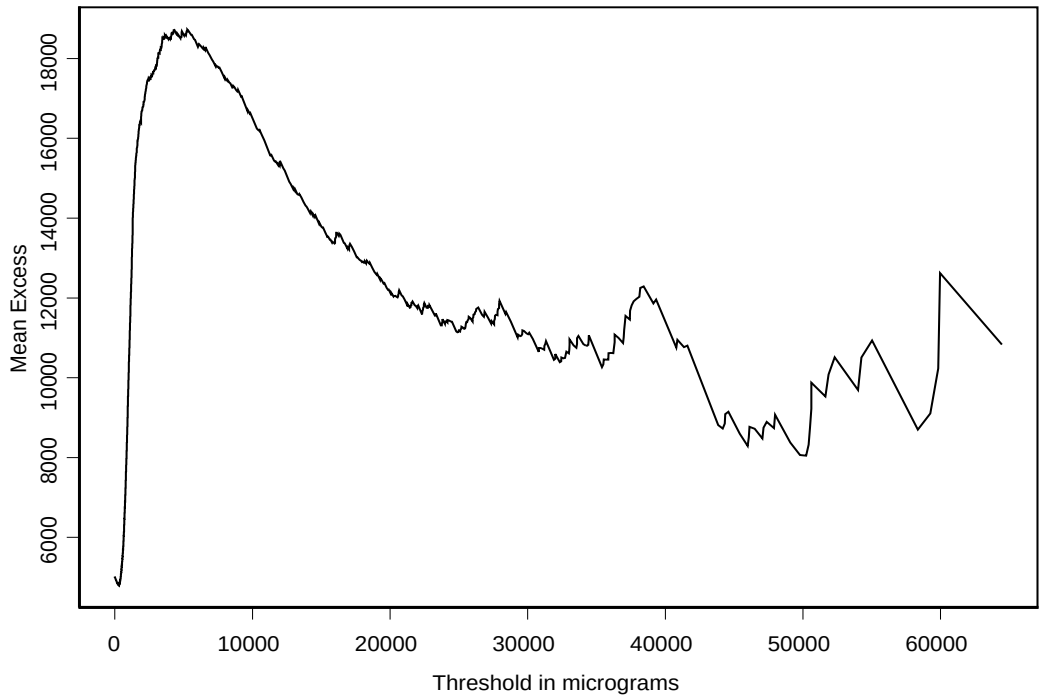

Figure 5.2: Mean residual life plot for maximum retinol intake over a week.

An possible choice for our threshold value would have been $u = 3000$ which is the safe level for daily retinol consumption but with this value we are not able to obtain the tail of the underlying distribution. This is because with such a high threshold we do not have enough points in our data to estimate the parameters, since almost 97% of the intakes are less than 3000. With $u = 1500$ the model is fitted and parameter estimates are obtained using the FinMetrics module in S-Plus: details of using this module can be found in Zivot & Wang (2003). We have 95% of the intakes less than 1500. Table 5.2 gives the maximum likelihood estimates for ξ and $\tilde{\sigma}$.

If we compare the $\hat{\xi}$ values from Tables 5.1 and 5.2 for the maximum retinol intake values, there is a large difference between the two. The estimate of ξ depends very much on the value of the threshold chosen, and for a lower value of u the difference between the estimated ξ values from the GEV distribution and GPD decreases. Figures 5.6 and 5.7 give the tail of the underlying distribution

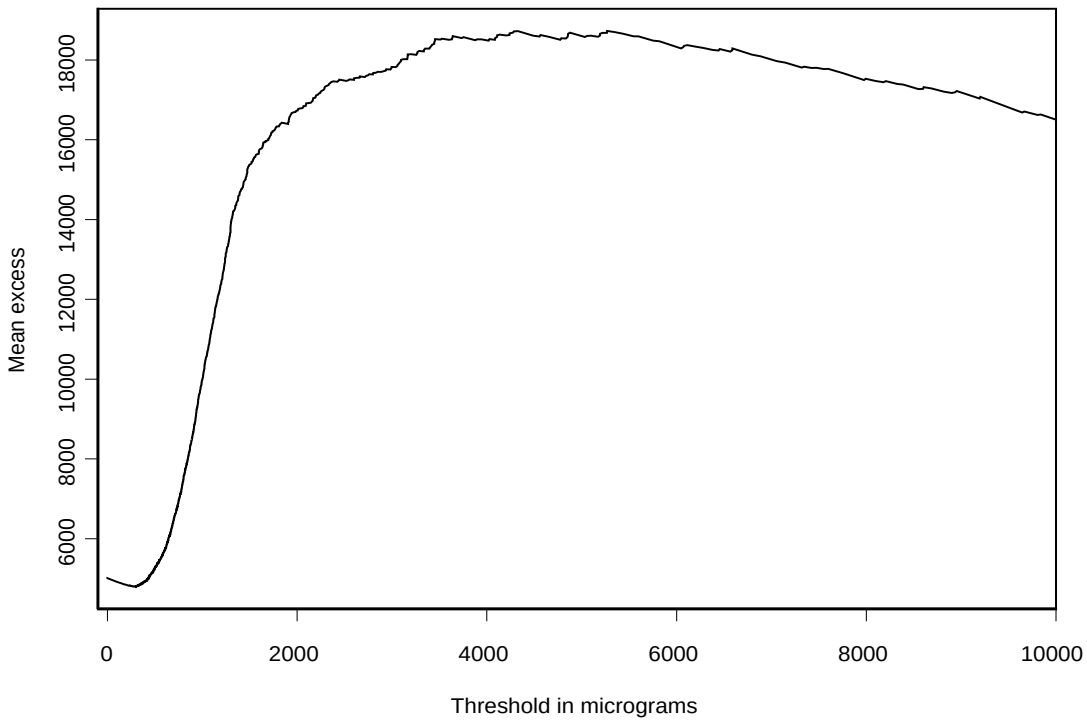

Figure 5.3: Mean residual life plot for maximum retinol intake over a week, truncated to a threshold of 10000 on the x-axis.

and the Exponential QQ plots for residuals for individual maximum intakes and for the whole data set. The QQ plot is better for the maximum retinol intakes. The QQ plots for residuals appear linear and the fit is acceptable.

The quantile z_p for a GPD can be determined using

$$\zeta_u \left[1 + \xi \left(\frac{z_p - u}{\tilde{\sigma}}\right)\right]^{-1/\xi} = 1/p$$

Rearranging,

$$z_p = u + \frac{\tilde{\sigma}}{\xi}\left\{(p\zeta_u)^{\xi} - 1\right\} \tag{5.7}$$

provided $z_p > u$ and $\zeta_u = Pr(z > u)$.

Estimation of return levels requires the replacement of parameter values by their estimates. A natural estimator of the exceedance probability $\hat{\zeta}_u$ is $\frac{k}{n}$, where k is the proportion of points exceeding u. Since the number of exceedances of u follows a Binomial distribution, $Bin(n, \zeta_u)$, $\hat{\zeta}_u$ is also the mle of ζ_u. As before we can estimate the standard errors for the return levels. The variance estimate is

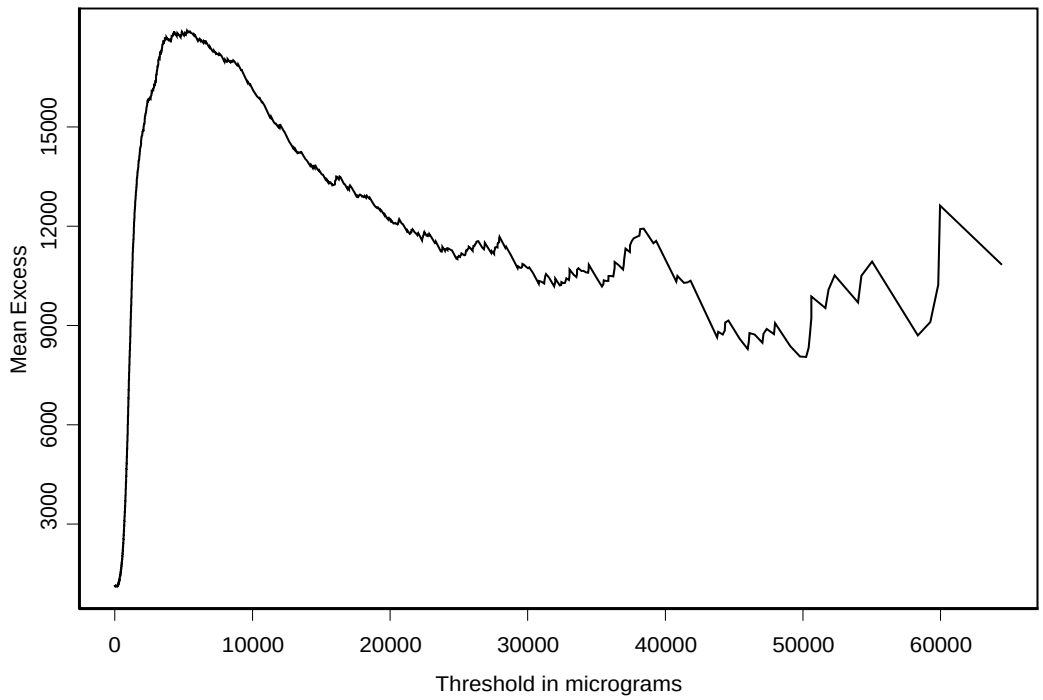

Figure 5.4: Mean residual life plot for daily retinol intakes for the whole data.

$$\begin{aligned} \mathrm{Var}(\hat{\nu}_{ij}) &\approx \nabla\nu_{ij}^T \Sigma \nabla\nu_{ij} \\ \nabla\nu_{ij}^T &= [\tilde{\sigma} N^{\xi} \zeta_u^{\xi-1}, \xi^{-1}(N\zeta_u)^{\xi} - 1, -\tilde{\sigma}\xi^{-2}(N\zeta_u)^{\xi} + \tilde{\sigma}\xi^{-1}(N\zeta_u)^{\xi}\log(N\zeta_u)]. \end{aligned}$$

The covariance matrix for $(\hat{\zeta}_u, \hat{\tilde{\sigma}}, \hat{\xi})$ is approximately

$$\Sigma = \begin{bmatrix} \hat{\zeta_u}(1-\hat{\zeta_u})/n & 0 & 0 \\ 0 & \tilde{\sigma}_{11} & \tilde{\sigma}_{12} \\ 0 & \tilde{\sigma}_{21} & \tilde{\sigma}_{22} \end{bmatrix},$$

where N represents the total number of observations, $\tilde{\sigma}_{ab}$ denotes the (a, b) term for the variance matrix of $\hat{\tilde{\sigma}}$ and $\hat{\xi}$. Table 5.3 gives various probabilities of exceeding high levels of retinol consumption for both the data sets. For the whole data set the maximum likelihood estimate of the exceedance probability is $\hat{\zeta}_u = 0.048$ with approximate standard deviation 0.0017. The estimate for the exceedance probability for the maxima for each individual is 0.263 and has approximate standard deviation 0.0094.

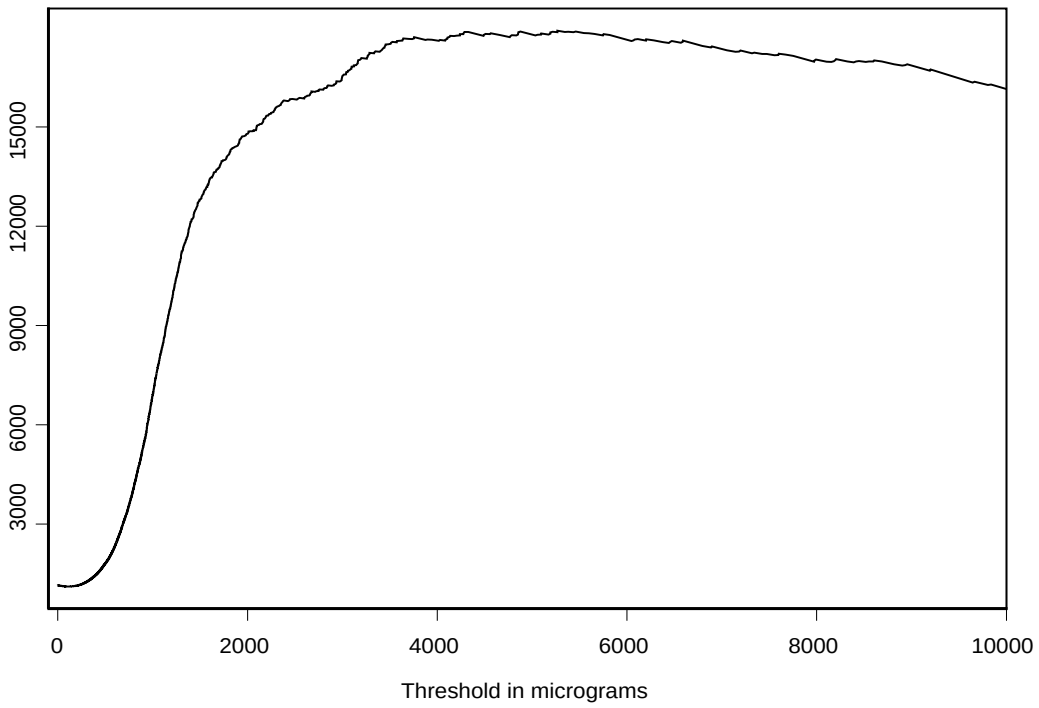

Figure 5.5: Mean residual life plot for daily retinol intakes for the whole data, truncated between 0 to 10000 on the x-axis.

From Table 5.3 we can say that the probability that the daily retinol intake for an individual is greater than 51890 μg is 0.01. Similarly, probability that the daily retinol intake is greater than 12803 μg is 0.17. At $p = 0.1$ the estimated daily intake is larger than the maximum intake, however at $p = 0.01$ this is just the opposite. The probability for an individual's maximum intake being greater than 5519 μg is 0.17. Under the GEV distribution we have a probability 0.17 of exceeding $3000\mu g$ whereas under the GPD model, we have a probability of 0.2 for the maximum retinol intake exceeding the safe level of $3000\mu g$. The probability estimates for the GPD depend upon the choice of threshold and for a different threshold value we will get a different set of return levels.

5.3 Discussion

In most cases extreme value theory is used to study the tail of a distribution so as to be able to extrapolate values larger than what has already been observed. Using this method to study retinol intakes restricts us to looking at the upper tail of the data. We are not able to comment on whether the intakes meet the

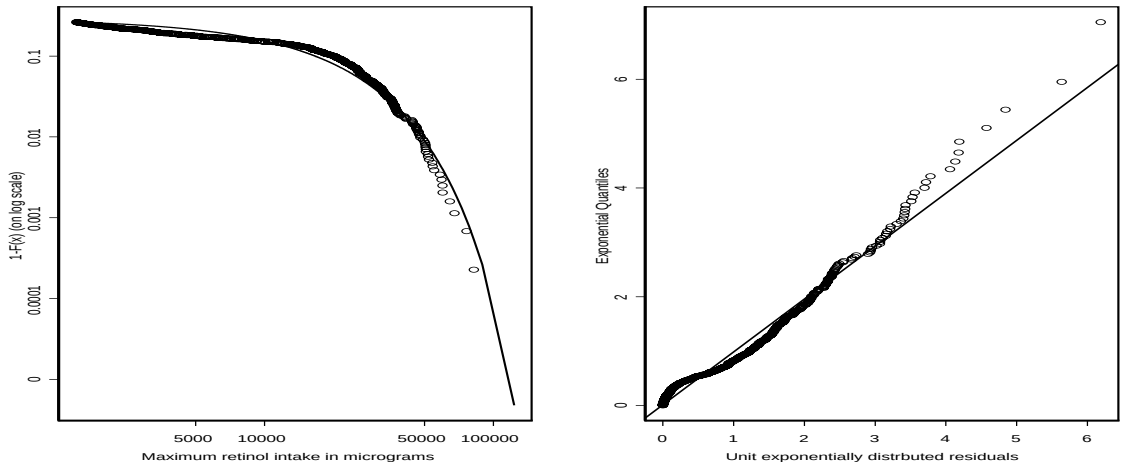

Figure 5.6: Plot of the tail of underlying distribution and the QQ plot with $u = 1500$ using the maximum retinol intake (X) for each individual.

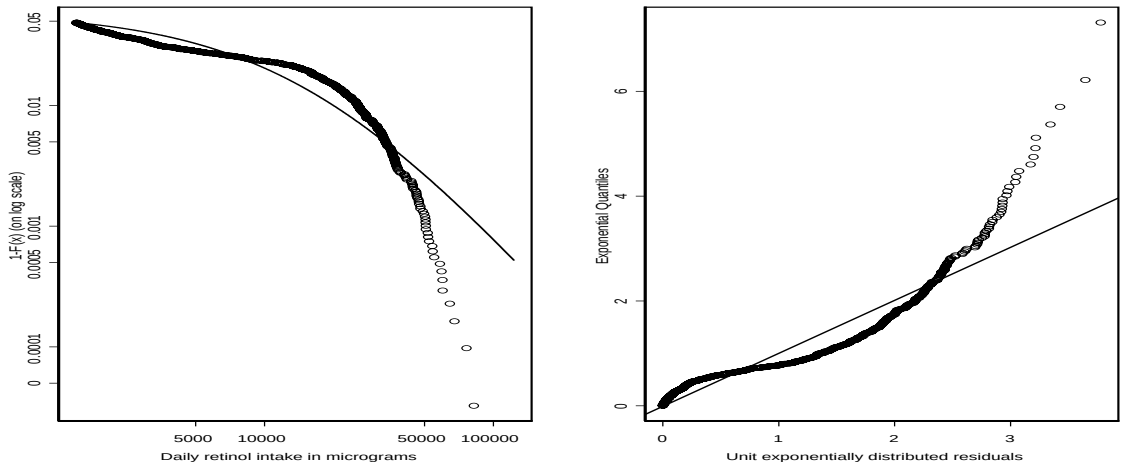

Figure 5.7: Plot of the tail of the underlying distribution and the QQ plot with $u = 1500$ for the whole data set, where X is the daily retinol intake.

Parameter	Maximum likelihood estimates		Confidence interval	
	Whole data	Individual maxima	Whole data	Individual maxima
ξ	0.4579	−0.0737	[0.2465,0.6693]	[-0.1770, 0.0296]
$\tilde{\sigma}$	16216	7950	[14119, 18312]	[6194, 9705]

Table 5.2: Mle of the GPD parameters for individual maximum intakes and whole data with the 95% confidence intervals.

Parameter	Whole data		Maximum per individual	
	p = 0.01	p = 0.1	p = 0.01	p = 0.1
$\hat{z}_p$	51890	20515	61678	11154
$sd(\hat{z}_p)$	7816	1874	4315	1695

Table 5.3: Return levels for GPD for the two data sets.

recommended daily allowances as these are not extreme values.

The choice of threshold for the the GPD is somewhat subjective. We may prefer to work with the safe level of $3000\mu g$ of retinol per day but that appears to be an unreasonable threshold choice as we are unable to obtain parameter estimates.

Here we have not used the fact that we have repeated observations on each individual. As a first step we model the distribution of maximum intakes for each individual, as using repeated measures which might be dependent is not straightforward. An alternative method is to fit a common GPD assuming all individuals to have a common shape parameter (ξ), but different scale parameters (σ). Seven observations per individual is not enough to fit a separate GPD to each individual. We can look at whether the variation between the scale parameters of the individuals can be explained with covariates; for example the scale parameter could be a function of sex and age. This may be based on the model of Smith (1989) to study air pollution using ground-level ozone.

Both the GEV distribution and the GPD assume an underlying series of independent and identically distributed random variables, though this is not a necessary condition for implementing extreme value distributions. This makes the arguments for the models simpler but might be an unrealistic assumption for some data sets. Also one can develop multivariate extreme value models as done by Paulo et al. (2004*b*) to look at multiple nutrient intake at a time.

Chapter 6

Conclusions

6.1 Summary

In this thesis we have developed novel methods for modelling dietary data. Past methods such as those described by Slob (1993) and Nusser et al. (1996) have mostly focussed on estimating the usual intakes for a nutrient, a toxin or a pesticide. However few papers provide methods which takes in to account all the problems associated with dietary data and allows us to determine the probability of a certain population to be at risk from large consumptions of a pesticide or nutrient. It needs to be mentioned here that though we give probability of exceeding certain levels of consumption for individuals in a sex-by -age group, some individuals within the same sex-by age group are more at risk of harm than others at any given consumption level. This may be due to the different health conditions of the individuals Nusser et al. (1996) focuses on providing a method to achieve Normality for the dietary data whereas many others such as Slob (1993) and Paulo et al. (2004*a*) use a log transformation for the same. These transformations do not achieve Normality when the data set is highly skewed. Also none of these methods make allowance for zeros in the data or correlation among consecutive intakes.

We provide various models for food risk assessment which are able to handle problems such as large proportions of zeros, skewness in the data, intake of multiple products inn the same period of observation and correlated intakes. Our models not only estimate the average intake of a product but also predict probabilities of exceedance and long-term intakes for any individual in a given sex-by-age group. On comparing results obtained from our models in Chapters 2, 3 and 4 with the data, we see that our models give accurate estimates. Most of the models in this thesis have been developed using Bayesian methodology. We use hierarchical Bayesian model for our data which allows us to account for the various sources of variability in the data and also reflect our prior knowledge of

the model parameters. The Bayesian models may appear complicated but the use of WinBUGS to fit the models allow us to obtain posterior predictive distributions for our model parameters quite easily. The methodologies illustrated in this these are applicable to other to other areas of dietary risk assessment.

Many dietary data have a large proportion of zero intakes. When modelling such data sets it is essential we account for these zeros. Few papers until now have demonstrated explicitly methods to model the presence of large proportions of zeros in dietary data. In the second chapter we have suggested two possible methods for modelling zeros in dietary data. The decision to consume a product on any day is modelled in the propensity model of Chapter 1. Using such an approach one can look at individual intakes over several days and model products which are consumed infrequently. Most dietary data do not give information about long term consumption patterns, for example the fact that a person does not consume a particular product during an observed week need not imply he or she never consumes that product. For such a person our model will give a very low propensity to consume that product but will not rule out completely the possibility of a non-zero consumption in the future.

We use data on alcohol intakes to illustrate the models in Chapter 2. Though for the alcohol intakes, the predicted cdf does not perfectly fit the data the predicted probability estimates agree well with the observed ones. As mentioned before in Section 2.8 alcohol intakes are atypical. One would rarely have a quarter of a glass of wine or of beer. The problem lies in the fact that we have very few small intakes of alcohol as one might consume at east 0.2 MJ of alcohol on one occasion. This problem may not occur with other food products.

The second approach for modelling zeros is using the latent Gaussian model. It is essential to achieve a suitable transformation to Normality for the model to perform well. We demonstrate this fact by providing results from a power transformation and then show the improvement in predictions using a two-part transformation. We treat the zeros in our data as censored observations. An added advantage of using such a Bayesian methods is that it allows us to model data with missing values and for a missing value and we can specify the possible range in which the intake can be.

In the third chapter we extend our univariate latent Gaussian model to a multivariate one. This chapter deals with exposure assessment to a pesticide Iprodione

through five food products. We work with a multivariate latent Gaussian model for consumption of multiple products simultaneously. Most past methods model the intake of a single product at a time however Paulo et al. (2004*a*) suggest a simple Bayesian approach to model consumption of multiple products using a multivariate Normal distribution. This model does not work in the presence of large number of zeros. We overcome this problem with our Bayesian multivariate latent Gaussian model.

Data on concentration of pesticides on food products also typically have large proportions of zero observations. These can be true zeros, when the pesticide is absent on that product, or can be false zeros, when the level of pesticide on the product is less than the level of detection (LOD). Either one assumes that all the zeros are true zeros or all the zeros are set to be at the LOD. This might over-estimate or under-estimate the proportion of zeros. One may also assign a value halfway between the LOD and zero. For our models in WinBUGS, using the $I(lower, upper)$ function we assume the upper limit is the LOD for censored observations, i.e. the zeros in the data. This allows the simulated intakes to be anything between $-\infty$ and the LOD. The negative simulated values are then set to zero. Though this method may not accurately estimate the true zeros, it at least gives a non-zero probability of randomly generating a non-zero concentration less than the LOD.

The above methods work well for looking at the bulk of the data or specifically for data sets with large number of non-consumption days. However food risk assessment involves at times looking at the upper tail of the data. In Chapter 4 we propose a model for dietary data with occasional large intakes and illustrate our method for retinol intakes. When dietary data sets are highly skewed such as the retinol intakes, it is difficult to determine a transformation to achieve Normality. For the retinol intakes we use a mixture of two Normal distributions for modelling intakes, one for moderate intakes and one for extreme ones. This allows us to model the extreme intakes observed. The mixture model applied to a square root transformation of retinol intakes predicts the exceedance probabilities well, even in the tail. This is an attractive feature as for many nutrients and food products we may be more interested in the high intakes and hence we concentrate on the upper tail of the distribution of intakes. A similar study can be conducted while looking at very low intakes.

All the above models assume some sort of individual effect on the consump-

tions. This causes repeated intakes on consecutive days for an individual to be positively correlated. We may want a model which allows for negative correlation also. The mixture model has been modified to account for possible correlated intakes on consecutive days. Individuals may not like to eat the same products on consecutive days or may eat the same product over a few meals. The intakes on a day can be positively or negatively dependent on the intakes on the previous day. This is a feature that many dietary models do not account for. When we have observations on consecutive days for an individual, it is quite possible that these intakes are correlated for certain products. We let a Markov chain govern the distribution from which the intakes come from on consecutive days. We have demonstrated this method using the retinol intakes, but for the retinol intakes the addition of a Markov chain to the mixture model does not improve the performance of the mixture model. However for products where one may expect correlation between intakes on consecutive days, this method can be useful to model this correlation. For our model we consider dependence only on the previous one day and hence use a first order Markov chain.

Another approach considered in Chapter 5 is the use of extreme value theory to study extreme intakes. Using the data on retinol intakes we fit standard distributions in extreme value theory and discuss certain pros and cons of using this approach. The methods we consider are able to deal with consecutive observations but only on a single individual. We also do not take into account any sex or age effects on the intakes.

6.2 Future Work

For our models we work with the Normal distribution for the convenience of modelling random effects at different levels. We mostly use simple power transformations to achieve normality for our data sets and this allows us to back-transform the predicted intakes to the original scale easily.

Some of the priors chosen for our models are arbitrary. This is mainly due to the absence of expert knowledge available to us about consumption and concentrations of the food products. In principle one would try to obtain relevant expert knowledge and incorporate these in the prior distributions for the models.

For the propensity model for alcohol intakes, to achieve Normality we work

with the eighth root. The fit of the predictive cdf does not give a very good fit to the data. This can be improved by using a better transformation to Normality. An extension for the latent Gaussian model with the two-part transformation, can be to assume the parameters in the transformation to achieve Normality to be unknown. In our Bayesian model we can assign a distribution to these parameters. However choosing prior distributions might be quite arbitrary.

Although in the third chapter we study exposure to only one pesticide, we can consider multiple pesticide concentrations. Certain agricultural practices follow a particular sequence in the usage of pesticides on a group of products, that is one pesticide will be followed by another particular pesticide. There may be a correlation between the concentrations of these pesticides on these products. In such a case we may model these pesticide concentrations together. Thus along with having a multivariate model for consumption we will have a multivariate model for concentrations provided the pesticide concentrations are correlated. The acute reference dose will have to be formulated for assessing cumulative risk, taking into account how humans react to multiple pesticides consumed at a time.

For the retinol intake data set, we ignore seasonal effects on retinol consumption. Seasonal effects can be accounted for by adding a factor along with the sex-by-age effects to the model. For data sets with a large proportion of zeros and also highly skewed, we can also extend the mixture model by including a third component with all values equal to zero.

We use maximum likelihood estimation for obtaining parameter estimates for our distributions using Extreme Value Theory. A future project may involve developing these Extreme Value Theory models in a Bayesian framework and extending the univariate models to a multivariate case to study extreme intakes of multiple products simultaneously.

Dietary data are collected so that one can investigate the eating habits of the whole population or a sub-population. Hence it is important that these data sets are large enough and representative of the population. The data set used for alcohol intakes is quite small in that we have about 10 individuals in each sex-by-age group. It will be better to have a larger data set with individuals who are not in an artificial environment. Such data will probably pick out trends in the age groups better and we can have smaller age categories. The results from a larger data set will be more representative of the whole population. Also we will

be able to observe if there is any day of the week effect for a data set which is from a real life setting.

The results from all our models are based on the assumption that the reported intakes are true. However no dietary data are free from errors. People in the study may make recording errors or simply mis-report their consumption. Some work has been done to model the mis-reporting in dietary data, see Stubbs et al. (Aug 2001). Future models can include the effect of mis-reporting and hence observe the effective change in the exceedance probabilities.

Appendix A

WinBUGS Codes for Models in Chapter 2

A.1 Propensity Model

```
{
for (i in 1 :59)
{
for (j in 1:12)
{
```
Propensity for individual i on day j
```
prop[i,j] ~ dnorm(bet[i],1)
```
Notional alcohol consumption for individual i on day j
```
al.notional[i,j] ~ dnorm(mu[i], deltaw[i])
al1[i,j] <- al.notional[i,j]*step(prop[i,j])
```
Actual alcohol consumption for individual i on day j
```
al[i,j] ~ dnorm(al1[i,j], 10000000)
```
Sex-by-age effect
```
c[i] <- gamma[gr[i]]
mu[i] ~ dnorm(c[i], deltab)
bet[i] ~ dnorm(rho[gr[i]],1)
log(deltaw[i]) <- lambda[i]
lambda[i] ~ dnorm(mulambda,deltalambda)
```
Within-inidvidual varinace
```
sigma2w[i] <- 1/deltaw[i]
}
```
Between-individual variance
```
sigma2b <- 1/deltab
```

For a new individual in sex-by-age group k

```
for( k in 1:6)
{
bet.n[k] ~ dnorm(rho[k],1)
log(deltaw.n[k]) <- lambda.n[k]
lambda.n[k] ~ dnorm(mulambda,deltalambda)
mu.n[k] ~ dnorm(gamma[k], deltab)
```

To simulate alcohol intake for the new individual at a future day d

```
for(d in 1:7)
{
prop.d[k,d] ~ dnorm(bet.n[k],1)
al.notional.d[k,d] ~ dnorm(gamma[k], deltaw.n[k])
```

Predicted alcohol intake

```
al.d[k,d] <- al.notional.d[k,d]*step(pi.d[k,d])
}
total[k] <- sum(al.d[k,])
}
```

PRIORS

```
deltab ~ dgamma(0.1,0.1)
omega ~ dnorm(1,0.25)
for( k in 1:6)
{
gamma[k] ~ dnorm(omega,0.25)
rho[k] ~ N(0, 1)
}
deltalambda ~ dgamma(0.0005,0.0005)

}
```

A.2 Chapter 2: Latent Gaussian Model

```
{

   for( i in 1:59)
{
for(j in 1:12)
```

```
{
```
Actual alcohol consumption for individual i on ay j
```
a[i,j] ~ dnorm(mu[i], deltaw[i])I( , a.cen[i,j] )
}
```
Probability of a zero intake
```
p[i] <- phi((-mu[i])/sqrt(sigma2w[i]))
```
Sex-by-age effect
```
eff[i] <- gamma[gr[i]]
mu[i] ~ dnorm(eff[i], deltab)
log(deltaw[i]) <- lambda[i]
lambda[i] ~ dnorm(wmu, wdelta)
```
Within-individual variance
```
sigma2w[i] <- 1/deltaw[i]
var[i] <- sigma2w[i] + sigma2b
}
```
PRIORS
```
omega~ dnorm(0, 0.005)
for( k in 1:6)
{
gamma[k] ~ dnorm(omega,0.005)
}
deltab ~ dgamma(0.01,0.01)
```
Between-individual variance
```
sigma2b <- 1/deltab
wmu ~ dnorm(0, 0.1)
wdelta ~ dgamma(1, 10)
```

Alcohol consumption for a new individual in sex-by-age group k
```
for(k in 1:6)
{
mu.n[k] ~ dnorm(group[k], deltab)
log(deltaw.n[k]) <- lambda.n[k]
lambda.n[k] ~ dnorm(wmu, wdelta)
```
To simulate alcohol intake for the new individual at a future day d
```
for(d in 1:7)
{
al.d[k,d] ~ dnorm(mu.n[k], deltaw.n[k])
```
Predicted alcohol intake

```
a.d[k,d] <- max(al.d[k,d],0)
}
```
Total predicted alcohol intake over a week
```
total[k] <- sum(a.d[k,])
}
}
```

Appendix B

WinBUGS Codes for Models in Chapter 3

B.1 Multivariate Latent Gaussian Model for Consumption of Multiple Products

```
{
for(i in 1:5756)
{
for(j in 1:2)
{
Intake for the individual i on day j for the five food products
y[i,j, 1:5] ~ dmnorm(mu[i,], prec.food[ , ]) I( , y.cen[i,j,1:5])
}
for(f in 1:5)
{
Additive effect of individual and food products on consumption
mu[i,f] <- mu.food[f] + mu.ind[i]
}
mu.ind[i] ~dnorm( a, prec.ind)
}
prec.ind ~ dgamma(1,50)
a ~ N(0,0.0001)
PRIORS
for(f in 1:5)
{
mu.food[f] ~ dnorm(0,2)
}
prec.food[1:5,1:5] ~ dwish(s[,], 5)
```

```
var.food[1:5,1:5] <- inverse(prec.food[ , ] )

Intake of the five food products for a new individual in the sex-by-age group k
for( k in 1:5)
{
mu.n[k] <- mu.food[k] + mu.ind.n
Setting the predicted negative intakes to zero
y.pred[k]<- max(0, y.pred1[k])
}
mu.ind.n ~ dnorm(a, prec.ind)
Obtaining the predicted intakes for the five food products for a new individual
y.pred1[ 1:5] ~ dmnorm( mu.n[], prec.food[,] )
}
```

B.2 Latent Gaussian Model for Concetrations

```
{
for( i in 1:700)
{
Concentration on the i^th sample of one of the five food products
f[i] ~ dnorm(mu, delta)I(, f.cen[i])
}
PRIORS
mu ~dnorm(0,0.1)
delta ~ dgamma(1,1)
sigma <- sqrt(1/delta)

To predict the concentrations on one of the food products
f.pred1 ~ dnorm(mu, delta)
f.pred <- max(0,c.pred1)
}
```

B.3 Latent t Model for Concentrations

```
{
for( i in 1:700)
{
```

Concentration on the i^{th} sample of one of the food products

```
c[i] ~ dt(mu,tau,2)I(, f.cen[i])
}
```

PRIORS

```
mu ~ dnorm( 0,0.1)
tau ~ dgamma(1,1)
```

To predict the concentrations on one of the food products

```
f.pred1 ~ dt(mu,tau,2)
f.pred <- max(0,c.pred1)
}
}
```

Appendix C

WinBUGS Codes for Models in Chapter 4

C.1 Mixture Model

```
{
for ( i in 1 :  2197)
{
for ( j in 1:7)
{
```

Response for the i^{th} individual on the j^{th} day

```
va[i,j] ~ dnorm(mu[i,j], tau[i,j])
mu[i,j] <- lambda[gr[i],s[i,j]]
tau[i,j] <- delta[s[i,j]]
```

To determine which Normal distribution the intakes come from using the mixing probability pi[i,]

```
s[i,j] ~ dcat (pi[i,])
}
y[i] ~ dnorm(muy, tauy)
```

Logit transformations for the mixing probabilities

```
pi[i,1] <- exp(y[i])/(1+exp(y[i]))
pi[i,2] <- pow(1+exp(y[i]), -1)
}
```

PRIORS

```
omega ~ dnorm(20,0.01)
for(g in 1:8)
{
theta[g] ~ dnorm(200, 0.0001)
```

We impose an identifiability constraint on the means of the Normal distribution

```
lambda[g,2] <- lambda[g,1] + theta[g]
lambda[g,1] ~ dnorm(omega, 0.02)
}
delta[1] ~ dgamma(1, 100)
ratio ~ dbeta(1.1, 10)
```

We impose an identifiability constraint on the variances of the Normal distribution

```
delta[2] <- delta[1]* ratio
muy ~ dnorm(1, 0.02)
tauy ~ dgamma(1,100)
```

To generate retinol intake for a new individual in the sex-by-age group k

```
for(g in 1:8)
{
max.n[g,1] <- 0
y.n[g] ~ dnorm(muy, tauy)
pi.n[g,1] <- exp(y.n[g])/(1+exp(y.n[g]))
pi.n[g,2] <- pow(1+exp(y.n[g]) , -1)
```

To simulate intakes in the future

```
for (n in 1:7)
{
s.n[g,n] ~ dcat(pi.n[g, ])
mu.n[g,n] <- lambda[g, s.n[g,n]]
tau.n[g,n] <- delta[s.n[g,n]]
va.n1[g,n] ~ dnorm(mu.n[g,n], tau.n[g,n])
```

Back-transformed predicted retinol intakes for a new individual

```
va.n[g,n] <- va.n1[g,n]*va.n1[g,n]
```

To obtain maximum retinol intake in a week

```
max.n[g,n+1] <- max(va.n[g,n] , max.n[g,n])
}
}
}
```

C.2 Mixture Model with Markov Dependence Between Days

```
{
for ( i in 1:2197)
{
```
Response for individual i on day j
```
va[i,1] ~ dnorm(mu[i,1], tau[i,1])
mu[i,1] <- lambda[gr[i],s[i,1]]
tau[i,1] <- delta[s[i,1]]
s[i,1] ~ dcat(pi[i, ])
```
Stationary probabilities for individual i
```
pi[i,1] <- (p[i,2,1])/(p[i,1,2]+p[i,2,1])
pi[i,2] <- pow(1+exp(pi[i,1], -1))
```
Transitional probabilities between the Normal distributions for individual i
```
for (k in 1:2)
{
p[i,k,1] <- exp(y[i,k])/(1+exp(y[i,k]))
p[i,k, 2] <- pow(1+exp(p[i ,k, 1]))
y[i,k] ~ dnorm(muy[k], tauy[k])
}
```
Response for individual i on day 1
```
for( j in 1:6)
{
s[i,j+1] ~ dcat(p[i, s[i,j], ])
mu[i,j+1] <- lambda[ gr[i],s[i,j+1]]
tau[i,j+1] <- delta[s[i,j+1]]
va[i,j+1] ~ dnorm( mu[i,j+1], tau[i,j+1])
}
}
```
PRIORS
```
delta[1] ~ dgamma(1, 100)
ratio ~ dbeta(1.1, 10)
delta[2] <- delta[1]*ratio
tauy[1] ~ dgamma(1,100)
tauy[2] ~ dgamma(1,100)
sigma[1] <- sqrt(1/delta[1])
sigma[2] <- sqrt(1/delta[2])
```

```
a0 ~ dnorm(20, 0.01)
muy[1] ~ dnorm(1,0.02)
muy[2] ~ dnorm(1,0.02)
```

To predict retinol intake for a new individual in sex-by-age group g

```
for(g in 1 :8)
{
lambda[g,1] ~ dnorm( a0, 0.02)
lambda[g,2] <- lambda[g,1] + theta[g]
theta[g] ~ dnorm( 200, 0.0001)
}
for(g in 1:8)
{
for(k in 1:2)
{
y.n[g,k] ~ dnorm(muy[k], tauy[k])
p.n[g,k,1] <- exp(y.n[g,k])/(1+exp(y.n[g,k]))
p.n[g,k,2] <- pow(1+exp(p.n[g,k,1]))
}
}
for (g in 1:8)
{
max.n[g,1] <- 0
```

To predict retinol intake over a week

```
for( n in 1:7)
{
s.n[g, n+1] ~ dcat(p.n[g, s.n[g,n], ])
mu.n[g, n+1] <- lambda[g, s.n[g,n+1] ]
tau.n[g, n+1] <- delta[s.n[g,n+1]]
va.n[g,n+1] ~ dnorm(mu.n[g,n+1], tau.n[g, n+1])
```

Back-transformed predicted retinol intakes for a new individual

```
ac.va[g,n+1] <- va.n[g,n+1]*va.n[g,n+1]
```

Maximum predicted retinol intake over a week

```
max.n[g,n+1] <- max(va.n[g, n], max.n[g,n])
}
```

Stationary probability for the new individual

```
pi.n[g,1] <- p.n[g,2,1]/(p.n[g,2,1]+p.n[g,1,2])
```

```
pi.n[g,2] <- pow(1+exp(pi.n[g,1]))
s.n[g,1] ~ dcat(pi.n[g, ])
mu.n[g,1] <- lambda[g,s.n[g,1]]
tau.n[g,1] <- delta[s.n[g,1]]
va.n[g,1] ~ dnorm(mu.n[g,1] , tau.n[g,1])
```
Predicted retinol consumption on the first day of the future time period
```
ac.va[g,1] <- va.n[g,1]*va.n[g,1]
}
}
```

Bibliography

Ahmed, F., Azim, A. & Akhtaruzzaman, M. (2002), 'Vitamin A deficiency in poor, urban, lactating women in Bangladesh: factors influencing Vitamin A status', *Public Health Nutrition* **6**, 447–452.

Alcohol Concern Factsheets (2003), 'Binge drinking'. Factsheet: 20.

Alcoholis (2002), 'Statistical notes', *Medical Council on Alcohol newsletter* **21**, Issue No 3.

Allcroft, D. J. & Glasbey, C. A. (2003), 'Analysis of crop lodging using a latent variable model', *Journal of Agricultural Science* **140**, 383–393.

Allcroft, D. J., Glasbey, C. A. & Paulo, M. J. (2005), 'A latent Gaussian model for multivariate consumption data'. Submitted.

Amemiya, T. (1984), 'Tobit models: A survey', *Journal of Econometrics* **24**, 3–61.

Anonymous & Nederlan, Z. (1998), 'Resultaten van de voedselconsumptiepeiling 1997-1998 [Result of Dutch national food consumption survey 1997-1998]', *Voedingscentrum, DenHaag*.

Arves, P. & Choquet, M. (1999), 'Regional variations in alcohol use among young people in France. Epidemiological approach to alcohol use and abuse by adolescents and conscripts', *Drug and Alcohol Dependence* **56**, 145–155.

BBC (2002), 'BBC news'. http://news.bbc.co.uk/2/hi/health/3402891.stm.

Berk, K. N. & Lachenbruch, P. A. (2002), 'Repeated measures with zeroes', *Statistical Methods in Medical Research* **11**, 303–316.

Bertillon (1863), 'Bulletin de la sociétè d'anthropologie'. Cited in Livi(1883).

Besag, J. & York, J. (1989), *Bayesian restoration of images. In Analysis of Statistical Information (ed. T. Matsunawa)*, Tokyo: Institute of Statistical Mathematics.

Boon, E. P., Lignell, S., van Klaveren, J. D. & Tjoe Nij, I. M. E. (2004), 'Estimation of acute dietary exposure to pesticides using the probabilistic approach and the point estimate methodology', *RIKILT Institute of Food Safety*.

Brooks, S. P. (1998), 'Markov chain Monte Carlo method and its application', *The Statistician* **47**, 69–100.

Buck, R. J., Hammerstorm, K. A. & Ryan, P. B. (1995), 'Estimating long-term exposures from short-term measurements', *Journal of Exposure Analysis of Environment Epidemiology* **5**, 359–373.

Carriquiry, A. L. (2003), 'Estimation of usual intake distribution of nutrients and foods', *American Society of Nutritional Sciences* **133**, 601S–608S.

Chavas, J. P. & Kim, K. (2004), 'A heteroskedastic multivariate tobit analysis of price dynamics in the presence of price floors', *American Journal of Agricultural Economics* **86**, 576–589.

Coles, S. (2001), *An Introduction to Statistical Modeling of Extreme Values*, London: Springer.

Committee on Toxicity of Chemicals in Food, Consumer Products and the Environment (2002), 'Risk assessment of mixtures of pesticides and similar substances'. http://www.food.gov.uk/multimedia/pdfs/report(indexed).pdf.

Cornick, J., Cox, T. L. & Gould, B. W. (1994), 'Fluid milk purchases: A multivariate tobit analysis', *American Journal of Agricultural Economics* **76**, 74–82.

Cowles, M. K. & Carlin, B. P. (1996), 'Markov chain Monte Carlo convergence diagnostics: A comparative review', *Journal of the American Statistical Association* **91**, 883–904.

Criqui, H. M., Cowman, D. L., Tyroler, A. H., Bangdiwala, S., Heiss, G., Wallace, R. B. & Cohn, R. (1987), 'Lipoproteins as mediators for the effects of alcohol consumption and cigarette smoking on cardiovascular mortality: results from the lipid research clinics follow-up study', *American Journal of Epidemiology* **126**, 629–637.

Davison, A. C. & Smith, R. L. (1990), 'Models for exceedances over high thresholds', *Journal of the Royal Statistical Society, B* **52**, 393–442.

Doll, R. (1997), 'Alcohol and the reduction of mortality', *Journal of Clinical Research* **3**, 8–9.

Expert Group on Vitamins and Minerals (2002), 'Revised review of Vitamin A'. http://www.food.gov.uk/multimedia/pdfs/reviewvita.pdf.

Extenion Toxicology Network, USA (1992), 'Iprodion'. http://pmep.cce.cornell.edu/profiles/extoxnet.

Ferrier, H., Nieuwenhuijsen, M., Boobis, A. & Elliott, P. (2002), 'Current knowledge and recent developments in consumer exposure assessment of pesticides: A UK perspective', *Food Additives and Contaminants* **19**, 837–852.

Food Standards Agency (2001), 'Vitamin A'. http://www.food.gov.uk/healthiereating/vitaminsminerals/vitsminaz/vitamin-a.

Frezza, M., di Padova, C., Pozzato, G., Terpin, M., Baraona, E. & Lieber, C. (1990), 'High blood alcohol levels in women: The role of decreased gastric alcohol dehydrogenase activity and firstpass metabolism', *New England Journal of Medicine* **322**, 95–99.

Gay, C. (2000), 'Estimation of population distributions of habitual nutrient intake based on a short-run weighed food diary', *British Journal of Nutrition* **83**, 287–293.

Gelman, A. & Rubin, D. (1992), 'Inference from iterative simulation using multiple sequences', *Statical Science* **7**, 457–511.

Gelman, A., Carlin, J., Stern, H. & Rubin, D. (2003), *Bayesian Data Analysis*, second edn, New York: Chapman & Hall.

Geman, S. & Geman, D. (1984), 'Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

Geyer, C. J. (1992), 'Practical Markov chain Monte Carlo', *Statistical Science* **7**, 473–511.

Guardian (Nov 2002), 'Distilling the truth about drinking', pp. 24–25.

Guenther, P. M., Kott, P. S. & Carriquiry, A. L. (1997), 'Development of an approach for estimating usual nutrient intake distributions at population level', *Journal of Nutrition* **127**, 1106–1112.

Hamey, P. Y. (2000), 'A practical application of probabilistic modelling in assessment of dietary exposure of fruit consumers to pesticide residues', *Food Additives and Contaminants* **17**, 601–610.

Hamilton, H. R. (1999), 'HMO selection and Medicare costs: Bayesian MCMC estimation of a robust panel data tobit model with survival', *Health Economics and Econometrics* **8**, 403–414.

Hastings, W. K. (1970), 'Monte Carlo sampling methods using Markov chains and their applications', *Biometrika* **57**, 97–109.

Hoffmann, K., Boeing, H., Dufour, A., Volatier, L., Telman, J., Virtanen, M., Becket, W. & De Henauw, S. (2002), 'Estimating the distribution of usual dietary intake by short-term measurements', *European Journal of Clinical Nutrition* **56**, S53–S62.

Insightful Corporation (2002), *S+FinMetrics, Reference Manual*, Version 1.0, Insightful Corporation.

Institute of Alcohol Studies (May 2004), 'Alcohol– Consumption and Harm in UK and EU'. http://www.ias.org.uk/factsheets/harm-ukeu.pdf.

Iowa State University (2002), 'Software for Intake Distribution Estimation'. http://cssm.iastate.edu/software/cside.html.

IPCS (1977), 'International Programme on Chemical Safety: Iprodione'. http://www.inchem.org.

Johnson, N. L., Kotz, S. & Kemp, A. W. (1992), *Univariate discrete distributions*, second edn, New York: John Wiley.

Karkeck, J. (1987), 'Improving the use of dietary survey methodology', *Journal of American Dietetic Association* **87**, 869–871.

Kijima, M. (1997), *Markov Processes for Stochastic Modeling*, London: Chapman & Hall.

Kim, W. W., Mertz, W., Judd, J. T., Marshall, M. W., Kelsay, J. L. & Prather, E. S. (1984), 'Effect of making duplicate food collections on nutrient intakes calculated from diet records.', *American Journal of Clinical Nutrition* **40**, 1333–1337.

Kistemaker, C., Bouman, M. & Hulshof, K. F. A. M. (1998), 'De consumptie van afzonderlijke producten door nederlandse bevolkingsgroepen voedselconsumptiepeiling', *TNO-Voelding, Zeist*.

Kroes, R., Müllet, D., Lambe, J., Löwik, M. R. H., van Klaveren, J., Kleiner, J., Massey, R., Mayer, S., Urieta, I., Verger, P. & Visconti, A. (2002), 'Assessment of intake from the diet', *Food and Chemical Toxicology* **40**, 327–385.

Kroke, A., Klipstein-Grobusch, K., Voss, S., Mösender, J., Thielecke, F., Noack, R. & Boeing, H. (1999), 'Validation of a self-administered food-frequency questionnaire administered in the European Prospective Investigation into Cancer and Nutrition (EPIC) Study: comparison of energy, protein and macronutrient intaks estimated with doubly labeled water, urinary nitrogen, and repeated 24-h dietary recall methods', *American Journal of Clinical Nutrition* **70**, 439–447.

Livi, R. (1883), 'Sulla statura degli italiani', *Archivio per l'antropologia e la etnologia* **13**, 243–290, 317–379.

Lunchick, C. (2001), 'Probabilistic exposure assessment of operator and residential non-dietary exposure', *Annals of Occupational Hygiene* **45**, S29–S42.

Marin, J., Mengersen, K. & Robert, J. P. (2005), *Bayesian modelling and inference on mixture distributions In Dey, D. and Rao, C.R., Eds. Handbook of Statistics Volume 25*, Elsevier Science.

McLachlan, G. & Peel, D. (2000), *Finite Mixture Models*, New York: John Wiley.

Melhus, H., Michaelsson, K. & Kindmark, A. (1998), 'Excessive dietary intake of Vitamin A is associated with reduced bone mineral density and increased risk from hip fracture', *Annals of Internal Medicine* **129**, 770–778.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M., Teller, A. H. & Teller, E. (1953), 'Equations of state ca;culations by fast computing machines', *Journal of Chemical Physics* **21**, 1087–1091.

Minitab (1972-2005), 'Release 14'. Statistical software.

Myles, J. P., Price, G. M., Hunter, N., Day, M. & Duffy, S. W. (2003), 'A potentially useful distribution model for dietary intake data', *Public Health Nutrition* **6**, 513–519.

National Institutes of Health (2001), 'Facts about dietary supplements: Vitamin A'. http://www.cc.nih.gov/ccc/supplements/intro.html.

Nusser, S. M., Carriquiry, A. L., Dodd, K. W. & Fuller, W. A. (1996), 'Estimating usual daily intake distributions', *Journal of the American Statistical Association* **91**, 1440–1449.

Nusser, S. M., Fuller, W. A. & Guenther, P. M. (1997), 'Estimating usual dietary intake distribution: Adjusting for measurement error and nonnormality in 24-hour food intake', *Survey Measurement and Process Quality* **eds. L. Lybergn, M. Collins, E. DeLeeuw, C. Dippo, W. Schwartz and D.Trewn, New York: Wiley**, 697–709.

Office of Dietary Supplements (2003), 'Vitamin A and Carotenoids'. http://ods.od.nih.gov/factsheets/cc/vita.html.

Palisade (2005), '@RISK 4.5'. www.palisade.com.

Paulo, M. J., van der Voet, H., Jansen, M., ter Braak, C. & van Klaveren, J. (2004*a*), 'Bayesian modelling of dietary intake of pesticides using monitoring and survey data'. Submitted.

Paulo, M. J., van der Voet, H., Wood, J., Marion, G. & van Klaveren, J. (2004*b*), 'Analysis of multivariate extreme intakes of food chemicals and nutrients'. in preparation.

Pearson, K. (1894), 'Contribution to the mathematical theory of evolution', *Proceedings Royal Society A* **185**, 71–110.

Petersen, J. B. (2000), 'Probabilistic modelling: theory and practice', *Food Additives and Contaminants* **17**, 591–599.

Sales, J., Duffy, J. C., Plant, M. & Peck, D. (1989), 'Alcohol consumption, cigarette sales and mortality in the United Kingdom: an analysis of the period 1970-1985', *Drug and Alcohol Dependence* **24**, 155–160.

Sempos, C. T., Johnson, N. E., Smith, E. & Gilligan, C. (1985), 'Effects of intra-individual and inter-individual variation in repeated dietary records', *American Journal of Epidemiology* **121**, 120–130.

Sieri, S., Agudo, A., Kesse, E., Klipstein-Grobusch, K. & San-Jose, B. (2002), 'Patterns of alcohol consumption in 10 European countries participating in the European Prospective Investigation into Cancer and Nutrition (EPIC) project', *Public Health Nutrition* **5**, 1287–1296.

Slob, W. (1993), 'Modelling long-term exposure of the whole population to chemicals in food', *Risk Analysis* **13**, 525–530.

Smith, R. L. (1989), 'Extreme value analysis of environmental time series: An application to trend detection in ground-level ozone', *Statistical Science* **4**, 367–393.

Spiegelhalter, D. J., Thomas, A., Best, N. & Lunn, D. (2004), 'Winbugs, version 1.4.1'. BUGS 1996–2004: Medical Research Council (MRC) UK, WinBUGS. http://www.mrc-bsu.ac.uk.bugs/winbugs/contents.html.

Spiegelhalter, D., Thomas, A., Best, N. & Lunn, D. (2003*c*), *Tutorial: How many iterations after convergence?*, WinBUGS User Manual.

Spiegelhalter, D., Thomas, A., Best, N. & Lunn, D. (Jan 2003*a*), 'Examples Volume I, stacks: robust regression'. WinBUGS user Manual, version 1.4.

Spiegelhalter, D., Thomas, A., Best, N. & Lunn, D. (Jan 2003*b*), 'Examples Volume II, jaw : repeated measure analysis of variance'. WinBUGS user Manual, version 1.4.

Stubbs, R. J., O'Reilly, L., Fuller, Z., Horgan, G. W., Mehere, C., Deary, I., Austin, E., Ritz, P., Miline, E. & Lames, W. P. T. (Aug 2001), 'Detecting and modelling mis-reporting of food intake with special reference to under reporting in the obese', *Rowett Research Institute*.

Suhre, F. B. (2000), 'Pesticide residues and acute risk assessment - the US EPA approach', *Food Additives and Contaminants* **17**, 569–573.

Titterington, D. M., Smith, A. F. M. & Makov, U. E. (1985), *Statistical Analysis of Finite Mixture Distributions*, New York: John Wiley.

Tobin, J. (1958), 'Estimation of relationships for limited dependent variables', *Econometrica* **26**, 24–36.

UK Data Archive (1987), 'SN 2836 - Dietary and nutritional survey of British adults, 1986 -1987'. http://data-archive.ac.uk.

United States Environmental Protection Agency (1998), 'Iprodione', *Reregistration eligibility decision facts*.

van Dooren, M. M. H., Boeijen, I., van Klaveren, J. D. & van Donkersgoed, G. (1995), 'Conversie van consumeerbare voedings middelen naar primaire agrarische produkten [Conversion of foods to primary agricultural products.]', *RIKILT, Wageningen*.

van Klaveren, J. D. (1999), 'Quality programme for agricultural products. Results residue monitoring in the Netherlands', *RIKILT Institue of Food Safety, Wageningen*.

Verger, P., Ireland, J., Møller, A., Abravicius, A. J., De Henauw, S. & Naska, A. (2002), 'Improvement of comparability of dietary intake assessment using current available individual food consumption surveys', *European Journal of Clinical Nutrition* **56**, S18–S24.

Voltier, L. J., Turrini, A. & Welten, D. (2002), 'Some statistical aspects of food intake assessment', *European Journal of Clinical Nutrition* **56**, S46–S52.

Wallace, L. A., Duan, N. & Ziegenfus, R. (1994), 'Can long-term exposure be predicted from short-term measurements?', *Risk Analysis* **14**, 75–85.

Webb, E., Ashton, C. H., Kelly, P. & Kamali, F. (1996), 'Alcohol and drug use in UK university students', *Lancet* **348**, 922–925.

Whittaker, J. (1990), *Graphical Models in Applied Multivariate Statistics*, Chichester: Wiley.

World Health Organization (1997), 'Guidelines for predicting dietary intake of pesticide residues'. GEMS/Food in collaboration with codex committee on pesticide residues. Document WHO/FSF/FOS/97.7.

Zivot, E. & Wang, J. (2003), *Modeling Financial Time Series with S-Plus*©, New York: Springer.