UNIVERSITY OF CAMBRIDGE

MULTIVARIATE ANALYSIS OF DIETARY DATA IN THE PRESENCE OF EXCESS ZEROS AND MEASUREMENT ERROR

This thesis is submitted for the degree of Doctor of Philosophy

---

Author:

Ms Yulia
CHERNOVA-CHERNAYA
HUGHES HALL

Supervisor:

Dr Ivonne SOLIS-TRAPALA

July 2019

Declaration

The work described in this thesis was carried out at the Medical Research Council Human Nutrition Research (later Medical Research Council Elsie Widdowson Laboratory until 2018) under the supervision of Dr Ivonne Solis-Trapala between October 2013 and June 2019. This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

Yulia Chernova-Chernaya

July 2019

Abstract
Multivariate analysis of dietary data in the presence of excess zeros and measurement error

Yulia Chernova-Chernaya

Nutritional epidemiology is a complex area of research plagued with bias and measurement error partly arising from the use of imperfect methods of dietary data measurement and inadequate data analyses. Some foods or nutrients are consumed habitually, while others are consumed occasionally. This thesis addresses three important public health problems: (1) Estimation of under and over consumption of foods consumed occasionally such as alcohol; (2) Investigation of the determinants of food intake, and (3) Estimation of the effect of the intake of occasionally-consumed foods on health outcomes.

The analysis of food intake is complicated because foods are often eaten in combination. This gives rise to multiple correlated habitually- and occasionally-consumed food intakes, characterised by a large proportion of zero observations. Modern statistical methods are required to deal with measurement error, excess zeros in the intake distributions, and correlated preferences for frequency of consumption and portion sizes across foods.

The thesis demonstrates the use of contemporary statistical methods, based on mixed-distributions and mixed-effects modelling approaches, for the analysis of a single and multiple correlated habitually and occasionally-consumed food intakes. These methods are complex due to the need to evaluate intractable integrals for parameter estimation.

Firstly, to describe under and over consumption, the thesis provides a new numerical approach, which is a quicker and simpler alternative to Monte Carlo simulation, to estimate the distributional quantiles of occasionally-consumed food intakes in predefined sub-populations.

Secondly, dietary data from the UK National Diet and Nutrition Survey Rolling Programme (NDNS RP), which provides the only source of current authoritative information on food

and nutrient intake in the UK, are analysed with a mixed-effects two-part model to estimate the associations between personal and socio-economic risk factors and the intake of several foods of current public health importance.

This is the first time this approach is applied to NDNS RP data to assess explicitly socio-economic and personal characteristics related to occasionally-consumed food intakes in the UK population.

Then, the thesis develops a novel multivariate joint model for several correlated occasionally-consumed food intakes utilising a pseudolikelihood approach and parametric bootstrap for parameter estimation. The approach is illustrated by modelling the intake of alcohol, jointly with the intake of other foods, and the resulting analysis is compared with that based on the two-part model for a single food and the traditional multivariable linear regression model widely used in nutritional epidemiology.

Finally, a regression calibration approach is applied to correct for the effect of excess zeros and multiple correlated person-specific preferences when several food intakes are investigated as predictors of health outcomes. Again the results are contrasted with those obtained by applying multivariable regression analysis which ignores excess zeros and correlated preferences, introducing potential bias in the effect estimates. As an example, the relationships between alcohol intake in a male sub-population of NDNS RP and haemoglobin A1c (HbA1c), a known predictor of type 2 diabetes mellitus, are investigated. Ideally, to obtain unbiased estimates of predictors' effects, all correlated unobserved person-specific effects should be accounted for. However, the task is incredibly complex and to tackle this the suggestion is to simplify the model by accounting only for the largest residual correlations. Even in this imperfect form, regression calibration produces markedly different results from multivariate linear regression.

Acknowledgment

Throughout the writing of this dissertation I have received a great deal of support and assistance. I would first like to thank my supervisor, Dr Ivonne Solis-Trapala, whose expertise was invaluable in the formulation of the research topic and methodology and who was patient, supportive and encouraging, which helped me greatly during the course of my work.

I would also like to thank Dr Angela Wood for her constructive feedback, which helped to improve the presentation of results.

I am also immensely grateful to my husband, Dr Rodion Skovoroda, for being always available for discussions and sharing generously his time and support with me.

Yulia Chernova-Chernaya

July 2019

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

What we eat might affect us in numerous ways. While the consequences of extreme cases of nutrient deficiencies or overload are well recognised (vitamin C deficiency and scurvy; extreme iodine deficiency and cretinism), the effects of more subtle variations in people's usual diets are subject to debate. Ioannidis (2013) argues that "Almost every single nutrient imaginable has peer reviewed publications associating it with almost any outcome". Statements like this reflect recognition that observational studies, the major source of information on diet and health outcomes, are prone to various biases (Fraser, 2003).

One of the major biases in observational research is confounding (Rochon et al., 2005; Mamdani et al., 2005; Normand et al., 2005), which, if not accounted for, can lead to erroneous conclusions regarding the effect of the risk factor of interest on a health outcome. Part of the answer to this problem lies in conducting more well designed and well executed randomised control trials (RCTs). A well known example of a well-conducted RCT in the area of nutrition is the original Dietary Approaches to Stop Hypertension (DASH) trial (Appel et al., 1997), which linked increase in fruits, vegetables and low-fat dairy products

consumption with lowering of blood pressure. However, often RCTs are not feasible: they are too expensive or not ethical to conduct (when the exposure is potentially harmful), then public health policy making has to rely on information obtained from observational studies. Observational studies, therefore, form the basis of information we rely on when linking health outcomes and diet. It is of utter importance that confounding should be thoroughly considered at a design stage of a study and all possible measures should be taken to record and then adjust for confounding factors at the analysis stage.

Missing values can also impact the inference drawn from research in several undesirable ways. For example, if a group of potential survey respondents with certain demographic or social characteristics refuse survey participation then, in case these characteristics are related to food consumption, the estimated food intake distribution will be biased away from the distribution specific to this group. Hence every step should be taken to acquire data from various demographic populations to minimise bias. Some degree of missing information may be unavoidable, in this case it is recommended to document the reasons for this and compare the characteristics of the individuals who have missing records with those from individuals with complete information. When the primary outcome has missing records and these are thought to be missing at random, a likelihood method of estimation will yield consistent and efficient estimates, whereas some form of multiple imputation may be appropriate if explanatory variables are thought to be missing at random.

Other sources of bias in the realm of Nutritional Epidemiology arise from limitations of current instruments available to capture usual dietary intake, which is defined by the Institute of Medicine as the individual's average intake over a long period of time (Institute of Medicine, Food and Nutrition Board, 2003). The most commonly used dietary assessment tools rely on self-report and these are food frequency questionnaires (FFQ), food diaries (FD) and 24 hour food recalls (24HR).

Data collected by self-reporting are prone to systematic bias that arise when people consistently under- or over-report their food intake, certain characteristics like age, education

and BMI are linked to misreporting (Braam et al., 1998; Freedman et al., 2014). Method-comparison and biomarker-validation studies indicate that FFQs are less reliable compared to multiple-days FD and multiple 24HR (Day et al., 2001; Burrows et al., 2010; Bingham et al., 1994, 1997; McKeown et al., 2001). Recent studies indicate that multiple web-based diet 24-hour recalls can provide a convenient and modern alternative to paper-based diaries with the reliability comparable to multiple 24HR and multiple days FD (Arab et al., 2011). Hereafter, FFQs are left out of the scope of this work and the methods developed in this work apply to data collected from multiple days food diaries and multiple 24HR recalls. This work focuses on minimising the impact of random measurement error in nutritional exposure which, unlike systematic measurement error, is a result of random variation in daily food consumption and short observational period.

Multiple days diaries and 24HRs records reflect information on both, long-term individual usual intake and daily within-subject variations in food consumption. When the observational period is reduced, as is the case in the majority of studies, observational error, defined as the difference between the measured diet and its true value can be quite large (Rutishauser and Black, 2002; Beaton et al., 1979; Nelson et al., 1989; Sempos et al., 1985; Borrelli et al., 1992). The presence of observational error, referred to as measurement error in the published literature, leads to certain implications in the analysis and interpretation of available data.

1.2 Measurement error

First, we describe the classical additive measurement error model for a single continuous intake T in the context of habitually consumed foods, *i.e.* those foods consumed most days. Suppose that T_i is the true long-term consumption of person i, $i = 1, ...m$. Due to the limitations in food intake measurement tools, on day j, $j = 1, ...n$, we are unable to observe T_i but instead observe R_{ij}, which contains information on both, T_i and daily within-person variation ϵ_{ij}. Under the assumptions of the classical additive measurement error model, the relation between the true intake T_i and daily variation ϵ_{ij} is additive and

$R_{ij} = T_i + \epsilon_{ij}$. Thus, ϵ_{ij} introduces random noise around the true long-term individual intake T_i. We assume that the random variable ϵ_{ij} does not depend on the true T_i and $\mathrm{E}(\epsilon_{ij}) = 0$. It can also be reasonable to assume that the $\epsilon_{ij}s$ are identically independently distributed (i.i.d.) and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_w^2)$. This assumption is called homogeneity of the errors. Therefore, it follows that R_{ij} is an unbiased estimate of T_i: $\mathrm{E}(R_{ij}) = T_i$ and $\mathrm{Var}(R_{ij}) = \sigma_w^2$. The mean of observed individual dietary records is often used as a measure of the true individual intake. Consider an individual average $\overline{R_i}$ obtained over the available number of diary days n so that $\overline{R_i} = \frac{1}{n}(R_{i1} + R_{i2} + ... + R_{in})$, it can be seen that $\overline{R_i}$ is also an unbiased estimate of the long-term personal true intake T_i with variance $\mathrm{Var}(\overline{R_i}) = \frac{1}{n}\sigma_w^2$. Increasing the number of observed days will decrease the variation of $\overline{R_i}$ around T_i, however, the required number of diary days to achieve certain precision will depend on the magnitude of the daily variation.

As an example, Basiotis et al. (1987) showed that in a study of 13 men over 1 year, in order to estimate the true individual usual niacin intake with 10 percent accuracy, it was required to collect 53 record days, and 249 days of intake data were needed to estimate usual vitamin C intake with the same precision. Clearly, this number of diary records is impossible to obtain in the majority of research studies and hence, the presence of measurement error needs to be considered and its consequences evaluated, and, when possible, minimised. We next consider how the presence of non-negligible daily variation in nutritional data can impact research findings in various contexts.

1.2.1 Monitoring food consumption in populations of interest

Monitoring habitual long-term dietary intake to assess nutrient adequacy of a population is of public health interest as it can highlight those subgroups of the population which are particularly prone to consuming a less healthy diet, similarly it helps to monitor the adherence of subpopulations to dietary recommendations. However, unaccounted measurement error can distort the estimates of food intake distributions in subgroups. Let's consider an example where the interest lies in estimating a food intake in a group of in-

dividuals with a true group mean intake μ and a true between-person variance σ_b. If we independently draw a representative sample of individuals from the target population and somehow learn their true habitual intakes $T_1, ..., T_m$ then the sample mean and the sample variance would give us unbiased estimates of μ and σ_b. If, instead, we draw our conclusions based on $\overline{R_1}, ..., \overline{R_m}$ as measured proxies for true habitual intake of these individuals, then we will obtain an unbiased estimate of the group mean $\mathrm{E}(\overline{R}) = \mu$, but the estimate of the group variance will be artificially inflated by σ_w^2/n, because $\mathrm{Var}(\overline{R}) = \sigma_b^2 + \frac{1}{n}\sigma_w^2$. Thus, utilising individual averages $\overline{R_i}$ to characterise the group distribution will produce an unbiased estimate of the group mean μ but increase the group variance reflecting both, between-person and daily within-person variation. The effect will be especially pronounced at the tails of the distribution leading to biased estimates of over- or under-consumers. This, potentially, can lead to implementing inefficient public health policies.

1.2.2 Current statistical approaches dealing with measurement error

Dodd et al. (2006) provided a review of statistical methods available to account for within-person variation when estimating the distribution of usual dietary intake within a population group using individual means. The methods make certain assumptions on the distribution of intake and the relationship between the true and observed intakes. First, unbiasedness of intake on the original scale or a transformed scale (to obtain a symmetric distribution) is assumed , so that $R_{ij} = T_i + \epsilon_{ij}$, or $R_{ij}^* = T_i^* + \epsilon_{ij}$, where R_{ij}^* and T_i^* are transformed recorded and true intakes respectively. Then the total group variance of the food intake distribution is decomposed into within-person (σ_ϵ^2, daily) and between-person (true) parts. Using the estimated between-person variance and assuming an approximately Gaussian distribution for food intake (on the original or transformed scale), we can obtain individual approximations to the true food intake, *i.e.* intermediary individual values, leaving out measurement error. These intermediary individual values are back-transformed to the original scale and used instead of observed intake. Their distribution on the original scale should have approximately the same mean and between-person vari-

ance as the original food intake distribution. To carry out these type of analysis, multiple records of food intake per individual are required. We briefly summarise the main points and methods of the review paper below.

The Institute of Medicine (Institute of Medicine, Food and Nutrition Board, 2003) suggested using a power or log transformation of the original data, estimating within– and between–person variances on the transformed data, constructing intermediary values on the transformed scale, where intermediary values are weighted averages of transformed individual means and transformed group means, and using the inverse of the original transformation to get a new set of values from which the distributional parameters can be assessed. Intermediary individual values are constructed so that more weight is given to the group mean if within–person variance is large and, vice versa, smaller within–person variance results in more weight given to individual means.

Iowa State University (ISU) method (Guenther et al., 1997) assumes unbiasedness on the original scale, relaxes the homogeneity assumption of measurement error and allows its variance to vary from individual to individual, adjusts for seasonality and allows adjustment for complex survey designs. It applies a two-stage transformation and the constructed intermediary values are based on theoretical quantiles of the normal distribution instead of individual means. The inverse to the initial two-stage transformation is applied and a new set of intermediary values on the original scale is used to make inference about the group intake distribution.

Nusser et al. (1996) also suggested **the Best Power method**, which is a simplified version of the Iowa State University method. It is also applicable to complex surveys, it does not allow the within-person variance to vary across individuals but applies only one transformation so that back-transformation and adjustment for transformation-induced bias is easier that in the ISU method.

More recently, Tooze et al. (2010) suggested utilising a **mixed-effects modelling approach** (Diggle et al., 2002) without requiring to reduce the data to individual averages.

Instead, the observed intake R_{ij} (or observed transformed intake R^*_{ij}) is modelled via a linear combination of observed covariates (and/or factors) such as age, education, day of the week or season $X^1_{ij}, ..., X^k_{ij}$, unobserved person-specific effects u_i that are impractical, impossible, very expensive to measure or simply unknown, and daily variation ϵ_{ij}, so that the observed intake can be recorded as $R_{ij} = \mathbf{X'_{ij}}\beta + u_i + \epsilon_{ij}$ or, on a transformed scale, $R^*_{ij} = \mathbf{X'_{ij}}\beta + u_i + \epsilon_{ij}$. For a specific person i unobserved preferences u_i are constant but in a group of people these unobserved preferences introduce between-person variation additional to and unaccounted for by observed measured covariates $X^1_{ij}, ..., X^k_{ij}$. In mathematical terms, u_i is represented by a random variable, typically, normally distributed $u_i \sim \mathcal{N}(0, \sigma^2_u)$. As above, ϵ_{ij} represents random daily variation and is assumed to have mean 0, variance σ^2_w, realisations of ϵ_{ij} are uncorrelated with each other, with the true individual intake T_i (T^*_i) and with u_i. A maximum likelihood approach is used to estimate the model parameters. When the estimates of the model parameters are obtained they can be utilised to simulate a new set of data that should be further back-transformed to the original scale. This new back-transformed dataset is then used to empirically estimate the distributional parameters of the true intake in the original group. Several applications of this method can be found in the literature (Tooze et al., 2010; Guenther et al., 2006) and it forms a part of **National Cancer Institute (NCI) method** (National Cancer Institute, 2015).

The methods described above are suitable for describing the intake distribution of habitually consumed foods. Each method has its own advantages and limitations, and these are summarised in Table 1.1. For example, **The Institute of Medicine** is easy to apply and easy to programme relative to **ISU methods**. However, occasionally consumed foods are characterised by a high frequency of zero intake records, which presents further challenges in analysis. The methods of dealing with within-person variance that assume that food intake can be transformed to be approximately Gaussian using a monotone function are not directly applicable to zero-inflated data. This distributional assumption is clearly violated for occasionally consumed foods. For food intakes with multiple zero-consumption

days, **The ISU modified method** should provide more reliable estimates of usual intake as it accounts for the probability of consumption, which is an important method's feature for food with many zero daily records. Chapter 2 of this thesis discusses a method to estimate the distribution of occasionally consumed foods in subgroups and proposes a new approach that can further simplify the application of the method.

Table 1.1: Summary of available methods adjusting for measurement error in estimation of usual food intake in a population

	METHODS			
	IOM	**ISU**	**BP**	**ISU Modified**
Transformation of the original observed values so that the distribution of transformed data is approximately normal	Power or log transformation	Two-stage transformation	Power or log transformation	Two-stage transformation Note: Transformation is applied to non-zero values of observed daily intake only
Adjustment for covariates such as age groups, season, day or the week	Not allowed	Allowed	Allowed	Allowed
Adjustment for probability of consumption (important for food intakes, which are not consumed daily)	No	No	No	Yes Estimates distribution of consumption probability based on observed consumption frequency. Does not allow for correlation between consumption probability and portion size
Assumption of unbiasedness of usually intake	On transformed scale	On untransformed scale	On untransformed scale	On untransformed scale Applies to a **non-zero** 24-hour recall only
Within-person variance	The same for all individuals	Can vary between individuals	The same for all individuals	Can vary between individuals
Intermediary values are constructed on	Transformed scale	Transformed scale	Transformed scale	Original scale
Back-Transformation function	Inverse of initial power or log transformation	Inverse of the initial two-stage normality transformation coupled with bias-adjustment before back-transformation	Inverse of initial power or log transformation coupled with bias-adjustment before back-transformation	Inverse of the initial two-stage normality transformation coupled with bias-adjustment Black-transformation applied to **non-zero** values of food intake only

Table 1.1 Summary of available methods adjusting for measurement error in estimation of usual food intake in a population (Continued)

	IOM	**ISU**	**BP**	**ISU Modified**
Estimation of intermediary values involves additional steps other than elimination of within-person variability	No	No	No	Yes Mathematically combines the estimated distribution of the original-scale daily usual intake with the estimated distribution of consumption probability to obtain a set of intermediary values
Estimated usual intake distribution is	Empirical distribution of back-transformed intermediary values	Empirical distribution of original-scale intermediary values	Empirical distribution of original-scale intermediary values	Empirical distribution of original-scale intermediary values **Assumption:** Usual intake is the probability to consume on a given day multiplied by the usual portion size for a day the food is consumed. Probability of consumption and portion size is assumed to be uncorrelated
IOM - The institute of Medicine **ISU** – Iowa State University **BP** – Best Power **ISU Modified** – Iowa State University Modified				

1.3 Determinants of food intake

Another extensive research area where the application of correct modelling techniques and accounting for measurement error is of great importance is the area of investigation of determinants of food intake.

Lifestyle and, in particular, food choices are important health determinants in today's world of plenty of food and reduced physical activity. The abundance of easily available high-fat, high-sugar, cheap and palatable food, and sedentary lifestyle have led to increasing rates in obesity in the developed world (Lobstein et al., 2004). Obesity has been implicated in a number of health conditions such as insulin resistance and metabolic syndrome, which increase the risk of cardiovascular disease and type 2 diabetes (T2D) (Cloostermans et al., 2015; Tuomilehto et al., 2001). The rising obesity rates lead to a larger burden on health care costs (Kent et al., 2017) to society. For individuals, obesity can lead to detrimental quality of life and reduced life expectancy (Katzmarzyk et al., 2003; Chan et al., 2015). Diet is an important target for public health interventions. However, to slow down the trend, more information is needed on the driving forces behind people's preferences for certain foods. This knowledge can inform the design and implementation of effective intervention strategies for healthier lives (Appleton et al., 2016).

Factors that can influence an individual's food consumption choices vary and include, but are not restricted to, genetics, pre-natal exposures, child feeding practices, cultural background and family habits, peers influence, surroundings, marketing and various socio-economic factors.

For example, human genetic variation can influence people's food choices (Feeney et al., 2011) and there is a wide variation in how well humans can detect bitter compounds phenylthiocarbamide (PTC) and propylthiouracil (PROP) (Bartoshuk et al., 1994). The population can be broadly described as super tasters (high sensitivity to PROP), medium tasters (medium sensitivity to PROP) and non-tasters (lowered sensitivity to PROP) making up 20, 50 and 30 percent of the population respectively (Tepper et al., 2009). This

taste sensitivity to PROP can be partially explained by genetic variations in the TAS2R genes family, which links people with certain genetic profiles to dislike of bitter tastes (Bufe et al., 2005; Kim and Drayna, 2005; Tepper et al., 2014).

This, in turn, can influence food choices with carriers avoiding certain bitter tasting foods like cruciferous vegetables, soya or green tea (Tepper et al., 2009), or increasing fat and sugar consumption (Feeney et al., 2011). However, it has been reported that this genetic influence is most prominent during childhood and becomes less important with age (Mennella et al., 2005; Navarro-Allende et al., 2008).

This suggests that childhood preferences can be modified through environmental factors. Summarising the available evidence, Mennella (2014) concluded that taste differentiation, unrelated to genetic profile, can start in the uterus, when foetuses are exposed to various taste components of amniotic waters, then during the lactating period through the taste of mother's milk and, during weaning time, through repeated exposures, provision of healthy snacks and mother's modelling behaviour (Nicklaus, 2011; Cooke and Fildes, 2011; McGowan et al., 2012).

Cultural norms further shape children's developing food preferences. For example, Rozin et al. (2011) shows that French and American cultures, broadly speaking, have different attitudes towards food. Americans, as opposed to French, prefer quantity to quality, comfort to uniqueness and variety and abundance to moderation when it comes to food choices. Furthermore, lower income and socio-economic status (SES) has been found to correlate with less healthy diet in a number of studies (Darmon and Drewnowski, 2008; Conklin et al., 2014; Galobardes et al., 2001; Groth et al., 2001; Lallukka et al., 2006; Pomerleau et al., 1997; Rydén and Hagfors, 2011; Turrell et al., 2003; Maguire and Monsivais, 2015). However, food is just one of the goods that people choose to buy and the same disposable income can be distributed differently by people with different social and cultural backgrounds, and different educational levels (Galobardes et al., 2001; Groth et al., 2001; Lallukka et al., 2006; Rydén and Hagfors, 2011; Ranjit et al., 2015).

Additionally to socio-economic and cultural backgrounds, psychology (Smeets et al., 2010) and physiology play their parts in food preferences, and stress and sleep deprivation can be related to food choice too (Yau and Potenza, 2013; Sinha and Jastreboff, 2013; Dashti et al., 2015; Tomiyama et al., 2012).

These and other potentially important factors need to be considered and investigated further to understand their role as determinants of food intake in specific populations. This will enable public health policy makers to develop and implement more targeted and, consequently, more effective programmes to help people live healthier lives.

Appleton et al. (2016) recently showed that the majority of published interventions with the aim to increase vegetable consumption had little significant change in vegetable consumption habits. This might indicate a lack of understanding of the motives behind people's food choices or inefficiently chosen target populations as well as misapplied statistical modelling methods.

The use of correct modelling techniques and statistical methods is of great importance when eliciting the determinants of food intake. Probable consequences of incorrect modelling include a loss of power to detect an important determinant or, vice versa, falsely highlighting the importance of some factors that might be just a correlate of unobserved true determinants.

Chapter 3 applies the most novel methods (mixed-distribution mixed-effects modelling approach described in detail in Chapter 2) of statistical methodology to analyse three waves (2009-2012) of the National Diet and Nutrition Survey Rolling Programme (NDNS RP) data. The models adjust for measurement error, correlation arising from the repeated measurements on the same person, excess-zeros and unobserved correlated preferences.

Chapter 4 extends the two-part model introduced in the previous chapters and, based on composite likelihood theory (Varin et al., 2011; Lele and L. Taper, 2002), develops

an approach that takes into account the unobserved preferences between probability of consumption and portion size of a single food as well as multiple correlated food intakes. In particular, the two-part model is extended to a triple-wise model with the focus of estimating the effect of various determinants of alcohol intake. The results demonstrate that the composite likelihood (or pseudolikelihood) approach, which is used to combine the estimated triple-wise parts models, and a subsequent bootstrap tool to make inference about uncertainty of model parameters is an attractive approach that accounts, to a certain degree, for a correlation between various food intakes.

1.4 Estimating the association of the intake of multiple foods with health outcomes

Naturally, the interest in peoples' nutritional preferences arises from the potential impact of dietary choices on people's health beyond weight change (Willett, 2012; Malik et al., 2010; Appel et al., 1997; Gadgil et al., 2013). One of the most recent examples, where cumulative research evidence suggests that a particular nutrient is on the causal pathway to an adverse health condition, is the increased risk of developing cardiovascular disease with increased consumption of trans fatty acids (TFA) (Mozaffarian and Willett, 2007). However, establishing these types of connections can be extremely difficult due many factors like confounding, selection mechanisms, short observational periods and measurement errors which are inherent to dietary assessment methods and observational research.

Regression dilution is a well known phenomenon when the estimated relationship between an outcome and a single risk factor measured with error is biased towards the null. Extensive literature exists to provide methods to correct for regression dilution including Fuller (1980), Lindley (1953), Madansky (1959) and more recently Carroll et al. (2006), Buonaccorsi (2010) and Keogh and White (2014) who provide an extensive overview on the subject in the field of nutrition.

However, the effect of measurement errors in multiple correlated exposures is less stud-

ied. It has been shown that when multiple correlated factors are measured with error, the estimates of their effects on health outcomes can be biased both downwards and upwards and the size and the direction of the bias depend on the covariance structure of the covariates. This observation is directly relevant to data analysis in nutritional epidemiology as the intakes of certain foods and/or nutrients are correlated through personal preferences. The picture becomes even more complex when we consider occasionally-consumed foods, such as fish, nuts, certain vegetables, and alcohol, as diet diary records of these foods contain excess zeros making them unsuitable for analysis with traditional statistical tools.

A potential approach to address this problem is to investigate the usability of the regression calibration approach described by Carroll et al. (2006). The following notation will be used.

1.4.1 Notation

Let Y_i represent a health outcome, X_i the true continuous unobserved exposure, W_i the observed exposure measured with error and $\boldsymbol{Z_i}$ a vector of covariates measured without error for a person i. We highlight vector notation in boldface: for example, for multiple unobserved and observed exposures with errors, we will use $\boldsymbol{X_i}$ and $\boldsymbol{W_i} = \boldsymbol{X_i} + \boldsymbol{\epsilon_i}$ respectively, where $\boldsymbol{X_i}$ is a k-vector of multiple exposures for participant i, $\boldsymbol{X_i} = (X_{1i}, X_{2i}, ..., X_{ki})$. Denote Σ_{ab} a covariance matrix of random variables a, b and S_{ab} the corresponding sample covariance, and the variance of a random variable a as σ_a^2 and its expectation as μ_a. It is reasonable to assume that measurement error ϵ is such that $\mathrm{E}(\epsilon_i|X_i) = 0$ and $\epsilon_i|X_i \sim \mathcal{N}(0, \Sigma_\epsilon)$.

For illustration of the impact of measurement error on regression modelling, it is convenient to start with the method of ordinary least squares (OLS) as it is one of the most widely used analytic tools, and it is easy to show where bias occurs with assumptions of the classical additive measurement error.

1.4.2 Ordinary least squares linear regression and additive non-differential measurement error

The exposure is assumed to have been measured with additive error $\boldsymbol{W_i} = \boldsymbol{X_i} + \boldsymbol{\epsilon_i}$ and the health outcome Y_i can be related to the true $\boldsymbol{X_i}$ through a linear function:

$$Y_i = \beta_0 + \boldsymbol{\beta_x'} \boldsymbol{X_i} + \boldsymbol{\beta_z'} \boldsymbol{Z_i} + \delta_i$$

where δ_i is an error arising from Y_i, uncorrelated with $\boldsymbol{\epsilon_i}$ i.e the assumption of non-differential measurement error is valid, and $\boldsymbol{\beta_x}, \boldsymbol{\beta_z}$ are the vectors of corresponding regression coefficients.

When estimating the multivariate regression coefficients $(\boldsymbol{\beta_x}, \boldsymbol{\beta_z})$ with the OLS method and using $\boldsymbol{W_i}$ as proxy for $\boldsymbol{X_i}$, we can observe that the estimates of $(\boldsymbol{\beta_x}, \boldsymbol{\beta_z})$ are biased. The OLS *naive* regression coefficients $(\boldsymbol{\beta_{x^*}}, \boldsymbol{\beta_{z^*}})$ satisfy:

$$\begin{aligned}
\begin{pmatrix} \beta_{x^*} \\ \beta_{z^*} \end{pmatrix} &= \begin{pmatrix} \Sigma_{xx} + \Sigma_{\epsilon\epsilon} & \Sigma_{xz} \\ \Sigma_{zx} & \Sigma_{zz} \end{pmatrix}^{-1} \left(\begin{pmatrix} \Sigma_{xy} \\ \Sigma_{zy} \end{pmatrix} + \begin{pmatrix} \Sigma_{\epsilon\delta} \\ 0 \end{pmatrix} \right) \\
&= \begin{pmatrix} \Sigma_{xx} + \Sigma_{\epsilon\epsilon} & \Sigma_{xz} \\ \Sigma_{zx} & \Sigma_{zz} \end{pmatrix}^{-1} \left(\begin{pmatrix} \Sigma_{xx} & \Sigma_{xz} \\ \Sigma_{zx} & \Sigma_{zz} \end{pmatrix} \begin{pmatrix} \beta_x \\ \beta_z \end{pmatrix} + \begin{pmatrix} \Sigma_{\epsilon\delta} \\ 0 \end{pmatrix} \right) \quad (1.1)
\end{aligned}$$

It can be seen from (1.1) that when only a single exposure X_i is measured with error, such that X_i is uncorrelated with the other exposures measured without errors Z_i, and ϵ_i is uncorrelated with δ_i, then the well-known expression for the bias in the regression coefficient for X_i is obtained, $\beta_{x^*} = \lambda\beta_x$, where the coefficient λ, commonly known as reliability ratio (Fuller, 1987), is less than one: $0 < \lambda = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_\epsilon^2} < 1$ so that the estimate of the regression coefficient β_x by OLS is attenuated and this effect is commonly referred to as regression dilution bias. If validation data or repeatedly measured data are available on a subset of study participants, the reliability coefficient can be estimated with various methods (Frost and Thompson, 2000; Keogh and White, 2014).

Interestingly, in the case, when X_i and $\boldsymbol{Z_i}$ are correlated, the regression coefficients for $\boldsymbol{Z_i}$

will also be biased (Carroll et al., 1985). This has important implications for the analysis of the treatment effect in clinical trials (Carroll, 1989).

In the multivariate case, when $\boldsymbol{X_i}$ are correlated, the estimated naive regression coefficients $\boldsymbol{\beta_{x^*}}$, are not necessary attenuated to 0 and can even change signs opposite to those of the true coefficients as their estimation depends on the covariance structure of $\boldsymbol{X_i}$ (Fuller, 1987; Carroll et al., 2006).

The method-of-moments estimation has been proposed in (Fuller, 1987) to correct for bias when data are available to estimate the required covariances and when ϵ and δ are uncorrelated (the assumption of non-differential measurement error), the method-of-moments estimator is:

$$\begin{pmatrix} S_{ww} - S_{\epsilon\epsilon} & S_{wz} \\ S_{zw} & S_{zz} \end{pmatrix} \begin{pmatrix} S_{wy} \\ S_{zy} \end{pmatrix}$$

Correcting the *naive* regression coefficients is one way to correct for the bias induced by measurement error. Carroll et al. (2006) and Buonaccorsi (2010) give an extensive overview of various measurement error models which are suitable for the application of the method. However, this approach can be too restrictive in cases with highly skewed covariates, multiplicative error or non-linear models.

The alternative method, suggested by Rosner et al. (1989) and Rosner et al. (1990) for the case of logistic regression but applicable to a broader range of models, is to find a replacement X_{RP} for the true but unobserved X, so that X_{RP} has certain properties, which allow it to be placed instead of X in the exposure-to-health-outcome model, yielding consistent estimates of $\boldsymbol{\beta_x}$.

1.4.3 Regression calibration

Carroll et al. (2006) describes the method of **regression calibration**, which is applicable to a broader range of models when internal validation data, replicate data or data collected with unbiased instrument are available for, at least, a subset of participants. The method relies on substituting the observed exposure measured with error by the esti-

mated expectation of unobserved true exposure conditional on the values of the observed proxy.

The general principle of regression calibration with application to nutritional data can be described as follows:

Suppose, HO_i is a health outcome for person i that we can relate to the true food intake T_i and the other relevant covariates measured without error $\mathbf{Z_i}$ through a linear combination: $\mathrm{E}(HO_i|T_i, \mathbf{Z_i}) = \alpha_0 + \alpha_T T_i + \mathbf{Z_i}'\alpha_Z$.

The regression calibration approach suggests that if we consistently estimate $\mathrm{E}(T_i|W_{i1}, \ldots, W_{in_i}, \mathbf{Z_i})$ then utilising $\mathrm{E}(T_i|W_{i1}, \ldots, W_{in_i}, \mathbf{Z_i})$ instead of T_i will produce a consistent estimate of α_T for ordinary least squares (OLS) linear regression or generalised linear models (GLM). We adjust the standard errors using bootstrap or a sandwich estimator.

Carroll et al. (2006) describes the algorithm of regression calibration in the following way:

• Estimate the conditional expectation of $\boldsymbol{X}$ given $\boldsymbol{Z}$ and $\boldsymbol{W}$: $\mathrm{E}(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{W}) = m_{\boldsymbol{X}}(\boldsymbol{Z}, \boldsymbol{W}, \theta)$, where $m_X(\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{\theta})$ denotes the calibration function, which is a function of additional parameters $\boldsymbol{\theta}$.

• Replace unobserved $\boldsymbol{X}$ with the estimate of $m_{\boldsymbol{X}}(\boldsymbol{Z}, \boldsymbol{W}, \hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ is the estimate of $\boldsymbol{\theta}$.

• Fit a standard regression analysis with $m_{\boldsymbol{X}}(\boldsymbol{Z}, \boldsymbol{W}, \hat{\boldsymbol{\theta}})$ instead of $\boldsymbol{X}$ to estimate the relationships between $\boldsymbol{Y}$ and $\boldsymbol{X}$, and adjust the standard errors using bootstrap or a sandwich estimator.

Several methods to obtain an estimate of the expectation of the unobserved true exposure conditioned on the observed proxy (i.e. regression calibration function) have been proposed. One of them, developed by Carroll and Stefanski (1990) and Gleser (1990), and described in detail in Keogh and White (2014), uses a linear approximation as the regression calibration function and multiple replicates of the proxy for the true exposure. In

the case of a Normally distributed exposure, the best linear predictor is also the best linear unbiased predictor, yielding consistent estimates of regression coefficients. Gleser (1990) gives the bound of the bias of ordinary least squares (OLS) regression coefficients under the best linear regression calibration model. Rosner et al. (1989) and Rosner et al. (1990) showed that if the exposure can be assumed approximately Normal then the regression calibration approach for logistic regression models can yield consistent estimates of regression parameters. However, this does not need to be true in a more general case, for example, a multiplicative error model and relaxed assumptions on the distribution of the health outcome can provide challenges for unbiased estimation of the effect of the risk factor. Lyles and Kupper (1997) provide an alternative to the regression calibration estimate in the form of a quasi-likelihood estimator where they show that the quasi-likelihood estimator is consistent under the specified model assumptions and behaves very similar to the regression calibration estimator.

Chapter 5 investigates the effect of alcohol consumption on HbA1C, a well-established marker of type II diabetes, using a subset of male sample of the NDNS RP survey Years 2-4. It compares the results obtained with various analysis: the traditional multivariate linear regression where alcohol intake is estimated with an observed individual average; the regression calibration approach, where alcohol intake is predicted based on the mixed-effect mixed-distribution model and available socio-demographic characteristics; and the regression calibration approach, where alcohol intake is predicted based on the mixed-effect mixed-distribution model and composite likelihood theory, aiming to take into account potential residual correlation between alcohol intake and consumption of various occasionally-consumed foods that was not accounted for by the observed personal and socio-economic characteristics.

Chapter 2

Estimation of the intake distribution of occasionally-consumed foods

This chapter provides an overview of methods available for the estimation of the intake distribution of occasionally-consumed foods, it develops a numerical approach for the estimation of the distributional quantiles of food intake as an alternative to Monte Carlo (MC) simulations. The methods are illustrated through the analysis of self-reported alcohol consumption collected during the screening phase of a randomised controlled trial investigating the effect of the types of fats and carbohydrates in diet on glucose and insulin metabolism.

2.1 Background

Monitoring usual or long-term dietary intake is of interest to health researchers and public health policy makers to assess nutrient adequacy of a group or population. Recent public-health programmes include monitoring of alcohol consumption by personal, social and demographic characteristics in the research programme “Reducing alcohol-related health harms in an English context” led by the School for Public Health Research of the UK National Institute for Health Research (School for Public Health Research, 2013).

The statistical analysis of dietary data presents several challenges due to limitations in dietary assessment tools and the presence of within-person variation in consumption. The most commonly used dietary assessment tools are food frequency questionnaires (FFQ), food diaries (FD) and 24 hour food recalls (24HR). Of these, methods comparison and biomarker validation studies suggest that multiple days FD and multiple 24HR are more reliable (Burrows et al., 2010; Bingham et al., 1994, 1997; McKeown et al., 2001).

These tools were developed to capture long-term habitual diet but due to reduced observation periods, they are subject to observational error, defined as the difference between the measured diet and its true value (Rutishauser and Black, 2002; Beaton et al., 1979). Moreover, the records of intake of occasionally-consumed dietary components (e.g. fish, alcohol, nuts) usually contain high frequencies of zeros, adding further complexity to the analysis of the distribution of these components. The mean and a measure of spread describe symmetrical distributions well, but not those with skewed shapes. The majority of occasionally-consumed food intake distributions have skewed shapes so the information contained in the mean and a measure of spread will not suffice to estimate, say, under- or over- consumption, which is often of major interest to public health policy makers. Therefore, in the evaluation of dietary intake, the tails of the population intake distribution are often as important as the mean or the median. Thus, quantile estimation provides a useful tool for monitoring diet and complements regression analysis of the mean.

This chapter provides an overview of methods available for the estimation of occasionally-consumed food intake distribution and proposes a numerical approach to estimate the quantiles of the distribution of occasionally-consumed foods in specified sub-populations as an alternative to Monte Carlo (MC) simulations. The method accounts for within-person variation, correlation arising from multiple measurements taken from the same person and the high frequency of zero observations of recorded food intake.

2.1.1 Within-person variation

Within-person variation arises from individual daily variation in food consumption and observational error. The mean of observed individual dietary records is often used as a measure of true individual intake; however, the mean contains information of both, the true long-term habitual intake and within-person variation. Although increasing the number of days in dietary records reduces observational error (Nusser et al., 1987), in practice, most FDs and 24HRs contain only 2 to 4 days of dietary intake records, which leads to a significant daily variation in individual means (Beaton et al., 1979; Nelson et al., 1989; Sempos et al., 1985). Therefore, using individual means to describe food intake in population groups artificially inflates the group variance estimate, which, in turn, results in biased estimates of upper and lower quantiles of food intake distribution and in biased estimates of compliance with respect to recommended intake guidelines (Tooze et al., 2010; Guenther et al., 2006). To illustrate this, consider the estimation of the 90th quantile of a normal distribution. If the mean is 0 and the standard deviation 1, the 90th quantile is 1.28. But, the same 90th quantile, for a distribution with the same mean, but 1.5 times larger standard deviation becomes 1.92.

Dodd et al. (2006) provided a review of statistical methods which account for within-person variation when estimating the distribution of usual dietary intake within a population group using individual means. More recently, Tooze et al. (2010) suggested utilising a mixed-effects modelling approach without reducing the data to individual averages. This method suggests that if a person i has true intake T_i (T_i^* on a transformed scale) then the individual daily food record R_{ij} (R_{ij}^* on a transformed scale), of a person i on day j, can be described as $R_{ij} = T_i + \epsilon_{ij}$ ($R_{ij}^* = T_i^* + \epsilon_{ij}$), where ϵ_{ij} represents random daily variation and is assumed to have mean 0 and variance σ_ϵ^2. This assumption can be described as the unbiasedness of the recorded individual intake either on the original or a transformed scale. Then the total group variance of food intake distribution is decomposed into a within-person (σ_ϵ^2, daily) and between-person (true) parts. Using the estimated between-person

variance and mean and assuming approximately Gaussian distribution of food intake (on the original or transformed scale), we can reconstruct the true food intake distribution within a specified group leaving out the estimated within-person variance. Several applications of this method can be found in the literature (Tooze et al., 2010; Guenther et al., 2006; Tooze et al., 2006).

2.1.2 Excess zeros

Occasionally consumed foods are further characterised by high frequency of zero intake records, which presents further challenges in analysis. Firstly, the methods of dealing with within-person variance described above are not directly applicable to zero-inflated data as they assume that food intake can be transformed to be approximately Gaussian using a monotone function. This distributional assumption is clearly violated for occasionally consumed foods. Secondly, the number of daily records needed to reliably estimate within-person and between-person variation, if consumption occurs only infrequently, exceeds the number of daily records typically available from food diaries or food recalls.

A preferred method for modelling occasionally-consumed food intake for a given individual, adopted in this chapter, looks at the data as generated by a two-step process: the first step (the *probability* step) generates the event of consumption (yes/no) on a given day and the second step (the *amount* step) generates the amount of food consumed on a consumption day. The probability part can be modelled by a mixed-effects logistic regression and the amount component by a mixed-effects linear regression model.

Importantly, as discussed by Olsen and Schafer (2001) and Tooze et al. (2002), consumption behaviours are complex and the outcomes of the first and the second steps are not, generally, independent. In particular, it is plausible that the more often someone consumes, the larger the amount consumed on any given consumption day: examples include fruits and vegetables, whole grains and alcohol (Ashfield-Watt et al., 2004; Guenther et al., 2006). Consequently, the *probability* and the *amount* parts are likely to be correlated.

The correlation can arise, *inter alia*, from personal preferences affecting the probability of consumption and the amount consumed simultaneously. When some of these personal preferences are unobserved, because they may be impractical, impossible, or very expensive to measure, the model needs to account for this unobserved heterogeneity. This can be done through inclusion of one random effect into each component of the model and allowing the two random effects to be correlated. Ignoring this correlation in the estimation of food intake distribution when, in fact, the correlation is positive, can lead to over-estimation of the amount consumed by people with low probability of consumption and under-estimation of the amount consumed by people with high probability of consumption. The magnitude of the bias can be especially pronounced when the between-person variation is quite large and there is not enough information to explain it and when the correlation between unobserved preferences is substantial (Albert, 2005; Kipnis et al., 2009; Su et al., 2009).

Monitoring dietary intake at a group level requires the estimation of distribution characteristics, such as quantiles. Obtaining these from the two-part mixed-effects model is not straightforward due to the presence of the random effects in the model. The current practice, suggested by Guenther et al. (2006), is to: i) estimate individual linear predictors from fitting the two-part model, ii) simulate 100 random effects, per individual, from a bivariate normal distribution, with mean zero and variance parameters estimated from the fitted model, iii) add the simulated random effects to the estimated linear predictors, and iv) obtain empirical quantile estimates from the simulated datasets. This method forms part of the NCI method (National Cancer Institute, 2015) for the estimation of usual dietary intake, recommended by the US National Institute of Health. However, the precision of MC estimates is affected by random sampling variation, and the size of the simulated data that is needed to achieve the required precision is population- and model-specific, which can hinder reproducibility of results. The simulations can also be time consuming with increasing number of sub-populations for which intake distribution is of interest.

We suggest an approach which is based on the two-part model (Olsen and Schafer, 2001; Tooze et al., 2002) and circumvents the need of simulation by use of numerical integration to estimate the distribution of occasionally-consumed food in specified sub-populations. The method is a quicker, easier to implement and more accurate alternative to the simulation-based method. Additionally, we illustrate the impact of ignoring the correlation between the *probability* and *amount* parts of the two-part model in the model specification, and compare the performance of our approach with that based on Monte Carlo (MC) simulations.

2.2 Methods

In this section, the two-part mixed-effects model (Olsen and Schafer, 2001; Tooze et al., 2002) for modelling individual intakes of occasionally-consumed foods is described. Then, the two-part model, along with the numerical method, are utilised for the estimation of habitual dietary intake of occasionally-consumed foods in sub-populations. Finally, the proposed numerical method for the quantile estimation of habitual dietary intake of occasionally-consumed foods is presented in Appendix 2.B.

Additionally, Appendix 2.D.1 contains code written in the statistical software R (R Core Team, 2017) which allows the estimation of the two-part model parameters in a special case when two days of food intake observations per person are available for analysis and this is the first freely available code, to our knowledge, which helps to analyse this type of data.

2.2.1 Two-part mixed-effects model

We briefly describe the two-part mixed-effects model for repeated positive continuous responses with excess zeros (cf. Olsen and Schafer (2001), Tooze et al. (2002), and Su et al. (2009) for full details). As discussed above, for each person, $i, i = 1, \ldots, m$ on day $j, j = 1, \ldots n_i$, the data consist of two parts: the occurrence of food consumption (yes/no),

which can be recorded as an indicator variable I_{ij} such that:

$$I_{ij} = \begin{cases} 1, & \text{if the food is consumed by person } i \text{ on day } j \\ 0, & \text{otherwise} \end{cases}$$

and the amount of food consumed if consumption took place, which we record as $A_{ij}, A_{ij} > 0$ if $I_{ij} = 1$.

Natural heterogeneity arise among subjects due to personal preferences for consumption. We denote unobservable person-specific information related to propensity to consume certain foods as v_i and unobservable person-specific information related to amount consumed on consumption day as u_i. Then, conditionally on v_i and u_i, responses I_{ij} and A_{ij} are independent. The indicator variable I_{ij} is assumed to follow a Bernoulli distribution with probability p_{ij}, and to allow for skewness, we assume $A_{ij}, A_{i,j} > 0$ to be log-normally distributed. In this paper, we suggest the following model specification: the first part response I_{ij} follows the logistic regression model:

$$\text{logit}\{\Pr(I_{ij} = 1|v_i)\} = x_{ij}^{'}\gamma + v_i$$

where $x_{ij}^{'}$ is the vector of relevant covariates, relating individual characteristics to propensity for food intake, and γ is the vector of corresponding regression coefficients. And, considering, $\log(A_{ij}) = Y_{ij}$ is approximately normal, we can write:

$$Y_{ij} = x_{ij}^{'}\beta + u_i + \epsilon_{ij}$$

where $\mathrm{E}(Y_{ij}|u_i) = x_{ij}^{'}\beta + u_i$ and $\mathrm{Var}(Y_{ij}|u_i) = \sigma_\epsilon^2$ (within-person daily variation); $x_{ij}^{'}$ is the vector of relevant covariates relating individual characteristics to the amount of food consumed, β is the vector of corresponding regression coefficients. The potential correlation between the *probability* and *amount* parts is linked through person-specific effects u_i and v_i , which are assumed to have a common bivariate normal distribution with means 0 and variance-covariance matrix:

$$\Sigma = \begin{pmatrix} \sigma_u^2 & \rho\sigma_u\sigma_v \\ \rho\sigma_u\sigma_v & \sigma_v^2 \end{pmatrix}$$

where ρ denotes the correlation between u_i and v_i, σ_u^2 and σ_v^2 are the variances of u_i and v_i respectively. These are called random effects and are assumed to be independent of ϵ_{ij}. The unknown model parameters $\theta = (\gamma, \beta, \sigma_u, \sigma_v, \sigma_\epsilon, \rho)$ can be estimated through maximising the full marginal likelihood function, where we utilise the conditional independence of responses I_{ij} and Y_{ij} and their distributional assumptions. Because the random effects u_i and v_i are unobserved, they need to be integrated out, so that the full marginal likelihood function is:

$$L(\theta) \propto \prod_{i=1}^{m} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \prod_{j=1}^{n_i} f_I(I_{ij} \mid v_i, \theta) f_Y(Y_{ij} \mid I_{ij} > 0, u_i, \theta) f_{UV}(u_i, v_i \mid \theta) du_i dv_i \tag{2.1}$$

where f_I, f_Y and f_{UV} denote the density functions of the binomial, normal and bivariate normal distributions, respectively. More precisely, taking into account the assumed distributions for Y_i, I_i, u_i, v_i the likelihood can be further re-written as

$$L \propto \prod_{i=1}^{m} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \exp(l_{V_i}) \exp(l_{U_i}) f_{U|V}(u_i \mid v_i; \theta) f_V(v_i; \theta) du_i dv_i \tag{2.2}$$

where $f_{U|V}$, f_U denote the conditional normal distribution for $U|V$ and the marginal normal distribution of U respectively, l_{V_i} is the loglikelihood contribution from I_{ij}, and l_{U_i} from Y_{ij} given by

$$\begin{aligned} l_{V_i} &= \sum_{j=1}^{n_i} (I_{ij}(X_{ij}^T \gamma + v_i) + \log(1 - p_{ij})) \\ l_{U_i} &= -\frac{n_i^*}{2} \log(\sigma_\epsilon^2) - \frac{1}{2\sigma_\epsilon^2} \sum_{j=1}^{n_i^*} (Y_{ij} - (X_{ij}\beta + u_i))^T (Y_{ij} - (X_{ij}\beta + u_i)) \end{aligned} \tag{2.3}$$

where n_i^* denotes the number of observed positive values of Y_{ij} for participant i.

The loglikelihood function does not have a closed form and needs to be evaluated numerically.

We note that if it is assumed that the random effects are independent, i.e. $\rho = 0$, estimation is considerably simplified as the two parts can be fitted separately using standard statistical software for generalised mixed-effects models. However, if this assumption does not hold, i.e. $\rho \neq 0$, then the estimation of the two-part model requires more specialised programming, for example, the SAS PROC NLMIXED procedure (SAS Institute, Cary, NC, Version 9.1, Littell et al 2006, SAS for mixed model) can be used in this case.

2.2.2 Distribution of habitual dietary intake

The expected individual habitual daily intake T_{ij} for a person i on a day j is calculated as the product of the individual daily probability of consuming the food, p_{ij}, and the expected individual consumed amount on a consumption day:
$T_{ij} = \Pr(I_{ij} = 1|v_i) \cdot \mathrm{E}(A_{ij}|A_{ij} > 0, u_i)$. Under the two-part model T_{ij} depends on the regression parameters β and γ, as well as the unobserved person-specific effects u_i and v_i, which may be correlated. Maximum likelihood estimates: $\tilde{\beta}, \tilde{\gamma}, \tilde{\Sigma}, \tilde{\sigma_\epsilon}$ can be obtained by fitting the two-part model, but the person-specific variation has to be accounted for when estimating a group distribution of dietary intake. One way to account for this variation is to perform MC simulations.

This method and its application in the present context have been described elsewhere (Guenther et al., 2006; Tooze et al., 2002; Freedman et al., 2010). Briefly, first, point estimates of the model parameters are obtained from fitting the two-part model. Secondly, for each combination of covariates of interest, fixed effect predictions are obtained using the estimated regression coefficients. Thirdly, N pairs (u_i, v_i) are generated from a bivariate normal distribution with the parameters of the distribution estimated earlier at the first step. Guenther et al. (2006) recommends to simulate 100 observations per original sample observation with the same covariate values but varying person-specific effects. Thus, for each combination of covariates we have a dataset containing N (e.g. 100 times the original sample size) simulated observations whose distribution characterises the distribution of occasionally-consumed dietary intake in a sub-population with the same covariate pattern as that of the observed sample. This dataset is then used to obtain empirical quantile estimates. If the intake is assumed to be unbiased on the original scale then back-transformation needs to be used (Nusser et al., 1996).

This work suggests the use of optimisation and numerical integration methods to estimate the quantiles of occasionally-consumed food intake distributions as an alternative to MC simulations. To compare the proposed approach with MC simulations, we undertook a

simulation study following the NCI method described above, up to the point where we needed to decide on the size of simulated data. One of the research questions we set to answer was to investigate the MC convergence in the context of the application of the two-part model, so it was decided to simulate data sets of varying size including 1000, 5000, 10000 and 50000 observations per fixed covariate values. The covariates we adjusted for in the model were gender and age, so for men and women, and for each of the following age values (years): 40, 45, 50, 55, 60, 65 we simulated 4 data sets of different sizes. In the Results section we compare how our MC simulated results compare with the results obtained from the proposed approach. The following section describes the proposed numerical method with further technical details provided in Appendix 2.A and Appendix 2.B provides code for implementation in R.

2.2.3 Quantiles of habitual dietary intake

Quite often, the distribution of the amount of food consumed on a consumption day appears to be skewed and a logarithmic transformation can be an appropriate choice to obtain a symmetric distribution (Xiao et al., 2011). If we assume that the individual transformed intake $Y_{ij}|I_{ij} > 0, u_i$ follows a normal distribution with expectation $x_{ij}^{'}\beta + u_i$ and variance σ_ϵ then $A_{ij}|I_{ij} > 0, u_i$ follows log-normal distribution with expected value $\exp(x_{ij}^{'}\beta + u_i + 0.5\sigma_\epsilon^2)$ so we can write down the individual expected daily marginal amount consumed as

$$\tilde{T_{ij}} = \exp(x_{ij}^{'}\tilde{\beta} + u_i + 0.5\sigma_\epsilon^2)\frac{\exp(x_{ij}^{'}\tilde{\gamma} + v_i)}{1 + \exp(x_{ij}^{'}\tilde{\gamma} + v_i)}$$

Dietary intake, alcohol consumption for example, is likely to vary between a week day and a weekend. To account for this, the expected weekly consumption $\tilde{T_i}$ is estimated as the weighted average of habitual daily consumption comprising 4 working-week days and 3 weekend days:

$$\tilde{T_i} = 4\exp(x_{i0}^{'}\tilde{\beta} + u_i + 0.5\sigma_\epsilon^2)\frac{\exp(x_{i0}^{'}\tilde{\gamma} + v_i)}{1 + \exp(x_{i0}^{'}\tilde{\gamma} + v_i)} + 3\exp(x_{i1}^{'}\tilde{\beta} + u_i + 0.5\sigma_\epsilon^2)\frac{\exp(x_{i1}^{'}\tilde{\gamma} + v_i)}{1 + \exp(x_{i1}^{'}\tilde{\gamma} + v_i)}$$

where $\tilde{\beta}$ and $\tilde{\gamma}$ are point estimates from the two-part model, $x_{i0}^{'}$ are covariates corresponding to a working-week day and $x_{i1}^{'}$ are covariates corresponding to a weekend. $\tilde{T_i}$ depends on the two random variables u_i and v_i. By definition of cumulative distribution function, for a given probability p and the corresponding quantile c_p, we can write:

$$\Pr(\tilde{T_i} \leq c_p) = p \tag{2.4}$$

which, when substituting $\tilde{T_i}$, is equivalent to

$$\Pr\Big(4\exp(x_{i0}^{'}\tilde{\beta} + u_i + 0.5\sigma_\epsilon^2)\frac{\exp(x_{i0}^{'}\tilde{\gamma} + v_i)}{1+\exp(x_{i0}^{'}\tilde{\gamma} + v_i)} + 3\exp(x_{i1}^{'}\tilde{\beta} + u_i + 0.5\sigma_\epsilon^2)\frac{\exp(x_{i1}^{'}\tilde{\gamma} + v_i)}{1+\exp(x_{i1}^{'}\tilde{\gamma} + v_i)} \leq c_p\Big) = p$$

After re-arranging the terms and taking natural logarithm, the above expression is equivalent to

$$P\Big(u_i \leq \ln(c_p) - \ln\Big\{4\exp(x_{i0}^{'}\tilde{\beta} + 0.5\sigma_\epsilon^2)\frac{\exp(x_{i0}^{'}\tilde{\gamma} + v_i)}{1+\exp(x_{i0}^{'}\tilde{\gamma} + v_i)} + 3\exp(x_{i1}^{'}\tilde{\beta} + 0.5\sigma_\epsilon^2)\frac{\exp(x_{i1}^{'}\tilde{\gamma} + v_i)}{1+\exp(x_{i1}^{'}\tilde{\gamma} + v_i)}\Big\}\Big) = p$$

Let $h(c, v_i)$ denote the function:

$$h(c, v_i) \equiv \ln(c) - \ln\Big(4\exp(x_{i0}^{'}\tilde{\beta} + 0.5\sigma_\epsilon^2)\frac{\exp(x_{i0}^{'}\tilde{\gamma} + v_i)}{1+\exp(x_{i0}^{'}\tilde{\gamma} + v_i)} + 3\exp(x_{i1}^{'}\tilde{\beta} + 0.5\sigma_\epsilon^2)\frac{\exp(x_{i1}^{'}\tilde{\gamma} + v_i)}{1+\exp(x_{i1}^{'}\tilde{\gamma} + v_i)}\Big)$$

Then under the distributional assumptions for v_i and u_i as bivariate normal $(0, \Sigma)$ we can re-write (2.4) as

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{h(c,v_i)} f_{BN}(u_i, v_i)du_i dv_i = p. \tag{2.5}$$

The solution of (2.5) with respect to c, is the quantile c_p, corresponding to a given probability p. Appendix 2.A shows why this solution exists and is unique under the given model assumptions. To find c_p the integral in (2.5) can be approximated numerically, e.g. by quadrature methods, and the solution to the equation found through optimisation.

2.2.4 Data

To illustrate the method, we analysed alcohol intake from the screening phase of the RISCK (Reading, Imperial, Surrey, Cambridge, and Kings) study (Jebb et al., 2010), which was a randomised controlled trial (RCT), investigating the effect of the types of fats and carbohydrates in diet on glucose and insulin metabolism. Participants were recruited from the general population and baseline measures were collected from August 2004 to April 2006. The participants were eligible if their weight was stable 3 months prior to enrolment, i.e. their energy intake and energy expenditure were in balance (Rosenbaum et al., 1996), and if they were at risk of developing metabolic syndrome with special emphasis on enrolling participants with impaired glucose tolerance. Initially, 7-day food diaries were collected from 531 participants. These yielded 2214 days of dietary records in total, with the majority (81%) providing 4 days of the foods records. However, to reduce potential bias in data analysis (Braam et al., 1998; Mendez et al., 2011; Tooze et al., 2004), this analysis excluded data from 209 (39%) participants due to extreme under-reporting, leaving for analysis data on 322 (61%) participants. The status of *under-reporters* was defined by the Goldberg cut-off (Black, 2000a,b) (see Appendix 2.C for further details). Ethical approval for the RISCK study was obtained from the National Research Ethics Service, and written informed consent was given by participants.

2.2.5 Simulation study to assess the impact of model misspecification

A simulation study was conducted based on the R code written by the author for the analysis of semi-continuous data with correlated random effects in the special case when information from only 2-day food diaries is available (Appendix 2.D.1). To obtain maximum likelihood estimates, the second-order Laplace approximation for the numerical integration in (2.2) was used, requiring the analytic calculation of the first and second derivatives of the log-likelihood function with respect to all model parameters. The aim of the simulation study was to examine the potential consequences of model misspecification on model

parameters estimates from assuming that the correlation of the random effects from the two model components is zero. The results of the simulation study are presented in Appendix 2.D.2.

The data were generated from the two-part model described earlier for consumption of a single occasionally-consumed food recorded over 2 days. The probability of consumption $p_{ij} = \Pr(I_{ij} = 1)$ for a person i on day $j, j = 1, 2$ followed the model

$$\text{logit}(p_{ij}) = 12.7 - 30 \cdot X_{ij1} + 2 \cdot X_{ij2} + v_i$$

The amount consumed when consumption took place ($I_{ij} = 1$) was simulated from the model

$$Y_{ij} = 900 - 1000 \cdot X_{ij1} + 5 \cdot X_{ij2} + u_i + \varepsilon_{ij}$$

X_1 was chosen to be a continuous variable but constant across the two observed time points. X_2 was a binary variable (0/1) so that it could take different values at different time points for the same person and be representative of a consumption pattern that depends on the day of a week.

The joint distribution of v_i and u_i was generated as bivariate normal with variance components $\text{Var}(u_i) = \sigma_u^2 = 50$ and $\text{Var}(v_i) = \sigma_v^2 = 25$. The covariance component varied to assess how the regression estimates change with a change of magnitude of correlation between u_i and v_i. The within-person measurement error ε_{ij} was simulated from $\mathcal{N}(0, 100)$.

Five hundred datasets were generated under the above model assumptions with varying sizes (N = 150, 500, 800) and analysed in two ways. First, the data sets were analysed with the two-part model, which takes correlation between the model parts into account. Secondly, the data were analysed by fitting two parts of the model separately not accounting for correlation of random effects. Then, the results of two analysis were compared to investigate the effect of model misspecification (Appendix 2.D.2).

2.3 Results

2.3.1 Descriptive analysis

The sample available for analysis consisted of 186 (58%) women, with the following characteristics summarised as mean (standard deviation) or frequency (%): age 52 years (10), body mass index (BMI) 27.5 (4.2), smoking status (yes) 31 (4%), degree of under-reporting 0.96 (0.15); and of 136 (42%) men: age 53 years (11), BMI 27.6 (3.4), smoking status (yes) 36 (6.3%) and degree of under-reporting 0.93 (0.13).

To describe the probability of consuming alcohol in the period of observation, the ratio of the number of reported alcohol consumption days over the total number of diary records available for each participant was calculated. Table 2.1 shows that men and women have significantly different consumption patterns (overall p-value from chi-squared test is 0.004): more women than men (70 (37.6%) versus 32 (23.5%)) reported no alcohol consumption, whereas, there are fewer women than men (26 (14.0%) versus 32 (23.5%)) whose estimated probability of consuming is greater than 0.75 on a given day.

Table 2.1: Proportion of days of recorded alcohol intake out of total recorded days available

Proportion of days with recorded alcohol consumption	Men, N (%)	Women, N (%)
0 records	32 (23.5)	70 (37.6)
>0 and ≤ 0.25	20 (14.7)	42 (22.6)
>0.25 and ≤ 0.5	27 (19.9)	26 (14.0)
>0.5 and ≤ 0.75	25 (18.4)	22 (11.8)
>0.75	32 (23.5)	26 (14.0)

The percentage of days of recorded alcohol intake was estimated as a ratio of the number of reported alcohol consumption days over the total number of diary record days available.

Despite different frequency patterns of alcohol consumption, both, men and women, tend to consume more alcohol on a given consumption day if their frequency of consumption is higher compared to those who consume less frequently (Figure 2.1).

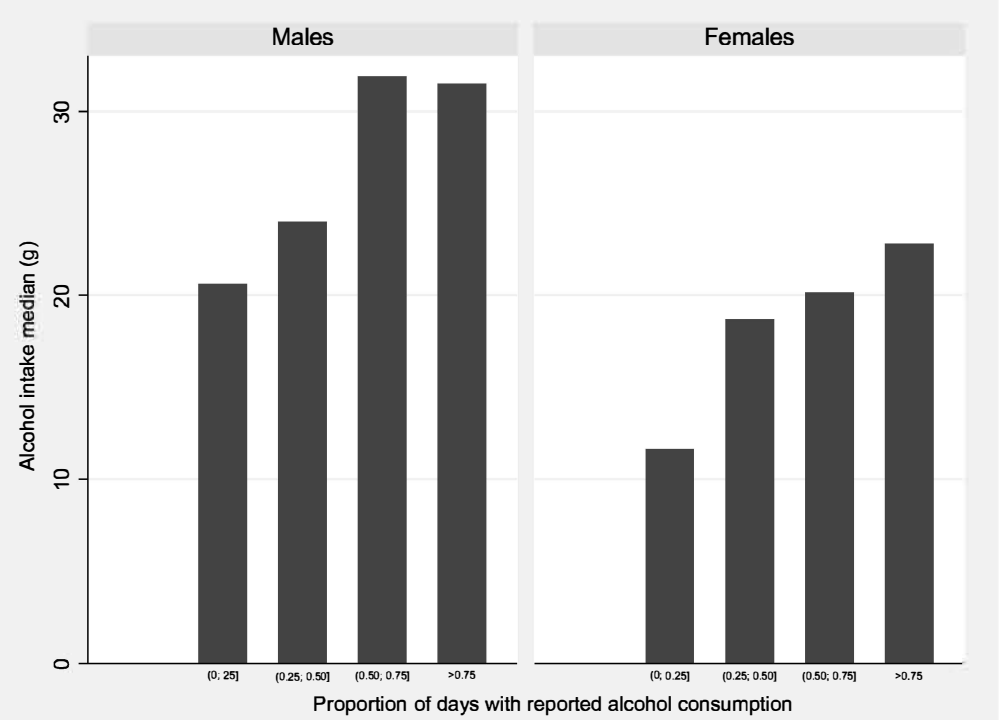


Figure 2.1: Median of alcohol intake by recorded frequency of alcohol consumption

2.3.2 Modelling alcohol intake

The statistical analyses were stratified by sex. After preliminary screenings of the sampling distributions of alcohol intake on consumption days, a logarithmic transformation was adopted to obtain a more symmetric distribution of the data. Figure 2.1 suggests that there might be a positive correlation between the probability of consuming alcohol and the amount of alcohol consumed on consumption day. We fitted the two-part mixed-effects model assuming that the correlation between the two parts is positive (Model A) and assuming that the correlation is zero (Model B). We compare the analysis results from Models A and B to assess the impact of model misspecification on both the estimation of parameters related to individual alcohol intake and the distribution of alcohol intake in specified sub-groups. We note that the regression parameters in the two parts of the models are person specific.

Correlation between probability and amount parts of the model

The estimated (adjusted for age and weekend) correlation between the model parts is 0.55 (p-value 0.004) in females and 0.30 (p-value 0.160) in males. This suggests that there exist some person-specific characteristics which simultaneously increase the probability to consume alcohol and the amount of alcohol consumed on consumption day.

Probability part

Estimates show no difference between models A and B in the estimation of the odds of daily alcohol consumption for both groups, men and women (Table 2.2).

Table 2.2: Effect of the covariates on daily probability of alcohol consumption and amount of alcohol consumed

	Males		**Females**	
Probability part	Odds ratio 95%CI	p-value	Odds ratio 95%CI	p-value
Model A				
Weekend	3.95 (2.37, 6.60)	<0.001	3.56 (2.27, 5.56)	<0.001
5 years increase in age	1.27 (1.01, 1.54)	0.042	1.29 (1.03, 1.55)	0.031
Model B				
Weekend	3.99 (2.39, 6.66)	<0.001	3.53 (2.25, 5.54)	<0.001
5 years increase in age	1.27 (1.01, 1.53)	0.043	1.30 (1.03, 1.56)	0.028
Amount part	Ratio of change 95% CI	p-value	Ratio of change 95% CI	p-value
Model A				
Weekend	1.48	<0.001	1.28	0.016

Table 2.2: Effect of the covariates on daily probability of alcohol consumption and amount of alcohol consumed (Continued)

	Males		**Females**	
Probability part	Odds ratio	p-value	Odds ratio	p-value
	95%CI		95%CI	
	(1.23, 1.79)		(1.04, 1.56)	
5 years increase in age	0.96	0.162	1.00	0.910
	(.90, 1.02)		(0.91, 1.08)	
Model B				
Weekend	1.45	0.001	1.23	0.062
	(1.19, 1.73)		(0.99, 1.48)	
5 years increase in age	0.95	0.105	0.98	0.695
	(0.89, 1.01)		(0.90, 1.06)	
Correlation between probability and amount parts	0.30	0.160	0.55	0.004

Amount part

The regression parameters are interpreted as percentage change in the amount of alcohol consumed on consumption day with a unit-change in the corresponding covariate, holding the other covariates fixed. For females, model B shows *weekend* as a non-significant predictor (at 5% significance level): 1.23 (95%CI (0.99, 1.48), p-value 0.062) times increase in the amount of alcohol consumed given consumption took place on weekend compared to a week day; whereas model A shows that, on weekend, women increase the amount of alcohol consumed (given it was consumed) by 1.28 times (95%CI (1.04, 1.56), p-value 0.016). Thus, under the wrong assumption of zero correlation between the model parts, a statistically significant predictor turns into non-significant.

For males the discrepancy between the results obtained from model A and model B is not as pronounced: 1.45 times increase on *weekend* in amount consumed if consumption takes place (95%CI (1.19, 1.73), p-value 0.001) for model B, and 1.48 times increase (95%CI (1.23, 1.79), p-value <0.001) for model A.

These findings show that when the zero correlation assumption between *probability* and *amount* parts is strongly violated, model A provides better estimates of regression coefficients. However, the greatest discrepancies between the results from model A and model B tend to be observed not around the (geometric) mean but around the tails of the distribution of alcohol intake.

Distribution of weekly alcohol consumption

Table 2.3 shows the estimated weekly alcohol intake quantiles (0.1, 0.25, 0.50, 0.75, 0.90 and 0.95) and the magnitude of discrepancies between weekly alcohol intake distributions estimated under model A and B assumptions, separately for males and females and for various ages. The difference between the models is most obvious at the tails of the distribution, where Model A, as expected, gives higher estimates than model B for higher quantiles. For example, our data show that, in men, model A estimates 0.90 quantile to be 321.8g versus 301.6g (model B) of weekly alcohol intake in 40-year-old participants. Since the detrimental effect of alcohol is believed to arise from excessive consumption, our results demonstrate that the application of the model with the correct assumptions provides a more accurate assessment of the potential public health burden.

Table 2.3: Weekly alcohol intake quantiles estimates under various model assumptions

		Quantiles					
Age, y	Model	0.1	0.25	0.5	0.75	0.9	0.95
Men							
40	A	8.2	29.6	91.1	194	321.8	419.1
	B	11.1	36	97.7	190	301.6	386

Table 2.3: Weekly alcohol intake quantiles estimates under various model assumptions (Continued)

		Quantiles					
Age, y	Model	0.1	0.25	0.5	0.75	0.9	0.95
	MC1	8.6	32	97.1	190.3	311.4	407.7
	MC5	8.2	29.5	91.3	191.5	312.4	410.9
	MC10	7.9	29.7	91	192	322.7	420.2
	MC50	8.3	30.1	90.6	194.1	321.8	416.2
45	A	9.8	34.2	97.9	198.1	321.3	414.9
	B	13.2	41.1	103.9	193.4	301.2	382.8
	MC1	7.8	32.6	86.1	190.4	306.4	391.5
	MC5	9.7	34.2	96	199.7	317.6	408.4
	MC10	9.4	33.9	97.3	196.1	317.3	416.4
	MC50	9.7	34.6	98	198.6	322.5	414
50	A	11.7	39	104	201	319.5	409.6
	B	15.5	46.3	109.2	195.6	299.6	378.5
	MC1	12.5	43.1	108	199.5	330.1	386.3
	MC5	13	41.3	104.9	203.1	325.1	416.6
	MC10	11.2	37.9	101.7	201.3	320.9	413.3
	MC50	11.4	38.5	103.9	201.7	317.4	407.8
55	A	13.9	43.9	109.2	202.7	316.6	403.2
	B	18.2	51.5	113.6	196.8	297.1	373.2

Table 2.3: Weekly alcohol intake quantiles estimates under various model assumptions (Continued)

		Quantiles					
Age, y	Model	0.1	0.25	0.5	0.75	0.9	0.95
	MC1	13.5	44.5	108.5	202	301.7	396.1
	MC5	13.3	42.4	107	205.4	318.7	406.4
	MC10	13.5	42.1	105.2	198.4	308.5	390.1
	MC50	13.7	43	108.5	204	319.7	406.8
60	A	16.3	48.9	113.5	203.3	312.7	395.9
	B	21.1	56.8	117	197	293.7	367.1
	MC1	14.5	43.9	107.6	200.4	306.7	375.8
	MC5	15.3	48.7	113.7	203.9	319	408
	MC10	16.4	47.7	111.2	203.2	311.1	392.1
	MC50	15.9	48.9	114.5	204.8	314.7	397.6
65	A	19	53.7	116.9	203	307.9	387.9
	B	24.5	61.5	119.6	196.4	289.4	360.2
	MC1	19.4	55	120.4	197.7	303.9	370.1
	MC5	18.5	52.6	115.9	201.8	304	392
	MC10	19.7	54.2	117.6	202.8	311.4	394.1
	MC50	19.2	54.3	118	204.3	308	387.1

Table 2.3: Weekly alcohol intake quantiles estimates under various model assumptions (Continued)

		Quantiles					
Age, y	Model	0.1	0.25	0.5	0.75	0.9	0.95
Women							
40	A	0	4	20.1	71.7	166	251
	B	1.7	6.5	25.2	70.7	143.5	207.4
	MC1	0.9	4	19.7	69.8	156.1	248.2
	MC5	0.8	4	20.2	73.8	167.4	252.8
	MC10	0.8	3.7	19.3	71.6	166.9	253.1
	MC50	0.8	4	19.7	70.5	163.2	249.8
45	A	1	5	24	79.7	176.8	262.7
	B	2.1	8.1	29.7	78.3	153.6	219.3
	MC1	1.2	5.2	26.5	90.9	186.6	256.1
	MC5	1	4.9	23.9	77.4	173.4	258.6
	MC10	1	5.2	24.3	80.9	174.4	255.5
	MC50	1	5	24.1	80.5	176	260.5
50	A	1.3	6.2	28.4	88.1	186.9	273.6
	B	2.6	10	34.6	85.8	163.4	230.6
	MC1	1.2	7	29.9	86	181.8	268.4
	MC5	1.3	6.1	28.2	86.4	182.1	272.3
	MC10	1.4	6.5	29.4	89.8	192.2	279.4

Table 2.3: Weekly alcohol intake quantiles estimates under various model assumptions (Continued)

		Quantiles					
Age, y	Model	0.1	0.25	0.5	0.75	0.9	0.95
	MC50	1.3	6.2	28.2	88.2	186.5	273.3
55	A	1.6	7.7	33.2	96.1	196.4	283.5
	B	3.3	12.3	39.7	93.2	172.8	241.3
	MC1	1.5	8.3	34.1	95.7	203.3	282.2
	MC5	1.6	7.9	34.2	95.7	195.6	300.1
	MC10	1.6	7.3	32.2	93.9	197.7	286.6
	MC50	1.6	7.6	33.1	95.5	195.7	282.8
60	A	2.1	9.5	38.3	103.8	205.1	292.6
	B	4.2	14.9	44.9	100.4	181.7	251.5
	MC1	2.5	11.3	40.9	109.9	216.9	286.2
	MC5	2.3	9.8	39.2	104.2	209.3	296
	MC10	2.3	10.2	40.3	110.6	213.8	301.5
	MC50	2.1	9.6	38.1	103.6	205.9	293
65	A	2.6	11.6	43.5	111.1	213.1	300.7
	B	5.2	17.8	50.2	107.2	190.1	261
	MC1	2.3	11	42.4	115	237.4	302.9
	MC5	2.8	12.5	45.4	112.6	208.4	305.2

Table 2.3: Weekly alcohol intake quantiles estimates under various model assumptions (Continued)

		Quantiles					
Age, y	Model	0.1	0.25	0.5	0.75	0.9	0.95
	MC10	2.7	11.9	44.9	111.4	214.5	299.9
	MC50	2.6	11.6	43.2	110.7	211.5	303

2.3.3 Comparison with Monte Carlo simulation

Table 2.3 shows the results of Monte Carlo simulation (model A only), based on 1000, 5000, 10 000 and 50 000 simulated datasets for a given covariate pattern.

Monte Carlo simulation estimates show better convergence to the estimates obtained via the numerical method with increasing number of simulations. The difference between results is more pronounced at the tails of the distribution. For example, for a group of 45-year-old men, the 0.95 quantile obtained from the Monte Carlo simulated dataset of 1000 observations is equal to 391.5g, which is considerably lower than 414.9g obtained from the suggested numerical approach and compared to 414.0g obtained when increasing the number of datasets to 50000.

Adherence to maximum recommended intake

The proposed method also allows the estimation of the percentage of participants who adhere to the current recommendations with respect to reference intakes. For example, the Department of Health (School for Public Health Research, 2013) recommends that maximum daily alcohol intake should not exceed 32g for men and 24g for women, which accumulates to weekly maximum intake of 224g for men and 168g for women. Applying the method described in this chapter we estimate that among 45-year-old participants 21% of males and 11% of females exceed the maximum recommended weekly alcohol intake.

2.4 Discussion

The chapter utilises the two-part mixed-effects model introduced by Olsen and Schafer (2001) and followed by Tooze et al. (2002), and extends the work by Guenther et al. (2006) by suggesting a concise numerical method, as an alternative to Monte Carlo simulations, for the estimation of the distribution of occasionally consumed foods in specified population sub-groups. We show that, although quantile estimates obtained with simulations converge to numerically obtained estimates, the number of simulated observations needed per covariate pattern cannot be known in advance and depends on the structure of the data at hand. With the differences between the estimates obtained from both methods most pronounced at the tails of the distribution, the method can be especially applicable when the focus of research is under- or over-consumption of certain nutrients, foods or beverages. Furthermore, since the method is faster than simulations, it is especially convenient when the number of covariate patterns is large.

There are several extensions to the two-part mixed-effects model, as Olsen and Schafer (2001) show, it may include random slopes in addition to the random intercepts used here, thus widening their application to more complex study designs, such as longitudinal studies. Guenther et al. (2006) suggested transforming the original recorded amount of food consumed based not only on the log-normal distribution, but also including Box-Cox power transformations. Consequently the back-transformations to the original scale of the continuous response is required (Nusser et al., 1996). Liu et al. (2010) suggested to use the generalised gamma distribution for continuous positive responses. Furthermore, Su et al. (2009) discussed in depth the bias, arising in regression coefficients, when the correlation between the model parts is not accounted for. Our results provide an illustration of the impact of this form of model specification on the estimated distribution of alcohol intake.

The multivariate Normal distribution for random effects is commonly used in the area of generalised mixed-effects models for computational convenience; however, it has also

been shown to be a robust assumption in many situations when using maximum likelihood estimation (McCulloch and Neuhaus, 2011). Su et al. (2011) suggested the bridge distribution for the joint random effects to relax this assumption for the two-part model and provided an extensive discussion on the interpretation of the marginal effects of the *probability* part. The Bridge distribution is similar in shape to the bivariate Normal distribution but has heavier tails. Its main advantage is that it links the conditional and marginal regression coefficients of the *probability* part; however, the estimation of the marginal coefficients for the linear part is not straightforward and requires numerical integration. Unlike the normal distribution, the Bridge distribution for correlated random effects is not widely known or implemented in standard statistical software.

It is possible to use the probit link to model the probability of consumption; however, the interpretation of the fixed-effects coefficients is not as straightforward as the interpretation of coefficients of logistic regression. The marginal effects have a convenient interpretation, but their estimation requires taking into account the potential non-zero correlation between the parts of the two-part model, making this model more difficult to interpret and use in practice.

Often, it is also of interest to investigate the relationship between predicted dietary intake and health outcomes. We have showed that the between-person variation of alcohol consumption can be substantial. Therefore, when utilising predicted values of intake in relationship with health outcomes, this variation should be taken into account.

There are several limitations of the described model and the proposed method. First, it is assumed that all consumed foods are reported (i.e. the reported intake is an unbiased measure of the true intake on the original or transformed scale), which might be unlikely for some subgroups of people as demonstrated by doubly labelled water studies (Tooze et al., 2004). We tried to minimise the potential bias by excluding those with high degree of energy under-reporting. However, if misreporting is present then the estimated intake distribution can also be biased.

Secondly, sometimes, a user might experience model convergence issues. The current optimisation routine starts with estimates obtained under an independence assumption of the two parts of the model. Varying the initial points from which the algorithm starts the optimisation routine is advised. Furthermore, residual diagnostics similar to those used for generalised linear mixed-effects models can be applied to assess goodness-of-fit.

The two-part model allows the probability to consume to be very small but not zero, so we cannot distinguish never- from rare-consumers. Keogh (2011) suggests a model extension to adjust for never-consumers.

We limited the applicability of the model to natural logarithm transformed data to obtain symmetry in the shape of the distribution, which might be too restrictive in some cases. This, along with the extension of the method to the estimation of intake of multiple correlated foods are areas of further research.

2.5 Conclusions

In summary, this chapter provides a new numerical method for the concise estimation of occasionally consumed food intake distribution within a specified sub-population. The method is based on estimates obtained from the two-part mixed-effects model and utilises numerical integration and optimisation techniques which can be readily implemented. It is less time consuming than a simulation based method, which is especially beneficial for when the number of the predictors of food intake is large. It does not rely on simulation so the precision of quantiles estimates does not depend on simulated data size. We hope that this work will encourage the application of the two-part mixed-effects model in the wider research community as it shows that the model is very flexible and can incorporate various explanatory factors such as seasonality, the day of the week, gender, age, behavioural and socio-economic status. Incorporating relevant explanatory factors reduces the between-person variation and thus can help uncover potential causal relationships between food intake and social, environmental, personal and behavioural predictors. This

is a very active area of current nutrition research.

List of abbreviations

24HR - 24 hour food recall

BMI - Body mass index

FD - Food diary

FFQ - Food frequency questionnaire

ISU - Iowa State University

MC - Monte Carlo

NCI - National Cancer Institute

Appendix

2.A Mathematical justification of proposed numerical approach

We show that the solution to equation (2.5), which defines the quantile c_p, is unique. This involves the integral

$$I(c) = \int_{-\infty}^{+\infty} \Big(\int_{-\infty}^{h(c,v_i)} f(u_i, v_i) du_i \Big) dv_i$$

where $f(u_i, v_i)$ is the probability density function of the bivariate normal distribution with mean 0 and covariance matrix $\Sigma = \begin{pmatrix} \sigma_u^2 & \rho\sigma_u\sigma_v \\ \rho\sigma_u\sigma_v & \sigma_v^2 \end{pmatrix}$ ρ is the correlation between u_i and v_i and σ_u, σ_v are the corresponding standard deviations, and consider $h(c, v_i)$

$$h(c, v_i) = \ln(c) - \ln\Bigg(4\exp(x_{i0}^{'}\tilde{\beta} + u_i + 0.5\sigma_\epsilon^2)\frac{\exp(x_{i0}^{'}\tilde{\gamma} + v_i)}{1 + \exp(x_{i0}^{'}\tilde{\gamma} + v_i)} + \\ + 3\exp(x_{i1}^{'}\tilde{\beta} + u_i + 0.5\sigma_\epsilon^2)\frac{\exp(x_{i1}^{'}\tilde{\gamma} + v_i)}{1 + \exp(x_{i1}^{'}\tilde{\gamma} + v_i)} \Bigg)$$

Then under the above assumptions the following holds true:

$h(c, v_i)$ is continuous for all $c > 0$ and v_i,

$h(c, v_i)$ is strictly decreasing in v_i, $\frac{\partial h(c,v_i)}{\partial v_i} < 0$ for all v_i,

$h(c, v_i)$ is strictly increasing in c, $\frac{\partial h(c,v_i)}{\partial c} > 0$ for all c,

As $f(u_i, v_i)$ is strictly positive and continuous by assumption, it follows that $I(c)$ is continuous and is strictly increasing in c. Moreover, as $f(u_i, v_i)$ is a p.d.f., the following holds: $0 < I(c) < 1$ for all $c > 0$ and both the upper and the lower bounds for $I(c)$ are tight in the sense that $\lim_{c\to 0} I(c) = 0$, and $\lim_{c\to\infty} I(c) = 1$. Hence, by applying the Intermediate value theorem to $I(c)$, it follows that for all $p \in (0, 1)$ there exists a *unique* consumption level c_p that solves equation $I(c_p) = p$ (A Course in Mathematical Analysis: Volume 1, Foundations and Elementary Real Analysis, D.J.H. Garling, 2013, Cambridge). The uniqueness follows from the strict monotonicity of $I(c)$.

2.B R code for the estimation of quantiles

This Appendix provides R code for the estimation of quantiles of occasionally-consumed food intake in sub-populations while taking correlations between the two parts of the model into account and avoiding Monte Carlo simulations to save estimation time and increase precision.

The function returns the daily amount consumed that corresponds to the specified percentile for a defined population. The function takes into account weekend and age differences in consumption when estimating percentiles and can be extended further to accommodate the necessary covariates.

```
#theta is a vector - c(, ..., ) of 13 inputs; percentile = [0, 1]; load "pracma"
    package for "integral2"
# package pracma allows 2 dimensional integration

##############################################################################################

# returns daily amount consumed that corresponds to the specified percentile
# if weekly amount needs to be estimated then 4/7 and 3/7 in the _Umin_ function
    should be replace by 4 and 3
```

```
# if Friday is not a part of weekend then 4/7 and 3/7 in the _Umin_ function
    should be replace by 5/7 and 2/7
#requires some starting value for the consumption to be estimated specified in the
    parameter _consumption0_

uncond_consum_amount = function(theta){

#the entries for the function - estimated previously model parameters

#fixed effect part related parameters

age=theta[1]
const_beta=theta[2]
beta=theta[3:4]
const_gamma=theta[5]
gamma=theta[6:7]

#variance part related parameters

sigmaUsq=theta[8]          # variance of random effect of amount part
sigmaVsq=theta[9]          # variance of random effect of probability part
sigmaEsq=theta[10]        # variance of measurement error
rho=theta[11]                 # correlation between the two parts of the model

#optimisation related parameters

percentile=theta[12]       # required percentile to be estimates, in the region
    [0, 1]
consumption0=theta[13] # initial value of the consumption (best guess)
```

```
# variance-covariance matrix of random effects

sigma =
    matrix(c(sigmaVsq,rho*(sigmaVsq*sigmaUsq)^(1/2),rho*(sigmaVsq*sigmaUsq)^(1/2),sigmaUsq),
    ncol=2)

# probability density function for the 2 dimensional normal distribution

jointerror=function(v,u){
   x=c(v,u)
   return(1/(2*pi)/((det(sigma))^0.5)*exp(-1/2*t(x)%*%solve(sigma)%*%x))
}

# need to vectorise the joint normal pdf so that integration can be performed

vecjointerror=Vectorize(jointerror)

# function _myintegral.func_ which needs optimising with respect to _consumption_

myintegral.func=function(consumption){

# _Umin_ is the function which estimates the limit of integration in the function
    h(c, v_i) where h(c, v_i) is specified in the Methods section of this chapter
# the proportion 4/7 or 3/7 in the estimated consumption function is referred to
    the number of week days and weekend days (including Friday)

   Umin=function(v){ return(log(consumption)-
                  log((4/7)*exp(const_gamma+c(age,0)%*%gamma+v)*exp(const_beta+
                                        c(age,0)%*%beta+sigmaEsq/2)/(1+
                                        exp(const_gamma+c(age,0)%*%gamma+v))+
```

```
                    (3/7)*exp(const_gamma+c(age,1)%*%gamma+v)*exp(const_beta+
                                        c(age,1)%*%beta+sigmaEsq/2)/(1+
                                        exp(const_gamma+c(age,1)%*%gamma+v))))
                }
   integral = integral2(vecjointerror, -6*(sigmaVsq)^0.5, 6*(sigmaVsq)^0.5, Umin,
       6*(sigmaUsq)^0.5,reltol = 1e-9)
   return(abs(integral$Q-(1-percentile)))
}

 # the optimisation routine for the output from _myintegral.func_ which takes
     _consumption0_ as the starting consumption value

consumption=nlminb(consumption0,myintegral.func,gradient=NULL,control=
                        list(iter.max=500),lower=0.00000001,upper=Inf)

# returns the estimated amount for the specified percentile

return(c(consumption))
}

##############################################################################################

# the function returns percentile of the daily consumption for the specified
    consumed amount
# the required function parameters align with the previous function
# the only new parameter is _consumption_ for which percentile needs to be
    estimated

uncond_consum_percent = function(theta){
age=theta[1]
```

```
const_beta=theta[2]
beta=theta[3:4]
const_gamma=theta[5]
gamma=theta[6:7]
sigmaUsq=theta[8]
sigmaVsq=theta[9]
sigmaEsq=theta[10]
rho=theta[11]
consumption=theta[12]

sigma =
    matrix(c(sigmaVsq,rho*(sigmaVsq*sigmaUsq)^(1/2),rho*(sigmaVsq*sigmaUsq)^(1/2),sigmaUsq),
    ncol=2)
jointerror=function(v,u){
   x=c(v,u)
   return(1/(2*pi)/((det(sigma))^0.5)*exp(-1/2*t(x)%*%solve(sigma)%*%x))
}
vecjointerror=Vectorize(jointerror)
myintegral.func1=function(amount){
   Umin=function(v){ return(log(amount)-
                  log((4/7)*exp(const_gamma+c(age,0)%*%gamma+v)*exp(const_beta+
                                        c(age,0)%*%beta+sigmaEsq/2)/(1+
                                        exp(const_gamma+c(age,0)%*%gamma+v))+
                      (3/7)*exp(const_gamma+c(age,1)%*%gamma+v)*exp(const_beta+
                                        c(age,1)%*%beta+sigmaEsq/2)/(1+
                                        exp(const_gamma+c(age,1)%*%gamma+v))))
                }
   integral = integral2(vecjointerror, -6*(sigmaVsq)^0.5, 6*(sigmaVsq)^0.5, Umin,
       6*(sigmaUsq)^0.5,reltol = 1e-9)
   return(1-integral$Q)
```

```
}

return(myintegral.func1(consumption))
}

##############################################################################

##  Examples of the code application ########

## Consider that the estimated two-part model parameters are the following ########

##  age          =                    -10, age value ( in this example age was
    centred on the sample mean so -10 means that this estimation is for the
    population 10 years younger than average)
##  const_beta =                   3.1 - intercept for the amount part
##  beta = theta [3:4] =       (-0.01, 0.39) are beta-coefficients for age and
    weekend correspondingly in the amount part
##  const_gamma=theta[5] = -0.72 - intercept for the probability part
## gamma=theta[6:7] =       (0.05, 1.37)  are gamma-coefficients for age and
    weekend correspondingly in the probability part
## sigmaUsq=theta[8] =       0.26 - estimated variance of the amount part of the
    random effect
## sigmaVsq=theta[9] =        4.71 - estimated variance of the probability part of
    the random effect
## sigmaEsq=theta[10] =       0.30 - estimated variance of measurement error
    (within-person variation)
## rho=theta[11] =              0.30 - estimated correlation between two model
    parts
## percentile=theta[12] =    0.90   percentile in the region (0, 1)
## consumption0=theta[13] = 100 - initial value of the consumption ( our best
```

```
    guess)

## Then the the function _uncond_consum_amount_ will estimate percentile based on
    the specified model parameters and specified age

## uncond_consum_amount(c(-10, 3.1, -0.001, 0.39, -0.72, 0.05, 1.37, 0.26, 4.71,
    0.30, 0.30, 0.9, 100)) ###

## parameter value of the output after the function run is the required daily
    consumption in the specified population for the specified percentile
## the output should be monitored for the convergence as if the initially supplied
    vlaue is too far from the estimated value convergence issues might arise


################################################################################
```

2.C Method of identification of under-reporters

The analysis required careful consideration of under-reporting. A brief summary of the information used to identify extreme under-reporters in this study is presented. The status of "under-reporter" was defined by the Goldberg cut-off (Black, 2000a) and based on the following information: the within-person coefficient of variation of the total Kcal intake of 24%, the number of dietary records of, on average, 4 days per person, Basal Metabolic Rate (BMR) estimated by Schofield equations (Schofield et al., 1985), Physical Activity Level (PAL) obtained from the WHO recommendations for energy requirements and assigned to be 1.53 to represent sedentary or lightly active lifestyle (Food and Agriculture Organization of the United Nations, United Nations University, World Health Organization, 2004). Reported energy intake (rEI) was calculated as individual average of reported total energy intake over available record days and estimated energy require-

ment was estimated as a product of estimated BMR and PAL. To exclude extreme cases of under-reporting, status of "under-reporter" was assigned to everyone whose degree of under-reporting (rEI:EER) was below 0.75 and ‘’adequate” reporters to those whose degree of under-reporting was greater than or equal to 0.75.

2.D R code for the joint estimation of two-part model parameters and simulations

2.D.1 Joint estimation of two-part model parameters

This Appendix provides R code to jointly estimate the two-part model parameters when two days observations are available per person while taking into account the correlation between the model parts. The notation used is based on Raudenbush (2000) and the code utilises Laplace approximations when integration is required. The code also provides a means to perform simulations to compare the effect of misspecification of the two-part model.

```
## Simulation part starts

simulate=function(m,sigmaEsq.true,sigmaUsq.true,sigmaVsq.true,rho.true,
                  beta0.true,beta1.true,beta2.true,gamma0.true,gamma1.true,gamma2.true){
library(lme4)
library(numDeriv)

loopi=0
coeff.mat=rep(0,17)
REreg.mat=rep(0,8)
logisticREreg.mat=rep(0,7)
pooledOLS.mat=rep(0,6)
```

```
while(loopi<500){

#for N = 500 generate data under two-model assumptions based on age, ethnicity and
    weekend for two days

ID=1:m
age=.45+.1*rnorm(m)
edinichny=rep(1,m)
weekend1=rep(0,m)
weekend2=rep(1,m)
oshibka1=sigmaEsq.true^0.5*rnorm(m)
oshibka2=sigmaEsq.true^0.5*rnorm(m)

tte1=rnorm(m)
tte2=rnorm(m)
tte3=rnorm(m)
u_i=sigmaUsq.true^0.5*((1-rho.true)^0.5*tte1+rho.true^0.5*tte2)
v_i=sigmaVsq.true^0.5*((1-rho.true)^0.5*tte3+rho.true^0.5*tte2)

p1=exp(gamma0.true*edinichny+gamma1.true*age+gamma2.true*weekend1+v_i)/
   (1+exp(gamma0.true*edinichny+gamma1.true*age+gamma2.true*weekend1+v_i))
p2=exp(gamma0.true*edinichny+gamma1.true*age+gamma2.true*weekend2+v_i)/
   (1+exp(gamma0.true*edinichny+gamma1.true*age+gamma2.true*weekend2+v_i))

indicator1=rbinom(m,1,p1)
indicator2=rbinom(m,1,p2)

amount1=indicator1*(beta0.true*edinichny+beta1.true*age+beta2.true*weekend1+u_i+oshibka1)
amount2=indicator2*(beta0.true*edinichny+beta1.true*age+beta2.true*weekend2+u_i+oshibka2)
```

```
# Combine generated data into one matrix

temp1.mat=cbind(ID,edinichny,age,weekend1,weekend2,amount1,amount2,
                indicator1,indicator2,u_i,v_i,oshibka1,oshibka2)

## Simulation part ends

#####################################################################

## Estimation part starts

# function that returns the gradient to my.OLS.likelihood

gradlike = function(theta){

sigmaEsq=theta[1]                      # variance of measurement error in amount
    part
sigmaUsq=theta[2]                     # variance of random effect in amount part
sigmaVsq=theta[3]                     # variance of random effect in probability
    part
rho=theta[4]                             # correlation between the two model parts
beta=theta[5:7]                         # b1 - intercept, b2 - age, b3 - weekend -
    vector of model parameters for amount part
gamma=theta[8:length(theta)]  # gamma1 - intercept, gamma2 - age, gamma3 - weekend
    - vector of model parameters for probability part

# Read data from the created matrix

X1= temp1.mat[,c(2,3,4)]
X2= temp1.mat[,c(2,3,5)]
```

```
I1= temp1.mat[,8]
I2= temp1.mat[,9]

u=seq_along(temp1.mat[, 1])

Day1=I1*(temp1.mat[,6]-X1%*%beta) # array for i = participants
Day2=I2*(temp1.mat[,7]-X2%*%beta) # array for i = participants

sigmaS12Usq=(I1/sigmaEsq+I2/sigmaEsq+1/sigmaUsq/(1-rho^2))^(-1)           # array

  # The remaining power function under the integral to be maximized to find Vhat
  # Notations are based on Raudenbush (2000)

  h.func = function(v){
  muS12U=(Day1/sigmaEsq+Day2/sigmaEsq+
          rho*v/(1-rho^2)/(sigmaUsq*sigmaVsq)^(0.5))*sigmaS12Usq
                                    # array
  sumlogS12U=-t(v)%*%v/2/sigmaVsq-
              1/2*(t(Day1)%*%(Day1)/sigmaEsq+t(Day2)%*%(Day2)/sigmaEsq+
              rho^2*t(v)%*%v/sigmaVsq/(1-rho^2)-sum(muS12U*muS12U/sigmaS12Usq))
  l_v1 = I1%*%(X1%*%gamma+v)-sum(log(1+exp(X1%*%gamma+v)))              # logit
      term, Day 1
  l_v2 = I2%*%(X2%*%gamma+v)-sum(log(1+exp(X2%*%gamma+v)))              # logit
      term, Day 2
  l_v=l_v1+l_v2
  return(-(l_v+sumlogS12U))
  }

 # gradient h
  hgrad.func=function(v){
```

```
  muS12U=(Day1/sigmaEsq+Day2/sigmaEsq+
          rho*v/(1-rho^2)/(sigmaUsq*sigmaVsq)^(0.5))*sigmaS12Usq
                                    # array
  return(-(I1+I2-exp(X1%*%gamma+v)/(1+exp(X1%*%gamma+v))-
           exp(X2%*%gamma+v)/(1+exp(X2%*%gamma+v))-v/sigmaVsq/(1-rho^2)+
           muS12U*rho/(1-rho^2)/(sigmaUsq*sigmaVsq)^(0.5)))
  }

# $par - the value of V_hat; $objective - the maximised value of h.func

V_hat=nlminb(rep(0,time=m),h.func,hgrad.func,control=list(iter.max=400))
muS12U=(Day1/sigmaEsq+Day2/sigmaEsq+
                                                          # array
        rho*V_hat$par/(1-rho^2)/(sigmaUsq*sigmaVsq)^(0.5))*sigmaS12Usq

p1=exp(X1%*%gamma+V_hat$par)/(1+exp(X1%*%gamma+V_hat$par))
# probability of consumption at V_hat, day 1
p2=exp(X2%*%gamma+V_hat$par)/(1+exp(X2%*%gamma+V_hat$par))
# probability of consumption at V_hat, day 2

d3h_dv3_Vhat=-(p1*(1-p1)*(1-2*p1)+p2*(1-p2)*(1-2*p2))
                                      # third derivative of h at Vhat
d2h_dv2_Vhat=-p1*(1-p1)-p2*(1-p2)-1/(1-rho^2)/sigmaVsq+
                                     # second derivative of h at Vhat
         rho^2*sigmaS12Usq/sigmaVsq/sigmaUsq/((1-rho^2)^2)

sumlogV_Vhat=sum(-1/2*log(-d2h_dv2_Vhat))
# sum log ((-d2hdv2_Vhat)^(-1/2))
com_coeff=1/2*d3h_dv3_Vhat/d2h_dv2_Vhat/d2h_dv2_Vhat
# array, common coefficient to the indirect term of the gradients,
```

```
     # is equal to -dl_dvhat/(d2h_d2vhat)^2

# derivates with respect to model parameters

dsigmaS12Usq_dsigmaEsq=(I1+I2)*sigmaS12Usq^2/sigmaEsq^2
dmuS12U_dsigmaEsq=-(Day1/sigmaEsq^2+Day2/sigmaEsq^2)*sigmaS12Usq+
                  (Day1/sigmaEsq+Day2/sigmaEsq+
                   rho*V_hat$par/(1-rho^2)/(sigmaUsq*sigmaVsq)^(0.5))*
                   dsigmaS12Usq_dsigmaEsq

d2h_dvdsigmaEsq=rho/(1-rho^2)/sigmaUsq^0.5/sigmaVsq^0.5*dmuS12U_dsigmaEsq      #
    array

d_dsigmaEsq=sum(-1/2*(I1+I2)/sigmaEsq+1/2*dsigmaS12Usq_dsigmaEsq/sigmaS12Usq+
    #scalar
            1/2*(Day1^2+Day2^2)/sigmaEsq^2+muS12U*dmuS12U_dsigmaEsq/sigmaS12Usq-
            1/2*muS12U^2*dsigmaS12Usq_dsigmaEsq/sigmaS12Usq^2-
            1/2*rho^2/(1-rho^2)^2*dsigmaS12Usq_dsigmaEsq/d2h_dv2_Vhat/sigmaVsq/sigmaUsq+
            com_coeff*d2h_dvdsigmaEsq)


dsigmaS12Usq_dsigmaUsq=sigmaS12Usq^2/sigmaUsq^2/(1-rho^2)
                                          #Y

d3h_dv2_dsigmaUsq_Vhat =
    (rho^2/(((1-rho^2)^2)*sigmaVsq))*((sigmaUsq*dsigmaS12Usq_dsigmaUsq -
    sigmaS12Usq)/sigmaUsq^2) #Y

dmuS12U_dsigmaUsq =
    ((rho*V_hat$par)/(sigmaVsq^0.5*(1-rho^2)))*((dsigmaS12Usq_dsigmaUsq*sigmaUsq^0.5-
```

```
                    0.5*sigmaS12Usq/sigmaUsq^0.5)/sigmaUsq)+
                  (Day1+Day2)*dsigmaS12Usq_dsigmaUsq/sigmaEsq

dh_dsigmaUsq_Vhat = (muS12U*dmuS12U_dsigmaUsq*sigmaS12Usq -
    0.5*muS12U^2*dsigmaS12Usq_dsigmaUsq)/sigmaS12Usq^2

d2h_dvdsigmaUsq = (rho*(dmuS12U_dsigmaUsq*sigmaUsq^0.5 -
    0.5*muS12U/sigmaUsq^0.5))/((1-rho^2)*sigmaVsq^0.5*sigmaUsq)

d_dsigmaUsq = sum(0.5*dsigmaS12Usq_dsigmaUsq/sigmaS12Usq - 1/2/sigmaUsq -
    0.5*d3h_dv2_dsigmaUsq_Vhat/d2h_dv2_Vhat+dh_dsigmaUsq_Vhat+
             com_coeff*d2h_dvdsigmaUsq)

dsigmaS12Usq_dsigmaVsq=0
dmuS12U_dsigmaVsq=-1/2*rho*V_hat$par/(1-rho^2)*sigmaS12Usq/(sigmaVsq^1.5)/(sigmaUsq^0.5)

d2h_dvdsigmaVsq=V_hat$par/(1-rho^2)/sigmaVsq^2+rho/(1-rho^2)/(sigmaUsq^0.5)*
               (dmuS12U_dsigmaVsq*sigmaVsq-muS12U/2)/sigmaVsq^1.5         # array

d_dsigmaVsq=sum(-1/2/sigmaVsq-1/2/d2h_dv2_Vhat/(1-rho^2)/sigmaVsq^2*    # scalar
            (1-rho^2*sigmaS12Usq/(1-rho^2)/sigmaUsq)+
             1/2*(V_hat$par)^2/sigmaVsq^2/(1-rho^2)+
             muS12U*dmuS12U_dsigmaVsq/sigmaS12Usq+com_coeff*d2h_dvdsigmaVsq)

dsigmaS12Usq_drho=-2*rho*sigmaS12Usq^2/sigmaUsq/(1-rho^2)^2
dmuS12U_drho=V_hat$par/(sigmaVsq^0.5)/(sigmaUsq^0.5)*(1+rho^2)/(1-rho^2)^2*sigmaS12Usq+
             dsigmaS12Usq_drho*(Day1/sigmaEsq+Day2/sigmaEsq+
             rho*V_hat$par/(1-rho^2)/(sigmaUsq*sigmaVsq)^(0.5))

d2h_dvdrho=-2*V_hat$par*rho/sigmaVsq/(1-rho^2)^2+1/(sigmaUsq*sigmaVsq)^(0.5)* #
```

```
    array
         (dmuS12U_drho*(1-rho^2)*rho+muS12U*(1+rho^2))/(1-rho^2)^2
dh_drho=-rho*V_hat$par^2/sigmaVsq/(1-rho^2)^2+
         (2*muS12U*dmuS12U_drho*sigmaS12Usq-muS12U^2*dsigmaS12Usq_drho)/2/sigmaS12Usq^2
dh3_d2vdrho=-2*rho/sigmaVsq/(1-rho^2)^2+1/(sigmaUsq*sigmaVsq)*
         (2*rho*(1-rho^4)*sigmaS12Usq+rho^2*(1-rho^2)^2*dsigmaS12Usq_drho)/(1-rho^2)^4

d_drho=sum(1/2*dsigmaS12Usq_drho/sigmaS12Usq+rho/(1-rho^2)-1/2*dh3_d2vdrho/d2h_dv2_Vhat+
    # scalar
        dh_drho+com_coeff*d2h_dvdrho)

dsigmaS12Usq_dbeta=0
dmuS12U_dbeta=-sigmaS12Usq/sigmaEsq*(cbind(I1,I1,I1)*X1+cbind(I2,I2,I2)*X2)
                    # matrix
d2h_dvdbeta=dmuS12U_dbeta*rho/(sigmaUsq*sigmaVsq)^(0.5)/(1-rho^2)
                                # matrix

d_dbeta=colSums(1/sigmaEsq*(cbind(Day1,Day1,Day1)*X1+cbind(Day2,Day2,Day2)*X2)+
    # array 3 elements
        cbind(muS12U,muS12U,muS12U)*dmuS12U_dbeta/sigmaS12Usq+
        cbind(com_coeff,com_coeff,com_coeff)*d2h_dvdbeta)

tt1=p1*(1-p1)*(1-2*p1)
tt2=p2*(1-p2)*(1-2*p2)
dh_dgamma=cbind(I1-p1,I1-p1,I1-p1)*X1+cbind(I2-p2,I2-p2,I2-p2)*X2
d3h_d2vdgamma=-cbind(tt1,tt1,tt1)*X1-cbind(tt2,tt2,tt2)*X2

d2h_dvdgamma=cbind(-p1*(1-p1),-p1*(1-p1),-p1*(1-p1))*X1+
             cbind(-p2*(1-p2),-p2*(1-p2),-p2*(1-p2))*X2
```

```
d_dgamma=colSums(-1/2*cbind(1/d2h_dv2_Vhat,1/d2h_dv2_Vhat,1/d2h_dv2_Vhat)*d3h_d2vdgamma+
                dh_dgamma+cbind(com_coeff,com_coeff,com_coeff)*d2h_dvdgamma)
                                          # array 3 elements



return(c(-d_dsigmaEsq,-d_dsigmaUsq,-d_dsigmaVsq,-d_drho,-d_dbeta,-d_dgamma))
}


my.OLS.likelihood = function(theta) {
sigmaEsq=theta[1]
sigmaUsq=theta[2]
sigmaVsq=theta[3]
rho=theta[4]
beta=theta[5:7]                           # b1 - intercept, b2 - age, b3 -
    weekend - vector of model parameters for amount part
gamma=theta[8:length(theta)]      # gamma1 - intercept, gamma2 - age, gamma3 -
    weekend - vector of model parameters for probability part

X1= temp1.mat[,c(2,3,4)]
X2= temp1.mat[,c(2,3,5)]
I1= temp1.mat[,8]
I2= temp1.mat[,9]

u=seq_along(temp1.mat[, 1])

Day1=I1*(temp1.mat[,6]-X1%*%beta)
Day2=I2*(temp1.mat[,7]-X2%*%beta)
```

```
sigmaS12Usq=(I1/sigmaEsq+I2/sigmaEsq+1/sigmaUsq/(1-rho^2))^(-1)           # array

  h.func = function(v){                                                    # the
     remaining power function under the integral to be maximized to find Vhat
  muS12U=(Day1/sigmaEsq+Day2/sigmaEsq+
         rho*v/(1-rho^2)/(sigmaUsq*sigmaVsq)^(0.5))*sigmaS12Usq          # array
  sumlogS12U=-t(v)%*%v/2/sigmaVsq-
             1/2*(t(Day1)%*%(Day1)/sigmaEsq+t(Day2)%*%(Day2)/sigmaEsq+
             rho^2*t(v)%*%v/sigmaVsq/(1-rho^2)-sum(muS12U*muS12U/sigmaS12Usq))
  l_v1 = I1%*%(X1%*%gamma+v)-sum(log(1+exp(X1%*%gamma+v)))
  # logit term, Day 1
  l_v2 = I2%*%(X2%*%gamma+v)-sum(log(1+exp(X2%*%gamma+v)))
  # logit term, Day 2
  l_v=l_v1+l_v2
  return(-(l_v+sumlogS12U))
  }

  hgrad.func=function(v){                                                  # gradient
     h
  muS12U=(Day1/sigmaEsq+Day2/sigmaEsq+
         rho*v/(1-rho^2)/(sigmaUsq*sigmaVsq)^(0.5))*sigmaS12Usq          # array
  return(-(I1+I2-exp(X1%*%gamma+v)/(1+exp(X1%*%gamma+v))-
           exp(X2%*%gamma+v)/(1+exp(X2%*%gamma+v))-v/sigmaVsq/(1-rho^2)+
           muS12U*rho/(1-rho^2)/(sigmaUsq*sigmaVsq)^(0.5)))
  }

 # $par - the value of V_hat; $objective - the maximised value of h.func

V_hat=nlminb(rep(0,time=m),h.func,hgrad.func,control=list(iter.max=450))
```

```
# Laplace approximation; (the second derivative of h) at Vhat

d2hdv2_Vhat= -exp(X1%*%gamma+V_hat$par)/(1+exp(X1%*%gamma+V_hat$par))^2-
             exp(X2%*%gamma+V_hat$par)/(1+exp(X2%*%gamma+V_hat$par))^2-
             1/sigmaVsq/(1-rho^2)+
             rho^2*sigmaS12Usq/sigmaUsq/sigmaVsq/((1-rho^2)^2)

sumlogV_Vhat=sum(-1/2*log(-d2hdv2_Vhat))                              # sum log
    ((-d2hdv2_Vhat)^(-1/2))
 # probability of consumption at V_hat, day 1
p1=exp(X1%*%gamma+V_hat$par)/(1+exp(X1%*%gamma+V_hat$par))
# probability of consumption at V_hat, day 2
p2=exp(X2%*%gamma+V_hat$par)/(1+exp(X2%*%gamma+V_hat$par))

w1=p1*(1-p1)
w2=p2*(1-p2)

m31=w1*(1-2*p1)
m32=w2*(1-2*p2)
m3=m31+m32

m41=w1*(1-6*w1)
m42=w2*(1-6*w2)
m4=m41+m42

m61=m41*(1-12*w1)-12*(m31)^2
m62=m42*(1-12*w2)-12*(m32)^2
m6=m61+m62
```

```
 # Taylor expansion

return(-(1/2*sum(log(sigmaS12Usq)-(I1+I2)*log(sigmaEsq)-log(sigmaUsq)-log(1-rho^2)-
        log(sigmaVsq))+sumlogV_Vhat-V_hat$objective))
}

resultsnograd = function(sigmaEsq_0, sigmaUsq_0, sigmaVsq_0, rho_0, beta0_0,
                         beta1_0, beta2_0, gamma0_0, gamma1_0, gamma2_0) {

theta=c(sigmaEsq_0, sigmaUsq_0, sigmaVsq_0, rho_0, beta0_0, beta1_0, beta2_0,
        gamma0_0, gamma1_0, gamma2_0)

return(nlminb(theta, my.OLS.likelihood,NULL,control=list(iter.max=500,trace=0),
       lower=c(0.001,0.001, 0.001, -1, -Inf,-Inf,-Inf, -Inf,-Inf,-Inf),
       upper=c(Inf,Inf,Inf, 1, Inf,Inf,Inf, Inf,Inf,Inf)))

}


results = function(sigmaEsq_0, sigmaUsq_0, sigmaVsq_0, rho_0,
                   beta0_0, beta1_0, beta2_0, gamma0_0, gamma1_0, gamma2_0){

theta= c(sigmaEsq_0, sigmaUsq_0, sigmaVsq_0, rho_0, beta0_0, beta1_0, beta2_0,
         gamma0_0, gamma1_0, gamma2_0)

#optimises likelihood

solution.results=nlminb(theta, my.OLS.likelihood,gradlike,
    control=list(iter.max=500,trace=0),
       lower=c(0.001,0.001, 0.001, -.999, -Inf,-Inf,-Inf, -Inf,-Inf,-Inf),
```

```
      upper=c(Inf,Inf,Inf, .999, Inf,Inf,Inf, Inf,Inf,Inf))

 # my.OLS.likelyhood returns -1*likelihood, returns hessian fo,

if(solution.results$convergence==0){
   hess=-jacobian(gradlike,solution.results$par)
   stand.error=(diag(solve(-hess)))^0.5}

else {
   stand.error=rep(0,length(solution.results$par))}

return(c(solution.results$par,stand.error,solution.results$convergence))
}
#########################################################################
indicator.long=c(indicator1,indicator2)
amount.long=c(amount1,amount2)
age.long=c(age,age)
weekend.long=c(weekend1,weekend2)
ID.long=c(ID,ID)

####  starting values for sigmaV and gamma are based on logistic RE regression
    ######################

#print("~~~~~~~~~~~~~~~~~~~~Logistic RE~~~~~~~~~~~~~~~~~~~~~~~~~~~~")

logistic.reg=glmer( indicator.long~ 1+age.long+weekend.long+(1|ID.long), family =
    binomial(),nAGQ=25)

#print("~~~~~~~~~~~~~~~~~Pooled OLS~~~~~~~~~~~~~~~~~~~~~~~~~~~")
```

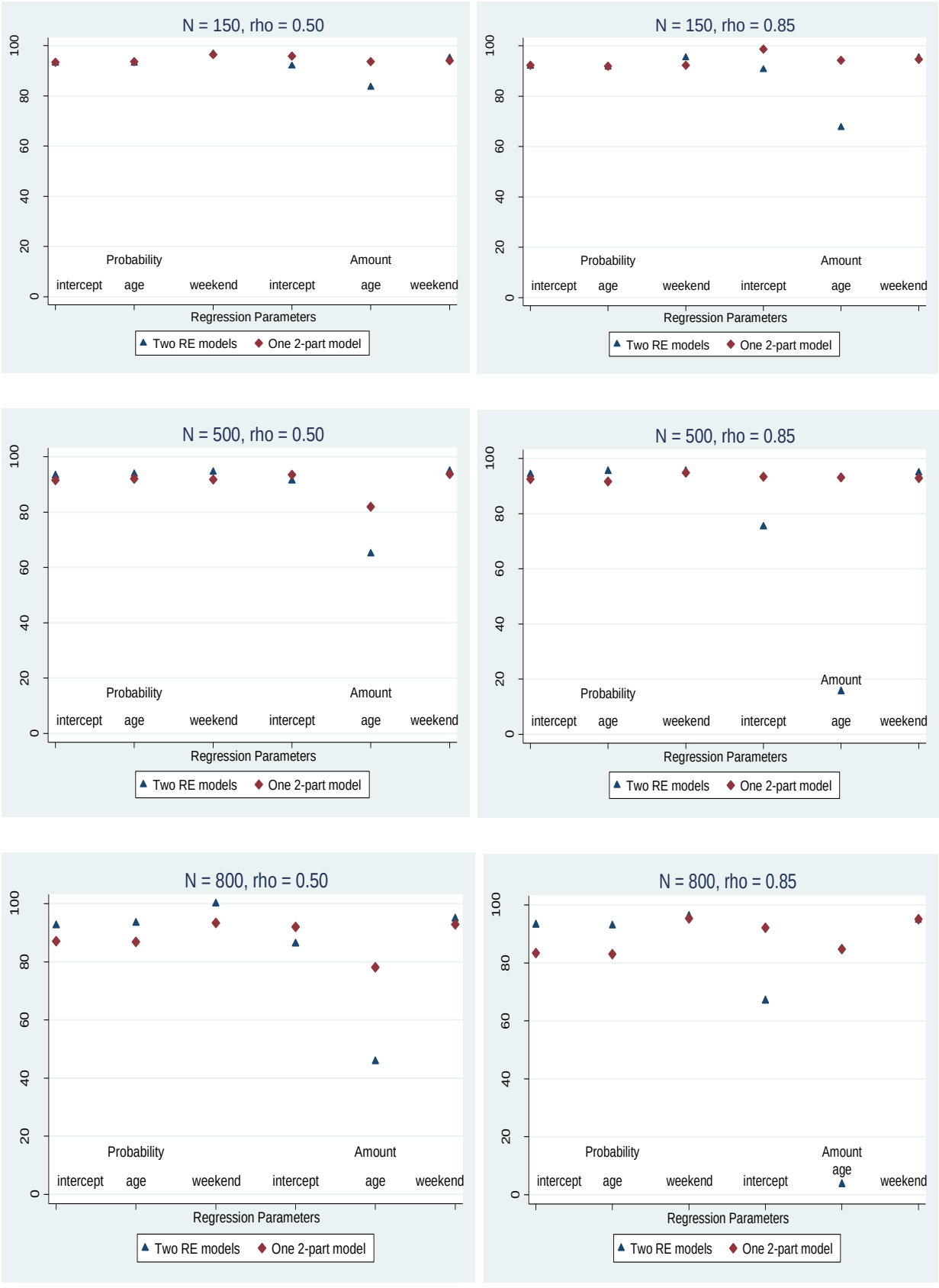
```
amount.long=amount.long[indicator.long!=0]
age.long=age.long[indicator.long!=0]
weekend.long=weekend.long[indicator.long!=0]
OLS.reg=glm( amount.long~ 1+age.long+weekend.long, family = gaussian())

#print("~~~~~~~~~~~~~~~~~~~~~~RE~~~~~~~~~~~~~~~~~~~~~~~~~~")
####  starting points for sigmaE, sigmaU, and betas are based on Random Effects
    regression ######################

indicator.re=apply(cbind(indicator1,indicator2),1,min)
amount.re=c(amount1[indicator.re!=0],amount2[indicator.re!=0])
age.re=c(age[indicator.re!=0],age[indicator.re!=0])
weekend.re=c(weekend1[indicator.re!=0],weekend2[indicator.re!=0])
ID.re=c(ID[indicator.re!=0],ID[indicator.re!=0])

RE.reg=lmer(amount.re~1+age.re+weekend.re+(1|ID.re))

solution=results((getME(RE.reg,"sigma"))^2,(getME(RE.reg,"sigma")*getME(RE.reg,"theta"))^2,
                         (getME(logistic.reg,"theta"))^2,
                         0,fixef(RE.reg)[1],fixef(RE.reg)[2],fixef(RE.reg)[3],
            fixef(logistic.reg)[1],fixef(logistic.reg)[2],fixef(logistic.reg)[3])

if(loopi==0) {coeff.mat=solution
   REreg.mat=
   c((getME(RE.reg,"sigma"))^2,(getME(RE.reg,"sigma")*getME(RE.reg,"theta"))^2,
                               fixef(RE.reg),sqrt(diag(vcov(RE.reg))))
   logisticREreg.mat=
   c((getME(logistic.reg,"theta"))^2,fixef(logistic.reg),sqrt(diag(vcov(logistic.reg))))
  pooledOLS.mat=
  c(OLS.reg$coefficients,sqrt(diag(vcov(OLS.reg))))}
```

```
else        {coeff.mat=cbind(coeff.mat,solution)
   REreg.mat=
   cbind(REreg.mat,c((getME(RE.reg,"sigma"))^2,
                  (getME(RE.reg,"sigma")*getME(RE.reg,"theta"))^2,
                   fixef(RE.reg),sqrt(diag(vcov(RE.reg)))))
   logisticREreg.mat=
   cbind(logisticREreg.mat,c((getME(logistic.reg,"theta"))^2,fixef(logistic.reg),
                                                        sqrt(diag(vcov(logistic.reg)))))
   pooledOLS.mat=
   cbind(pooledOLS.mat,c(OLS.reg$coefficients,sqrt(diag(vcov(OLS.reg)))))}

loopi=loopi+1

}
return(list("POLS"=pooledOLS.mat,"REREG"=REreg.mat,"LOGITRE"=logisticREreg.mat,
                                                    "TWOPART"=coeff.mat))
}
```

2.D.2 Results of simulation

Figure 2.D.1 shows the coverage rate of 95% confidence intervals of model parameters in relation to the model specification and dataset size. Convergence rates differed depending on the sample size of individual dataset and correlation. The lowest convergence rate was 57% ($N = 150, \rho = 0.50$); the highest convergence rate was 73% ($N = 800, \rho = 0.85$ and $N = 500, \rho = 0.50$). The figure demonstrates that the model parameters related to the amount part can have a certain degree of bias under model misspecification. Table 2.D.1 shows the results of model parameters estimation under various assumptions. The variance components parts have less bias when estimation takes the true correlation structure into account. The model parameters related to the probability part are better estimated with the Gaussian quadrature method than the Laplace approximation. However,

Figure 2.D.1: Coverage probability of 95% confidence intervals of model parameters estimated with two separate random effects models and a two-part model with correlated random effects.



the amount part model parameters are biased when the correlation of random effects are not taken into account.

Table 2.D.1: Estimated model parameters from two 2-part models allowing correlated ($\rho \neq 0$) and assuming uncorrelated ($\rho = 0$) random effects, under various scenarios from simulation Study 2

ρ	N	Convergence (%)	True parameters		Model parameters estimates			
					Model allows $\rho \neq 0$		Model assumes $\rho = 0$	
					Mean	SE	Mean	SE
0.85	150	59	σ_ϵ	100	96.3	16.2	99.6	.
			σ_u	50	56.4	24.9	29.5	.
			σ_v	25	38.8	22.5	23.1	.
			ρ	0.85	0.85	0.24	NA	NA
			γ_0	12.7	14.4	4.8	12	3.6
			γ_1	-30	-34.2	11.3	-28.1	8.2
			γ_2	2	1.8	0.6	1.8	0.6
			β_0	900	900.8	5.2	897	5.2
			β_1	-1000	-1003.2	13.8	-981	12.4
			β_2	5	5	1.7	4.9	1.8
	500	70	σ_ϵ	100	98.9	9.3	99.6	.
			σ_u	50	57.2	13.8	29.8	.
			σ_v	25	37.1	13.6	24.2	.
			ρ	0.85	0.9	0.1	NA	NA
			γ_0	12.7	15.4	2.9	12.5	2.1
			γ_1	-30	-36.9	7	-29.6	4.7
			γ_2	2	2.1	0.4	1.9	0.33
			β_0	900	901.2	2.9	896.3	2.8
			β_1	-1000	-1005.3	7.4	-979.4	6.8

Table 2.D.1: Estimated model parameters from two 2-part models allowing correlated ($\rho \neq 0$) and assuming uncorrelated ($\rho = 0$) random effects, under various scenarios from simulation Study 2 (Continued)

ρ	N	Convergence (%)	True parameters		Model parameters estimates			
					Model allows $\rho \neq 0$		Model assumes $\rho = 0$	
					Mean	SE	Mean	SE
			β_2	5	5.2	0.9	5	1
	800	73	σ_ϵ	100	99.6	7.4	99.8	.
			σ_u	50	57.8	11	29.8	.
			σ_v	25	32.5	10.2	25.2	.
			ρ	0.85	0.91	0.1	NA	NA
			γ_0	12.7	15	2.3	12.5	1.6
			γ_1	-30	-36.1	5.5	-29.7	3.7
			γ_2	2	2	0.3	2	0.3
			β_0	900	901.2	2.3	896.3	2.2
			β_1	-1000	-1005.7	5.9	-979.7	5.3
			β_2	5	5.2	0.7	5	0.8
0.5	150	57	σ_ϵ	100	98.4	17.1	99.8	.
			σ_u	50	58.4	24.5	43.2	.
			σ_v	25	31.5	23	35.1	.
			ρ	0.5	0.53	0.29	NA	NA
			γ_0	12.7	14.7	5.1	14.2	4.5
			γ_1	-30	-34.9	11.9	-33.7	10.2
			γ_2	2	1.9	0.7	2.3	0.7
			β_0	900	900.6	5.5	897.8	5.7

Table 2.D.1: Estimated model parameters from two 2-part models allowing correlated ($\rho \neq 0$) and assuming uncorrelated ($\rho = 0$) random effects, under various scenarios from simulation Study 2 (Continued)

ρ	N	Convergence (%)	True parameters		Model parameters estimates			
					Model allows $\rho \neq 0$		Model assumes $\rho = 0$	
					Mean	SE	Mean	SE
			β_1	-1000	-1003.2	14.8	-988	13.5
			β_2	5	5.1	1.7	4.9	1.8
	500	73	σ_ϵ	100	100.5	9.6	100.3	.
			σ_u	50	57.6	13.3	42.9	.
			σ_v	25	32.9	10.9	24.4	.
			ρ	0.5	0.65	0.14	NA	NA
			γ_0	12.7	15	2.8	12.4	2.1
			γ_1	-30	-35.6	6.5	-29.5	4.7
			γ_2	2	2	0.37	2	0.33
			β_0	900	901.5	3	898.1	3.1
			β_1	-1000	-1006.6	8.1	-988.5	7.3
			β_2	5	5.2	0.9	4.9	1
	800	70	σ_ϵ	100	100.2	7.6	99.5	.
			σ_u	50	61.3	11	44.1	.
			σ_v	25	35	10.6	25.5	.
			ρ	0.5	0.7	0.1	NA	NA
			γ_0	12.7	15.3	2.3	12.6	1.7
			γ_1	-30	-36.4	5.4	-29.9	3.8
			γ_2	2	2	0.31	2	0.27

Table 2.D.1: Estimated model parameters from two 2-part models allowing correlated ($\rho \neq 0$) and assuming uncorrelated ($\rho = 0$) random effects, under various scenarios from simulation Study 2 (Continued)

ρ	N	Convergence (%)	True parameters		Model parameters estimates			
					Model allows $\rho \neq 0$		Model assumes $\rho = 0$	
					Mean	SE	Mean	SE
			β_0	900	901.4	2.4	897.7	2.4
			β_1	-1000	-1007.8	6.3	-987.6	5.8
			β_2	5	5.3	0.7	5	0.8

Chapter 3

Socio-economic and personal determinants of food intake in the UK using a two-part mixed-effects model with correlated random effects

This chapter demonstrates the application of the two-part model introduced in Chapter 2 to estimate the effect of various personal and socio-economic factors on consumption of various foods of public health importance. The analyses account for excess zeros, measurement errors and unobserved correlation between probability of consumption and portion size. The analysis are based on dietary data from a sub-sample of population from the UK National Diet and Nutrition Survey Rolling Programme.

3.1 Background

Understanding diet choice is complex due to the intricate interplay of personal and environmental factors affecting dietary intake (Lusk et al., 2013; Shepherd and Raats, 2013). Additionally, records of dietary intake often have substantial natural within-person varia-

tion, and measurement error arising from imperfect measurement instruments (Beaton et al., 1979; Nusser et al., 1987; Rutishauser and Black, 2002; Basiotis et al., 1987; Nelson et al., 1989; Kipnis et al., 2002). Further, most diet diaries and recalls collect records over 2-4 days only and, consequently, dietary data often contain a high proportion of zero records as consumption may not occur during the measurement period. Hence, careful statistical modelling to account for various variability sources and excess zeros is required to minimise potential bias in analysis and interpretation. This chapter demonstrates the application of novel statistical methods in observational research to provide reliable inferences on food intake determinants in a free-living population in an uncontrolled environment. We analyse data collected as a part of the UK National Diet and Nutrition Survey Rolling Programme (NDNS RP), which provides information on food intake, collected with a 4-day food diary, along with various personal, lifestyle and socio-economic determinants of nutritional behaviour of a representative sample of adults drawn from the UK population. An extensive literature exists on the potential determinants of people's food choices (Bufe et al., 2005; Kim and Drayna, 2005; Mennella et al., 2005; Navarro-Allende et al., 2008; Nicklaus, 2011; McGowan et al., 2012; Ranjit et al., 2015; Rasmussen et al., 2006; Kröller and Warschburger, 2009; Turrell et al., 2003; Maguire and Monsivais, 2015; Novaković et al., 2014; Galobardes et al., 2001; Rozin et al., 2011). However, despite growing understanding behind people's diet preferences, governmental dietary programmes to encourage healthier eating habits in the general population are not reaching their target audience. This is reflected in recent findings from the NDNS RP Years 1-4 which showed that a large proportion of the UK population do not adhere to the nutritional recommendations by Public Health UK. These findings were echoed by a review of various diet interventions aimed at increasing fruits and vegetables consumption that showed small or inconsistent effects on the desired behaviours (Appleton et al., 2016). A recent review of interventions aiming to modify discretionary food choices highlighted some potentially useful strategies but no single most effective strategy was identified (Grieger et al., 2016). Overall, current research suggests a need for more targeted interventions and policies

addressing specific lifestyles and food preferences. To achieve this, better understanding of the determinants, driving forces and constraints behind people's long-term usual food choices in every-day life, outside experimental settings, is required. The task is not trivial and requires reliable analytical and measurement tools. Currently, the most reliable and widely available long-term food intake measurement tools are multiple-day food diaries or multiple 24-hour recalls (Bingham et al., 1994; Kipnis et al., 2003; Arab et al., 2011). These food records, however, are subject to a large within-person daily variation, which is observed along with the true personal long-term intake (Nelson et al., 1989). The prevailing statistical approach is to compute individual averages which are used as proxies for long-term true personal intake in ordinary least squares (OLS) regression. However, using proxies in the analysis leads to artificially inflated food intake variance, which, in turn, reduces the chance of detecting an important predictor. The statistical analysis of intake of occasionally-consumed foods, such as alcohol, fruits, fish, and certain vegetables, is further complicated by a high frequency of zero intake records making OLS or logistic regression of limited inferential value (Tooze et al., 2006). Therefore, nutritional data analytical tools should account for within-person variation, measurement error and excess zeros to avoid violations of model assumptions which can lead to incorrect analyses, data misinterpretation or inefficient use of available information. This chapter investigates the effect of personal, lifestyle and socio-economic determinants on the consumption of certain nutrients and food groups, shown to be of public health importance (Guenther et al., 2013; Department of Health, 2014; US Department of Health and Human Services, 2015), in an adult population drawn from NDNS RP data (Years 2-4). We utilise a two-part statistical model with random effects (Olsen and Schafer, 2001; Tooze et al., 2002) which accounts for within-person variation and excess zeros in occasionally-consumed foods. The analysis simultaneously adjusts for multiple correlated factors, such as education, income and occupation, to minimise bias arising from confounding. This analysis is the first, to the best of the author's knowledge, to take advantage of contemporary statistical advances and a unique data source to provide robust and detailed evidence on the

determinants of food intake in a sample representative of the general adult UK population.

3.2 Methods

3.2.1 The UK National Diet and Nutrition Survey Rolling Programme

The NDNS RP is an annual cross-sectional survey, jointly funded by Public Health England and the UK Food Standards Agency, undertaken since 2008. The survey aims to assess diet, nutrient intake and nutritional status of the UK population aged 18 months and older, living in private households. A nationally representative sample of the UK households is selected by a multistage sampling procedure: firstly, the Postcode Address File (PAF), which contains all the addresses in the UK, is accessed to sample Primary Sampling Units (PSUs). These are small geographical areas formed by neighbouring postcodes. Secondly, twenty seven addresses are sampled from a selected PSU at random, where either an adult or a child is selected. The 27 addresses are randomly allocated to one of two groups to determine whether an adult (aged 19 years or over) and a child (aged 1.5 to 18 years), or a child only, are selected for interview. At nine of the selected addresses the interviewer selects one adult and, where present, one child for inclusion in the survey. The remaining 18 addresses form a "child boost" where only households with children are selected. Where more than one person is eligible the participants are selected using a random selection procedure.

The data collection included an estimated four-day food diary, life-style factors, socio-economic measurements and demographics. Diary response rate was 56%. Further details can be found in Public Health England (2014).

3.2.2 Dietary data collection

Estimated four-day food diaries were supplemented with pictures of 15 frequently consumed foods in small, medium and large portions. Records included portion sizes (household measures), brand names and label/wrapper information for unusual or ready-made meals. Diaries were coded by trained coders and editors. Food intakes were entered into

an NDNS RP specific version of the dietary assessment system DINO (Diet In Nutrients Out), an all-in-one dietary recording and analysis system with the food composition data used from the Department of Health's (DH NDNS) Nutrient Databank (Fitt et al., 2010).

3.2.3 Data sample

The data sample analysed in this chapter comprises individuals older than 18 years drawn from NDNS RP Years 2, 3, 4 (2009-2012). The initial sample available were 702 males and 899 females. Of them 45 (6.4%) males and 113 (12.6%) females were excluded due to dieting, further 65 (9.3%) and 79 (8.8%) females were excluded due to potential extreme under-reporting. Extreme under-reporting was defined as the reported energy intake of less than 80% of the estimated resting energy expenditure which was based on Mifflin equations that adjust for weight, height, age and gender (Mifflin et al., 1990). For some participants, information on food intake predictors was missing providing the final sample of 509 males and 618 females.

3.2.4 Food groups and nutrients

Key macronutrients and food groups were selected for analysis according to current nutritional guidelines (Department of Health, 2014; US Department of Health and Human Services, 2015) and public health importance. These comprised alcohol (g), energy (Kcal), protein (g), saturated fatty acids (SFA) (g), monounsaturated fatty acids (MUFA) (g), Omega-3 and Omega-6 fatty acids (g), trans-fatty acids (TFA) (g), fibre (g), starch (g), extrinsic sugars (Kcal), fruits (g), cooked vegetables (g), raw and salad vegetables (g), processed meat (g), oily fish (not canned tuna) (g) and sugary beverages (g).

3.2.5 Demographics, socio-economic and life-style factors

Demographic factors comprised age; sex; ethnicity (white; non-white); body mass index (BMI); lipid and blood pressure lowering medications taken (yes; no); self-assessed general health problems (no; yes, but no impact on mobility; yes, with impact on mobility); partner status (married or in partnership; never married or lived in partnership; previously

married or lived in partnership but now single).

Life-style factors comprised take away shopping habits (rarely or never; once or twice per month; every week or more often); fruits and vegetables shopping habits (less than weekly; weekly and more often); being a non-meat-eater (yes; no); smoking (never smoked; quit >10 years ago; quit $\leq$ 10 years ago; current smoker; occasional smoker); alcohol consumption (never drink; rarely drink; the rest); moderate-to-vigorous physical activity (MVPA) assessed through a recent self-completed Physical Activity Questionnaire, expressed in min per day and combined from four domains: home, commuting, work and leisure (0 min ; 0–10 min; 10–20 min; 20–40 min; 40–60 min; $\geq$ 60 min) (Mindell, 2014).

Socio-economic factors comprised education as the highest obtained degree (bachelor degree and above; unfinished degree; current student; A levels; GCSE grades A-C; GCSE grades below C or no qualifications; foreign degree); socio-economic status (never worked and others; routine; semi-routine; lower supervisory and technical; small employers and own account; intermediate; lower managerial and professional; higher managerial and professional); income over the previous 12 months as assessed through self-report and equalised to take into account the household composition by a rescaled version of the Organisation for Economic Development modified equivalence scale (Anyaegbu, 2010) (<£15,000; £15,000–£25,000; £25,000–£35,000; £35,000–£50,000; $\geq$ £50,000). For some people (71 (14%) for men and 85 (13.8%) for women) information was not collected so they were assigned into a separate category; tenure (own or mortgage; renting privately; renting from local authority).

3.2.6 Statistical analysis

Statistical models were developed separately for occasionally- and habitually-consumed foods intakes. Occasionally-consumed food intake is characterised by a high frequency of zero values. Histograms were used to distinguish these from habitually-consumed foods which are consumed most days. All models were adjusted for survey year, weekend and age. Male and female records were analysed separately. Habitually-consumed food

intakes were analysed with linear mixed-effect regression models (Diggle et al., 2002). The intakes from occasionally-consumed foods were analysed with two-part mixed-effects models with correlated random effects. The model consists of the probability part which estimates the probability of food consumption for a given individual (modelled with logistic regression with a random intercept) and the amount part which estimates the amount of food consumed when consumption took place (modelled with linear mixed-effects regression model with a random intercept). The parts are linked by allowing the two random intercepts to be correlated. This allows portion sizes and consumption frequency correlate through unobserved individual preferences. Further details on the two-part model and its application in epidemiology can be found in Appendix 3.B.1 and Tooze et al. (2002); Olsen and Schafer (2001); Smith et al. (2015); Chernova and Solis-Trapala (2016); Liu et al. (2010). Natural log-transformation was applied when the distributions of continuous variables were skewed.

Model selection was based on a backward selection process. This strategy was chosen because the potential risk factors are correlated and it is unknown a priory, which of these factors will present themselves as most significant when modelling a particular food intake. This strategy allows us to find a combination of correlated risk factors, which might show significant model improvement together but not when fitted separately.

An initial model A for food intake contained all the potential predictors. The predictors with p-values larger than 0.20 were removed one at a time. Secondly, an intermediate model B contained the predictors retained at the first step. Thirdly, the remaining predictors stayed in the final model C if they were significant at 10% significance level. The choice of moderate significance levels at each step reflects the large number of model parameters, medium size dataset and potential correlation between the predictors. Likelihood ratio tests were used at each step of model selection. The Akaike information criterion (AIC) was used if a selection between non-nested models had to be made and the model with the smallest AIC was retained (Burnham and Anderson, 2002). The results of the final

model are presented.

Model interpretation: the regression parameters of the linear mixed-effects model in the two-part model are subject-specific. For example, the difference in consumption at weekend compared to a week day is a within-person difference, whereas a between-person difference would have to be assessed taking into account the value of the unobserved person-specific effects. A marginal interpretation of regression coefficients, i.e. the effect averaged over the whole population may be of interest in some instances. Examples include ethnicity or parental background, and cases where a predictor's effects go in the opposite directions for the probability and amount model parts. Numerical integration is required to estimate marginal effects of covariates in non-linear models and further technical details can be found in Appendix 3.B.2 and Su et al. (2011).

The residuals from the portion part of the two-part model were estimated on the logarithmic scale for model diagnostics plots. The portion sizes were predicted based on empirical Bayes rules and then residuals were estimated as the difference between observed and predicted intakes.

The complex survey design includes individual combined selection weights to adjust for different probabilities of dwelling unit, catering unit and individuals' selection, with key variables for creating weights including age, sex and Government Region Office. The presented analysis is model-based and is not meant to provide inferences on population statistics therefore, it does not adjust for weights or clustering (Muthen and Satorra, 1995; Chambers et al., 2012).

This decision is discussed in more detail in Appendix 3.B.3. Standard errors were estimated by using a Huber-White-sandwich estimator. Stata 14 software was utilised to analyse the data (StataCorp. 2015. Stata Statistical Software: Release 14. College Station, TX: StataCorp LP).

3.3 Results

Personal, socio-demographic and lifestyle sample characteristics are shown in Table 3.1. The study sample comprised 509 men, mean (sd) age 48.3 (17.5), BMI 27.4 (4.5) and 610 women with mean (sd) age 48.2 (18.1), BMI 26.7 (5.7). The majority of participants are of white ethnicity (over 92%), self-reported healthy (over 60%) and have a wide spread of lifestyle and socio-economic characteristics.

Table 3.1: Personal, lifestyle and socio-demographic characteristics of the sample population selected for analysis from the NDNS RP Years 2-4

	Males (N = 509)		Females (N = 618)	
	N (mean)	% (SD)	N (mean)	% (SD)
Age, years (mean (SD))	(48.3)	(17.5)	(48.2)	(18.1)
BMI (mean(SD))	(27.4)	(4.5)	(26.7)	(5.7)
Survey Year				
2009/2010	170	33.4	202	32.7
2010/2011	146	28.7	178	28.8
2011/2012	193	37.9	238	38.5
Ethnicity				
white	471	92.5	571	92.4
non-white	38	7.5	47	7.6
Health, self-reported				
no health problems	341	67.0	381	61.7
health problems, no mobility restrictions	86	16.9	110	17.8
health problems, mobility restrictions	82	16.1	127	20.6
Lipid lowering drug taken				
no	451	88.6	560	90.6
yes	58	11.4	58	9.4
Blood pressure lowering drug taken				
no	454	89.2	544	88.0
yes	55	10.8	74	12.0
Partner				
married or live in partnership	273	53.6	266	43.0
never married or lived in partnership	148	29.1	168	27.2
previously married or lived in partnership but now single	88	17.3	184	29.8
Lifestyle				
Smoking				
never smoked	257	50.5	355	57.4
ex-smoker, quit >10 year ago	77	15.1	73	11.8
ex-smoker, quit <=10 years ago	46	9.0	65	10.5
current	115	22.6	116	18.8
occasional	14	2.8	9	1.5

Table 3.1 Personal, lifestyle and socio-demographic characteristics of the sample population selected for analysis from the NDNS RP Years 2-4 (Continued)

	Males (N = 509)		Females (N = 618)	
	N (mean)	% (SD)	N (mean)	% (SD)
Moderate to vigorous physical activity, min/day				
0	14	2.8	15	2.4
0-10	62	12.2	138	22.3
10-20	33	6.5	76	12.3
20-40	71	14.0	116	18.8
40-60	65	12.8	70	11.3
>60	264	51.9	203	32.9
Drinking				
yes	444	87.2	524	84.8
rarely	28	5.5	37	6.0
never	37	7.3	57	9.2
Fruits and vegetables buying habits				
weekly or more often	472	92.7	575	93.0
less often than weekly	37	7.3	43	7.0
Non-meat eaters	9	1.8	23	3.7
Take away habit				
rarely or never	200	39.3	291	47.1
less than once a week	198	38.9	204	33.0
once a week and more	111	21.8	123	19.9
Socio-economic				
Tenure				
mortgaged or owned	355	69.7	433	70.1
rented privately	90	17.7	89	14.4
rented from local authority	64	12.6	96	15.5
Qualifications				
bachelor degree and above	133	26.1	136	22.0
unfinished degree	50	9.8	64	10.4
students	20	3.9	32	5.2
A levels	83	16.3	98	15.9
GCSE A_C	86	16.9	114	18.5
GCSE below C and no qualifications	110	21.6	146	23.6
foreign qualifications	27	5.3	28	4.5

Table 3.1 Personal, lifestyle and socio-demographic characteristics of the sample population selected for analysis from the NDNS RP Years 2-4 (Continued)

	Males (N = 509)		Females (N = 618)	
	N (mean)	% (SD)	N (mean)	% (SD)
Social Status				
higher managerial and professional occupation	104	20.4	86	13.9
lower managerial and professional occupation	134	26.3	174	28.2
intermediate occupations	45	8.8	63	10.2
small employers and own account workers	50	9.8	74	12.0
lower supervisory and technical occupation	49	9.6	47	7.6
semi-routine occupations	60	11.8	93	15.1
routine occupations	56	11.0	58	9.4
never worked or other	11	2.2	23	3.7
Equalised household income, £ 1000				
<=15	70	13.8	134	21.7
15-25	99	19.5	124	20.1
25-35	95	18.7	110	17.8
35-50	79	15.5	87	14.1
>50	95	18.7	78	12.6
missing	71	14.0	85	13.8

Table 3.2 displays the median and interquantile range of intake of macronutrients which are habitually consumed while Table 3.3 shows the sample distributions of daily intake of occasionally- consumed foods.

Table 3.2: The sample distribution of macronutrients intake in males (N = 509) and females (N=618)

	Males		**Females**	
Macro nutrients	Median	IQR	Median	IQR
Energy, Kcal	2200	1900 - 2500	1660	1460 - 1860
Protein, g	87	73 - 100	67	58 - 75
SFA, g	30.4	24.5 - 36.4	23.9	18.9 - 28.8
Fibre, g	15.6	12.7 - 18.5	13.1	10.7 - 15.6
Starch, g	151	125 - 177	113	97 - 129
MUFA, g	30	25 - 36	23	19 - 27
TFA, g	1.4	1.1 - 1.9	1.1	0.8 - 1.5
NMES, g	68	44-106	48	30-74
Omega 3 FA, g	2.2	1.8 - 2.8	1.8	1.5 - 2.1
Omega 6 FA, g	11	9 - 14	9	7 - 11
Total sugar, g	114	85 - 143	89	68 - 110

A description of food intake by number of consumption (Table 3.4) helps illustrate the need to use a two-part model for estimation of intake of occasionally-consumed foods to minimise inferential bias. The percentage of participants who reported zero intakes on all four days of food diaries ranged from 6% for fruits to 73% for oily fish intake. In men, the total number of zero records is 1252 (61.7%) for alcohol, 703 (34.6%) for fruits, 1255 (61.8%) for salad vegetables, 967 (47.6%) for cooked vegetables, 1063 (52.4%) for processed meat and 1446 (71.2%) for soft drinks. We also notice that the portion sizes can relate to consumption frequency and tend to increase for the majority of the foods. For example, sugary drinks median portion size increases by 70%(women) and 85% (men)

Table 3.3: The sample distributions of intake of occasionally-consumed foods in males (N = 509) and females (N=618). Correlation presented in the table shows the extent to which probability of consumption and portion size is correlated

Occasionally-consumed food groups	Median	IQR	Corr	p-value
Males				
Alcohol, g	12.4	0-29.8	0.23	<0.001
Fruits, g	73	31 -163	0.67	<0.001
Cooked vegetables, g	85	53 - 131	0.43	0.005
Salad and raw vegetables, g	25	9 - 60	0.39	0.002
Processed meat, g	39	23 - 62	0.12	0.390
Oily fish, g	6	2 - 15	-0.45	0.016
Sugary drinks (not juice), g	50	7 - 212	0.41	0.003
Females				
Alcohol, g	6.1	0-15.3	0.25	0.030
Fruits, g	86	36 -180	0.55	<0.001
Cooked vegetables, g	76	52 - 108	0.29	0.083
Salad and raw vegetables, g	39	19 - 70	0.24	0.095
Processed meat, g	26	15 - 41	-0.03	0.830
Oily fish, g	5	2 - 11	-0.14	0.663
Sugary drinks (not juice), g	34	7 - 126	0.44	<0.001

when consumption frequency increases from one to four recorded consumption days. This is further supported by the estimated correlation of random effects (Table 3.3). For example, healthy eating habits like fruits and vegetables consumption show a high positive correlation of random effects indicating that those with higher consumption frequency also consume bigger portions. It also supports the presence of person-specific characteristics which drive both, the consumption frequency and the consumption portions.

Due to its high public health significance, the results from modelling alcohol intake are

Table 3.4: Median portion sizes (g) of occasionally-consumed foods relative to consumption frequency and the number of non-consumption days (N(%)) as reported in food diaries

	The number of consumption days				
	0	1	2	3	4
	N (%)	Median daily portion size, g			
Males					
Alcohol	183 (36.0)	28.2	37.9	36.3	37.9
Fruits	55 (10.8)	19	83	112	144
Cooked vegetables	52 (10.2)	121	149	171	171
Salad and raw vegetables	147 (28.9)	60	73	87	105
Processed meat	83 (16.3)	75	70	87	96
Oily fish	368 (72.3)	89	72	72	111
Sugary drinks (not juice)	257 (50.5)	301	388	522	558
Females					
Alcohol	277 (44.8)	22.5	26.8	28.9	25.8
Fruits	39 (6.3)	65	99	104	153
Cooked vegetables	62 (10.0)	112	125	136	147
Salad and raw vegetables	143 (23.1)	75	88	90	105
Processed meat	129 (20.9)	50	67	55	68
Oily fish	452 (73.1)	61	62	62	60
Sugary drinks (not juice)	316 (51.1)	260	300	333	440

reported in the following dedicated section, prior to the presentation of results for the intake of macronutrients and foods. The tables presented in the remainder of this chapter display estimates from the final model for the intake of each food, *i*.e., the statistically significant regression parameters for clarity. To aid interpretation of results forest plots of effect sizes are shown in Figures 3.1-3.4.

3.3.1 Alcohol intake

Tables 3.5 and 3.6 present the results from modelling alcohol intake in men and women respectively. Alcohol consumption increases during the weekend (including Friday) compared to weekdays: odds of consumption increase by 3.72, 95% CI: (2.77, 4.99) for men and 3.00, 95% CI: (2.32, 3.89) for women (both p-values <0.001) and portion size increases by 1.31 95% CI: (1.18, 1.46) times in men and 1.21 95% CI: (1.21, 1.49) times in women (both p-values<0.001). In men, the increase was associated with being a smoker or having recently quit, being a student, being divorced or widowed, and frequent takeaways consumption. Decrease in alcohol consumption was associated with being other than white ethnicity, privately renting compared to owning or paying mortgage and equalised income of less than £25,000.

In women, increase in alcohol intake was associated with occasional/regular takeaways consumption, being in lower managerial or intermediate occupations, being divorced or widowed. Decrease was related to taking lipid lowering medicine, and having lower education and lower socio-economic status.

Marginal effects. The expected daily alcohol consumption for groups was estimated for base groups in men and women (see Methods section for details) and found to be 21g and 13g of alcohol respectively. Additionally, in men, current smokers show increased daily alcohol consumption by 14.7g compared to the base group. In women, occasional or regular takeaway consumption is associated with daily increase of 4.4g of alcohol intake.

Table 3.5: Alcohol intake predictors in men in the sample population selected for analysis from the NDNS RP Years 2-4

	SUBJECT-SPECIFIC TWO-PART MODEL EFFECTS					
	PROBABILITY (N = 1882)			CONDITIONAL AMOUNT (N = 778)		
Predictors	Odds Ratio (95%CI)	*Wald p-value*	*LR p-value*[2]	Relative Change (95% CI)	*Wald p-value*	*LR p-value*[2]
Weekend[1]	3.72 (2.77, 4.99)		*<0.001*	1.33 (1.18, 1.50)		*<0.001*
Survey Year (base: 2009/2010)						
2010/2011	0.63 (0.35, 1.15)	*0.134*		1.20 (0.98, 1.48)	*0.076*	
2011/2012	0.76 (0.44, 1.31)	*0.328*		0.97 (0.78, 1.19)	*0.750*	
Age, years	1.036 (1.019,1.054)		*<0.001*	0.99 (0.99, 1.00)		*0.046*
Ethnicity (non-white)	0.26 (0.09, 0.75)		*0.013*			
Partner (base: married)						*0.005*
never married				1.14 (0.92, 1.43)	*0.237*	
previously married				1.39 (1.13, 1.72)	*0.002*	
Smoking (base: never smoked)			*0.095*			*0.002*
ex, quit >10 year ago	1.79 (0.86, 3.74)	*0.119*		1.00 (0.76, 1.32)	*0.987*	
ex, quit <=10 years ago	1.45 (0.64, 3.27)	*0.369*		1.37 (1.01, 1.86)	*0.042*	
current	2.35 (1.24, 4.46)	*0.009*		1.47 (1.18, 1.83)	*0.001*	
occasional	1.12 (0.27, 4.64)	*0.873*		2.10 (1.49, 2.96)	*<0.001*	
Alcohol habit consumption (base: regular)						
rarely	0.02 (0.01, 0.10)		*<0.001*	0.09 (0.04, 0.20)		*<0.001*
never	NA			NA		
Take away (base: rarely or never)						*0.033*
occasionally				0.99 (0.81, 1.22)	*0.944*	
regularly				1.28 (1.01, 1.61)	*0.040*	
Tenure (base: mortgaged or owned)			*0.073*			
privately rented	0.47 (0.25, 0.91)	*0.025*				
local authority rented	0.98 (0.44, 2.18)	*0.964*				

Table 3.5 Alcohol intake predictors in men in the sample population selected for analysis from the NDNS RP Years 2-4 (Continued)

McClement equivalence score, £ 1000			*0.028*	
<=15	0.39 (0.17, 0.93)	*0.032*		
15-25	0.32 (0.15, 0.67)	*0.003*		
(base) 25-35	1			
35-50	0.76 (0.34, 1.66)	*0.487*		
>50	0.85 (0.41, 1.78)	*0.674*		
Qualifications	1		*0.031*	
(base) bachelor degree and above				
unfinished degree	0.63 (0.27, 1.45)	*0.281*		
current students	1.51 (0.42, 5.40)	*0.524*		
A levels	0.43 (0.21, 0.88)	*0.020*		
GCSE A_C	0.51 (0.25, 1.05)	*0.066*		
GCSE below C and no qualifications	0.29 (0.14, 0.63)	*0.002*		
foreign qualifications	0.63 (0.21, 1.90)	*0.413*		
Residual between-person SD	1.95 (1.68, 2.26)			0.62 (0.53, 0.71)
Residual within-person SD				0.66 (0.62, 0.71)
Residual within-person correlation				0.46 (0.37, 0.55)
Residual correlation between model parts	0.37 (0.17, 0.58) *(p-value< 0.001)*			

[1]Weekend includes Friday.

[2]p-value from likelihood ratio (LR) test of joint significance in case when a factor variable has more than two levels.

Table 3.6: Alcohol intake predictors in women in the sample population selected for analysis from the NDNS RP Years 2-4

	SUBJECT-SPECIFIC TWO-PART MODEL EFFECTS					
	PROBABILITY (N = 2234)			CONDITIONAL AMOUNT (N = 728)		
Predictors	Odds Ratio (95%CI)	*Wald p-value*	*LR p-value*[2]	Relative Change (95%CI)	*Wald p-value*	*LR p-value*[2]
Weekend[1]	2.97 (2.26, 3.92)		*<0.001*	1.34 (1.21, 1.49)	*<0.001*	
Survey Year						
(base) 2009/2010	1			1		
2010/2011	0.73 (0.45, 1.19)	*0.201*		1.08 (0.89, 1.32)	*0.427*	
2011/2012	0.83 (0.53, 1.31)	*0.419*		1.17 (0.98, 1.40)	*0.082*	
Age, years	1.02 (1.01, 1.04)		*0.001*	0.99 (0.98, 0.99)		*<0.001*
BMI, kg/m^2				1.02 (1.00, 1.03)		*0.067*
Lipid lowering medicine	0.41 (0.18, 0.94)		*0.036*	0.73 (0.50, 1.08)		*0.118*
Partner						*0.034*
(base) married				1		
never married				1.08 (0.89, 1.32)	*0.439*	
previously married				1.28 (1.07, 1.52)	*0.006*	
Alcohol habit consumption			*<0.001*			*0.002*
(base) regular	1			1		
rarely	0.08 (0.03, 0.21)			0.26 (0.11, 0.62)		
never	N/A			N/A		
Take away shopping			*0.024*			*0.078*
(base) rarely or never	1			1		
Occasionally/regularly	1.64 (1.07, 2.53)			1.18 (0.98, 1.43)		
Qualifications			*0.027*			
(base) bachelor degree and above	1					
unfinished degree	1.10 (0.55, 2.19)	*0.783*				
current students	0.35 (0.14, 0.88)	*0.026*				
A levels	0.55 (0.28, 1.06)	*0.074*				
GCSE A_C	0.69 (0.38, 1.24)	*0.216*				
GCSE below C and no qualifications	0.60 (0.32, 1.14)	*0.120*				
foreign qualifications	0.27 (0.08, 0.94)	*0.039*				

Table 3.6 Alcohol intake predictors in women in the sample population selected for analysis from the NDNS RP Years 2-4 (Continued)

Socio-economic status			*0.003*	
(base) higher managerial and professional occupation	1			
lower managerial and professional occupation	1.83 (1.05, 3.20)	*0.034*		
Intermediate occupations	1.35 (0.62, 2.93)	*0.450*		
Small employers and own account workers	1.26 (0.62, 2.58)	*0.526*		
Lower supervisory and technical occupation	0.45 (0.17, 1.21)	*0.116*		
Semi-routine occupations	0.73 (0.36, 1.50)	*0.390*		
Routine occupations	0.51 (0.22, 1.16)	*0.106*		
Never worked and Other	0.60 (0.17, 2.09)	*0.426*		
Residual between-person SD	1.68 (1.44, 1.93)			0.57 (0.49, 0.64)
Residual within-person SD				0.57 (0.53, 0.61)
Residual within-person correlation				0.50 (0.41, 0.58)
Residual correlation between model parts	0.31 (0.09, 0.53) *(p-value 0.006)*			

[1]Weekend includes Friday.

[2]p-value from likelihood ratio (LR) test of joint significance in case when a factor variable has more than two levels.

3.3.2 Personal, socio-economic and lifestyle characteristics in relation to food intake in men

Below is a summary of the most interesting trends in relationships between determinants and food intake described by predictor. The full results of modelling intake of macronutrients and foods are presented in Appendix 3.A. The trends and overview of relationships between the considered predictors and considered occasionally- consumed food intakes, along with p-values, are summarised in Table 3.7.

Weekend is associated with increase in energy, protein, MUFA, SFA, TFA, Omega-6 FA, total sugars, NMES intake and the consumption of alcohol, cooked vegetables, processed meat and sugary drinks. It is also associated with decrease in fruits and raw vegetables consumption.

Ethnicity (non-white) showed increased consumption of starch, omega-6 FA, cooked and raw vegetables and decreased consumption of SFA, total sugars and NMES, alcohol and processed meat.

Lipid lowering medication is related to increase in raw vegetables, oily fish and processed meat intake and decrease in sugary beverages.

Fruits/Vegetables shopping less than weekly is associated with increase in sugary beverages and decrease in raw vegetable intake.

Partner status Being divorced or widowed is associated with higher alcohol consumption, lower raw vegetables and fibre intake and somewhat lower intake of Omega-3 FA and Omega-6 FA; never married is associated with higher intake of TFA and raw vegetables.

Smoking Being a current smoker is related to increase in alcohol intake, processed meat and MUFA, and decrease in fibre, fruits, cooked and raw vegetables. Being a recent ex-smoker ($\leq$10 years) is still associated with higher alcohol and processed meat and lower fruit and fibre intake, but also with higher intake of raw vegetables. Ex-smoker (>10 years ago) showed only higher intake in extrinsic sugars and occasional smokers increase in

energy and MUFA compared to never-smokers.

MVPA increase is related to increased intake of all macronutrients except NMES, and increase in fruit, vegetable and oily fish intake and decrease in sugary beverages.

Alcohol habit Those who rarely drink alcohol consumed more energy, protein, MUFA, SFA, TFA, fibre, starch, total sugars, fruit and less oily fish than regular drinkers, but those who said they never drink alcohol did not statistically differ from regular drinkers in their food preferences except they consumed less processed meat and less oily fish.

Regular takeaway consumption is related to more energy, MUFA, SFA, TFA and alcohol and to less fibre, total sugars, NMES and oily fish, and fewer fruit and cooked vegetable intake.

Non-meat-eaters consumed less protein, fruits and raw vegetables but more fibre, starch, total sugars and NMES, and cooked vegetables and oily fish.

Socio-economic indicators

Tenure Renting is related to increase in SFA, Omega-6 FA and processed meat, and decrease in alcohol, fruit and cooked vegetable intake. Those renting from local authority indicated lower intake of sugary drinks.

Qualifications. Lower qualifications are associated with lower energy, protein, Omega-3 FA, TFA, SFA, fibre, fruit, raw vegetable and oily fish intake. Being a student is somewhat related to increased consumption of SFA, TFA and omega-3 FAI, and lower intake of protein and fibre, but no estimates (except fibre) reached statistical significance, possibly, due to very small numbers of participants in the students group.

Socio-economic status (occupation) Those involved in routine occupations consumed less raw vegetables and more processed meat.

Equalised income, when adjusted for the other food intake predictors, was related to lower alcohol intake (<£25,000) and decrease in TFA, SFA and MUFA intake (for those in <£15,000 and for those in $\geq$ £50,000 groups).

Table 3.7: Summary of the overall effects of personal, lifestyle and socio-economic predictors on the consumption of occasionally-consumed foods presented as p-values and effect directions in men.

Predictors	Alcohol Pr	Alcohol Portion	Fruits Pr	Fruits Portion	Veg cooked Pr	Veg cooked Portion	Veg raw Pr	Veg raw Portion	Process meat Pr	Process meat Portion	Oily fish Pr	Oily fish Portion	Sugary drinks Pr	Sugary drinks Portion
Weekend	**<0.001**	**<0.001**	0.020	<0.001		**0.060**	0.003	**0.067**	**0.064**	**0.010**				**0.022**
Survey Year 2010/2011														
Survey Year 2011/2012														
Age, years	**<0.001**	0.039	**<0.001**	**0.032**	**<0.001**		**0.013**	**0.069**	<0.001		**0.010**	**0.092**	<0.001	<0.001
BMI, kg/m^2			0.115						**0.032**			0.023		
Ethnicity, non-white	0.013					**0.003**			<0.001					
Health problems (self-reported)									**0.061**					
Lipid lowering drug									**0.009**		**0.024**		0.081	
Blood pressure lowering drug														
Fruits/veg buy < weekly							0.097				0.076	**0.039**	**0.121**	
Non-meat-eaters				0.023		**0.037**		0.089			**0.012**			
Partner, not married		**0.005**												
Smoking	**0.095**	**0.002**	<0.001	0.005	0.051		0.017	0.175		**0.101**				
Physical activity, increase min/day			**0.004**	**0.001**			**0.036**				**0.006**	**0.064**	0.096	0.002
Alcohol consumption, rare/never	<0.001	<0.001	0.069						0.092		0.006			
Take away shopping, regular		**0.033**		0.052							0.022			
Tenure, renting	0.073			0.098						**0.092**			0.122	0.151
Qualifications, lower levels[1]			<0.001	0.006	0.019						0.029			
SES, lower levels[1]							0.017	0.063		**0.348**				
Equalised income, increase £ 1000	**0.028**													

[1]For the individual effects of "Qualification" variable strata on alcohol and "SES" variable strata on raw vegetables please refer to Table 3.5 and Table Table A 3.2.3. P-values highlighted in bold indicate associated intake increase, underlined p-values indicate that the effects of the predictor depends on its category. Only significant effects are shown.

Personal, socio-economic and lifestyle characteristics in relation to food intake in women

Below is the summary of the most interesting trends in relationships between determinants and food intakes described by predictor. The full results of modelling intake of macronutrients and foods are presented in Appendix 3.A. The trends and overview of relationships between the considered predictors and considered occasionally- consumed food intakes, along with p-values, are summarised in Table3.8.

Weekend showed increase in intake of alcohol, cooked vegetables, processed meat, energy, protein, SFA, MUFA, TFA, Omega-3 and Omega-6 FA, total sugars and NMES intake and small decrease in fruit intake.

Ethnicity (non-white) is associated with increased intake of raw vegetables, oily fish, sugary beverages, MUFA, Omega-3 and Omega-6 FA intake increase, and decrease in processed meat, SFA, TFA and total sugar intake.

BMI is associated with increase in MUFA and sugary drinks.

Lipid lowering medication - increase in Omega-6 FA and decrease in alcohol, raw vegetables, fibre, sugary drinks consumption.

Blood pressure medication - decrease in fish and Omega-6 FA.

Fruits/Veg shopping less than weekly - decrease in cooked vegetables, oily fish, energy, protein, fibre and NMES.

Non-meat-eaters consumed more cooked and raw vegetables, more fibre, starch and sugary drinks; and less protein, SFA, MUFA and TFA.

Partner status Being divorced or widowed is associated with increase in alcohol, cooked vegetables and decrease in Omega-3 FA. Never married is associated with decreased intake of protein and starch, but increased intake of processed meat and sugary drinks.

Smoking Being a current smoker is associated with lower fruit, energy, protein, fibre, starch, total sugars, Omega-3 FA and Omega-6 FA intakes.

MVPA - increase in cooked and raw vegetables, oily fish, Omega-3 FA, and sugary drinks; and decrease in starch.

Alcohol habit Those who rarely drink alcohol consume less cooked and raw vegetables, but more total sugars.

Occasional and regular takeaway consumption is associated with more alcohol, processed meat and sugary drinks but less fruit, cooked and raw vegetables, less oily fish and fibre.

Socio-economic indicators

Tenure Renting is associated with less cooked and raw vegetables, sugary drinks and fibre.

Qualifications Lower qualifications level is associated with less alcohol, fruits (A levels and below), raw vegetables, oily fish, omega-3 FA and Omega-6 FA and slightly less TFA, and with more starch (A level and GCSE A-C grades) and slightly more sugary drinks, extrinsic sugars and processed meat. Being a student is associated with less alcohol, processed meat and fibre and fewer raw vegetables.

Socio-economic status Those involved in semi-routine and routine occupations consume more starch; those on lower managerial and professional positions – more alcohol and less raw vegetables; SES below small employers status indicate fewer raw vegetables consumption.

Equalised income, when adjusted for the other food intake predictors, is related to less fruit and fish intake (<£15,000), more cooked vegetables (£15,000 - £25,000) and more MUFA and less sugary drinks ($\geq$ £50,000).

Table 3.8: Summary of the overall effects of personal, lifestyle and socio-economic predictors on the consumption of occasionally-consumed foods presented as p-values and effect directions in women.

Predictors	Alcohol Pr	Alcohol Portion	Fruits Pr	Fruits Portion	Veg cooked Pr	Veg cooked Portion	Veg raw Pr	Veg raw Portion	Process meat Pr	Process meat Portion	Oily fish Pr	Oily fish Portion	Sugary drinks Pr	Sugary drinks Portion
Weekend	**<0.001**	**<0.001**		0.052	**0.005**		0.113		**<0.001**					
Age, years	**0.001**	<0.001	**<0.001**	**0.001**	**<0.001**		**0.008**	**0.015**	0.013		**<0.001**	**<0.001**	<0.001	<0.001
BMI, kg/m^2		**0.025**	0.002				**0.032**							
Ethnicity, non-white							**0.023**		<0.001		**0.024**		**0.088**	
Health problems (self-reported)			0.086				0.046	**0.091**		**0.144**				
Lipid lowering drug	0.017	0.062					0.114							
Blood pressure lowering drug												0.023		0.024
Fruits/veg buy < weekly			0.027	0.114	0.033	0.092			0.025	**0.099**	0.012			
Non-meat-eaters					**<0.001**	**<0.001**	**0.042**							**0.026**
Partner[1], not married		**0.034**				0.011							**0.085**	
Smoking			0.047							**0.012**			0.103	
Physical activity, increase min/day			**0.028**	**0.001**	**0.006**	**0.086**	**0.164**	**0.006**			**0.201**		**0.053**	**0.079**
Alcohol preference, rare/never	<0.001	<0.001	0.069		0.035			0.097						
Take away shopping, regular	**0.028**	**0.064**	0.008	<0.001	0.008		0.096		**0.106**		0.011			**0.002**
Tenure, renting			<0.001		0.001		0.023							0.032
Qualifications, lower levels[1]	0.027		0.096	0.003			0.014				0.055			**0.041**
SES, lower levels[1]	0.003						0.052							
Equalised income[1], decrease				0.088		0.007					0.005		**0.208**	**0.173**

[1]For the individual effects of "SES" strata on alcohol, "Partner" status and equalised income on cooked vegetables intakes please refer to Tables 3.6 and Table A 3.3.2. as the effects are non-linear and strata-specific. P-values highlighted in bold indicate associated intake increase, underlined p-values indicate that the effects of the predictor depends on its category. Only significant effects are shown.

The results of residual diagnostics are presented in Figures 3.5-3.10. The distribution of intake across the foods seems reasonably symmetric (Figures 3.5 and 3.8). There appears to be a few outlier observations which are prevalent in national survey data, otherwise the distribution of the residuals can be reasonably assumed to be normally distributed (Figures 3.6 and 3.9). It appears that further model development, which would take into account potential heterogeneity of measurement errors might be beneficial (Figures 3.7 and 3.10).

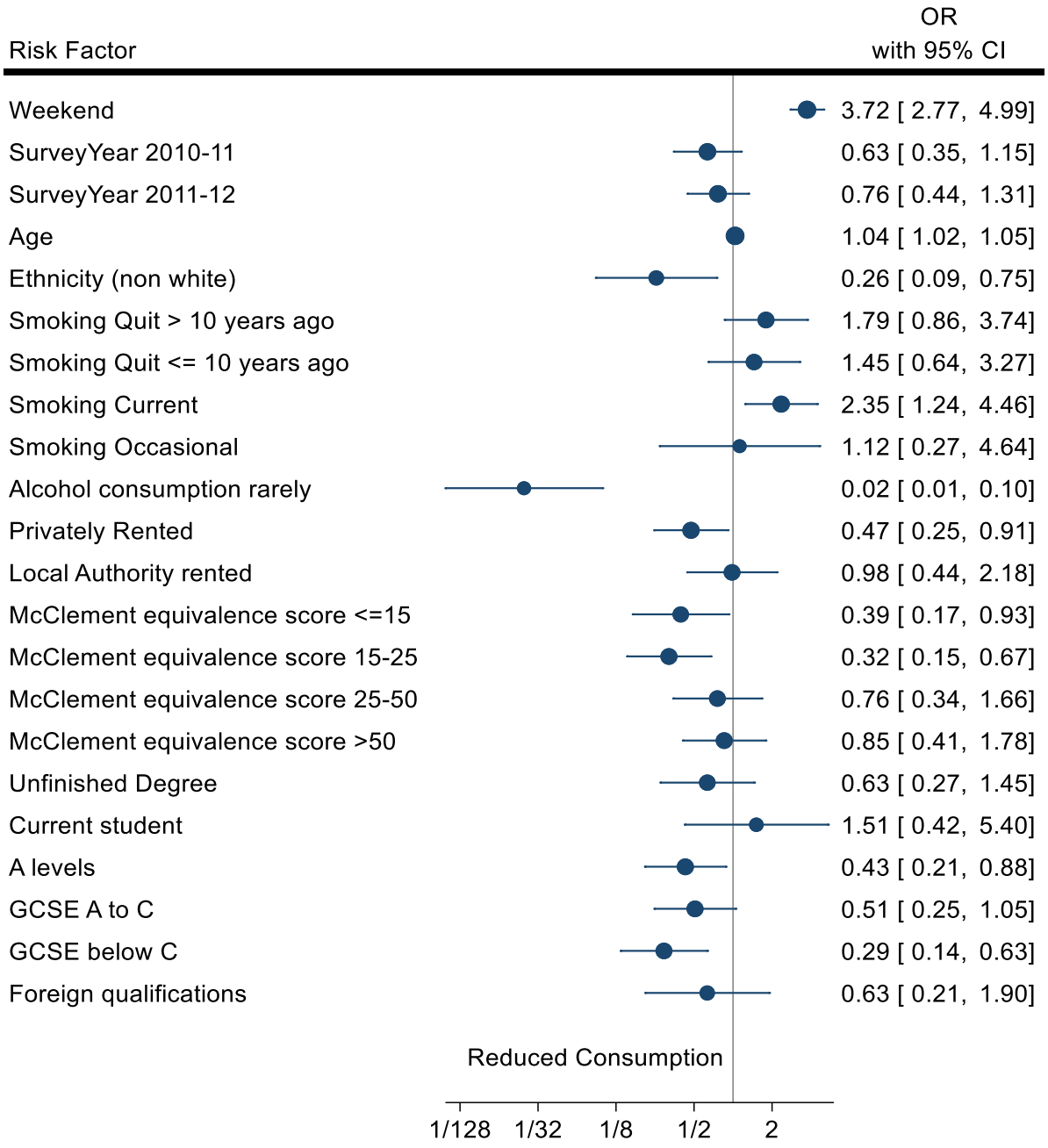


Figure 3.1: Forest plot of odds ratios for various risk factors of alcohol consumption in probability part of the two-part model in male sub-population of NDNS RP

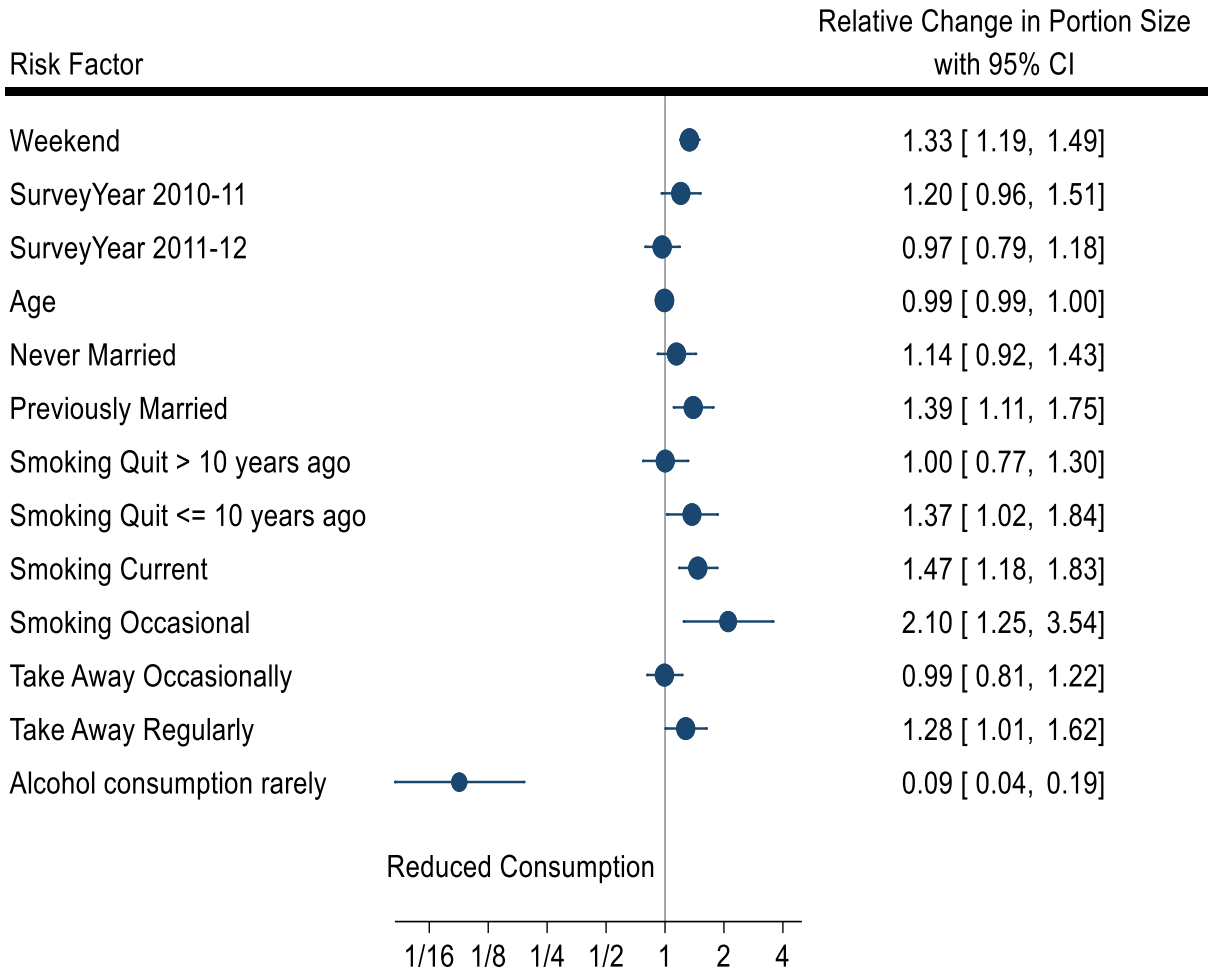


Figure 3.2: Forest plot of relative change in portion size for various risk factors of alcohol consumption in portion part of the two-part model in male sub-population of NDNS RP

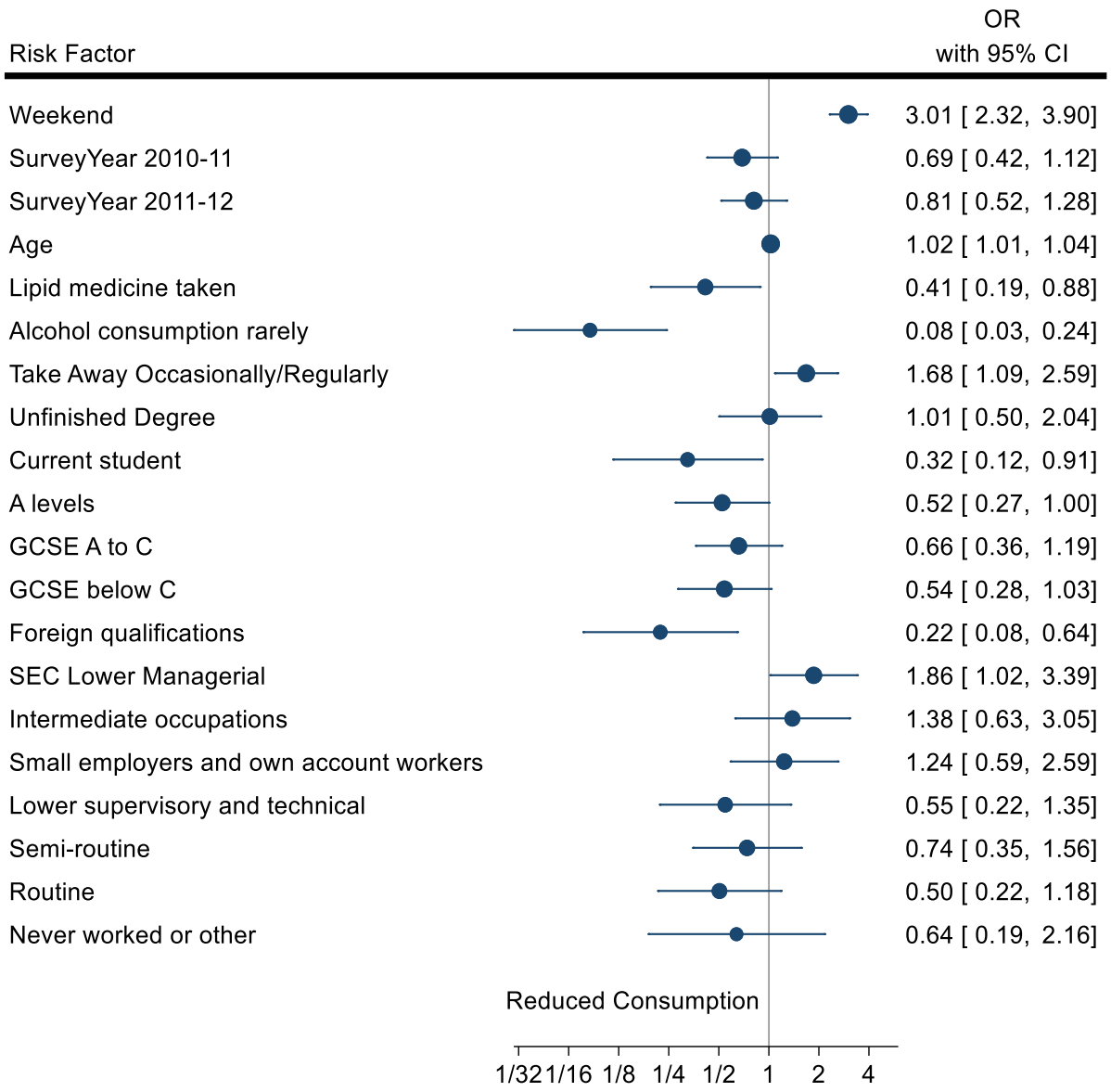


Figure 3.3: Forest plot of odds ratios for various risk factors of alcohol consumption in probability part of the two-part model in female sub-population of NDNS RP

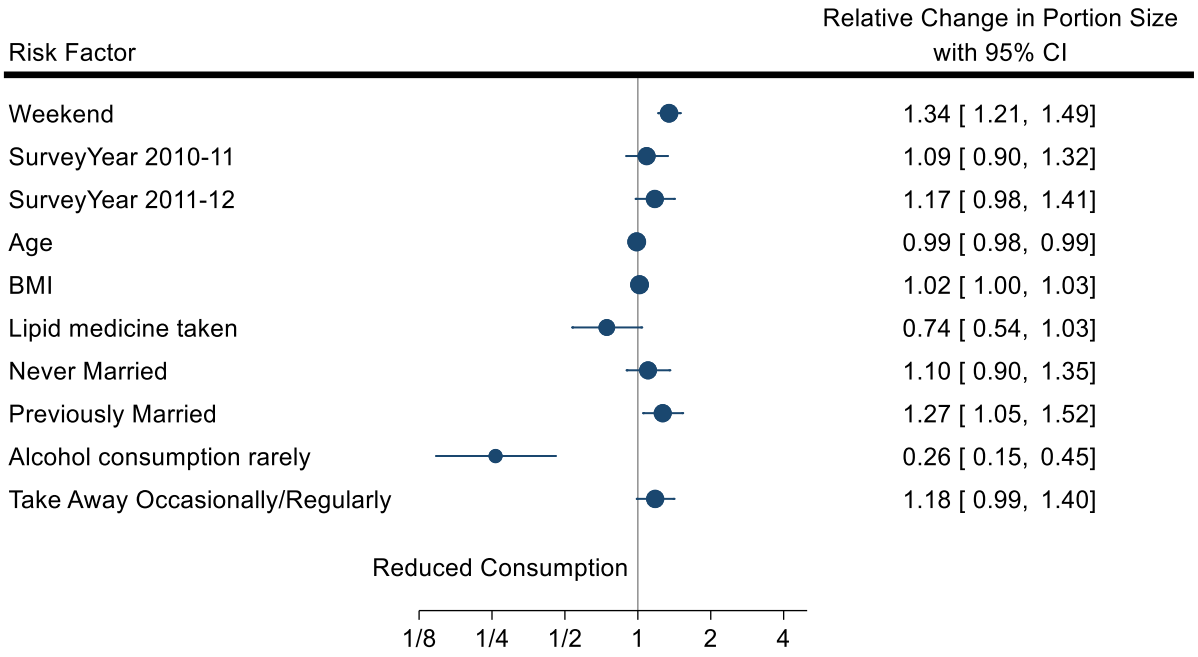


Figure 3.4: Forest plot of relative change in portion size for various risk factors of alcohol consumption in portion part of the two-part model in female sub-population of NDNS RP

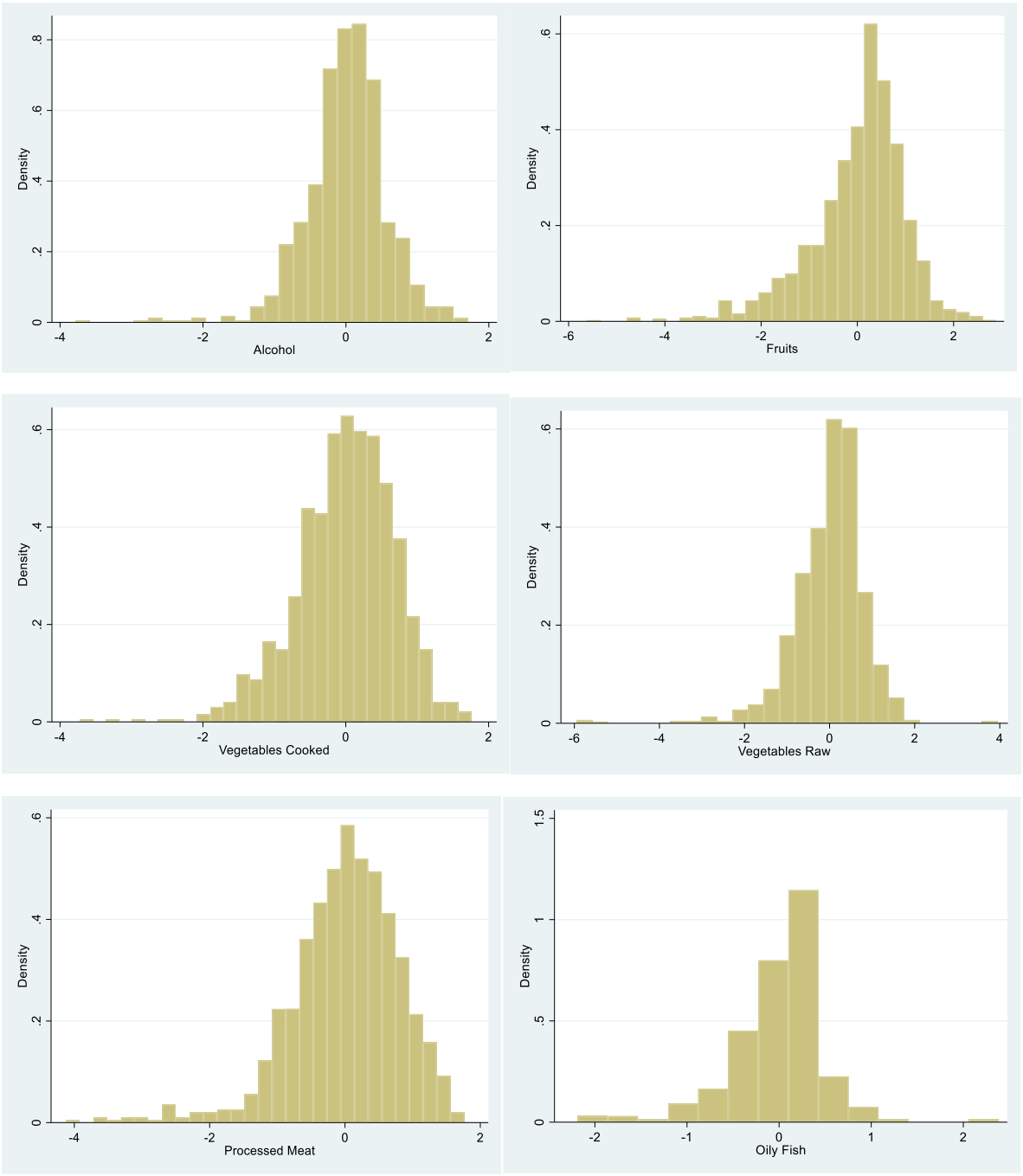


Figure 3.5: Histograms of residuals from portion size part of the two-part model in male sub-population

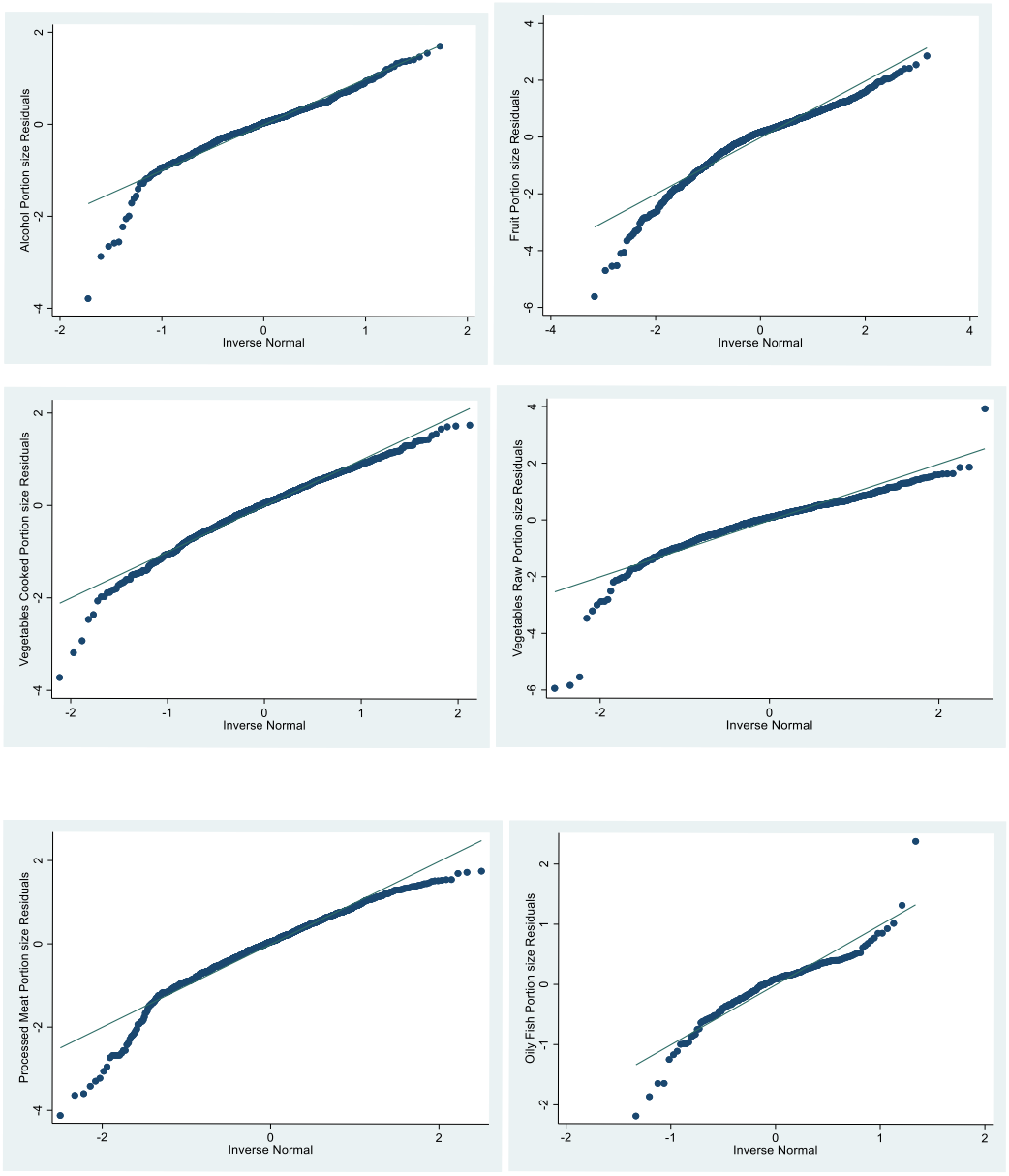


Figure 3.6: QQ plots of residuals from portion size part of the two-part model in male sub-population

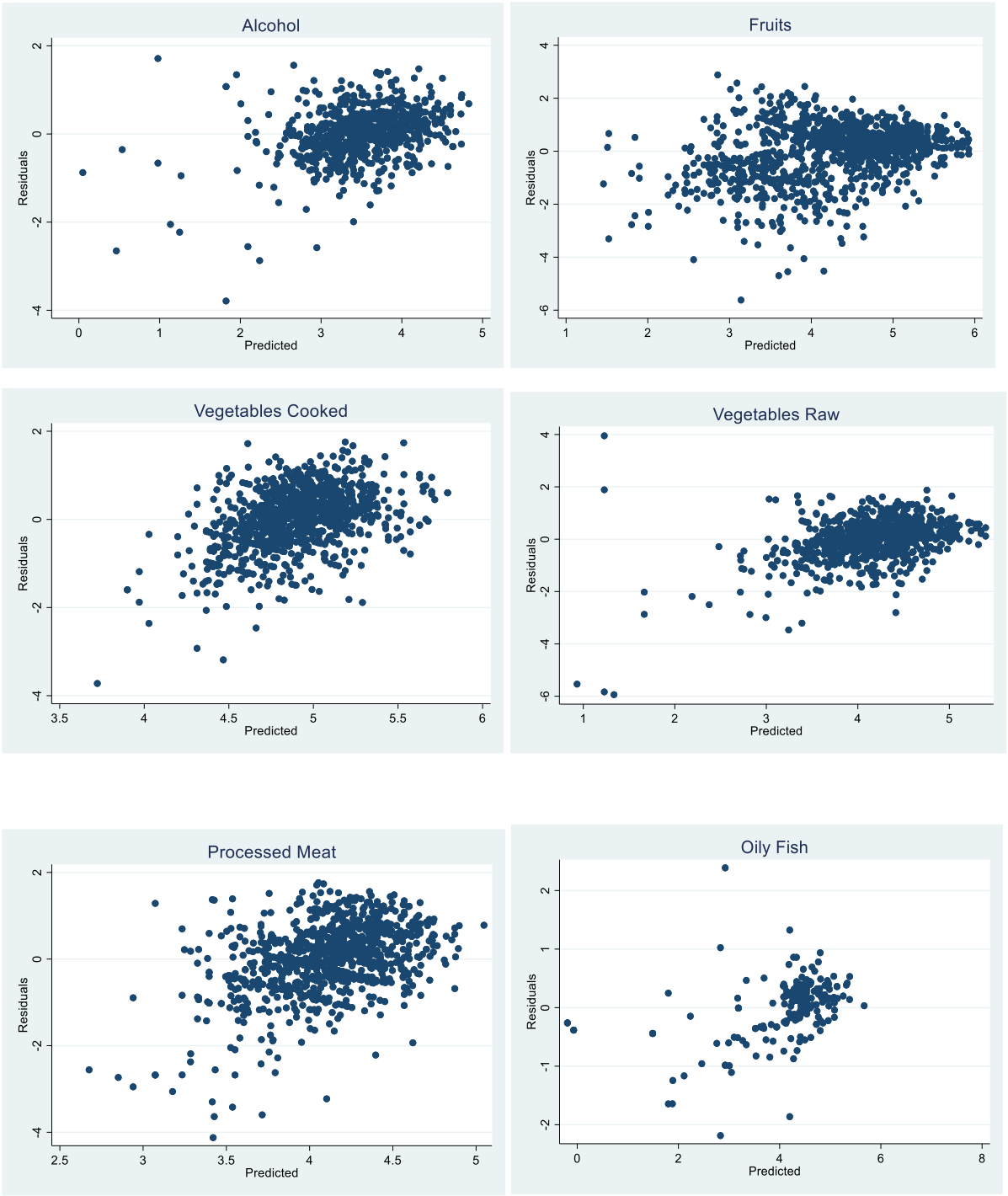


Figure 3.7: Residuals versus fitted values from portion size part of the two-part model in male sub-population

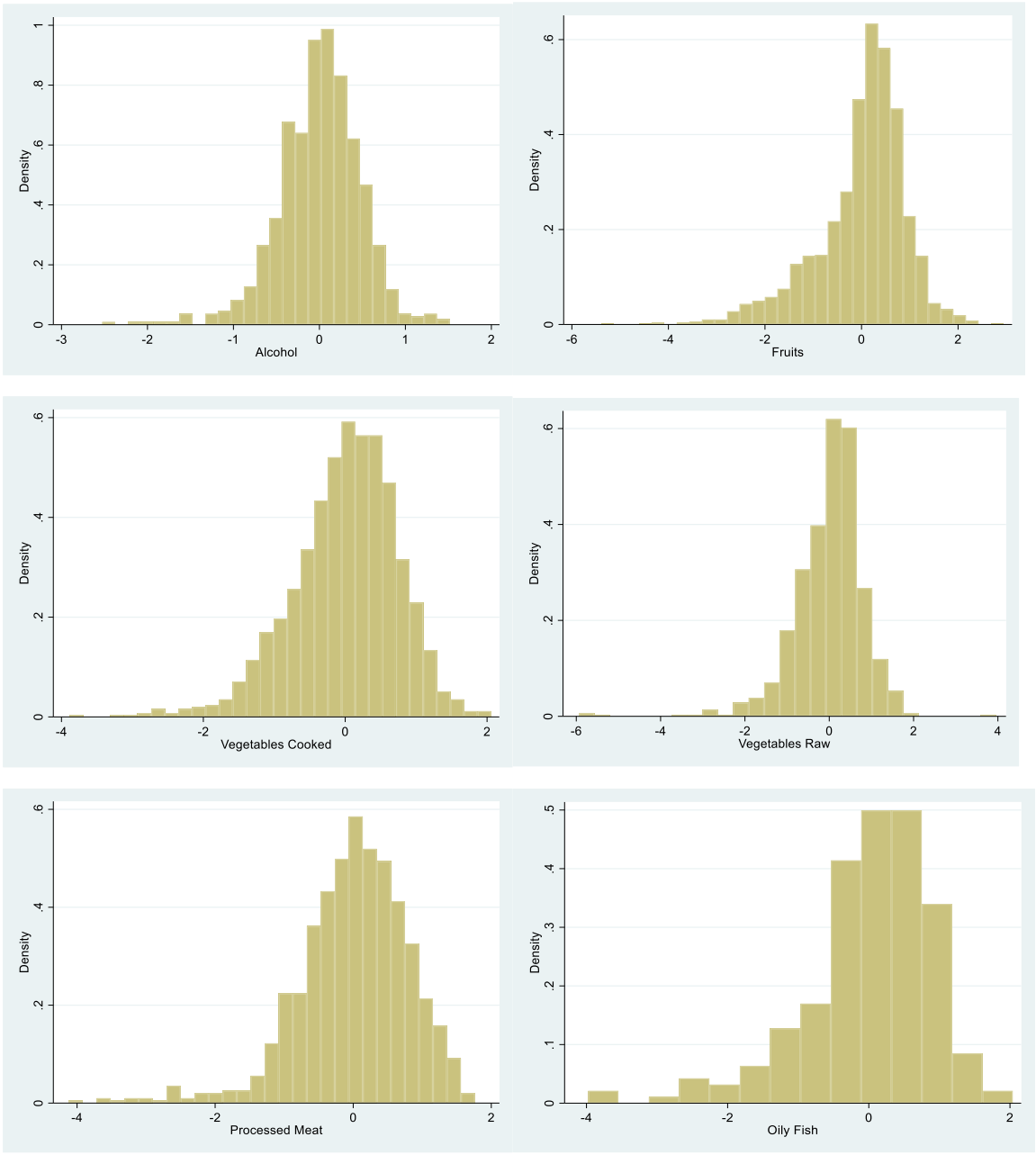


Figure 3.8: Histograms of residuals from portion size part of the two-part model in female sub-population

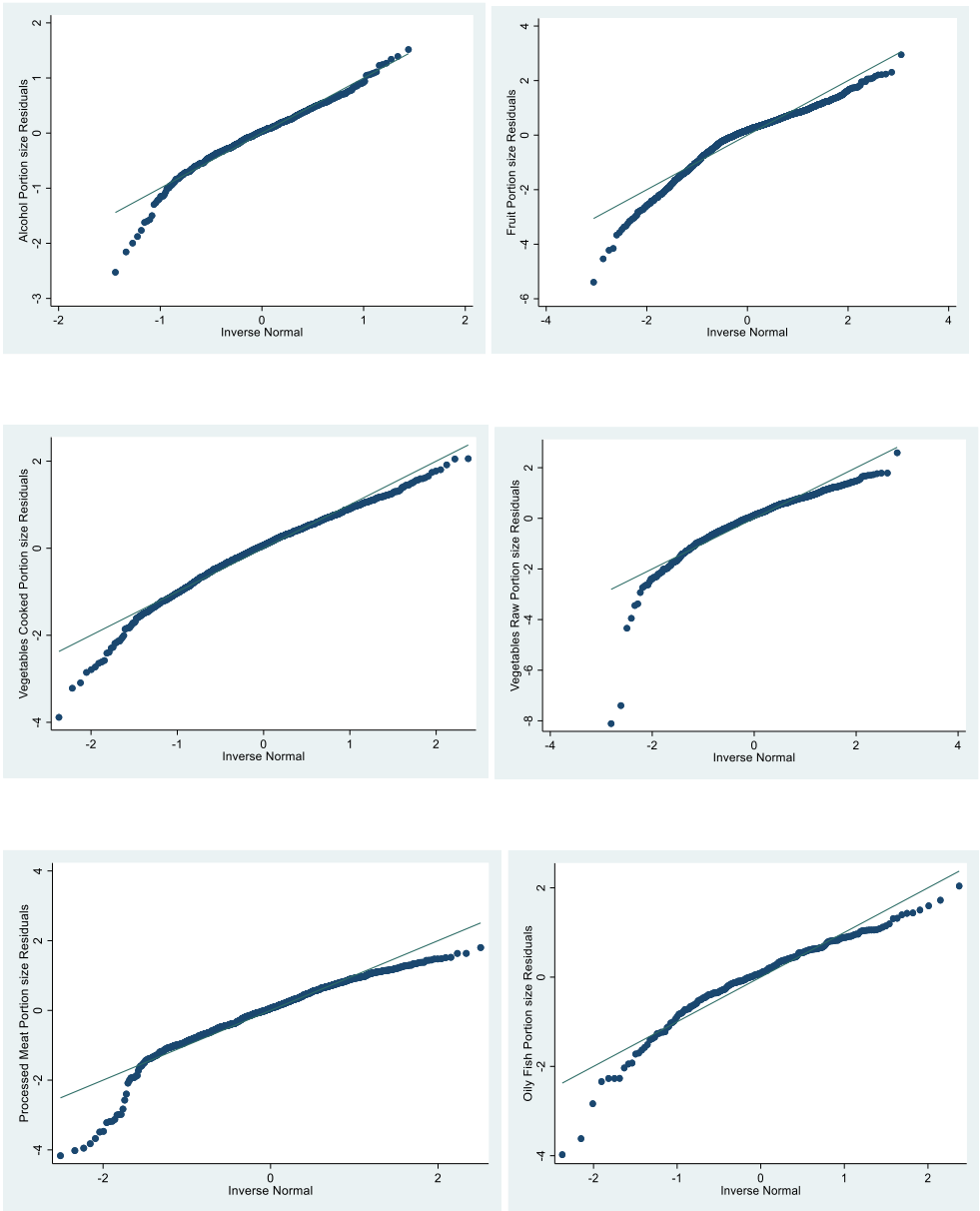


Figure 3.9: QQ plots of residuals from portion size part of the two-part model in female sub-population

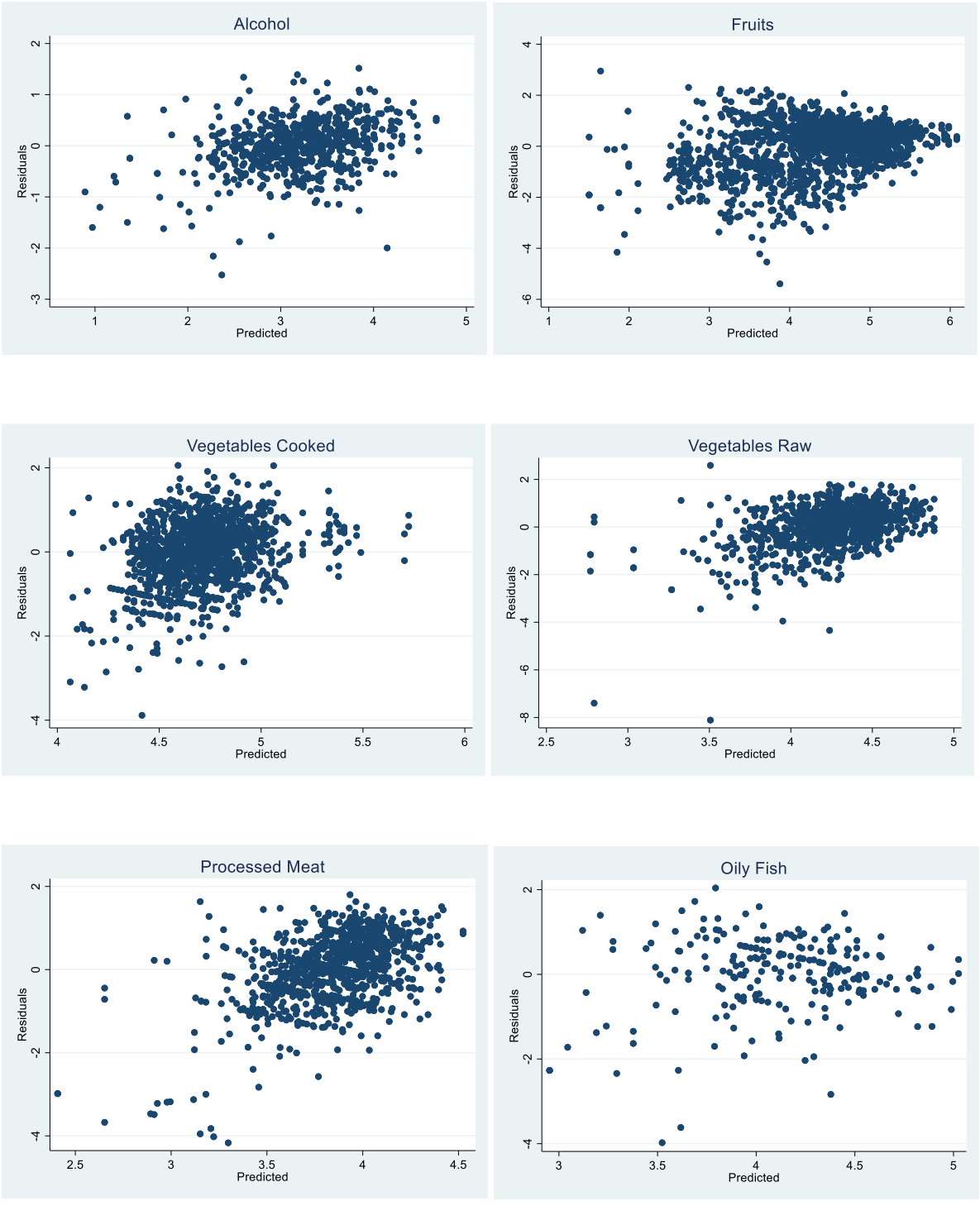


Figure 3.10: Residuals versus fitted values from portion size part of the two-part model in female sub-population

3.4 Discussion

We demonstrated how a modern statistical approach, based on the two-part mixed-effects model, can be applied to obtain valid inferences and provide in-depth and detailed description of food intake determinants when analysing inherently complex nutritional data from observational studies. In agreement with prior research, our results show that various socio-economic indicators, although related, contribute additional information to explain food preferences (Galobardes et al., 2001; Groth et al., 2001; Lallukka et al., 2006; Darmon and Drewnowski, 2008). In line with Pomerleau et al. (1997) and Groth et al. (2001) and, unlike Turrell et al. (2003), we show that the highest gained qualification level was the strongest consistent socio-economic determinant in this sample of the UK population. Socio-economic indicators are, necessarily, related to a myriad of various other factors. Qualifications are related to family background and peer influence. Occupation, in turn, is also related to family background and peer influence, but, additionally, it is related to the availability of workplace health promotion, working hours, job satisfaction, commuting time and food coping mechanism, which might be better indicators of food choices than occupation stratum per se. This suggests that more careful consideration should be paid to the choice of predictors, informed by potential mechanisms, during study design and data collection and, necessarily, adjustment for potential confounding should be carried out in data analysis when analysing food preferences in a socio-economic context. Interestingly, fats such as SFA, MUFA and TFA show the attributes of inferior goods (Mankiw and Taylor, 2007) in males (but not in females) and sugary drinks in the female population as they are consumed less with higher income. On the other hand, alcohol in men and fruit and oily fish in women show the attributes of normal consumer goods in the sub-population of below average earners as their consumption increases with income. The results also suggest that people with the same level of income can choose to distribute it differently according to their perceived priorities. For example, with the same income and qualifications level, men in rented accommodation consume less alcohol and fruit and more SFA and sugary beverages, whilst women who rent consume less vegeta-

bles. This indicates that promotion for healthier lifestyle should be developed more in line with people's background, expenses and time spending priorities. Overall, the data suggest that public health policy makers may want to shift their focus from income inequalities as a whole to education and environmental inequalities, especially, in the sub-populations with McClement equivalent score of above £15,000. This suggestion is in line with McKinnon et al. (2014) who showed that diet-disease association knowledge were the biggest factors attenuating the relationships between higher SES and better food choice. Ranjit et al. (2015) showed similar results in a population of children when SES became a non-significant predictor of a Healthy Eating Index score after adjustment for home food environment. A sub-population of those who rarely drink alcohol deserves some discussion as men in this category report increased consumption of energy, protein, MUFA, SFA, TFA, fibre, starch and fruit, and decreased intake of oily fish, and women report lower vegetable intake. This might partially explain the J-shaped relationships consistently observed between alcohol intake and cardiovascular disease (CVD) risk (Kloner and Rezkalla, 2007; Roerecke and Rehm, 2012) as the increase in energy and TFA intake (and/or decrease in oily fish intake) can be on the causal pathway between seldom alcohol consumption and increased CVD risk (Willett, 2012; Djoussé et al., 2012). We recommend that future analysis very clearly differentiate between never-drinkers, ex-drinkers for health reasons, and those who rarely drink alcohol out of personal preference and, importantly, energy, fats, fish intake and vegetable consumption should be accounted for when assessing the relationship between CVD risk and alcohol intake to disentangle this association. Interestingly, men, but not women, taking lipid powering medicine showed some healthy food choices like increase in raw vegetables, oily fish and processed meat intake and decrease in sugary beverages and extrinsic sugars consumption. First, this finding indicate potential reverse causality, and, second, potential gender differences. Longitudinal data are needed to address this question in detail.

Additionally, we showed that the combination of factors such as lower qualification, smoking, reduced MVPA, regular takeaways and being divorced or widowed is related to un-

healthy dietary patterns such as lower consumption of fruit, vegetables, oily fish and fibre and excess intake of alcohol intake in men. The results suggest that monitoring sub-populations with clustering of the mentioned lifestyle characteristics could be of public health significance in tackling diet-related health conditions. The analysis' strengths include the application of a model suitable for the analysis of occasionally-consumed foods taking into account excess-zeros and within-person variation including measurement error. Furthermore, the model accounts for the correlation between the probability of consumption and the portion size to minimise bias. The analysis also benefits from the availability of multiple interrelated SES predictors: qualifications, occupation, income and tenure, and lifestyle habits like smoking, physical activity and take away shopping habits. The application of the two-part model is not restricted to dietary intake but can be applied to any data that exhibit similar behaviour. For example, physical activity data, collected with actigraph devices or multiple-day diaries, have a similar complex structure and issues associated with an imperfect measurement process. Study limitations include the cross-sectional nature of the data making it impossible to examine the effect of change in predictors on food preferences. This, however, equally applies to many studies on determinants of health behaviour. Additionally, more detailed lifestyle and personal information would be desirable to help differentiate between the effects of correlated predictors. For instance, additional information on parental qualification could help to clarify the effect of a person's highest qualifications gained on food choice controlling for parental background. The model application also requires a relatively large sample size and is not particularly suitable for small datasets. However, more data will allow to capture more subtle preferences between food subgroups such as brassicas, salads and root vegetables. Additionally, fitting two-part model requires some statistical expertise to deal with potential convergence problems inherent to all maximum-likelihood based methods. Response rate of just above 50%, although typical of survey data, leaves a chance that the most disadvantaged strata of the population was not fully represented. If this is the case, we would expect that some associations found in the data between less healthy food pref-

erences and unfavourable circumstances would become even stronger. We encourage to conduct more research on the determinants of diet behaviour in these sub-populations.

3.5 Conclusion

The presented work shows that the benefits provided by this modern statistical approach available for the analysis of habitually- and occasionally-consumed foods and the uniqueness of the data (i.e. food intake records quality, data size and quite detailed information on food choice predictors) overweight the potential limitations and provide in-depth and detailed insight into the food intake habits of the UK adult population. We hope that this work not only adds to the current research on food intake determinants, which is a topic of great public health importance, but encourages nutritional scientists to use the most modern statistical tools in their research, preferably in collaboration with a statistician, to obtain robust inferences that can be reliably interpreted and used for public health policy making.

List of abbreviations

24HR - 24 hour food recall

BMI - Body mass index

DINO - Diet In Nutrients Out

FD - Food diary

FFQ - Food frequency questionnaire

ISU - Iowa State University

MC - Monte Carlo

MUFA - Monounsaturated fatty acids

MVPA - Moderate to vigorous physical activity

NCI - National Cancer Institute

NDNS - National Diet and Nutrition Survey

RCT - Randomised control trial

SES - Socio-economic status

SFA - Saturated fatty acids

TFA - Trans fatty acids

Appendix

3.A Estimated regression parameters of two-part models for food intake

3.A.1 Determinants of macronutrients intake

This appendix shows the tables that describe the relationships between predictors and macronutrients in men (Table 3.A.1) and women (Table 3.A.3) . Only significant associations are shown. SES, frequency of shopping for fruits and vegetables and being on lipid or blood pressure medication were not found to be significant predictors in men and health-related self report in women for any macronutrients so were omitted for clarity of presentation. Weekend includes Friday for Total sugars and NEMS. All macronutrients are presented in grams except energy, which is presented in Kcal.

Table 3.A.1: Macronutrient intake predictors in men in the sample selected for analysis from the NDNS RP Years 2-4

	Linear models Regression coefficients indicate absolute effects (95%CI)							*Log-linear* models Regression coefficients indicate relative effects (95%CI)			
Predictors	Energy	Protein	SFA	Fibre	Starch	MUFA	Total Sugars	TFA	NMES	Omega3 FA	Omega6 FA
Weekend[1]	149***	5.3***	2.8***	-0.2	-2.1	2.4***	4.8*	1.08*	1.15***	1.05	1.06*
	(88, 210)	(2.6, 8.0)	(1.4, 4.2)	(-0.7, 0.4)	(-7.2, 2.9)	(1.1, 3.7)	(0.5, 9.1)	(1.00, 1.16)	(1.07, 1.24)	(0.98, 1.11)	(1.01, 1.12)
Survey Year											
(2009/2010)											
2010/2011	-23	-1.0	-2.0†	0.1	2.8	0.04	-1.2	0.73***	1.00	0.94	0.94
	(-134, 88)	(-6.0, 4.0)	(-4.3, 0.4)	(-1.0, 1.1)	(-6.7, 12.3)	(-2.1, 2.2)	(-11.3, 8.9)	(0.65, 0.82)	(0.86, 1.17)	(0.85, 1.03)	(0.86, 1.02)
2011/2012	15	-2.2	-1.6	0.6	11.1*	-0.5	5.7	0.62***	1.04	0.98	0.96
	(-88, 117)	(-6.8, 2.5)	(-3.8, 0.6)	(-0.3, 1.6)	(2.2, 20.0)	(-2.5, 1.5)	(-4.1, 15.4)	(0.56, 0.69)	(0.90, 1.20)	(0.89, 1.07)	(0.88, 1.04)
Age, years	-3*	-0.1	0.05	0.01	-0.6***	-0.09**	-0.4**	1.004*	0.991***	0.999	0.997**
	(-7, -0.1)	(-0.2, 0.03)	(-0.02, 0.12)	(-0.02, 0.05)	(-0.8, -0.4)	(-0.15, -0.03)	(-0.7, -0.1)	(1.000, 1.008)	(0.987, 0.995)	(0.997, 1.002)	(0.994, 0.999)
BMI, kg/m^2		0.4†					-1.1*		0.98**		
		(-0.04, 0.9)					(-2.1, -0.01)		(0.95, 0.99)		
Ethnicity, non-white			-4.3*		18.6*		-18.5*		0.74*		1.15*
			(-8.1, -0.6)		(3.9, 33.3)		(-33.6, -3.4)		(0.59, 0.94)		(1.00, 1.31)
Health (self-reported)											
(base: no problems)											
health problems, no		-6.0*									
mobility restrictions		(-11.6, -0.5)									
health problems,		-0.7									
mobility restrictions		(-6.6, 5.2)									
Non-meat eaters	127	-6.2†		2.1*	11.3		13.1†		1.14†		
	(-37, 291)	(-13.6, 1.25)		(0.6, 3.7)	(-2.8, 25.4)		(-2.6, 28.8)		(0.99, 1.32)		

Table 3.A.1 Macronutrient intake predictors in men in the sample selected for analysis from the NDNS RP Years 2-4 (Cont.)

	Linear models Regression coefficients indicate absolute effects (95%CI)							*Log-linear* models Regression coefficients indicate relative effects (95%CI)			
Predictors	Energy	Protein	SFA	Fibre	Starch	MUFA	Total sugars	TFA	NMES	Omega3 FA	Omega6 FA
Partner (base: married/partner)											
never married				0.6				1.15*			
				(-0.5, 1.7)				(1.02, 1.29)			
previously married				-1.1†				1.05			
				(-2.3, 0.04)				(0.92, 1.19)			
Smoking (base: never)											
ex, quit >10 year ago	40			-0.2		0.6					
	(-99, 179)			(-1.5, 1.1)		(-2.1, 3.3)					
ex, quit <=10 years	108			-1.3†		0.1					
ago	(-47, 264)			(-2.8, 0.1)		(-2.9, 3.1)					
current	87			-2.0***		2.1†					
	(-27, 201)			(-3.1, -0.9)		(-0.1, 4.2)					
occasional	237†			-0.2		4.5†					
	(-30, 504)			(-2.7, 2.3)		(-0.6, 9.7)					
MVPA, min/day, (base: 0)											
0-10	264†	16.2*	7.6*	2.8*	18.5	3.7		1.43*		1.30*	1.32*
	(-28, 557)	(2.8, 29.5)	(1.3, 13.8)	(0.1, 5.6)	(-6.7, 43.8)	(-2.0, 9.4)		(1.05, 1.94)		(1.01, 1.67)	(1.05, 1.66)
10-20	332*	19.5**	8.1*	3.7*	22.4	5.4†		1.51*		1.31†	1.40**
	(13, 651)	(4.9, 34.1)	(1.2, 15.0)	(0.7, 6.6)	(-5.2, 49.9)	(-0.8, 11.6)		(1.08, 2.12)		(1.00, 1.73)	(1.09, 1.80)
20-40	390**	20.0**	9.7**	4.5**	25.0†	5.7*		1.50*		1.45**	1.46**
	(97, 684)	(6.5, 33.5)	(3.4, 16.0)	(1.8, 7.3)	(-0.4, 50.5)	(0.03, 11.5)		(1.10, 2.04)		(1.12, 1.86)	(1.16, 1.84)
40-60	409**	22.9**	9.3**	4.0**	22.8†	7.3*		1.43*		1.57**	1.52***
	(113, 705)	(9.4, 36.3)	(2.9, 15.7)	(1.3, 6.8)	(-2.7, 48.4)	(1.6, 13.1)		(1.04, 1.95)		(1.22, 2.02)	(1.21, 1.92)
>60	473**	24.9***	10.9**	4.9***	31.3*	7.1*		1.59**		1.40**	1.50***
	(198, 748)	(12.2, 37.5)	(5.0, 16.9)	(2.4, 7.5)	(7.5, 55.2)	(1.7, 12.5)		(1.19, 2.12)		(1.11, 1.78)	(1.21,1.86)

Table 3.A.1 Macronutrient intake predictors in men in the sample selected for analysis from the NDNS RP Years 2-4 (Cont.)

	Linear models							*Log-linear* models			
	Regression coefficients indicate absolute effects (95%CI)							Regression coefficients indicate relative effects (95%CI)			
Predictors	Energy	Protein	SFA	Fibre	Starch	MUFA	Total sugars	TFA	NMES	Omega3 FA	Omega6 FA
Alcohol habit consumption											
rarely	221*	12.0**	7.9***	2.3*	22.9**	5.6**	22.7†	1.29*			
	(31, 411)	(3.3, 20.6)	(3.8, 11.9)	(0.5, 4.1)	(6.5, 39.4)	(1.9, 9.2)	(-0.8, 46.1)	(1.06, 1.57)			
never	-93	-3.1	1.8	-0.2	1.7	-0.8	9.0	0.93			
	(-260, 74)	(-10.7, 4.5)	(-1.9, 5.4)	(-1.7, 1.4)	(-13.1, 16.6)	(-4.1, 2.4)	(-12.1, 30.0)	(0.78, 1.11)			
Take away shopping											
occasionally	-25		-0.4	-1.1*		-0.9	-10.7*	1.01	0.87†		
	(-134, 83)		(-2.7, 1.9)	(-2.1, -0.1)		(-2.9, 1.2)	(-20.9, -0.5)	(0.90, 1.13)	(0.75, 1.02)		
regularly	153*		2.8*	-0.9		2.8*	-0.5	1.13†	1.08		
	(29, 278)		(0.2, 5.5)	(-2.1, 0.2)		(0.4, 5.2)	(-12.3, 11.3)	(1.00, 1.29)	(0.91, 1.29)		
Tenure (base: owned/mortgaged)											
privately rented			3.3*								1.09†
			(0.8, 5.9)								(0.99, 1.20)
LA rented			2.3								0.95
			(-0.8, 5.3)								(0.86, 1.06)
Qualifications											
unfinished degree	5	3.0	-1.4	-1.4†			12.5†	0.91			
	(-157, 167)	(-4.3, 10.3)	(-4.9, 2.1)	(-2.9, 0.1)			(-1.8, 26.8)	(0.77, 1.08)			
students	89	-3.7	3.2	-2.8*			18.9	1.16			
	(-155, 332)	(-14.7, 7.3)	(-2.1, 8.6)	(-5.1, -0.5)			(-9.3, 47.1)	(0.89, 1.51)			

Table 3.A.1 Macronutrient intake predictors in men in the sample selected for analysis from the NDNS RP Years 2-4 (Cont.)

A levels	-118	-8.5**	-3.6*	-2.9***			-7.2	0.85*			
	(-254, 18)	(-14.7, -2.4)	(-6.5, -0.7)	(-4.2, -1.6)			(-19.2, 4.8)	(0.74, 0.98)			
GCSE A_C	-166	-6.9*	-4.3**	-2.4***			-4.3	0.87†			
	(-303, -28)	(-13.0, -0.7)	(-7.3, -1.3)	(-3.7, -1.1)			(-15.8, 7.2)	(0.75, 1.00)			
GCSE below C and no qualifications	-137	-9.7**	-2.4	-3.1***			-0.1	0.85*			
	(-273, -0.2)	(-15.8, -3.7)	(-5.4, 0.7)	(-4.4, -1.8)			(-13.7, 13.5)	(0.73, 0.98)			
foreign qualifications	-115	-3.8	-0.9	-2.8**			-3.4	0.96			
	(-326, 96)	(-13.3, 5.6)	(-5.4, 3.7)	(-4.8, -0.8)			(-19.9, 13.0)	(0.77, 1.20)			
	Linear models							*Log-linear* models			
	Regression coefficients indicate absolute effects (95%CI)							Regression coefficients indicate relative effects (95%CI)			
Predictors	Energy	Protein	SFA	Fibre	Starch	MUFA	Total sugars	TFA	NMES	Omega3 FA	Omega6 FA
McClement equivalence score, £ 1000											
<=15			-1.9			-2.5		0.85†			
			(-5.3, 1.5)			(-5.5, 0.5)		(0.72, 1.00)			
15-25			-0.13			-1.0		0.94			
			(-3.1, 2.8)			(-3.7, 1.7)		(0.81, 1.09)			
(base) 25-35								1			
35-50			0.6			1.7		0.95			
			(-2.6, 3.8)			(-1.2, 4.6)		(0.81, 1.11)			
>50			-3.4*			-2.0		0.83*			
			(-6.5, -0.3)			(-5.0, 1.0)		(0.72, 0.97)			
Variability	Standard deviation (SD) on absolute scale							Standard deviation (SD) on Log scale			
Between-person SD	396	18.3	8.0	3.7	35.4	7.2	40.3	0.38	0.61	0.31	0.29
Within-person SD	567	25.2	13.1	5.1	47.1	12.1	42.4	0.68	0.70	0.60	0.50
Within-person correlation	0.33	0.34	0.27	0.35	0.36	0.26	0.47	0.24	0.43	0.21	0.26

Statistical significance: † < 0.10; * < 0.05; ** < 0.01; *** < 0.001

Table 3.A.2: P-values of macronutrient intake predictors in men in the sample selected for analysis from the NDNS RP Years 2-4

Predictors	Energy	Protein	SFA	Fibre	Starch	MUFA	Total sugars	TFA	NMES	Omega3 FA	Omega6 FA
Weekend	<0.001	<0.001	<0.001	0.537	0.412	<0.001	0.025	0.042	<0.001	0.160	0.031
Survey Year 2010/2011	0.684	0.690	0.100	0.896	0.567	0.969	0.823	<0.001	0.973	0.197	0.145
Survey Year 2011/2012	0.776	0.363	0.144	0.190	0.014	0.637	0.242	<0.001	0.612	0.617	0.292
Age, years	0.044	0.137	0.160	0.437	<0.001	0.002	0.003	0.034	<0.001	0.531	0.001
BMI, kg/m^2		0.075					0.026		0.003		
Ethnicity, non-white			0.022		0.013		0.025		0.012		0.043
Health (self-reported)		0.097									
Non-meat eaters	0.130	0.103		0.006	0.117		0.092		0.066		
Partner				0.062				0.066			
Smoking	0.238			0.001		0.203					
Physical activity, min/day	0.002	0.001	0.005	<0.001	0.051	0.031		0.040		0.009	0.004
Alcohol habit consumption	0.035	0.016	<0.001	0.037	0.024	0.010	0.028	0.028			
Take away shopping	0.009		0.300	0.085		0.005	0.059	0.115	0.025		
Tenure			0.027								0.099
Qualifications	0.122	0.005	0.025	<0.001			0.141	0.063			
Equalised income, £ 1000			0.012			0.052		0.150			

Table 3.A.3: Macronutrient intake predictors in women in the sample selected for analysis from the NDNS RP Years 2-4

	Linear models Regression coefficients indicate absolute effects (95%CI)							*Log-linear* models Regression coefficients indicate relative effects (95%CI)			
Predictors	Energy	Protein	SFA	Fibre	Starch	MUFA	Total sugars	TFA	NMES	Omega3 FA	Omega6 FA
Weekend[1]	140***	4.6***	2.3***	-0.2	1.1	2.5***	2.3	1.19***	1.13***	1.10**	1.10***
	(97, 182)	(2.5, 6.6)	(1.3, 3.4)	(-0.7, 0.2)	(-2.5, 4.8)	(1.5, 3.4)	(-0.9, 5.4)	(1.11, 1.28)	(1.05, 1.20)	(1.04, 1.17)	(1.04, 1.16)
Survey Year (2009/2010)											
2010/2011	-34	0.7	-0.8	0.5	3.0	0.2	-7.7*	0.72***	0.84*	0.92†	0.95
	(-107, 38)	(-2.4, 3.8)	(-2.6, 1.0)	(-0.3, 1.3)	(-2.7, 8.7)	(-1.2, 1.7)	(-14.5, -0.9)	(0.65, 0.80)	(0.72, 0.97)	(0.85, 1.00)	(0.88, 1.03)
2011/2012	-9	1.3	-0.4	0.4	0.8	0.4	-4.0	0.68***	0.85*	0.91*	0.97
	(-76, 58)	(-1.6, 4.3)	(-2.0, 1.3)	(-0.4, 1.1)	(-4.5, 6.1)	(-1.0, 1.8)	(-10.5, 2.4)	(0.62, 0.76)	(0.75, 0.99)	(0.84, 0.99)	(0.91, 1.04)
Age, years	-3**	-0.01	0.00	0.03*	-0.56***	-0.06**	-0.1	1.003*	0.998	1.003*	0.999
	(-4, -1)	(-0.10, 0.08)	(-0.04, 0.04)	(0.01, 0.06)	(-0.73, -0.38)	(-0.09, -0.02)	(-0.2, 0.1)	(1.001, 1.006)	(0.995, 1.001)	(1.001, 1.005)	(0.997, 1.002)
BMI, kg/m^2							-0.5†				
							(-1.0, 0.0)				
Ethnicity (non-white)			-2.8*			2.1†	-14.4**	0.86†		1.13†	1.19**
			(-5.5, -0.1)			(-0.1, 4.3)	(-24.1, -4.7)	(0.73, 1.02)		(0.99, 1.28)	(1.05, 1.33)
Lipid lowering drug				-1.5*							1.10
				(-2.6, -0.3)							(0.97, 1.24)
Blood pressure lowering drug											0.91†
											(0.81, 1.01)
Fruits/veg buy < weekly	-106†	-7.4**		-1.1†			-6.4*				
	(-217, 5)	(-12.3, -2.5)		(-2.3, 0.1)			(-11.8, -0.9)				
Non-meat eaters		-9.2***	-1.7†	1.1**	3.9	-1.5†		0.86**		0.93	
		(-12.4, -5.9)	(-3.6, 0.2)	(0.3, 1.9)	(-2.0, 9.8)	(-3.0, 0.02)		(0.77, 0.96)		(0.85, 1.02)	
Partner (base: married)											
never married		-3.8*			-10.6**					1.03	
		(-7.1, -0.4)			(-17.1, -4.2)					(0.93, 1.14)	

Table 3.A.3 Macronutrient intake predictors in women in the sample selected for analysis from the NDNS RP Years 2-4 (Cont.)

previously married		-1.5			-4.5					0.91†	
		(-4.6, 1.6)			(-10.2, 1.2)					(0.82, 1.01)	
	Linear models							*Log-linear* models			
	Regression coefficients indicate absolute effects (95%CI)							Regression coefficients indicate relative effects (95%CI)			
Predictors	Energy	Protein	SFA	Fibre	Starch	MUFA	Total sugars	TFA	NMES	Omega3 FA	Omega6 FA
Smoking											
(base: never smoked)											
ex, quit >10 year ago	-34	-0.02		-0.1	-4.4		-1.8			0.98	0.99
	(-99, 179)	(-4.03, 3.99)		(-1.1, 0.9)	(-11.6, 2.8)		(-10.9, 7.2)			(0.88, 1.09)	(0.90, 1.09)
ex, quit <=10 years ago	12	3.2		0.2	-1.3		-13.6**			1.09	1.07
	(-47, 264)	(-1.0, 7.3)		(-0.8, 1.2)	(-8.9, 6.2)		(-21.5, -5.7)			(0.98, 1.22)	(0.97, 1.18)
current	-124**	-6.4***		-2.5***	-19.6***		-7.8†			0.91†	0.92†
	(-27, 201)	(-9.8, -3.1)		(-3.4, -1.6)	(-25.9, -13.2)		(-15.8, 0.2)			(0.83, 1.00)	(0.85, 1.00)
occasional	155	4.3		2.3†	-4.6		4.9			1.07	1.19
	(-30, 504)	(-5.9, 14.6)		(-0.2, 4.9)	(-23.3, 14.0)		(-15.4, 25.3)			(0.81, 1.41)	(0.93, 1.53)
MVPA, min/day, (base: 0)											
0-10					-8.4					1.08	
					(-23.4, 6.5)					(0.87, 1.35)	
10-20					-7.6					1.21	
					(-23.3, 8.0)					(0.96, 1.53)	
20-40					-10.8					1.17	
					(-26.0, 4.5)					(0.93, 1.47)	
40-60					-14.5†					1.10	
					(-30.2, 1.2)					(0.87, 1.39)	
>60					-14.8†					1.20	
					(-29.6, 0.1)					(0.97, 1.50)	

Table 3.A.3 Macronutrient intake predictors in women in the sample selected for analysis from the NDNS RP Years 2-4 (Cont.)

Alcohol habit consumption (base: regular)											
rarely							12.2†				
							(-1.2, 25.6)				
never							7.0				
							(-4.6, 18.5)				
	Linear models							*Log-linear* models			
	Regression coefficients indicate absolute effects (95%CI)							Regression coefficients indicate relative effects (95%CI)			
Predictors	Energy	Protein	SFA	Fibre	Starch	MUFA	Total sugars	TFA	NMES	Omega3 FA	Omega6 FA
Take away shopping											
occasionally				-0.9*							
				(-1.6, -0.1)							
regularly				-1.2**							
				(-2.1, -0.4)							
Tenure (base: owned/mortgaged)											
privately rented				-0.7							
				(-1.6, 0.3)							
LA rented				-1.0*							
				(-2.0, -0.1)							
Qualifications											
unfinished degree				-0.2	-5.8			0.82*		0.86*	0.89*
				(-1.4, 0.9)	(-14.6, 2.9)			(0.69, 0.96)		(0.76, 0.97)	(0.80, 1.00)
students				-1.4†	-6.7			1.02		1.02	0.96
				(-3.0, 0.1)	(-18.6, 5.3)			(0.83, 1.27)		(0.86, 1.20)	(0.83, 1.12)
A levels				-0.1	6.6†			1.00		0.96	0.98
				(-1.1, 0.9)	(-1.1, 14.2)			(0.87, 1.15)		(0.86, 1.07)	(0.89, 1.08)
GCSE A - C				-0.4	5.9			0.89†		0.93	0.92†
				(-1.4, 0.6)	(-1.5, 13.3)			(0.78, 1.02)		(0.84, 1.04)	(0.83, 1.01)

Table 3.A.3 Macronutrient intake predictors in women in the sample selected for analysis from the NDNS RP Years 2-4 (Cont.)

GCSE below C and no qualifications				-1.7** (-2.7, -0.7)	1.1 (-6.6, 8.8)			0.89† (0.78, 1.02)		0.87* (0.78, 0.97)	0.86** (0.78, 0.95)
foreign qualifications				-1.4† (-3.1, 0.2)	-9.4 (-21.3, 2.5)			0.78* (0.62, 0.97)		0.85† (0.71, 1.01)	0.83* (0.71, 0.97)
	Linear models Regression coefficients indicate absolute effects (95%CI)							*Log-linear* models Regression coefficients indicate relative effects (95%CI)			
Predictors	Energy	Protein	SFA	Fibre	Starch	MUFA	Total sugars	TFA	NMES	Omega3 FA	Omega6 FA
Socio-economic status											
(base) higher managerial											
lower managerial					-0.7						
and professional occupation					(-8.1, 6.7)						
Intermediate occupations					-0.5						
					(-10.0, 9.1)						
Small employers					2.7						
and own account workers					(-6.3, 11.7)						
Lower supervisory					0.6						
and technical occupation					(-9.8, 11.0)						
Semi-routine occupations					11.3*						
					(2.5, 20.2)						
Routine occupations					11.4*						
					(1.2, 21.5)						
Never worked and Other					14.8*						
					(1.4, 28.2)						

Table 3.A.4: P-values of macronutrient intake predictors in women in the sample selected for analysis from the NDNS RP Years 2-4

Predictors	Energy	Protein	SFA	Fibre	Starch	MUFA	Total sugars	TFA	NMES	Omega3 FA	Omega6 FA
Weekend	<0.001	<0.001	<0.001	0.274	0.540	<0.001	0.138	<0.001	<0.001	0.002	<0.001
Survey Year 2010/2011	0.348	0.663	0.381	0.199	0.306	0.746	0.031	<0.001	0.016	0.052	0.190
Survey Year 2011/2012	0.796	0.372	0.676	0.336	0.767	0.542	0.228	<0.001	0.030	0.24	0.451
Age, years	0.001	0.799	0.977	0.010	<0.001	0.001	0.431	0.019	0.139	0.014	0.644
BMI, kg/m^2						0.106	0.057				
Ethnicity, non-white			0.044			0.060	0.010	0.077		0.071	0.005
Health (self-reported)											
Lipid lowering drug				0.010							0.126
Blood pressure lowering drug											0.081
Fruits/veg buy rarely	0.062	0.003		0.044			0.024				
Non-meat eaters		<0.001	0.076	0.005	0.194	0.053		0.009		0.114	
Partner		0.083			0.004						
Smoking	0.009	<0.001		<0.001	<0.001		0.028			0.082	0.074
Physical activity, min/day		0.001			0.117					0.129	
Alcohol habit consumption							0.061				
Take away shopping				0.011							
Tenure				0.061							
Qualifications				0.006	0.013			0.056		0.057	0.029
SES					0.009						
Equalised income, £ 1000						0.049					

3.A.2 Determinants of occasionally-consumed food intake

This appendix presents the results of applying the two-part model to estimate associations between various personal and socio-economic predictors and the intake of occasionally-consumed foods in the men and women sample populations of NDNS RP Years 2-4. The analysis is shown for the following occasionally-consumed food intakes: fruits, vegetable (raw and cooked), processed meat, oily fish and sugary beverages.

Male sample

The results below correspond to the male sub-sample.

Table 3.A.5: Fruit intake predictors in men in the sample population selected for analysis from the NDNS RP Years 2-4

	SUBJECT-SPECIFIC TWO-PART MODEL EFFECTS					
	PROBABILITY (N = 2030)			CONDITIONAL AMOUNT (N = 1327)		
Predictors	Odds Ratio (95%CI)	*Wald p-value*	*LR p-value*[1]	Relative Change (95%CI)	*Wald p-value*	*LR p-value*[1]
Weekend	0.64 (0.49, 0.85)		*0.020*	0.74 (1.18, 1.46)		*<0.001*
Survey Year (base: 2009/2010)						
2010/2011	1.06 (0.67, 1.68)	*0.817*		0.88 (0.89, 1.39)	*0.365*	
2011/2012	1.39 (0.90, 2.13)	*0.133*		1.00 (0.76, 1.12)	*0.998*	
Age, years	1.04 (1.03, 1.06)		*<0.001*	1.009 (1.001, 1.017)		*0.032*
BMI, kg/m^2	0.97 (0.93, 1.01)		*0.115*			
Smoking (base: never smoked)			*<0.001*			*0.005*
ex, quit >10 year ago	0.91 (0.50, 1.67)	*0.765*		0.97 (0.71, 1.33)	*0.835*	
ex, quit <=10 years ago	0.80 (0.42, 1.52)	*0.503*		0.71 (0.49, 1.03)	*0.074*	
current	0.25 (0.16, 0.40)	*<0.001*		0.68 (0.50, 0.93)	*0.015*	
occasional	0.84 (0.27, 2.64)	*0.768*		2.07 (1.12, 3.81)	*0.020*	
Alcohol habit consumption			*0.069*			
rarely	2.16 (0.98, 4.76)	*0.057*				
never	0.69 (0.36, 1.31)	*0.252*				
MV physical activity, min/day (base: 0)			*0.004*			*0.001*
0-10	1.47 (0.43, 5.07)	*0.539*		0.86 (0.42, 1.77)	*0.683*	
10-20	2.29 (0.61, 8.61)	*0.222*		1.31 (0.60, 2.84)	*0.496*	
20-40	4.24 (1.23, 14.70)	*0.022*		1.66 (0.82, 3.38)	*0.159*	
40-60	3.48 (1.00, 12.12)	*0.051*		1.21 (0.59, 2.49)	*0.604*	
>60	4.05 (1.26, 13.06)	*0.019*		1.75 (0.89, 3.41)	*0.103*	
Non-meat eaters				0.65 (0.45, 0.94)		*0.023*

Table 3.A.5 Fruit intake predictors in men in the sample population selected for analysis from the NDNS RP Years 2-4 (Continued)

Take away (base: rarely or never)						*0.052*
occasionally				0.83 (0.65, 1.06)	*0.138*	
regularly				0.70 (0.53, 0.94)	*0.016*	
Tenure (base: mortgaged or owned)						*0.098*
privately rented				0.73 (0.55, 0.97)	*0.031*	
LA rented				0.91 (0.64, 1.30)	*0.609*	
Qualifications (base: bachelor and above)			*<0.001*			*0.006*
unfinished degree	0.90 (0.45, 1.80)	*0.762*		0.89 (0.61, 1.28)	*0.527*	
current students	0.55 (0.21, 1.48)	*0.240*		1.75 (0.94, 3.27)	*0.078*	
A levels	0.60 (0.34, 1.06)	*0.077*		0.80 (0.58, 1.10)	*0.165*	
GCSE A_C	0.35 (0.19, 0.61)	*<0.001*		0.57 (0.41, 0.79)	*0.001*	
GCSE below C and no qualifications	0.30 (0.17, 0.54)	*<0.001*		0.73 (0.52, 1.02)	*0.068*	
foreign qualifications	0.90 (0.35, 2.35)	*0.834*		1.06 (0.65, 1.72)	*0.824*	
Correlation between model parts is 0.59, p-value *<0.001*						
Within-person SD					1.11	
Between-person SD		1.48			0.87	
Within-person correlation					0.38	

[1]p-value from likelihood ratio test of the joint significance if a factor variable has more than two levels

Table 3.A.6: Cooked vegetables intake predictors in men in the sample population selected for analysis from the NDNS RP Years 2-4

	SUBJECT-SPECIFIC TWO-PART MODEL EFFECTS					
	PROBABILITY (N = 2030)			CONDITIONAL AMOUNT (N = 1063)		
Predictors	Odds Ratio (95% CI)	*Wald p-value*	*LR p-value*[1]	Relative Change (95% CI)	*Wald p-value*	*LR p-value*[1]
Weekend	1.01 (0.83, 1.23)		*0.936*	1.10 (1.00, 1.21)		*0.060*
Survey Year (base: 2009/2010)						
2010/2011	0.93 (0.70, 1.24)	*0.610*		0.97 (0.83, 1.13)	*0.708*	
2011/2012	1.16 (0.89, 1.52)	*0.269*		1.04 (0.90, 1.20)	*0.599*	
Age, years	1.02 (1.01, 1.03)		*<0.001*	1.001 (0.998, 1.005)		*0.448*
Ethnicity (non-white)				1.40 (1.12, 1.76)		*0.003*
Smoking (base: never smoked)			*0.051*			
ex, quit >10 year ago	1.01 (0.71, 1.44)	*0.955*				
ex, quit <=10 years ago	1.01 (0.68, 1.51)	*0.957*				
current	0.66 (0.49, 0.88)	*0.005*				
occasional	0.74 (0.37, 1.47)	*0.385*				
Non-meat eaters				1.25 (1.01, 1.53)		*0.037*
Qualifications (base: bachelor and above)			*0.019*			
unfinished degree	0.67 (0.44, 1.00)	*0.053*				
current students	0.96 (0.52, 1.78)	*0.898*				
A levels	0.66 (0.47, 0.94)	*0.020*				
GCSE A_C	0.73 (0.51, 1.03)	*0.074*				
GCSE below C and no qualifications	0.51 (0.36, 0.73)	*<0.001*				
foreign qualifications	0.69 (0.40, 1.18)	*0.175*				
Correlation between model parts is 0.50, p-value *0.005*						
Within-person SD					0.74	
Between-person SD		0.69			0.40	
Within-person correlation					0.47	

[1]p-value from likelihood ratio test of the joint significance if a factor variable has more than two level

Table 3.A.7: Raw and salad vegetables intake predictors in men in the sample population selected for analysis from the NDNS RP Years 2-4

	SUBJECT-SPECIFIC TWO-PART MODEL EFFECTS					
	PROBABILITY (N = 2030)			CONDITIONAL AMOUNT (N = 775)		
Predictors	Odds Ratio (95%CI)	*Wald p-value*	*LR p-value*[1]	Relative Change (95% CI)	*Wald p-value*	*LR p-value*[1]
Weekend	0.71 (0.56, 0.89)		*0.003*	1.16 (0.99, 1.36)		*0.067*
Survey Year (base: 2009/2010)						
2010/2011	0.76 (0.51, 1.15)	*0.190*		0.85 (0.65, 1.11)	*0.226*	
2011/2012	0.91 (0.62, 1.33)	*0.609*		0.78 (0.61, 1.00)	*0.047*	
Age, years	1.016 (1.003, 1.029)		*0.013*	1.006 (1.000, 1.013)		*0.069*
Ethnicity (non-white)	2.11 (1.16, 3.86)		*0.015*			
Lipid lowering drug	2.05 (1.20, 3.53)		*0.009*			
Partner (base: married)			*0.004*			
never married	1.55 (1.02, 2.35)	*0.038*				
previously married	0.59 (0.38, 0.94)	*0.026*				
Smoking (base: never smoked)			*0.017*			*0.175*
ex, quit >10 year ago	1.12 (0.67, 1.86)	*0.664*		0.75 (0.54, 1.03)	*0.072*	
ex, quit <=10 years ago	1.60 (0.91, 2.83)	*0.104*		0.83 (0.59, 1.19)	*0.313*	
current	0.63 (0.41, 0.97)	*0.036*		0.74 (0.55, 1.00)	*0.048*	
occasional	0.44 (0.16, 1.22)	*0.114*		0.70 (0.36, 1.38)	*0.309*	
MV Physical activity, min/day (base: 0)			*0.036*			
	1					
0-10	1.53 (0.51, 4.59)	*0.443*				
10-20	2.31 (0.71, 7.51)	*0.164*				
20-40	2.78 (0.92, 8.35)	*0.069*				
40-60	3.22 (1.06, 9.77)	*0.039*				
>60	3.17 (1.12, 8.96)	*0.030*				

Table 3.A.7 Raw and salad vegetables intake predictors in men in the sample population selected for analysis from the NDNS RP Years 2-4 (Continued)

Non-meat eaters				0.75		*0.089*
				(0.53, 1.05)		
Fruit/veg buy < weekly	0.58		*0.097*			
	(0.31, 1.10)					
Socio-economic status			*0.017*			*0.063*
(base) Higher managerial and professional occupation						
Lower managerial and professional occupation	0.81	*0.386*		1.04	*0.802*	
	(0.51, 1.30)			(0.78, 1.39)		
Intermediate occupation	0.77	*0.419*		0.68	*0.069*	
	(0.40, 1.46)			(0.45, 1.03)		
Small employers and own account workers	0.71	*0.293*		0.71	*0.095*	
	(0.38, 1.34)			(0.47, 1.06)		
Lower supervisory and technical occupation	0.96	*0.895*		0.90	*0.611*	
	(0.51, 1.79)			(0.61, 1.33)		
Semi-routine occupation	0.31	*<0.001*		1.30	*0.215*	
	(0.17, 0.58)			(0.86, 1.96)		
Routine occupation	0.54	*0.052*		1.30	*0.213*	
	(0.29, 1.00)			(0.86, 1.95)		
Never worked and other	0.64	*0.448*		1.11	*0.781*	
	(0.20, 2.03)			(0.53, 2.32)		
Correlation between model part is 0.50, p-value *<0.001*						
Within-person SD					0.95	
Between-person SD		1.32			0.72	
Within-person correlation					0.36	

[1]p-value from likelihood ratio test of the joint significance if a factor variable has more than two levels

Table 3.A.8: Processed meat intake predictors in men in the sample population selected for analysis from the NDNS RP Years 2-4

	SUBJECT-SPECIFIC TWO-PART MODEL EFFECTS					
	PROBABILITY (N = 1994)			CONDITIONAL AMOUNT (N = 966)		
Predictors	Odds Ratio (95%CI)	*Wald p-value*	*LR p-value*[1]	Relative change (95% CI)	*Wald p-value*	*LR p-value*[1]
Weekend	1.25 (0.99, 1.58)		*0.064*	1.19 (1.04, 1.36)		*0.010*
Survey Year (base: 2009/2010)						
2010/2011	1.27 (0.91, 1.77)	*0.163*		1.02 (0.84, 1.22)	*0.861*	
2011/2012	1.19 (0.87, 1.62)	*0.271*		0.99 (0.83, 1.17)	*0.876*	
Age, years	0.98 (0.97, 0.99)		*<0.001*	1.000 (0.996, 1.005)		*0.885*
Ethnicity (non-white)	0.28 (0.16, 0.50)		*<0.001*			
BMI, kg/m^2	1.034 (1.003, 1.066)		*0.032*			
Health (base: no problems)			*0.061*			
health problems, no mobility restrictions	1.12 (0.78, 1.62)	*0.538*				
health problems, mobility restrictions	1.59 (1.08, 2.34)	*0.018*				
Smoking (base: never smoked)						*0.101*
ex, quit >10 year ago				0.88 (0.70, 1.12)	*0.299*	
ex, quit <=10 years ago				1.21 (0.94, 1.55)	*0.140*	
current				1.14 (0.94, 1.38)	*0.188*	
occasional				1.43 (0.90, 2.29)	*0.131*	
Alcohol habit consumption			*0.092*			
rarely	0.83 (0.47, 1.46)	*0.516*				
never	0.56 (0.32, 0.96)	*0.035*				
Tenure (base: mortgaged or owned)						*0.092*
privately rented				1.15 (0.94, 1.41)	*0.165*	
local authority rented				1.28 (1.00, 1.64)	*0.048*	
Socio-economic status						*0.348*
(base) Higher managerial						

Table 3.A.8 Processed meat intake predictors in men in the sample population selected for analysis from the NDNS RP Years 2-4 (Continued)

and professional occupation			
Lower managerial and professional occupation		1.13 (0.92, 1.40)	*0.248*
Intermediate occupation		1.12 (0.84, 1.50)	*0.447*
Small employers and own account workers		1.00 (0.76, 1.32)	*0.993*
Lower supervisory and technical occupation		1.19 (0.90, 1.58)	*0.224*
Semi-routine occupation		1.08 (0.82, 1.43)	*0.564*
Routine occupation		1.34 (1.00, 1.78)	*0.047*
Never worked and other		1.74 (0.96, 3.14)	*0.068*
Between-person SD	0.91		21.5
Within-person SD			69.4
Within-person correlation			0.09

[1]p-value from likelihood ratio test of the joint significance if a factor variable has more than two levels

Table 3.A.9: Oily fish intake predictors in men in the sample population selected for analysis from the NDNS RP Years 2-4

	SUBJECT-SPECIFIC TWO-PART MODEL EFFECTS					
	PROBABILITY (N = 2030)			CONDITIONAL AMOUNT (N = 203)		
Predictors	Odds Ratio (95%CI)	*Wald p-value*	*LR p-value*[1]	Relative change (95% CI)	*Wald p-value*	*LR p-value*[1]
Weekend	0.85 (0.57, 1.27)		*0.425*	1.13 (0.82, 1.56)		*0.182*
Survey Year (base: 2009/2010)						
2010/2011	0.83 (0.49, 1.39)		*0.474*	0.78 (0.49, 1.22)		*0.267*
2011/2012	0.98 (0.61, 1.58)		*0.938*	0.98 (0.64, 1.50)		*0.933*
Age, years	1.021 (1.005, 1.037)		*0.010*	1.009 (0.998, 1.020)		*0.092*
BMI, kg/m^2				0.95 (0.91, 0.99)		*0.023*
Lipid lowering drug	2.03 (1.10, 3.75)		*0.024*			
Fruit/veg buy < weekly	0.42 (0.16, 1.10)		*0.076*	2.49 (1.05, 5.95)		*0.039*
Non-meat eaters	2.50 (1.22, 5.14)		*0.012*			
MV Physical activity, min/day (base: 0)			*0.006*			*0.064*
	1			1		
0-10	0.50 (0.13, 2.01)	*0.331*		6.84 (1.89, 24.71)	*0.003*	
10-20	1.31 (0.30, 5.71)	*0.721*		3.00 (0.79, 11.41)	*0.107*	
20-40	2.01 (0.54, 7.50)	*0.297*		4.93 (1.49, 16.28)	*0.009*	
40-60	2.20 (0.59, 8.22)	*0.242*		5.70 (1.75, 18.61)	*0.004*	
>60	1.05 (0.30, 3.64)	*0.939*		4.31 (1.37, 13.57)	*0.012*	
Alcohol habit consumption			*0.006*			
rarely	0.16 (0.03, 0.74)	*0.019*				
never	0.46 (0.18, 1.21)	*0.117*				
Take away shopping			*0.022*			
occasionally	0.94 (0.58, 1.52)	*0.796*				
regularly	0.44 (0.23, 0.84)	*0.012*				
Fish dislike	0.12 (0.04, 0.36)		*<0.001*			

Table 3.A.9 Oily fish intake predictors in men in the sample population selected for analysis from the NDNS RP Years 2-4 (Continued)

Qualifications (base: bachelor and up)			*0.029*	
unfinished degree	0.78 (0.40, 1.49)	*0.447*		
students	0.89 (0.29, 2.74)	*0.842*		
A levels	0.60 (0.33, 1.12)	*0.108*		
GCSE A_C	0.41 (0.21, 0.78)	*0.006*		
GCSE below C and no qualifications	0.37 (0.20, 0.69)	*0.002*		
foreign qualifications	0.45 (0.17, 1.17)	*0.100*		
Within-person SD				0.69
Between-person SD		1.07		0.90
Within-person correlation				0.63

[1]p-value from likelihood ratio test of the joint significance if a factor variable has more than two levels

Table 3.A.10: Sugary beverages (excluding juice) intake predictors in men in the sample population selected for analysis from the NDNS RP Years 2-4

	SUBJECT-SPECIFIC TWO-PART MODEL EFFECTS					
	PROBABILITY (N = 2030)			CONDITIONAL AMOUNT (N = 584)		
Predictors	Odds Ratio (95%CI)	*Wald p-value*[2]	*LR p-value*[1]	Relative change, g (95% CI)	*Wald p-value*[2]	*LR p-value*[1]
Weekend	0.79 (0.57, 1.10)		*0.165*	1.14 (1.02, 1.28)		*0.022*
Survey Year						
(base) 2009/2010						
2010/2011	0.91 (0.45, 1.84)		*0.791*	1.04 (0.85, 1.28)		*0.687*
2011/2012	1.22 (0.63, 2.36)		*0.561*	1.11 (0.91, 1.34)		*0.315*
Age, years	0.95 (0.93, 0.97)		*<0.001*	0.979 (0.974, 0.985)		*<0.001*
Lipid lowering drug	0.38 (0.13, 1.13)		*0.081*			
Fruit/veg buy < weekly	2.24 (0.81, 6.25)		*0.121*			
MV Physical activity, min/day[2] (base:0)			*0.096*			*0.002*
>0	0.21 (0.03, 1.32)			0.42 (0.24, 0.73)		
Tenure						
(base) mortgaged or owned	1		*0.122*	1		*0.151*
privately rented	1.58 (0.76, 3.30)	*0.224*		1.02 (0.84, 1.25)	*0.831*	
LA rented	0.54 (0.22, 1.33)	*0.179*		0.76 (0.57, 1.02)	*0.066*	
Within-person SD					0.51	
Between-person SD		2.53			0.55	
Within-person correlation					0.54	

[1]p-value from likelihood ratio test of the joint significance if a factor variable has more than two levels

[2]The effect of moderate to vigorous physical activity (MVPA) is modelled as two categories: none MVPA versus any MVPA as it showed better model fit (AIC criteria). When MVPA was fitted as 6 categories the effects of each category on the sugary drinks consumption in both parts of the model were all negative and almost identical in size.

Female sample

The results below correspond to the female sub-sample.

Table 3.A.11: Fruit intake predictors in women in the sample selected for analysis from the NDNS RP Years 2-4

	SUBJECT-SPECIFIC TWO-PART MODEL EFFECTS					
	PROBABILITY (N = 2461)			CONDITIONAL AMOUNT (N = 1780)		
Predictors	Odds Ratio (95%CI)	*Wald p-value*	*LR p-value*[1]	Relative Change (95% CI)	*Wald p-value*	*LR p-value*[1]
Weekend	0.86 (0.66, 1.12)		*0.255*	0.88 (0.78, 1.00)		*0.052*
Survey Year (base: 2009/2010)						
2010/2011	0.81 (0.54, 1.21)	*0.308*		0.96 (0.78, 1.19)	*0.741*	
2011/2012	0.91 (0.62, 1.33)	*0.608*		1.06 (0.87, 1.29)	*0.577*	
Age, years	1.05 (1.03, 1.06)		*<0.001*	1.01 (1.00, 1.02)		*0.001*
BMI, kg/m^2	0.96 (0.93, 0.98)		*0.002*			
Health (base: no problems)			*0.086*			
health problems, no mobility restrictions	0.67 (0.44, 1.02)	*0.059*				
health problems, mobility restrictions	1.15 (0.75, 1.78)	*0.516*				
Smoking (base: never smoked)			*0.047*			
ex, quit >10 year ago	0.79 (0.46, 1.34)	*0.378*				
ex, quit <=10 years ago	0.78 (0.47, 1.28)	*0.319*				
current	0.52 (0.34, 0.80)	*0.003*				
occasional	1.53 (0.36, 6.39)	*0.564*				
MV Physical activity, min/day (base: 0)			*0.028*			*0.001*
0-10	0.63 (0.20, 2.0)	*0.433*		0.87 (0.51, 1.50)	*0.625*	
10-20	0.75 (0.22, 2.49)	*0.634*		0.93 (0.52, 1.65)	*0.804*	
20-40	1.26 (0.39, 4.11)	*0.701*		1.19 (0.69, 2.07)	*0.531*	
40-60	1.36 (0.40, 4.59)	*0.620*		1.34 (0.76, 2.36)	*0.318*	
>60	1.11 (0.35, 3.53)	*0.854*		1.42 (0.83, 2.43)	*0.204*	
Fruit/veg buy <weekly	0.51 (0.37, 0.80)		0.027	0.76 (0.54, 1.07)		*0.114*

Table 3.A.11 Fruit intake predictors in women in the sample selected for analysis from the NDNS RP Years 2-4 (Continued)

Take away shopping			0.008			*<0.001*
occasionally	0.54 (0.37, 0.80)	*0.002*		0.70 (0.57, 0.86)	*0.001*	
regularly	0.66 (0.42, 1.02)	*0.063*		0.57 (0.44, 0.72)	*<0.001*	
Tenure (base: mortgaged or owned)			*<0.001*			
privately rented	0.83 (0.53, 1.30)	*0.418*				
LA rented	0.38 (0.24, 0.59)	*<0.001*				
Qualifications (base: bachelor degree and above)			*0.096*			*0.003*
unfinished degree	0.83 (0.46, 1.50)	*0.537*		1.19 (0.87, 1.62)	*0.287*	
current students	0.72 (0.34, 1.50)	*0.376*		0.70 (0.45, 1.11)	*0.128*	
A levels	1.21 (0.72, 2.04)	*0.479*		0.73 (0.55, 0.96)	*0.026*	
GCSE A_C	0.70 (0.42, 1.18)	*0.184*		0.77 (0.58, 1.01)	*0.059*	
GCSE below C and no qualifications	0.55 (0.33, 0.93)	*0.026*		0.65 (0.49, 0.85)	*0.002*	
foreign qualifications	0.96 (0.37, 2.50)	*0.939*		0.89 (0.58, 1.37)	*0.588*	
McClement equivalence score, £ 1000						*0.088*
<=15				0.77 (0.59, 1.01)	*0.061*	
15-25				0.97 (0.74, 1.27)	*0.832*	
(base) 25-35				1		
35-50				1.14 (0.86, 1.52)	*0.367*	
>50				1.02 (0.76, 1.37)	*0.896*	
Correlation between model parts is 0.41, p-value < *0.001*						
Within-person SD					1.03	
Between-person SD		1.24			0.75	
Within-person correlation					0.35	

[1]p-value from likelihood ratio test of the joint significance if a factor variable has more than two levels

Table 3.A.12: Cooked vegetables intake predictors in women in the sample selected for analysis from the NDNS RP Years 2-4

	SUBJECT-SPECIFIC TWO-PART MODEL EFFECTS					
	PROBABILITY (N = 2461)			CONDITIONAL AMOUNT (N = 1322)		
Predictors	Odds Ratio (95% CI)	*Wald p-value*[2]	*LR p-value*[1]	Relative Change (95% CI)	*Wald p-value*[2]	*LR p-value*[1]
Weekend	1.34 (1.09, 1.64)		*0.005*	1.02 (0.93, 1.12)		*0.688*
Survey Year (base :2009/2010)						
2010/2011	1.20 (0.94, 1.54)		*0.146*	0.95 (0.84, 1.08)		*0.458*
2011/2012	0.97 (0.77, 1.23)		*0.822*	0.98 (0.87, 1.10)		*0.728*
Age, years	1.014 (1.007, 1.020)		*<0.001*	0.998 (0.995, 1.001)		*0.214*
Partner[3](base: married/in partnership)						*0.011*
never married				0.91 (0.90, 1.04)	*0.170*	
previously married				1.15 (1.02, 1.30)	*0.022*	
Fruit/veg buy < weekly	0.66 (0.45, 0.97)		*0.033*	0.84 (0.68, 1.03)		*0.092*
Physical activity, min/day (base: 0)			*0.006*			*0.086*
0-10	2.26 (1.15, 4.45)	*0.018*		1.01 (0.69, 1.48)	*0.968*	
10-20	3.70 (1.82, 7.54)	*<0.001*		1.00 (0.67, 1.48)	*0.996*	
20-40	2.73 (1.37, 5.46)	*0.005*		1.09 (0.74, 1.60)	*0.678*	
40-60	2.42 (1.19, 4.93)	*0.015*		1.20 (0.81, 1.79)	*0.358*	
>60	2.69 (1.37, 5.28)	*0.004*		1.19 (0.81, 1.74)	*0.373*	
Vegetarian	1.70 (1.27, 2.27)		*<0.001*	1.27 (1.13, 1.41)		*<0.001*
Alcohol habit consumption			*0.035*			
rarely	0.58 (0.38, 0.88)	*0.010*				
never	0.97 (0.69, 1.37)	*0.872*				
Take away shopping			*0.008*			
occasionally	0.78 (0.61, 0.99)	*0.041*				
regularly	0.77 (0.58, 1.01)	*0.063*				

Table 3.A.12 Cooked vegetables intake predictors in women in the sample selected for analysis from the NDNS RP Years 2-4 (Continued)

Tenure (base: mortgaged or owned)			*0.001*			
privately rented	0.64	*0.004*				
	(0.48, 0.87)					
local authority rented	0.63	*0.001*				
	(0.48, 0.83)					
McClement equivalence score, £ 1000						*0.007*
<=15				1.01	*0.934*	
				(0.86, 1.17)		
15-25				1.29	*0.001*	
				(1.10, 1.50)		
(base) 25-35				1		
35-50				1.00	*0.992*	
				(0.84, 1.19)		
>50				1.05	*0.565*	
				(0.88, 1.26)		
Correlation between model parts is 0.24, p-value *0.350*						
Within-person SD					0.78	
Between-person SD		0.58			0.25	
Within-person correlation					0.09	

[1]p-value from likelihood ratio test of the joint significance if a factor variable has more than two levels

Table 3.A.13: Raw and salad vegetables intake predictors in women in the sample selected for analysis from the NDNS RP Years 2-4

	SUBJECT-SPECIFIC TWO-PART MODEL EFFECTS					
	PROBABILITY (N = 2461)			CONDITIONAL AMOUNT (N = 1024)		
Predictors	Odds Ratio (95%CI)	*Wald p-value*[2]	*LR p-value*[1]	Relative Change	*Wald p-value*[2]	*LR p-value*[1]
Weekend	0.83 (0.67, 1.04)		*0.113*	1.06 (0.91, 1.23)		*0.447*
Survey Year (base: 2009/2010)						
2010/2011	0.90 (0.65, 1.25)	*0.544*		1.06 (0.88, 1.27)	*0.458*	
2011/2012	0.88 (0.65, 1.20)	*0.418*		1.02 (0.85, 1.22)	*0.842*	
Age, years	1.014 (1.004, 1.025)		*0.008*	1.006 (1.001, 1.010)		*0.015*
BMI, kg/m^2	1.03 (1.00, 1.05)		*0.032*			
Ethnicity (non-white)	1.80 (1.08, 3.00)		*0.023*			
Health (self-reported)			*0.046*			*0.101*
health problems, no mobility restrictions	0.73 (0.90, 1.04)	*0.078*		0.91 (0.74, 1.12)	*0.375*	
health problems, mobility restrictions	0.68 (0.48, 0.97)	*0.032*		1.20 (0.97, 1.47)	*0.091*	
Lipid lowering drug	0.68 (0.42, 1.10)		*0.114*			
Physical activity, min/day (base: 0)			*0.164*			*0.006*
0-10	1.45 (0.58, 3.60)	*0.425*		1.06 (0.59, 1.90)	*0.837*	
10-20	1.96 (0.76, 5.08)	*0.166*		1.37 (0.75, 2.51)	*0.303*	
20-40	1.85 (0.73, 4.69)	*0.193*		1.35 (0.74, 2.43)	*0.318*	
40-60	2.12 (0.81, 5.51)	*0.123*		1.35 (0.74, 2.46)	*0.335*	
>60	2.28 (0.92, 5.66)	*0.077*		1.64 (0.92, 2.94)	*0.095*	
Vegetarian	1.42 (1.01, 2.00)		*0.042*			
Alcohol habit consumption						*0.097*
rarely				0.73 (0.54, 0.98)	*0.039*	
never				1.06 (0.82, 1.37)	*0.667*	

Table 3.A.13 Raw and salad vegetables intake predictors in women in the sample selected for analysis from the NDNS RP Years 2-4 (Continued)

Take away shopping			*0.096*	
occasionally	0.81 (0.59, 1.12)	*0.198*		
regularly	0.67 (0.46, 0.97)	*0.034*		
Tenure			*0.023*	
privately rented	0.65 (0.43, 0.98)	*0.042*		
LA rented	0.62 (0.41, 0.93)	*0.022*		
Qualifications (base: bachelor and up)			*0.014*	
unfinished degree	0.63 (0.38, 1.02)	*0.063*		
students	0.45 (0.23, 0.89)	*0.023*		
A levels	0.73 (0.47, 1.12)	*0.153*		
GCSE A_C	0.59 (0.38, 0.90)	*0.014*		
GCSE below C and no qualifications	0.45 (0.29, 0.71)	*<0.001*		
foreign qualifications	0.62 (0.31, 1.23)	*0.173*		
Socio-economic status (base: higher managerial and professional)			*0.052*	
Lower managerial and professional occupation	0.66 (0.43, 0.99)	*0.047*		
Intermediate occupation	0.99 (0.58, 1.70)	*0.976*		
Small employers and own account workers	0.90 (0.54, 1.50)	*0.684*		
Lower supervisory and technical occupation	0.56 (0.31, 1.03)	*0.064*		
Semi-routine occupation	0.53 (0.32, 0.88)	*0.014*		
Routine occupation	0.54 (0.30, 0.97)	*0.038*		
Never worked and other	0.95 (0.43, 2.07)	*0.893*		
Correlation between the model parts is 0.06, *p-value 0.691*				
Between-person SD	1.04		0.43	
Within-person SD			0.97	
Within-person correlation			0.16	

[1]p-value from likelihood ratio test of the joint significance if a factor variable has more than two levels

Table 3.A.14: Processed meat intake predictors in women in the sample selected for analysis from the NDNS RP Years 2-4

	SUBJECT-SPECIFIC TWO-PART MODEL EFFECTS					
	PROBABILITY (N = 2370)			CONDITIONAL AMOUNT (N = 1000)		
Predictors	Odds Ratio (95% CI)	*Wald p-value*[2]	*LR p-value*[1]	Relative change (95% CI)	*Wald p-value*[2]	*LR p-value*[1]
Weekend	1.50 (1.21, 1.85)		*<0.001*	1.02 (0.90, 1.16)		*0.730*
Survey Year (base: 2009/2010)						
2010/2011	0.82 (0.61, 1.09)	*0.178*		1.15 (0.96, 1.37)	*0.133*	
2011/2012	1.02 (0.78, 1.33)	*0.908*		1.02 (0.87, 1.20)	*0.836*	
Age, years	0.991 (0.984, 0.998)		*0.013*	1.000 (0.996, 1.004)		*0.588*
Ethnicity (non-white)	0.43 (0.27, 0.68)		*<0.001*			
Fruits/veg buy < weekly	0.60 (0.38, 0.94)		*0.025*	1.28 (0.96, 1.71)		*0.099*
Health (self-reported)						*0.144*
health problems, no mobility restrictions				1.18 (0.98, 1.43)	*0.009*	
health problems, mobility restrictions				0.97 (0.80, 1.17)	*0.729*	
Take away shopping			*0.106*			
occasionally	1.27 (0.95, 1.68)	*0.105*				
regularly	1.38 (1.00, 1.90)	*0.050*				
Current smoker[2]				1.25 (1.05, 1.48)		*0.012*
	Correlation between the model parts is 0.03, *p-value 0.856*					
Between-person SD	0.85			0.42		
Within-person SD				0.88		
Within-person correlation				0.19		

[1]p-value from likelihood ratio test of the joint significance if a factor variable has more than two levels

[2]Current smoker status compared to the other smoking related history showed the best prediction in the model. No differences were observed between the other levels of smoking history.

Table 3.A.15: Oily fish intake predictors in women in the sample selected for analysis from the NDNS RP Years 2-4

	SUBJECT-SPECIFIC TWO-PART MODEL EFFECTS					
	PROBABILITY (N = 2416)			CONDITIONAL AMOUNT (N = 219)		
Predictors	Odds Ratio (95%CI)	*Wald p-value*[2]	*LR p-value*[1]	Absolute change, g (95% CI)	*Wald p-value*[2]	*LR p-value*[1]
Weekend	0.87 (0.61, 1.25)		*0.460*	0.83 (0.62, 1.12)		*0.213*
Survey Year(base) 2009/2010						
2010/2011	0.90 (0.59, 1.37)	*0.624*		0.79 (0.58, 1.08)	*0.145*	
2011/2012	0.69 (0.46, 1.03)	*0.072*		0.65 (0.47, 0.88)	*0.006*	
Age, years	1.026 (1.013, 1.039)		*<0.001*	1.02 (1.01, 1.03)		*<0.001*
Ethnicity (non-white)	2.03 (1.10, 3.75)		*0.024*			
BP lowering drug				0.62 (0.42, 0.94)		*0.023*
Fruit/veg buy < weekly	0.28 (0.10, 0.75)		*0.012*			
Physical activity, min/day(base) 0			*0.201*			
0-10	3.39 (0.68, 16.93)	*0.136*				
10-20	4.78 (0.93, 24.64)	*0.062*				
20-40	4.68 (0.93, 23.52)	*0.061*				
40-60	3.99 (0.77, 20.72)	*0.100*				
>60	5.04 (1.02, 24.97)	*0.048*				
Take away shopping			*0.011*			
occasionally	0.58 (0.37, 0.90)	*0.015*				
regularly	0.52 (0.30, 0.89)	*0.018*				
Qualifications (base: bachelor and up)			*0.055*			
unfinished degree	0.63 (0.34, 1.16)	*0.135*				
students	0.65 (0.23, 1.83)	*0.417*				
A levels	0.49 (0.27, 0.89)	*0.020*				
GCSE A_C	0.44 (0.25, 0.78)	*0.005*				

Table 3.A.15 Oily fish intake predictors in women in the sample selected for analysis from the NDNS RP Years 2-4 (Continued)

GCSE below C and no qualifications	0.59 (0.35, 1.00)	*0.052*			
foreign qualifications	0.36 (0.15, 0.87)	*0.023*			
McClement equivalence score, £ 1000			*0.005*		
<=15	0.47 (0.26, 0.87)	*0.015*			
15-25	0.78 (0.45, 1.34)	*0.368*			
(base) 25-35	1				
35-50	1.26 (0.73, 2.17)	*0.410*			
>50	1.34 (0.76, 2.37)	*0.318*			
Fish dislike	0.20 (0.09, 0.43)		<0.001		
			Correlation[3] between the model parts is 0		
Within-person SD				0.93	
Between-person SD		0.77		0.22	
Within-person correlation				0.05	

[1]p-value from likelihood ratio test of the joint significance if a factor variable has more than two levels

Correlation between the probability and the conditional amount model parts was set to 0 based on the preliminary analysis described in Table 2 of the Results section in the main text.

Table 3.A.16: Sugary beverages (excluding juice) intake predictors in women in the sample selected for analysis from the NDNS RP Years 2-4

	SUBJECT-SPECIFIC TWO-PART MODEL EFFECTS					
	PROBABILITY (N = 2461)			CONDITIONAL AMOUNT (N = 616)		
Predictors	Odds Ratio (95%CI)	*Wald p-value*	*LR p-value*[1]	Relative Change (95% CI)	*Wald p-value*	*LR p-value*[1]
Weekend	1.14 (0.85, 1.53)		*0.361*	1.08 (0.96, 1.20)		*0.222*
Survey Year (base: 2009/2010)	1			1		
2010/2011	0.60 (0.35, 1.04)	*0.068*		0.93 (0.77, 1.12)	*0.464*	
2011/2012	0.75 (0.45, 1.24)	*0.259*		1.05 (0.88, 1.25)	*0.587*	
Age, years	0.97 (0.95, 0.98)		*<0.001*	0.99 (0.98, 0.99)		*<0.001*
Ethnicity (non-white)	1.96 (0.91, 4.26)		*0.088*			
BMI, kg/m^2				1.010 (0.998, 1.024)		*0.094*
Lipid lowering medicine				0.71 (0.53, 0.96)		*0.024*
Partner (base: married)	1		*0.085*			
never married	1.88 (1.07, 3.30)	*0.029*				
previously married	1.29 (0.75, 2.21)	*0.356*				
Smoking (base: never smoked)	1		*0.103*			
ex, quit >10 year ago	1.18 (0.59, 2.38)	*0.636*				
ex, quit <=10 years ago	0.39 (0.19, 0.83)	*0.014*				
current	1.13 (0.65, 1.97)	*0.604*				
occasional	1.19 (0.22, 6.54)	*0.842*				
MV Physical activity[2], min/day (base:0)	5.56 (0.98, 31.50)		*0.053*	1.90 (0.93, 3.89)		*0.079*
Vegetarian				1.22 (1.02, 1.46)		*0.026*
Take away shopping (base: never)				1		*0.002*
occasionally				1.20 (1.01, 1.43)	*0.043*	
regularly				1.44 (1.17, 1.76)	*<0.001*	

Table 3.A.16 Sugary beverages (excluding juice) intake predictors in women in the sample selected for analysis from the NDNS RP Years 2-4 (Continued)

Tenure (base: mortgaged or owned)				1		*0.032*
privately rented				0.84 (0.68, 1.03)	*0.096*	
LA rented				0.79 (0.64, 0.99)	*0.041*	
Qualifications[3] (base: bachelor degree and above)						
Lower than bachelor degree				1.21 (1.01, 1.44)		*0.041*
McClement equivalence score, £1000[4]			*0.208*			*0.173*
<=15	0.85 (0.43, 1.67)	*0.637*		0.92 (0.73, 1.16)	*0.474*	
15-25	0.92 (0.46, 1.83)	*0.816*		0.92 (0.73, 1.15)	*0.452*	
(base) 25-35	1			1		
35-50	0.51 (0.24, 1.11)	*0.090*		0.95 (0.73, 1.24)	*0.695*	
>50	0.46 (0.20, 1.03)	*0.060*		0.70 (0.52, 0.93)	*0.014*	
Correlation between the model parts is 0.42 , *p-value 0.003*						
Within-person SD					0.54	
Between-person SD		1.99			0.47	
Within-person correlation					0.43	

[1]p-value from likelihood ratio test of the joint significance if a factor variable has more than two levels

[2]MV Physical activity is analysed as 2 categories (0 min MVPA /day vs. >0 min MVPA/day)

[3]Qualifications is analysed as 2 categories (base: bachelor degree and above and lower than bachelor degree)

[4]McClement's equivalence score was jointly significant in both parts of the model at p-value 0.109.

3.B Detailed statistical methods

3.B.1 Mixed-effect mixed-distribution model

We briefly describe the two-part mixed-effects model suitable for repeated positive continuous responses with excess zeros (cf. Olsen and Schafer (2001), Tooze et al. (2002), and Su et al. (2009) for full details). For each person, $i, i = 1, \ldots, m$ on day $j, j = 1, \ldots n_i$, the data consist of two parts: the occurrence of food consumption (yes/no), which can be recorded as an indicator variable I_{ij} such that:

$$I_{ij} = \begin{cases} 1, & \text{if the food is consumed by person i on day j} \\ 0, & \text{otherwise} \end{cases}$$

and the amount of food consumed if consumption took place, which we record as $A_{ij}, A_{ij} > 0$ if $I_{ij} = 1$. Natural heterogeneity arise among subjects due to personal preferences for consumption. We denote unobservable person-specific information related to propensity to consume certain foods as v_i and unobservable person-specific information related to amount consumed on consumption day as u_i. Then, conditionally on v_i and u_i, responses I_{ij} and A_{ij} are independent. The indicator variable I_{ij} is assumed to follow a Bernoulli distribution with probability p_{ij}, and to allow for skewness, we assume $A_{ij}, A_{i,j} > 0$ to be log-normally distributed. In this chapter, we suggest the following model specification: the first part response I_{ij} follows the logistic regression model:

$$\text{logit}\{\Pr(I_{ij} = 1|v_i)\} = x_{ij}^{'}\gamma + v_i$$

where $x_{ij}^{'}$ is the vector of relevant covariates, relating individual characteristics to propensity for food intake, and γ is the vector of corresponding regression coefficients. And, considering, $\log(A_{ij}) = Y_{ij}$ is approximately normal, we can write:

$$Y_{ij} = x_{ij}^{'}\beta + u_i + \epsilon_{ij}$$

where $\text{E}(Y_{ij}|u_i) = x_{ij}^{'}\beta + u_i$ and $\text{Var}(Y_{ij}|u_i) = \sigma_\epsilon^2$ (within-person daily variation); $x_{ij}^{'}$ is the vector of relevant covariates relating individual characteristics to the amount of food consumed, β is the vector of corresponding regression coefficients. The potential correlation

between the *probability* and *amount* parts is induced through person-specific effects u_i and v_i, which are assumed to have a common bivariate normal distribution with means 0 and variance-covariance matrix:

$$\Sigma = \begin{pmatrix} \sigma_u^2 & \rho\sigma_u\sigma_v \\ \rho\sigma_u\sigma_v & \sigma_v^2 \end{pmatrix}$$

where ρ denotes the correlation between u_i and v_i, σ_u^2 and σ_v^2 are the variances of u_i and v_i respectively. These are called random effects and are assumed to be independent of ϵ_{ij}. The unknown model parameters $\theta = (\gamma, \beta, \sigma_u, \sigma_v, \sigma_\epsilon, \rho)$ can be estimated through maximising the full marginal likelihood function, where we utilise the conditional independence of responses I_{ij} and Y_{ij} and their distributional assumptions. Because the random effects u_i and v_i are unobserved, they need to be integrated out, so that the full marginal likelihood function is:

$$L(\theta) \propto \prod_{i=1}^{m} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \prod_{j=1}^{n_i} f_I(I_{ij} \mid v_i, \theta) f_Y(Y_{ij} \mid u_i, \theta) f_{UV}(u_i, v_i \mid \theta) du_i dv_i \tag{3.1}$$

where f_I, f_Y and f_{UV} denote the density functions of the binomial, normal and bivariate normal distributions, respectively. The likelihood function does not have a closed form and needs to be evaluated numerically.

We note that if it is assumed that the random effects are independent, i.e. $\rho = 0$, estimation is considerably simplified as the two parts can be fitted separately using standard statistical software for generalised mixed-effect models.

Due to non-linearity the estimates of the regression coefficients β and γ are subject-specific. To obtain estimates, which could be generalised to a group of participants, the marginal effects should be obtained.

3.B.2 Marginal effects

In certain cases, when the interpretation of subject-specific regression coefficients is problematic, marginal effects of predictors can be obtained. This section explains how to es-

timate the marginal effects of predictors in the two-part model, the expected individual intake and the expected group intake.

The expected individual habitual daily intake T_{ij} for a person i on a day j is calculated as the product of the expected individual daily probability of the food consumption, p_{ij}, and the expected individual consumed amount on a consumption day:
$T_{ij} = P(I_{ij} = 1|v_i) \cdot E(A_{ij}|A_{ij} > 0, u_i)$. Under the two-part model T_{ij} depends on the set of parameters θ as well as on the unobserved person-specific effects u_i and v_i, which may be correlated. The point model estimates $\tilde{\theta}$ of θ can be obtained through likelihood maximisation but the person-specific u_i and v_i are unobserved. Using the the notation introduced in Appendix 3.B.1, the estimated expected individual habitual daily intake can be then described as

$$\tilde{T_{ij}} = \exp(x_{ij}^{'}\tilde{\beta} + u_i + 0.5\tilde{\sigma}_{\epsilon}^2)\frac{\exp(x_{ij}^{'}\tilde{\gamma} + v_i)}{1 + \exp(x_{ij}^{'}\tilde{\gamma} + v_i)} \tag{3.2}$$

A weekend may change the expected daily intake (for example, this is true for alcohol consumption). To account for this, we assign weights to the estimated expected daily weekend and weekdays intakes accordingly, so that the expected individual daily intake $\tilde{T_i}$ now becomes:

$$\tilde{T_i} = \frac{4}{7}\exp(x_{i0}^{'}\tilde{\beta} + u_i + 0.5\tilde{\sigma}_{\epsilon}^2)\frac{\exp(x_{i0}^{'}\tilde{\gamma} + v_i)}{1 + \exp(x_{i0}^{'}\tilde{\gamma} + v_i)} + \frac{3}{7}\exp(x_{i1}^{'}\tilde{\beta} + u_i + 0.5\tilde{\sigma}_{\epsilon}^2)\frac{\exp(x_{i1}^{'}\tilde{\gamma} + v_i)}{1 + \exp(x_{i1}^{'}\tilde{\gamma} + v_i)} \tag{3.3}$$

where $x_{i0}^{'}$ and $x_{i1}^{'}$ are the vector of explanatory variables corresponding to week days and weekends.

From the above quantity and under the distributional assumptions for v_i and u_i as bivariate normal $(0, \tilde{\Sigma})$ we can estimate *the expected group intake* $\tilde{T}_A = \mathrm{E}(\tilde{T_i}|i \in A)$, where A denotes the subgroup of interest with certain fixed characteristics. The expected intake for this subgroup is given by

$$\tilde{T}_A = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \tilde{T}_i f_{UV}(u_i, v_i) du_i dv_i \tag{3.4}$$

Typically, the integration of the above quantity requires numerical methods.

The marginal effects of a predictor can be found via the estimated expected group intakes (Su et al., 2009). Consider participants in *group 1* that differ from participants from *group 2* only in one characteristic, the effect of which we are trying to estimate. The expected group intakes can be obtained as described above and, then, the difference between the intakes can be attributed to the effect of the characteristic in question. Clearly, the estimated marginal group intakes are non-linear with respect to the group participants characteristics, so, to make the interpretation of the results easier, we set a baseline group and compare the effect of predictors against this group. This chapter utilises the baseline group with the following characteristics: white, 48 years old, having BMI of 27.4 (men) and 26.7 (women), with no health and mobility problems, not on lipid or BP lowering drug, married or live in partnership, never-smoker, involved in moderate-to-vigorous physical activity more than 60 min daily, not vegetarian, regularly buying fruits and vegetables, rarely buying takeaways, owned or mortgaged accommodation, qualified to bachelor degree and above, being in the lower managerial and professional occupation, with equalised income of £25000 - £35000 (men) and £15000 - £25000 (women).

3.B.3 Complex survey design

Generally speaking, two methodological issues should be addressed when survey data are being analysed: firstly, standard errors can be misleading if clustering of the participants from the same sampling unit is not adjusted for. Secondly, the generalisability of the results need to be considered. Typically, the latter is addressed by applying individual weights to adjust for different selection probabilities when inference about such population parameters as means or frequencies is required.

However, as discussed by Carle (2009) and Pfeffermann (2011), when using survey data

to elicit determinants of participants' behaviour in a model-based approach, no clear-cut strategy on applying individual weights in the analysis currently dominates.

This chapter utilises National Diet and Nutrition Survey Rolling Programme (NDNS RP) data to make inference about the predictors of participants' food-related behaviour by applying a mixed-effect mixed-distribution modelling approach and it does not incorporate individual weights in the analysis for the reasons listed below.

Firstly, NDNS RP data come from a selection procedure that is non-informative because the primary sampling units (PSUs) were chosen at random and the inclusion probabilities, including non-response, are not explicitly conditioned on the examined outcomes. This implies that the point estimates of the effects of the predictors on participants' food choice should be unbiased.

Secondly, to minimise the potential selection bias associated with non-response, we adjust for possible predictors of both, non-response and food intake: age, tenure, income and other socio-demographic characteristics.

Thirdly, clustering of the responses is addressed in two of ways. Correlation of individual responses within a participant is explicitly incorporated into the model variance-covariance structure. Clustering within PSU is adjusted through taking into account certain neighbourhood's characteristics such as clustering of participants with similar social occupations and incomes through estimation of expected values at individual level.

Gelman (2007) suggests that under the above conditions within the framework of mixed-effect modelling the model-based point estimates derived from complex design surveys can serve as appropriate estimates of the predictors' effects. Carle (2009) shows that unweighted model-based inference on point estimates is very similar to the inference based on scaled weighting and lead to the same conclusions regarding the importance of the predictors.

There remains a possibility that residual clustering exists that is not accounted for by the measured predictors or by modelling of variance-covariance responses structure. This is addressed by running the analysis of habitually consumed foods (continuous responses only) taking into account both levels of correlation: individual (Level 1) and PSU (Level 2) through multi-level modelling. The analysis showed that, after adjusting for within-person correlation and various personal and socio-economic predictors the residual within-PSU correlation is negligible and, consequently, should not affect the results.

Chapter 4

Joint modelling of multiple correlated habitually- and occasionally-consumed food intakes with application to alcohol intake

This chapter extends the two-part model presented in Chapters 2 and 3 to model the intake of several occasionally consumed foods jointly. Parameter estimation is carried out using a pseudolikelihood approach coupled with parametric bootstrap. The method is illustrated by analysing alcohol intake, jointly with the intake of other foods, from a subsample population of the UK National Diet and Nutrition Survey Rolling Programme.

4.1 Background

We showed in previous chapters how to analyse occasionally-consumed food intake using a two-part mixed-effects model which allows person-specific preferences for consumption frequency and portion size to be correlated. This chapter models the intake of two or more occasionally-consumed foods, allowing for unobserved preferences to be correlated

across various food intakes. A full specification of the underlying correlation structure in a joint model allows the investigation of the residual correlation among various foods that remains after adjusting for observed predictors, and is expected to provide more robust inferences for the model parameters. The joint modelling of two occasionally-consumed food intakes consists of four parts, with four potentially correlated random effects, making maximum likelihood estimation challenging due to the required computation of a high-dimensional integral. The extension to model three or more occasionally-consumed foods jointly, quickly increases the dimensionality of the integration space. An approach to this problem found in the literature is to resource to Markov Chain Monte Carlo (MCMC) (Zeger and Karim, 1991; Hamra et al., 2013; Zhang et al., 2011), or Monte Carlo expectation maximisation (Guolo, 2011) within a Bayesian framework. However, these methods are not widely spread among practitioners, are time-consuming and difficult to reproduce. This chapter proposes an approach based on a pseudolikelihood approach, also known as composite likelihood (Lindsay, 1988; Cox and Reid, 2004; Bellio and Varin, 2005; Fieuws and Verbeke, 2006), to reduce the complex high-dimensional likelihood function arising from modelling multiple correlated occasionally-consumed food intakes jointly.

The composite likelihood (CL) approach arises from the asymptotic theory of estimating equations and misspecified likelihoods (White, 1982) with the method of generalised estimating equations developed for longitudinal data being a prime example of its successful application (Liang and Zeger, 1986) in medical research. CL assumes independence for individual conditional or marginal likelihood contributions which are pulled together to form a full composite likelihood, in cases where optimisation of the original full likelihood function is not feasible. The key properties of the method are, generally, its consistency and asymptotic normality of the estimated model parameters (White, 1982). The area of research on composite likelihood is rapidly developing due to its potential to reduce the complexity of estimation associated with high dimensions (Geys et al., 1999; Bellio and Varin, 2005; Fieuws and Verbeke, 2006; Varin et al., 2011). Recent applications can be found in the areas of spacial processes (Caragea and Smith, 2007), health care (Ivanova

et al., 2017), questionnaire data (Fieuws et al., 2006), statistical genetics (Larribe and Fearnhead, 2011), chemical exposures (Zhang et al., 2016) and meta-analysis (Chen et al., 2015) to name a few. For the most recent extensive review of the subject please refer to Varin et al. (2011).

The focus of the analysis presented here is on modelling alcohol intake, considering the intake of other dietary components jointly. The analysis is based on dietary data from the UK National Diet and Nutrition Survey Rolling Programme (NDNS RP) and it draws from the work developed in previous chapters. This work is, to the best of the author's knowledge, the first analysis of complex nutritional data utilising a composite likelihood approach to provide more robust inference of the determinants of alcohol intake in a representative sample drawn from the population of adults residing in United Kingdom (UK).

4.2 Methods

4.2.1 The UK National Diet and Nutrition Survey Rolling Programme Data

The NDNS RP is an annual cross-sectional survey, jointly funded by Public Health England and the UK Food Standards Agency, undertaken since 2008. The survey aims to assess diet, nutrient intake and nutritional status of the UK population aged 18 months and older, living in private households. A nationally representative sample of the UK households is selected by a multistage sampling procedure: firstly, the Postcode Address File (PAF), which contains all the addresses in the UK, is accessed to sample Primary Sampling Units (PSUs). These are small geographical areas formed by neighbouring postcodes. Secondly, twenty seven addresses are sampled from a selected PSU at random, where either an adult or a child is selected. The 27 addresses are randomly allocated to one of two groups to determine whether an adult (aged 19 years or over) and a child (aged 1.5 to 18 years), or a child only, are selected for interview. At nine of the selected

addresses the interviewer selects one adult and, where present, one child for inclusion in the survey. The remaining 18 addresses form a "child boost" where only households with children are selected. Where more than one person is eligible the participants are selected using a random selection procedure.

The collection included demographics, an estimated four-day food diary, life-style factors and socio-economic measurements. Diary response rate was 56%. Further details can be found elsewhere (Public Health England, 2014).

Demographics, socio-economic and life-style factors

The following patient characteristics were available as potential determinants of alcohol intake.

Demographic factors comprised age; sex; ethnicity (white; non-white); body mass index (BMI); lipid and blood pressure lowering medications taken (yes; no); self-assessed general health problems (no; yes, but no impact on mobility; yes, with impact on mobility); partner status (married or in partnership; never married or lived in partnership; previously married or lived in partnership but now single).

Life-style factors comprised take away shopping habits (rarely or never; once or twice per month; every week or more often); fruits and vegetables shopping habits (less than weekly; weekly and more often); being a non-meat-eater (yes; no); smoking (never smoked; quit >10 years ago; quit $\leq$ 10 years ago; current smoker; occasional smoker); alcohol consumption (never drink; rarely drink; the rest); moderate-to-vigorous physical activity (MVPA) assessed through a recent self-completed Physical Activity Questionnaire, expressed in min per day and combined from four domains: home, commuting, work and leisure (0 min ; 0–10 min; 10–20 min; 20–40 min; 40–60 min; $\geq$ 60 min) (Mindell, 2014).

Socio-economic factors comprised education as the highest obtained degree (bachelor degree and above; unfinished degree; current student; A levels; GCSE grades A-C; GCSE grades below C or no qualifications; foreign degree); socio-economic status (never

worked and others; routine; semi-routine; lower supervisory and technical; small employers and own account; intermediate; lower managerial and professional; higher managerial and professional); income over the previous 12 months as assessed through self-report and equalised to take into account the household composition by a rescaled version of the Organisation for Economic Development modified equivalence scale (Anyaegbu, 2010) (<£15,000; £15,000–£25,000; £25,000–£35,000; £35,000–£50,000; $\geq$ £50,000). For some people (71 (14$\%$) for men and 85 (13.8$\%$) for women) information was not collected so they were assigned into a separate category; tenure (own or mortgage; renting privately; renting from local authority).

4.2.2 Extension of the two-part model to model intake of two occasionally-consumed foods

Chapter 2 introduced a two-part mixed-effects model (Olsen and Schafer, 2001; Tooze et al., 2002) to model the individual intake of one occasionally-consumed food. This section extends the model for the joint modelling of the individual intake of two occasionally-consumed foods.

To extend the notation introduced in previous chapters, we use the sub-indeces h to denote each food, $h = 1, 2$, i for each individual, $i = 1, \ldots, m$ and j for day on which consumption took place $j = 1, \ldots n_{hi}$. The data consist of two parts: the occurrence of food consumption (yes/no), which can be coded as an indicator variable I_{hij} such that:

$$I_{hij} = \begin{cases} 1, & \text{if the food } h \text{ is consumed by person } i \text{ on day } j \\ 0, & \text{otherwise} \end{cases}$$

and the amount of food consumed if consumption took place, which we record as A_{hij}, with $A_{hij} > 0$ if $I_{hij} = 1$.

Natural heterogeneity arise among subjects due to personal preferences for consumption. We denote unobservable person-specific information related to propensity to consume certain foods as v_{hi} and unobservable person-specific information related to amount con-

sumed on consumption day as u_{hi}. Then, conditionally on v_{hi} and u_{hi}, responses I_{hij} and A_{hij} are independent. The indicator variable I_{hij} is assumed to follow a Bernoulli distribution with probability $p_{hij} = \Pr(I_{hij} = 1|v_{hi})$, and to allow for skewness, we assume $A_{hij}|A_{hij} > 0$ to be log-normally distributed. We assume a logistic regression model for I_{hij}

$$\text{logit}\{\Pr(I_{hij} = 1|v_{hi})\} = x_{hij}^{'}\gamma_h + v_{hi}$$

where $x_{hij}^{'}$ is the vector of relevant covariates, relating individual characteristics to propensity for food intake, and γ_h is the vector of corresponding regression coefficients. Further, under the assumption that $\log(A_{hij}|A_{hij} > 0) = Y_{hij}$ is approximately normal, we specify

$$Y_{hij} = x_{hij}^{'}\beta_h + u_{hi} + \epsilon_{hij},$$

where $\mathrm{E}(Y_{hij}|u_{hi}) = x_{hij}^{'}\beta_h + u_{hi}$ and $\mathrm{Var}(Y_{hij}|u_{hi}) = \sigma^2_{\epsilon_h}$ (within-person daily variation); $x_{hij}^{'}$ is the vector of relevant covariates relating individual characteristics to the amount of food consumed, β_h is the vector of corresponding regression coefficients.

The potential correlation between the *probability* and *amount* parts of both foods is linked through person-specific effects u_{hi} and v_{hi}, which are assumed to have a joint four-dimensional normal distribution with means 0 and variance-covariance matrix Σ. These are called random effects and are assumed to be independent of ϵ_{hij}.

The unknown model parameters $\boldsymbol{\theta} = (\gamma_h, \beta_h, \Sigma, \sigma_{\epsilon_h^2})$ can be estimated through maximising the full marginal likelihood function, where we utilise the conditional independence of responses I_{hij} and Y_{hij} and their distributional assumptions. Because the random effects u_{hi} and v_{hi} are unobserved, we take expectation of the above function over the random effects, so that the expected full marginal likelihood function becomes:

$$L(\theta) \propto \prod_{i=1}^{m} \iiiint \prod_{j=1}^{n_{1,i}} f_{I_1}(I_{1ij} \mid v_{1i}, \theta) f_{Y_1}(Y_{1ij} \mid u_{1i}, \theta) \prod_{j=1}^{n_{2,i}} f_{I_2}(I_{2ij} \mid v_{2i}, \theta) f_{Y_2}(Y_{2ij} \mid u_{2i}, \theta)$$
$$f_{U_1V_1U_2V_2}(u_{1i}, v_{1i}, u_{2i}, v_{2i} \mid \theta) du_{1i} dv_{1i} du_{2i} dv_{2i} \quad (4.1)$$

where f_{I_h}, f_{Y_h} and $f_{U_1V_1U_2V_2}$ denote the density functions of the Bernoulli, normal and multivariate normal distributions, respectively. The likelihood function does not have a closed form and needs to be evaluated numerically.

It can be seen that as the number of foods increases, the numerical integration computation becomes more challenging to evaluate.

4.2.3 Pseudolikelihood

The pseudolikelihood approach has been previously used to reduce the complexity of the likelihood function of mixed-effect models involving multiple correlated random effects (Geys et al., 1999; Bellio and Varin, 2005; Fieuws and Verbeke, 2006; Varin et al., 2011). We denote the loglikelihood contribution of individual i to the full loglikelihood by $\ell_i(\boldsymbol{\theta}; \boldsymbol{Y_{1i}}, \boldsymbol{Y_{2i}}, \ldots, \boldsymbol{Y_{qi}}, \boldsymbol{I_{1i}}, \boldsymbol{I_{2i}}, \ldots, \boldsymbol{I_{qi}})$, the pseudolikelihood approach suggests that instead of maximising the full likelihood function, we maximise likelihoods built on simpler models which do not involve more than four random effects simultaneously and, typically, are bivariate, so called *pairwise likelihoods*, so that the individual pairwise loglikelihood contribution becomes $\ell_{l,r,i}(\boldsymbol{\theta_{l,r}}; \boldsymbol{Y_{li}}, \boldsymbol{I_{ri}})$ where, $l = 1, \ldots, q-1, r = l+1, \ldots, q$ and $\boldsymbol{\theta_{l,r}}$ is the vector of parameters of the corresponding reduced simpler model. Then the following loglikelihood contributions can be maximised separately

$$\ell_{l,r} = \sum_{i=1}^{m} \ell_{l,r,i}(\boldsymbol{\theta_{l,r}} \boldsymbol{Y_{li}}, \boldsymbol{I_{ri}}),$$

and the log-pseudolikelihood function can then be obtained as the sum of the above $q(q-1)/2$ pairwise loglikelihoods. The estimates obtained from maximising pairwise likelihoods have properties grounded in the theory of estimating equations and are, generally, consistent and approximately normal provided the full joint model is correct (Cox and Reid, 2004). This follows from considering the estimating equations $U_{l,r}(\boldsymbol{\theta}; \boldsymbol{Y}) = \partial \ell_{l,r}(\boldsymbol{\theta}; \boldsymbol{Y}) / \partial \boldsymbol{\theta} = 0$, which are the first derivatives of $\ell_{l,r}(\boldsymbol{\theta}; \boldsymbol{Y})$ with respect to $\boldsymbol{\theta}$ and which are unbiased for $\boldsymbol{\theta}$. For some elements of $\boldsymbol{\theta}$ there will be more than one estimate coming from pairwise log likelihoods. In this case, using the Central Limit Theorem, the available estimates are averaged to get a single estimate of the model parameter.

Confidence intervals for model parameters based on pseudolikelihood

The distribution of the estimates of $\boldsymbol{\theta}$ obtained from maximising the pseudolikelihood differs from the distribution of estimates obtained from maximising the full likelihood; there-

fore, the Fisher information matrix cannot longer be used to estimate their variance. Instead, the variance of estimates is estimated by the inverse of the Godambe information matrix, $G(\boldsymbol{\theta})$, defined by $G(\boldsymbol{\theta}) = H(\boldsymbol{\theta})J(\boldsymbol{\theta})^{-1}H(\boldsymbol{\theta})$, where $H(\boldsymbol{\theta}) = \mathrm{E}(-\partial U(\boldsymbol{\theta};\boldsymbol{Y})/\partial\boldsymbol{\theta})$ and $J(\boldsymbol{\theta}) = \mathrm{E}(U(\boldsymbol{\theta};\boldsymbol{Y})U(\boldsymbol{\theta};\boldsymbol{Y})^T)$. The evaluation of this matrix is complex due to the joint structure of the random effects (Bellio and Varin, 2005; Varin et al., 2011). However, Aerts and Claeskens (1999) showed that parametric bootstrap will provides consistent estimates of confidence intervals for model parameters estimated from a log pseudolikelihood.

To apply the method of parametric bootstrap, P bootstrap dataset samples from a fully specified likelihood model where the true model parameters are replaced by our estimates obtained from maximising the pseudolikelihoods are generated. Each bootstrap sample is then analysed by applying pseudolikelihoods and, for each bootstrap sample, a corresponding maximum pseudolikelihood estimate of $\boldsymbol{\theta}$ is obtained. The procedure results in a set of P pseudolikelihood estimates, the sample variance of which is a bootstrap variance of $\boldsymbol{\theta}$ (Zhang et al., 2016).

The two-part models (pairwise likelihoods) from Chapter 3 set the basis for the inference on the predictors of alcohol intake. To maximise the utilisation of available information, and, simultaneously, whilst keeping the computational aspect of data analysis at a reasonable level of complexity, this work analyses the data in triple-wise likelihoods so that each simpler likelihood contains a model specifying the probability of alcohol intake, the amount of alcohol consumed if consumption takes place (i.e. portion size) and another model part for occasionally-consumed food intake (probability or amount part). This approach ensures that the parameters of the model for alcohol intake, probability and amount parts, are always analysed together, so that, the underlying correlated random effects related to alcohol intake are always fitted together. On top of that, they are also always adjusted for one more potential correlation coming from a random effect of another model, probability or amount, representing another occasionally-consumed food intake distribution.

More precisely, for each sub-population, male or female, twelve models are fitted where each model contains the part which models the probability of alcohol consumption, the part which models the portion size of alcohol consumed and a part which models either the probability of consumption of one of the following foods: fruits, cooked vegetables, raw vegetables, processed meat, oily fish and sugary drinks or the portion size of the corresponding occasionally-consumed foods. The results are stored and the model parameters, which appear in the pseudolikelihoods more than once, are averaged to obtain a single estimate. The resulting covariance matrix of the fourteen random effects is constructed and, if necessary, converted to the nearest positive-semi-definite matrix to have the attributes of variance-covariance matrices. This allows to specify the full likelihood function which consists of fourteen components describing the consumption probability and the portion size of the specified occasionally-consumed food intake distributions, via the combination of fixed effect parts with available observed food intake predictors and random effect parts. To build one bootstrap sample the actual available data are utilised so that for each person with observed food intake predictors such as age, gender, social and lifestyle characteristics, fourteen possible intakes are generated based on the available information on predictors, fixed effect model parameters estimates (obtained at the previous step from pseudolikelihood) and on person-specific effects, simulated from a fourteen-dimensional multivariate normal distribution with variance-covariance matrix obtained at the previous step. Each of the generated bootstrap samples is then analysed in the same manner as the original data by applying the psuedolikelihood approach to 12 triple-likelihoods and the results of model parameters point-estimates are stored. The results presented in this chapter are based on 10 bootstrap samples applied to the full likelihood, resulting in 120 point estimates of each vector of model parameters for alcohol intake. These 120 point estimates form the sample distribution of the model parameters and the corresponding ranges of values lying within the lower and the upper 5th percentiles represent 95% bootstrap-constructed confidence intervals to draw statistical inference for the model parameters of alcohol intake distribution.

4.2.4 Comparison of the pseudolikelihood approach with multivariable linear regression

We compared the proposed modern methodology to the traditional approach in nutritional epidemiology in which multivariable linear regression is applied to individual averages of alcohol intake as outcome. Individual averages are calculated as the average of all available daily alcohol intake per person. To correct for skewness and zero records, the individual average alcohol intake was transformed by applying natural logarithm to individual average increased by 1. The selection of risk factors when applying multivariable regression follows a similar process as that utilised in Chapter 3 for the two-part model. First, a model with the full set of predictors was fitted. Then, in a step-down manner, each predictor was tested for significance. In the first run, the predictors with p-values above 0.20 were removed and the process repeated. In the second and consequent runs, the variables were left in the model if the corresponding p-values were less than 0.10. All models were adjusted for age, survey year and weekend (including Friday) and excluded non-drinkers. Robust (Sandwich-Huber-White) standard errors were computed to correct for potential model mis-specification. An F-test was utilised to test the significance of factor variables and a Wald test was utilised to test significance of continuous variables.

Stata 14 software was utilised to analyse the data.

4.3 Results

Table 4.1 displays personal, lifestyle and socio-demographic characteristics of the sample population selected for analysis from the NDNS RP Years 2-4.

Table 4.1: Personal, lifestyle and socio-demographic characteristics of the sample population selected for analysis from the NDNS RP Years 2-4

	Males (N = 509)		Females (N = 618)	
	N (mean)	% (SD)	N (mean)	% (SD)
Age, years (mean (SD))	(48.3)	(17.5)	(48.2)	(18.1)
BMI (mean(SD))	(27.4)	(4.5)	(26.7)	(5.7)
Survey Year				
2009/2010	170	33.4	202	32.7
2010/2011	146	28.7	178	28.8
2011/2012	193	37.9	238	38.5
Ethnicity				
white	471	92.5	571	92.4
non-white	38	7.5	47	7.6
Health, self-reported				
no health problems	341	67.0	381	61.7
health problems, no mobility restrictions	86	16.9	110	17.8
health problems, mobility restrictions	82	16.1	127	20.6
Lipid lowering drug taken				
no	451	88.6	560	90.6
yes	58	11.4	58	9.4
Blood pressure lowering drug taken				
no	454	89.2	544	88.0
yes	55	10.8	74	12.0
Partner				
married or live in partnership	273	53.6	266	43.0
never married or lived in partnership	148	29.1	168	27.2
previously married or lived in partnership but now single	88	17.3	184	29.8
Lifestyle				
Smoking				
never smoked	257	50.5	355	57.4
ex-smoker, quit >10 year ago	77	15.1	73	11.8
ex-smoker, quit <=10 years ago	46	9.0	65	10.5
current	115	22.6	116	18.8
occasional	14	2.8	9	1.5

Table 4.1 Personal, lifestyle and socio-demographic characteristics of the sample population selected for analysis from the NDNS RP Years 2-4 (Continued)

	Males (N = 509)		Females (N = 618)	
	N (mean)	% (SD)	N (mean)	% (SD)
Moderate to vigorous physical activity, min/day				
0	14	2.8	15	2.4
0-10	62	12.2	138	22.3
10-20	33	6.5	76	12.3
20-40	71	14.0	116	18.8
40-60	65	12.8	70	11.3
>60	264	51.9	203	32.9
Drinking				
yes	444	87.2	524	84.8
rarely	28	5.5	37	6.0
never	37	7.3	57	9.2
Fruits and vegetables buying habits				
weekly or more often	472	92.7	575	93.0
less often than weekly	37	7.3	43	7.0
Non-meat eaters	9	1.8	23	3.7
Take away habit				
rarely or never	200	39.3	291	47.1
less than once a week	198	38.9	204	33.0
once a week and more	111	21.8	123	19.9
Socio-economic				
Tenure				
mortgaged or owned	355	69.7	433	70.1
rented privately	90	17.7	89	14.4
rented from local authority	64	12.6	96	15.5
Qualifications				
bachelor degree and above	133	26.1	136	22.0
unfinished degree	50	9.8	64	10.4
students	20	3.9	32	5.2
A levels	83	16.3	98	15.9
GCSE A_C	86	16.9	114	18.5
GCSE below C and no qualifications	110	21.6	146	23.6
foreign qualifications	27	5.3	28	4.5

Table 4.1 Personal, lifestyle and socio-demographic characteristics of the sample population selected for analysis from the NDNS RP Years 2-4 (Continued)

	Males (N = 509)		Females (N = 618)	
	N (mean)	% (SD)	N (mean)	% (SD)
Social Status				
higher managerial and professional occupation	104	20.4	86	13.9
lower managerial and professional occupation	134	26.3	174	28.2
intermediate occupations	45	8.8	63	10.2
small employers and own account workers	50	9.8	74	12.0
lower supervisory and technical occupation	49	9.6	47	7.6
semi-routine occupations	60	11.8	93	15.1
routine occupations	56	11.0	58	9.4
never worked or other	11	2.2	23	3.7
Equalised household income, £ 1000				
<=15	70	13.8	134	21.7
15-25	99	19.5	124	20.1
25-35	95	18.7	110	17.8
35-50	79	15.5	87	14.1
>50	95	18.7	78	12.6
missing	71	14.0	85	13.8

Table 4.2 and Table Table 4.3 present the results of applying the pseudolikelihood approach to the estimation of alcohol intake predictors in male and female sub-populations of NDNS RP correspondingly. The 95 $\%$ confidence intervals of model parameters were obtained by applying the parametric bootstrap method described in the previous section.

Table 4.2: Predictors of alcohol intake in the sample population of men selected for analysis from the NDNS RP Years 2-4, based on the pseudolkelihood approach

	Subject specific regression coefficient from two-part model	
Predictors	Odds Ratio (95%CI) (N = 1882)	Portion size Relative Change (95% CI) (N = 778)
Weekend[1]	3.59 (2.54, 4.45)	1.33 (1.27, 1.44)
Survey Year (base: 2009/2010)		
2010/2011	0.64 (0.30, 1.37)	1.25 (1.02, 1.45)
2011/2012	0.78 (0.41, 1.18)	1.00 (0.86, 1.11)
Age, years	1.04 (1.01, 1.06)	0.992 (0.987, 0.998)
Ethnicity (non-white)	0.29 (0.13, 0.75)	
Partner (base: married)		
never married		1.11 (0.97, 1.30)
previously married		1.40 (1.13, 1.59)
Smoking (base: never smoked)		
ex, quit >10 year ago	1.64 (0.86, 3.40)	0.99 (0.81, 1.13)
ex, quit <=10 years ago	1.70 (0.59, 6.07)	1.40 (1.16, 1.64)
current	2.21 (1.15, 6.05)	1.39 (1.13, 1.54)
occasional	1.33 (0.14, 4.20)	2.15 (1.37, 2.82)
Alcohol habit consumption (base: regular)		
rarely	0.04 (0.00, 0.26)	0.09 (0.07, 0.12)
never	NA	NA
Take away (base: rarely or never)		
occasionally		1.04 (0.96, 1.22)
regularly		1.30 (1.16, 1.51)

Table 4.2 Predictors of alcohol intake in the sample population of men selected for analysis from the NDNS RP Years 2-4, based on the pseudolkelihood approach (Continued)

Tenure (base: mortgaged or owned)		
privately rented	0.52	
	(0.21, 1.14)	
local authority rented	1.15	
	(0.37, 2.46)	
McClement equivalence score, £ 1000		
<=15	0.46	
	(0.23, 0.99)	
15-25	0.39	
	(0.16, 0.62)	
(base) 25-35	1	
35-50	0.90	
	(0.39, 1.86)	
>50	0.86	
	(0.50, 1.92)	

[1]Weekend includes Friday.

Table 4.3: Predictors of alcohol intake in the sample population of women selected for analysis from the NDNS RP Years 2-4, based on the pseudolkelihood approach

	Subject specific regression coefficient from two-part model	
Predictors	Odds Ratio (95%CI) (N = 2234)	Portion size Relative Change (95% CI) (N = 728)
Weekend[1]	3.11 (2.38, 3.86)	1.35 (1.34, 1.40)
Survey Year		
(base) 2009/2010	1	1
2010/2011	0.83 (0.59, 1.18)	1.11 (1.07, 1.19)
2011/2012	0.82 (0.58, 1.41)	1.20 (1.15, 1.31)
Age, years	1.027 (1.020, 1.040)	0.988 (0.987, 0.990)
BMI, kg/m^2		1.017 (1.015, 1.022)
Lipid lowering medicine	0.54 (0.26, 0.99)	0.74 (0.68, 0.86)
Partner		
(base) married		1
never married		1.09 (1.08, 1.15)
previously married		1.25 (1.18, 1.43)
Alcohol habit consumption		
(base) regular	1	1
rarely	0.09 (0.00, 0.17)	0.26 (0.24, 0.33)
never	N/A	N/A
Take away shopping		
(base) rarely or never	1	1
occasionally/regularly	1.69 (1.06, 2.56)	1.15 (1.11, 1.23)

Table 4.3 Predictors of alcohol intake in the sample population of women selected for analysis from the NDNS RP Years 2-4, based on the pseudolkelihood approach (Continued)

Predictors	Odds Ratio (95%CI)	Portion size Relative Change (95% CI)
Qualifications		
(base) bachelor degree and above	1	
unfinished degree	1.26 (0.52, 3.47)	
current students	0.40 (0.17, 0.85)	
A levels	0.58 (0.31, 1.28)	
GCSE A_C	0.70 (0.45, 1.65)	
GCSE below C and no qualifications	0.64 (0.34, 1.41)	
foreign qualifications	0.24 (0.05, 0.76)	
Socio-economic status		
(base) higher managerial and professional occupation	1	
lower managerial and professional occupation	2.00 (1.19, 3.10)	
Intermediate occupations	1.15 (0.59, 1.73)	
Small employers and own account workers	1.22 (0.73, 1.74)	
Lower supervisory and technical occupation	0.60 (0.32, 1.04)	
Semi-routine occupations	0.71 (0.37, 1.29)	
Routine occupations	0.51 (0.26, 0.95)	
Never worked and Other	0.63 (0.16, 1.26)	

[1]Weekend includes Friday.

Overall, the point estimates of model parameters obtained by fitting only pairwise likelihoods presented in Chapter 3 (Tables 3.5 and 3.6) are in line with the results obtained by fitting triple pseudolikelihoods and parametric bootstrap presented here. However, the bootstrap confidence intervals change the inference for some model parameters. For example, in the male sub-sample, the effect of renting privately on alcohol intake diminishes when confidence intervals are estimated with parametric bootstrap compared to pairwise likelihood based inference. In contrast, for some variables, like smoking history, the inference became even stronger when analysed by pseudolikelihood and parametric bootstrap. Theoretically, the difference between inference based on pairwise likelihoods and pseudolikelihood might depend on how similar the components of the Fisher information matrix and the Godambe information matrix are. In our particular case of alcohol intake predictors, the conclusions drawn from the analysis by triple pseudolikelihood and parametric bootstrap are similar to those drawn from the reduced pairwise likelihoods of the two-part model. All predictors, except tenure, remain important for the estimation of the alcohol intake distribution. The conclusions are similar for the female sub-population.

It is interesting to contrast the results from the two-part model approach and pseudolikelihood approach with the results obtained by applying the traditional multivariable linear regression. When estimating alcohol consumption utilising individual averages, the unadjusted distribution of alcohol intake in men had mean of 18.3g, median of 9.3g, interquartile range (0.0g - 29.6g), standard deviation of 23.8g, minimum of 0.0g, maximum of 149.4 g. In women, the alcohol intake distribution had a mean of 9.9 g, median of 2.9 g, interquartile range (0.0g - 15.0g), standard deviation of 15.4g, minimum of 0.0g and maximum of 96.7g. This under-estimates the median consumption both, in men and in women, by around 3 g daily, which is a bit more than a third of an alcohol portion (Department of Health, UK, 2016), compared to when the correlation between portion size and consumption probability is taken into account as shown in Table 3.3.

Interestingly, some predictors, for example, being married or having a partner, which

showed significant association when analysed with two-part model, became no longer significant in the results of analysis with linear regression. Conversely, income showed no association in two-part model but came out as significant in linear regression analysis. This highlights the importance of potential model misspecification with respect to inference on predictors of interest even when potential adjustment for confounding is available and carried out.

The results of full adjustments when multivariate linear regression tool is applied are shown in Table 4.4 for women and in Table 4.5 for men and presented as a change relative to a median portion.

Table 4.4: Predictors of alcohol intake in the sample population of women selected for analysis from the NDNS RP Years 2-4, based on multivariable linear regression

Predictors	Relative change	95%CI		*p-value F test*	*p-value Wald test*
Weekend	1.23	1.07	1.39		*0.002*
Survey Year (base: 2009/2010)					
2010/2011	0.86	0.71	1.00		*0.069*
2011/2012	1.07	0.90	1.23		*0.428*
Age, y	1.01	1.00	1.01		*0.042*
Lipid lowering drug	0.54	0.41	0.67		*0.000*
Alcohol consumption (base: regular) (rarely)	0.22	0.17	0.28		*0.000*
Smoking (base: never)				*0.003*	
ex, quit >10 year ago	1.04	0.82	1.26		*0.709*
ex, quit <=10 years ago	1.20	0.95	1.46		*0.095*
Current	1.37	1.12	1.63		*0.001*
Occasional	1.76	0.88	2.64		*0.032*
MV Physical activity, min/day (base: 0)				*0.005*	
0-10	1.35	0.88	1.82		*0.100*
10-20	1.11	0.69	1.54		*0.581*
20-40	1.26	0.81	1.72		*0.212*
40-60	1.28	0.81	1.75		*0.194*
>60	1.59	1.05	2.13		*0.009*
Take away shopping				*<0.001*	
Occasionally	1.33	1.12	1.54		*0.000*
	1.39	1.12	1.65		*0.001*
BMI, kg/m^2	1.01	1.00	1.02		*0.078*

Table 4.4 Predictors of alcohol intake in the sample population of women selected for analysis from the NDNS RP Years 2-4, based on multivariable linear regression (Continued)

Qualifications (base: bachelor degree and above				*<0.001*	
Unfinished degree	1.13	0.84	1.43		*0.355*
Current student	0.60	0.38	0.82		*0.003*
A levels	0.71	0.54	0.88		*0.004*
GCSE A_C	0.84	0.65	1.02		*0.114*
GCSE below C and no qualifications	0.82	0.63	1.01		*0.084*
foreign qualifications	0.46	0.27	0.64		*0.000*
Socio-economic status (base: higher managerial and professional)				*<0.001*	
Lower managerial	1.43	1.12	1.73		*0.001*
Intermediate occupation	1.33	0.97	1.69		*0.045*
Small employers	1.20	0.87	1.52		*0.200*
Lower supervisory	0.65	0.42	0.88		*0.013*
Semi-routine occupation	0.81	0.58	1.04		*0.133*
Routine occupation	0.67	0.44	0.89		*0.014*
Never worked and other	0.82	0.48	1.17		*0.360*
McClement equivalence score, £ 1000				*<0.001*	
<=15	1.14	0.90	1.39		*0.227*
15-25	1.27	1.00	1.53		*0.028*
Base (25-35)					
35-50	1.43	1.11	1.74		*0.002*
>50	2.28	1.76	2.81		*0.000*

Table 4.5: Predictors of alcohol intake in the sample population of men selected for analysis from the NDNS RP Years 2-4, based on multivariable linear regression

Predictors	Relative change	95%CI		*p-value F test*	*p-value Wald test*
Weekend	1.17	1.00	1.33		*0.033*
Survey Year (base: 2009/2010)					
2010/2011	0.94	0.76	1.12		*0.507*
2011/2012	0.97	0.81	1.14		*0.735*
Age, y	1.01	1.00	1.02		*0.002*
Ethnicity (non-white)	0.41	0.27	0.55		*<0.001*
Partner (base: married)				*0.012*	
never married	0.88	0.69	1.08		*0.267*
previously married	1.31	1.03	1.60		*0.014*
Blood pressure lowering drug (yes)	0.62	0.46	0.78		*<0.001*
Alcohol consumption (base: regular) (rarely)	0.08	0.05	0.10		*<0.001*
Smoking (base: never)				*<0.001*	
ex, quit >10 year ago	1.31	1.04	1.59		*0.012*
ex, quit <=10 years ago	1.57	1.17	1.97		*0.001*
Current	1.77	1.42	2.12		*<0.001*
Occasional	1.82	1.13	2.51		*0.002*
MV Physical activity, min/day (base: 0)				*0.035*	
0-10	1.56	0.71	2.41		*0.112*
10-20	2.03	0.82	3.24		*0.022*
20-40	1.74	0.82	2.65		*0.042*
40-60	2.23	1.04	3.42		*0.004*
>60	1.85	0.91	2.79		*0.020*

Table 4.5 Predictors of alcohol intake in the sample population of men selected for analysis from the NDNS RP Years 2-4, based on multivariable linear regression (Continued)

Take away shopping (base: rarely or never)				*<0.001*	
occasionally	1.39	1.15	1.63		*<0.001*
regularly	1.64	1.30	1.98		*<0.001*
Non-meat eaters	1.68	1.39	1.98		*<0.001*
BMI, kg/m^2	1.04	1.02	1.05		*<0.001*
Health, self-reported (base: no problems)				*0.005*	
Health problems, no mobility restrictions	1.36	1.10	1.62		*0.002*
Health problems, mobility restrictions	1.21	0.96	1.46		*0.075*
Tenure (base: mortgaged or owned)				*<0.001*	
privately rented	0.57	0.44	0.69		*<0.001*
local authority rented	0.95	0.70	1.21		*0.728*
Qualifications (base: bachelor degree and above				*<0.001*	
Unfinished degree	0.56	0.40	0.72		*<0.001*
Current student	1.24	0.67	1.81		*0.366*
A levels	0.92	0.70	1.14		*0.478*
GCSE A_C	0.67	0.51	0.82		*0.001*
GCSE below C and no qualifications	0.54	0.39	0.69		*<0.001*
foreign qualifications	0.79	0.51	1.08		*0.198*
Socio-economic status (base: higher managerial and professional)				*<0.001*	
Lower managerial	1.15	0.91	1.38		*0.188*
Intermediate occupation	1.20	0.83	1.57		*0.250*
Small employers	1.66	1.20	2.12		*<0.001*
Lower supervisory	0.60	0.40	0.80		*0.002*
Semi-routine occupation	1.29	0.90	1.68		*0.107*
Routine occupation	1.64	1.11	2.18		*0.003*
Never worked and other	1.73	0.85	2.60		*0.037*

Table 4.5 Predictors of alcohol intake in the sample population of men selected for analysis from the NDNS RP Years 2-4, based on multivariable linear regression (Continued)

McClement equivalence score, £ 1000				*<0.001*	
<=15	0.73	0.51	0.94		*0.033*
15-25	0.78	0.60	0.97		*0.045*
Base (25-35)					
35-50	0.78	0.60	0.97		*0.526*
>50	1.34	1.03	1.66		*0.015*

It is also worth noting that the residual correlation between model parts presented in Tables 4.6 and 4.7, generally, is not very big. It is possible that the robustness of the results based on fitting only a two-part model to estimate alcohol intake, can be partially attributed to the extensive specification of the fixed part of the two-part model, which explicitly modelled and explained some person-specific preferences, thus reducing unexplained variation. This, in turn, emphasises the importance of collecting and analysing all the available information, including confounders, that can be relevant and important to the estimation of food intake. It is worth noting that the two-part model might potentially account for the biggest source of correlation (after the fixed effect are taken into account) between the probability of consumption and portion size so that the remaining correlation might have less impact over the estimation of model parameters.

Table 4.6: Estimated correlation structure of random effects in the male sub sample of the NDNS RP Years 2-4

		Alcohol		Fruits		Veg Cooked		Veg Raw		Processed Meat		Oily Fish		Soft regular drinks	
		Pr[1]	Am[2]	Pr[1]	Am[2]	Pr[1]	Am[2]	Pr[1]	Am[2]	Pr[1]	Am[2]	Pr[1]	Am[2]	Pr[1]	Am[2]
Alcohol	Pr	1	0.37	-0.09	0.01	0.08	0.05	0.15	0.07	0.03	-0.13	0.07	-0.18	-0.05	-0.04
	Am		1	-0.30	0.02	-0.06	0.09	-0.07	0.00	-0.04	0.28	-0.12	0.02	-0.10	0.22
Fruits	Pr			1	0.59	0.22	0.06	0.32	0.16	-0.12	-0.13	0.19	0.13	-0.10	-0.10
	Am				1	-0.03	0.00	0.18	0.23	-0.15	0.03	0.36	0.07	-0.12	-0.16
Veg Cooked	Pr					1	0.35	-0.05	-0.15	-0.04	0.39	0.37	-0.09	-0.08	-0.12
	Am						1	-0.09	-0.14	-0.21	-0.09	0.08	0.08	-0.10	0.00
Veg Raw	Pr							1	0.52	0.11	-0.06	0.60	-0.46	0.01	-0.14
	Am								1	0.07	0.15	0.14	-0.04	-0.18	-0.22
Processed	Pr									1	0.03	-0.07	-0.35	0.18	0.01
Meat	Am										1	0.09	-0.13	0.09	0.18
Oily Fish	Pr											1	-0.54	-0.27	-0.05
	Am												1	-0.10	0.10
Soft regular	Pr													1	0.41
drinks	Am														1

[1]Pr stands for Probability part of the two-part model, [2]Am stands for Amount part of the two-part model

Table 4.7: Estimated correlation structure of random effects in the female sub sample of the NDNS RP Years 2-4

		Alcohol		Fruits		Veg Cooked		Veg Raw		Processed Meat		Oily Fish		Soft regular drinks	
		Pr[1]	Am[2]	Pr[1]	Am[2]	Pr[1]	Am[2]	Pr[1]	Am[2]	Pr[1]	Am[2]	Pr[1]	Am[2]	Pr[1]	Am[2]
Alcohol	Pr	1	0.31	-0.16	-0.09	0.06	-0.17	0.17	0.08	0.12	0.13	0.30	-0.89	0.02	-0.07
	Am		1	-0.24	-0.17	-0.11	-0.26	0.29	0.33	0.14	0.06	0.13	0.03	0.03	0.18
Fruits	Pr			1	0.40	0.32	-0.07	0.35	0.10	-0.06	-0.14	0.51	0.40	0.08	-0.10
	Am				1	0.34	0.19	0.16	0.29	-0.28	0.20	0.22	0.20	-0.06	-0.08
Veg Cooked	Pr					1	0.30	0.09	0.27	-0.03	0.17	0.59	0.51	0.05	-0.08
	Am						1	-0.49	0.36	-0.36	0.39	-0.57	0.53	-0.11	0.10
Veg Raw	Pr							1	0.07	0.19	-0.22	0.64	-0.97	0.02	0.09
	Am								1	-0.21	0.13	0.20	0.54	0.11	-0.12
Processed	Pr									1	0.01	-0.60	N/A[3]	0.17	0.03
Meat	Am										1	0.27	-0.12	-0.02	0.24
Oily Fish	Pr											1	-0.98	0.01	0.04
	Am												1	-0.15	0.13
Soft regular	Pr													1	0.44
drinks	Am														1

[1]Pr stands for Probability part of the two-part model, [2]Am stands for Amount part of the two-part model, [3] – estimate not available as the model did not achieve convergence

4.4 Discussion

This work, to the author's knowledge, is the first to utilise a novel pseudolikelihood approach and parametric bootstrap for the analysis of multiple occasionally-consumed food intakes distributions, taking into account measurement error, excess zeros and correlation between consumption frequency and portion sizes across various multiple foods. It shows that complex multivariate analysis of multiple correlated occasionally-consumed food intakes can be performed utilising modern statistical methods, current computer powers and existing statistical software. The method is useful to get insights into people's consumption preferences accounting for available demographic and personal information and unobserved person-specific preferences that can correlate across various multiple food intakes.

In line with the results obtained in Chapter 3, the results of Chapter 4 suggest that inclusion of fixed effects of both, probability and portion size parts of the two-part model, is very important to get robust inference on the potential predictors of people's health related behaviours, and careful consideration to confounding and available predictors should be given at this stage of the data analysis. The results also show the consequences of model misspecification regarding the choice of food intake predictors and the conclusions drawn from a misspecified model might be misleading.

The author hopes that despite considerable time efforts required to carry out the described method, the community of researchers in nutritional epidemiology will see the many benefits the modern methodology can bring to increase inference robustness and, with increasing computer power and readily available software, will be willing to apply the suggested method in their research practice.

List of abbreviations

Composite Likelihood (CL)

Markov Chain Monte Carlo (MCMC)

National Diet and Nutrition Survey Rolling Programme (NDNS RP)

Postcode Address File (PAF)

Primary Sampling Units (PSUs)

Chapter 5

Modelling multiple correlated habitually- and occasionally-consumed food intakes applied to the evaluation of the relationship between alcohol intake and glycosylated haemoglobin A1C

This chapter combines the joint modelling approach of multiple foods intake introduced in Chapter 4 with regression calibration to evaluate the relationship of alcohol intake, which has excess zeros and is correlated with the intake of other foods, and glycosylated haemoglobin A1C in a subsample population of the UK National Diet and Nutrition Survey Rolling Programme.

5.1 Background

This chapter, investigates the effect of alcohol consumption on glycosylated haemoglobin A1C (HbA1C), a biomarker of type 2 diabetes mellitus (T2D), which is a well-recognised adverse health condition affecting the human body's natural ability to process blood glucose by regulating the amount of insulin production (Polonsky, 2012). The recent increase in T2D incidence (World Health Organization, 2016) has been linked to obesity (Garber et al., 2008; Cloostermans et al., 2015), reduced physical activity (PA) (Jelleyman et al., 2015; Jeon et al., 2007; Cloostermans et al., 2015) and certain nutrition habits (Gadgil et al., 2013; Malik et al., 2010). In particular, sweet beverages and high glycaemic load meal consumption have been consistently associated with the development of insulin resistance, which increases the risk of T2D (Malik et al., 2010). However, the effects of intake of many other foods remain unclear. In particular, the reported effects of alcohol consumption on the risk of developing insulin resistance and T2D are inconsistent. Several cohort studies report no relationships with insulin resistance, the precursor of T2D, (Schrieks et al., 2015), while others report a J-shaped relationship between alcohol intake and T2D incidence (Baliunas et al., 2009; Koloverou et al., 2015; Koppes et al., 2005; Huang et al., 2017; Knott et al., 2015). The lack of robust results might be partially attributed to residual confounding and the presence of measurement error in correlated multiple exposures inherent to observational nutritional research.

Measurement error, defined as a difference between recorded intake and the true long-term intake, is particularly pronounced in nutrition research (Nelson et al., 1989) and can be broadly divided into systematic and random. Systematic error results from intentional or unintentional misreporting of certain foods and nutrients which correlates with certain personal characteristics such as social desirability, fear of negative evaluation, body mass index (BMI) and dieting history to name a few (Braam et al., 1998; Tooze et al., 2004; Neuhouser et al., 2008; Subar et al., 2003). Random or measurement error, on the other hand, comes naturally from individual daily variation of food consumption and a shortage

of measurement tools to capture the true long-term intake. Currently, multiple-day food diary or multiple 24 hour recalls are found to be the most reliable diet measurement tools to capture long-term nutritional habits in the general population (Burrows et al., 2010; Subar et al., 2003). However, due to short observation period and natural personal variation in diet, they are prone to large measurement error. Additionally, nutritional preferences can result in certain consumption patterns leading to the situation when multiple exposures measured with error are correlated. It has been shown that when measurement errors are present in multiple correlated predictors a bias arises when estimating the predictors' effects on outcome. Unlike in the case of a single predictor measured with error, the size and direction of the bias cannot be predicted in advance as it depends on the covariance structure of the predictors (Carroll et al., 2006) leading to potentially false positive results. Carroll et al. (2006) and Keogh and White (2014) provide suggestions to correct regression coefficients for bias in the case of multiple continuous correlated predictors measured with error.

The situation is more complex when occasionally-consumed foods like fish, nuts, certain vegetables or processed meat is considered in relation to a health outcome. The distribution of occasionally-consumed food intake is characterised not only by the presence of measurement error but also by the presence of excess zeros and potential correlation between the frequency of consumption and portion sizes (Ashfield-Watt et al., 2004), which makes the application of traditional methods problematic.

Kipnis et al. (2009) suggested a combination of several modern statistical methods, namely regression calibration (Armstrong, 1985), a two-part mixed-effects model with correlated random effects (Olsen and Schafer, 2001; Tooze et al., 2002) and empirical Bayes prediction (Casella, 1985), to correct for bias in the case of a single semi-continuous predictor when assessing the relationship between fish intake (exposure) and blood mercury levels (outcome). First, they applied a two-part model to estimate the distribution of fish intake based on a few covariates. Secondly, they utilised the results obtained from the two-part

model to estimate the joint posterior distribution of random effects given the observed intakes, utilising Bayes' theorem. Then, they predicted individual fish intake conditional on the covariates and observed intake utilising the estimated posterior distribution of random effects and, last, by applying the regression calibration principle, they used the empirical Bayes' individual fish intake predictions instead of the observed individual records to relate fish intake to blood mercury levels. Their simulation results suggest that in the case of a single semi-continuous predictor uncorrelated with other predictors this approach yields minimal estimation bias. It is important to notice that their choice of food exposure and health outcome involves an assumption, most probably valid in their particular case, that only a single semi-continuous predictor measured with error is related to the health outcome and there is no need to adjust for other nutritional predictors measured with errors.

In many other cases this assumption might not be valid. For example, health conditions as complex as cardiovascular disease or T2D might have multiple nutritional exposures relevant to disease risk which should be taken into account in the statistical analysis. The task is not trivial and is computationally challenging. Firstly, the covariance structure of the random effects of all relevant correlated predictors measured with error should be estimated. Secondly, the resulting covariance structure will be high-dimensional and efforts should be made to reduce its dimensions as the prediction of individual intake involves numerical integration over a high-dimensional space which is computationally challenging. Finally, once the number of dimensions is reduced, the joint posterior distribution of random effects and the prediction of individual food intake using the posterior distribution of the random effects can be carried out.

This chapter applies the regression calibration method to investigate alcohol intake exposure in relation to HbA1C, a biomarker of type 2 diabetes mellitus in a sub-sample of adult male participants of NDNS RP survey waves 2-4 (Public Health England, 2014). Alcohol intake as a potential predictor of HbA1C was chosen due to its high public health importance and inconsistent results found in published literature. For example, Stampfer

et al. (1988) report no relationships while Koloverou et al. (2015) report protective effect of alcohol. Attempts are carried out to adjust for unexplained residual correlation between alcohol and other various occasionally-consumed food intakes to minimise inferential bias.

5.2 Methods

We assume the classical additive non-differential measurement error, we use the regression calibration approach described in Chapter 1, and the two-part model for modelling of occasionally-consumed food intakes described in detail in Chapters 2 and 3 combined with the composite likelihood approach for multiple occasionally-consumed food intakes described in detail in Chapter 4.

5.2.1 Regression calibration for two occasionally-consumed foods

The regression calibration function to obtain an estimate of the expectation of the unobserved true food intake conditioned on the observed food intake suggested by Kipnis et al. (2009) is based on the combination of the application of two-part model and Bayes' empirical predictions.

Recall that *the individual expected habitual daily intake* T_{hij} for a food h, person i on a day j is calculated as the product of the individual daily probability of the food consumption, p_{hij}, and the individual expected consumed amount on a consumption day: $T_{hij} = (p_{hij} \mid v_i) \cdot E(A_{hij} \mid A_{hij} > 0, u_{hi})$. Under the two-part model T_{hij} depends on a set of parameters θ and the unobserved person-specific effects u_{hi} and v_{hi}, which may be correlated. The point model estimates $\tilde{\theta}$ of θ can be obtained through likelihood maximisation but the person-specific u_{hi} and v_{hi} are unobserved. Using the the notation introduced in Appendix 3.B.1, the expected estimated individual habitual daily intake can be then described as

$$\widetilde{T}_{hij} = \exp(x_{hij}^{'}\tilde{\beta} + u_{hi} + 0.5\tilde{\sigma}_{\epsilon h}^{2}) \frac{\exp(x_{hij}^{'}\tilde{\gamma} + v_{hi})}{1 + \exp(x_{hij}^{'}\tilde{\gamma} + v_{hi})} \tag{5.1}$$

Daily intake of alcohol consumption may vary by day of the week. To account for this, we assign weights to the expected estimated daily weekend and weekdays intakes accord-

ingly, so that the expected individual daily intake $\widetilde{T}_{hi}$ now becomes:

$$\widetilde{T}_{hi} = \frac{4}{7}\exp(x_{hi0}^{'}\tilde{\beta} + u_{hi} + 0.5\tilde{\sigma}_{\epsilon h}^{2})\frac{\exp(x_{hi0}^{'}\tilde{\gamma} + v_{hi})}{1+\exp(x_{hi0}^{'}\tilde{\gamma} + v_{hi})} + \frac{3}{7}\exp(x_{hi1}^{'}\tilde{\beta} + u_{hi} + 0.5\tilde{\sigma}_{\epsilon h}^{2})\frac{\exp(x_{hi1}^{'}\tilde{\gamma} + v_{hi})}{1+\exp(x_{hi1}^{'}\tilde{\gamma} + v_{hi})} \quad (5.2)$$

where $x_{hi0}^{'}$ and $x_{hi1}^{'}$ are the vectors of explanatory variables corresponding to week days and weekends respectively.

The above expression is a function of random variables u_{hi}, v_{hi} and hence a random variable itself. The expectation of this function can be estimated if the joint distribution f_{UV} of $(u_{1i}, u_{2i}, v_{1i}, v_{2i})$ is known.

For the estimation of the regression calibration predictor, $\widehat{T}_{hi}(\tilde{\theta}) = E(T_{hi} \mid I_{hij}, A_{hij}, x_{hij}, \tilde{\theta})$, the joint posterior distribution $f_{UV}(u_{1i}, u_{2i}, v_{1i}, v_{2i} \mid \{I_{hij}, A_{hij}, x_{hij}\}, \tilde{\theta})$ which is the joint distribution of the random effects conditional on the observed values of food intakes, can be estimated and utilised for the estimation of the regression calibration function.

Recall that under Bayes' theorem, the joint posterior distribution of $(u_{1i}, u_{2i}, v_{1i}, v_{2i})$ can be expressed through the joint conditional distribution of the observed intakes $f(I_{1ij}, I_{2ij}, A_{1ij}, A_{2ij} \mid u_{1i}, u_{2i}, v_{1i}, v_{2i})$ and the joint marginal distribution $f_{UV}(u_{1i}, u_{2i}, v_{1i}, v_{2i})$

$$f_{UV}(u_{1i}, u_{2i}, v_{1i}, v_{2i} \mid I_{1ij}, I_{2ij}, A_{1ij}, A_{2ij}, \tilde{\theta}, \{x_{hij}\}) = \frac{f(I_{1ij}, I_{2ij}, A_{1ij}, A_{2ij} \mid \tilde{\theta}, \{x_{hij}\}, u_{1i}, u_{2i}, v_{1i}, v_{2i}) f_{UV}(u_{1i}, u_{2i}, v_{1i}, v_{2i} \mid \{x_{hij}\}, \tilde{\theta})}{\iiiint f(I_{1ij}, I_{2ij}, A_{1ij}, A_{2ij} \mid \tilde{\theta}, \{x_{hij}\}, u_{1i}, u_{2i}, v_{1i}, v_{2i}) f_{UV}(u_{1i}, u_{2i}, v_{1i}, v_{2i} \mid \{x_{hij}\}, \tilde{\theta}) du_{1i} dv_{1i} du_{2i} dv_{2i}} \quad (5.3)$$

The joint conditional distribution of the observed intakes $f(I, A|u, v, \tilde{\theta})$ is the product of corresponding conditional distributions $f(I_{hij} \mid v_{hi}, \tilde{\theta})$ and $f(A_{hij} \mid u_{hi}, \tilde{\theta})$. In our model specification, the joint conditional distribution of the observed intakes $f(I, A|u, v, \tilde{\theta})$ can

be described as

$$f(I_i, A_i|u_i, v_i, \tilde{\theta}) = \prod_{j=1}^{n_{1,i}} f_{I_1}(I_{1ij} \mid v_{1i}, \tilde{\theta}) f_{Y_1}(Y_{1ij} \mid u_{1i}, \tilde{\theta}) \prod_{j=1}^{n_{2,i}} f_{I_2}(I_{2ij} \mid v_{2i}, \tilde{\theta}) f_{Y_2}(Y_{2ij} \mid u_{2i}, \tilde{\theta}) \quad (5.4)$$

where $f_{I_h}(I_h \mid \tilde{\theta}, x, v)$, $f_{Y_h}(Y_h \mid \tilde{\theta}, x, u)$ denote the density functions of the binomial and normal distributions and respectively, and $f_{I_h}(I_h \mid \tilde{\theta}, x, v)$ is parametrised here, via the logit link function (please refer to section 2.2.1 for model specification details).

The parameters of the marginal joint multivariate normal distribution $f_{UV}(u_{1i}, u_{2i}, v_{1i}, v_{2i} \mid \tilde{\theta})$ are the point estimates obtained from the covariance structure of random effects which forms part of $\tilde{\theta}$ and was obtained at the step of maximising the likelihood function.

Conditional on a food having been reported consumed if and only if it was consumed, the regression calibration predictor of an individual food intake $E(T_{hi} \mid I_{hij}, A_{hij}, x_{hij}, \tilde{\theta})$ becomes

$$\widehat{T}_{hi}(\tilde{\theta}) = E(T_{hi} \mid I_{hij}, A_{hij}, x_{hij}, \tilde{\theta}) = \iiiint \widetilde{T}_{hi} f_{UV}(u_{1i}, v_{1i}, u_{2i}, v_{2i} \mid I_{1ij}, I_{2ij}, A_{1ij}, A_{2ij}, \tilde{\theta}) du_{1i} dv_{1i} du_{2i} dv_{2i} \quad (5.5)$$

The results shown in the next section adjust the predicted intakes for the relevant predictors of T2D additionally to the predictors specific to their consumption.

To estimate multiple integrals Mathematica 11.2 software was utilised. Numerical integration was performed using the Adaptive Gauss-Hermit quadrature method with 7 quadrature points.

5.2.2 Fitted models to examine the relationship between alcohol intake and HbA1C

The outcome variable in the models was HbA1C levels split into three categories: HbA1C <5.5 %, which was defined as a reference category (N = 107 (39.9%)) ; HbA1C ≥ 5.5 % and < 6.0 % (normal-elevated, N = 114 (42.5%)); HbA1C ≥ 6.0 % (elevated and high, N = 47 (17.5%)). The choice of these cut-offs was based on the results from Bonora and

Jaakko (2011), where these cut-offs showed the importance of elevated HbA1C levels in progressing to diagnosed T2D. The numbers in the category with HbA1C levels above 6.5% required to diagnose T2D were small (N = 14 (5.2%)) and not feasible to analyse as a separate category.

Multinomial logistic regression was utilised to examine the relationship between HbA1C levels and alcohol intake. Three models were considered which differ in the way alcohol intake was predicted. Model 1 and Model 2 used the regression calibration approach to predict individual alcohol intake, where Model 1 used a two-part model to predict alcohol intake and Model 2 used a four-part model to predict alcohol intake. Model 3 used the traditional approach of individual averages (all observations available for individual averaged across diary days recorded) as a predictor of individual alcohol and other food intakes. The analysis started with fitting models with a full set of potential predictors or confounders: age (years), body mass index (BMI) (kg/m2), moderate to vigorous physical activity (min) (MVPA), history of smoking, qualifications and lipid medicine. Additionally, all the models were adjusted for the following food intakes: fruit (g), cooked vegetables (g), raw vegetables (g), regular soft drinks (g) and total energy intake (Kcal). For Models 1 and 2, the food intakes were predicted using the regression calibration approach based on the two-part models previously specified in Chapter 3 and taking into account the residual correlation between consumption probability and portion size for a given food only. It was decided not to adjust for blood test results to avoid over-adjustment. The predictors were chosen based on a likelihood ratio test and only those which remained significant (p-value <0.05) for at least one of the response levels were left in the final models and presented. To diagnose the models' fit Akaike information criterion (AIC) and Bayesian information criterion (BIC) were provided.

5.3 Results

5.3.1 Descriptive analysis

The sample available for analysis consisted of a 268 male sub-sample of the NDNS RP Years 2-4 population for whom HbA1C values were available. The sample's characteristics by the three HbA1C categories are summarised in Table 5.1

Table 5.1: Sample characteristics in the male sub-population of NDNS RP Years 2 - 4 available for analysis of Haemoglobin A1C (N = 268). Means (SD), Medians (IQR) or N(%) are shown

	HbA1C < 5.5 % (N = 107 (39.9%))	HbA1C ≥ 5.5 % and < 6.0 % (N = 114 (42.5%))	HbA1C ≥ 6.0 % (N = 47 (17.5%))	p-value
Age, y	41 (14)	51(16)	60 (17)	<0.001
HDL, mmol/l	1.37 (1.12 - 1.53)	1.35 (1.09 -1.59)	1.22 (0.98 -1.36)	0.020
HDL/LDL Ratio	0.45 (0.36 - 0.57)	0.36 (0.29 - 0.49)	0.47 (0.33 - 0.60)	0.003
Cholesterol, mmol/l	5.0 (1.0)	5.4 (1.1)	4.7 (1.1)	0.071
Triglicerides, mmol/l	1.1 (0.8 - 2.0)	1.2 (0.8 - 1.6)	1.6 (1.1 - 2.2)	0.029
BMI, kg/m2	26.3 (3.3)	27.6 (4.4)	29.9 (5.4)	<0.001
Lipid Medication (Yes)	5 (4.7 %)	10 (9.1 %)	19 (41.3 %)	<0.001
Smoking				0.005
Smoking (No, never smoked)	67 (63.2%)	53 (48.2%)	16 (34.8 %)	
Smoking (No, quit more 10 years ago)	10 (9.4%)	17 (15.5%)	16 (34.8 %)	
Smoking (No, quit less 10 years ago)	9 (8.5%)	11 (10.0%)	6 (13.0 %)	
Smoking (Yes, current smoker)	19 (17.9%)	25 (22.7%)	7 (15.2 %)	
Smoking (Occasional smoker)	1 (1.0%)	4 (3.6%)	1 (2.2 %)	
Moderate-vigorous physical activity, h/day	1.2 (0.5 - 3.0)	1.3 (0.6 - 3.3)	0.7 (0.1 - 2.0)	0.098
Qualifications				0.006
Degree level	34 (31.8%)	25 (21.9%)	8 (17.0%)	
Uninished degree, students or A levels	43 (40.2%)	35 (30.7 %)	13 (27.7 %)	
GCSE grades A-C	14 (13.0%)	26 (22.8 %)	6 (12.8 %)	
GCSE grades below C or no qualifications	13 (12.2 %)	20 (17.5 %)	17 (36.2%)	
Foreign degree	3 (2.8 %)	8 (7.0 %)	3 (6.4 %)	

p-values are from Kruskal-Wallis test for continuous predictors or from chi-squared test for categorical predictor

In line with previous research, age and BMI strongly correlate with elevated levels of HbA1C. Interestingly, significant unadjusted correlation was observed between ex- or current smoking, obtained qualifications and elevated HbA1C levels, but no significant correlation observed between HbA1C and moderate-vigorous physical activity.

5.3.2 Regression calibration applied to alcohol intake predictions

We utilise the results from Chapter 4 where the composite likelihood approach was applied to estimate the two-part model parameters from a multivariate model, including the covariance structure of random effects presented in Table 4.6. We reproduce this correlation structure in Table 5.2 to describe the estimated correlation structure of residual correlations between random effects after adjusting for numerous observed predictors of various food intakes in a male sub-sample of NDNS RP Years 2-4. The table shows that even after adjusting for various observed patients' characteristics related to food intakes, there remained unexplained part of personal preferences that shows high correlation for some food intakes. In particular, unobserved personal preferences for alcohol intake are negatively correlated with unobserved personal preferences for fruit consumption (probability part) and positively correlated with the unobserved personal preferences for processed meat (portion size). There are other remaining correlations between unobserved preferences for alcohol and unobserved preferences for other food intakes but these two, along with the correlation between alcohol intake probability and alcohol intake portion size, surfaced as the largest correlations between the estimated random effects related to alcohol intake and will be taken into account in Model 2.

Table 5.2: Estimated correlation structure of random effects in the male sub sample of the NDNS RP Years 2-4, refer to Chapter 4 for details on estimation

		Alcohol		Fruits		Veg Cooked		Veg Raw		Processed Meat		Oily Fish		Soft regular drinks	
		Pr[1]	Am[2]	Pr[1]	Am[2]	Pr[1]	Am[2]	Pr[1]	Am[2]	Pr[1]	Am[2]	Pr[1]	Am[2]	Pr[1]	Am[2]
Alcohol	Pr	1	0.37	-0.09	0.01	0.08	0.05	0.15	0.07	0.03	-0.13	0.07	-0.18	-0.05	-0.04
	Am		1	-0.30	0.02	-0.06	0.09	-0.07	0.00	-0.04	0.28	-0.12	0.02	-0.10	0.22
Fruits	Pr			1	0.59	0.22	0.06	0.32	0.16	-0.12	-0.13	0.19	0.13	-0.10	-0.10
	Am				1	-0.03	0.00	0.18	0.23	-0.15	0.03	0.36	0.07	-0.12	-0.16
Veg Cooked	Pr					1	0.35	-0.05	-0.15	-0.04	0.39	0.37	-0.09	-0.08	-0.12
	Am						1	-0.09	-0.14	-0.21	-0.09	0.08	0.08	-0.10	0.00
Veg Raw	Pr							1	0.52	0.11	-0.06	0.60	-0.46	0.01	-0.14
	Am								1	0.07	0.15	0.14	-0.04	-0.18	-0.22
Processed	Pr									1	0.03	-0.07	-0.35	0.18	0.01
Meat	Am										1	0.09	-0.13	0.09	0.18
Oily Fish	Pr											1	-0.54	-0.27	-0.05
	Am												1	-0.10	0.10
Soft regular	Pr													1	0.41
drinks	Am														1

[1]Pr stands for Probability part of the two-part model, [2]Am stands for Amount part of the two-part model

Table 5.3 shows the estimated regression parameters of Models 1 and 2 where alcohol intake was predicted based on a two- part model (taking into account the correlation between unobserved preferences for alcohol intake probability and alcohol intake portion size) and a four-part model (taking into account the correlation between unobserved preferences for alcohol intake probability, alcohol intake portion size, fruit intake probability and processed meat portion size) respectively. The estimated effects of factors on the probability of falling in one of the two groups with elevated HbA1C levels are presented as relative risk ratios (RRR) compared to the reference category (HbA1C $< 5.5\ \%$). The effects presented are for the increase in alcohol consumption by 1g per day and by 100g per day for the other foods. The results show that out of the available food intake predictors, neither alcohol nor cooked vegetables, total fruits, soft drinks nor total energy intakes were related to HbA1C levels. However, raw and salad vegetable intake was negatively and statistically significantly correlated with risk of being in the group with HbA1C $\geq 5.5\%$ and $< 6.0\%$ (RRR 0.55 (95% CI (0.33, 0.90), p-value 0.017) but the effect does not reach statistical significance for the risk of being in the group with the highest HbA1C levels $\geq 6.0\%$ (RRR 0.68 95% CI(0.37, 1.24), p-value 0.206). Interestingly, the effect of elevated BMI showed significance only for risk of being in the group with HbA1C with the highest levels. Prediction of alcohol intake based on the two-part or four-part model does not change the inference that can be drawn from the data, and the AIC is only marginally lower for Model 2 compared to Model 1.

Table 5.4 shows the results from Model 3 when alcohol intake, as a predictor, is estimated as an individual average. Interestingly and in line with the theory, the effect of alcohol intake in Model 3, when fitted as estimated individual average, is inclined in the direction of zero and exhibit high statistical significance. If we draw inferences from this model we would conclude that alcohol intake has a very strong significant protective effect on HbA1C. Additionally, the effect of the other potentially correlated predictors, in our case, raw vegetable intake, becomes more statistically significant and also tends in the direction of zero, probably indicating residual correlation between alcohol and raw vegetable intake

that was not taken into account when predicting vegetable intake.

Unexpectedly, in the adjusted analysis, neither physical activity nor smoking were found to be significantly related to HbA1C levels.

Table 5.3: The estimated effects of patients characteristics and food intakes on elevated Haemoglobin A1C from Model 1 and Model 2 where HbA1C $< 5.5\ \%$ is a reference category

Predictors	RRR	(95% CI)	p-value
HbA1C $\geq 5.5\%$ and $< 6.0\%$			
Model 1			
Alcohol, 1 g	1.000	(0.993, 1.008)	0.947
Veg Raw, 100g	0.55	(0.33, 0.90)	0.017
Age, year	1.05	(1.02, 1.07)	< 0.001
BMI , kg/m2	1.04	(0.97, 1.11)	0.315
Lipid reduction medicine (yes)	1.47	(0.45, 4.78)	0.524
Model 2			
Alcohol, 1 g	0.998	(0.988, 1.009)	0.779
Veg Raw, 100g	0.55	(0.34, 0.90)	0.018
Age, year	1.05	(1.02, 1.07)	< 0.001
BMI , kg/m2	1.04	(0.97, 1.11)	0.317
Lipid reduction medicine (yes)	1.47	(0.45, 4.78)	0.525
HbA1C $\geq 6.0\ \%$			
Model 1			
Alcohol, 1 g	0.992	(0.968, 1.015)	0.480
Veg Raw, 100g	0.68	(0.37, 1.24)	0.206
Age, year	1.07	(1.04, 1.11)	< 0.001
BMI , kg/m2	1.18	(1.07, 1.30)	0.001
Lipid reduction medicine (yes)	5.70	(1.68, 19.31)	0.005
Model 2			
Alcohol, 1 g	0.988	(0.963, 1.013)	0.346
Veg Raw, 100g	0.68	(0.37, 1.23)	0.200
Age, year	1.07	(1.04, 1.11)	< 0.001
BMI , kg/m2	1.18	(1.07, 1.30)	0.001
Lipid reduction medicine (yes)	5.72	(1.69, 19.40)	0.005

Model 1: Akaike Information criteria 493.23, Bayesian information criteria 536.32, Model 2: Akaike Information criterion 492.89, Bayesian Information criterion 535.98.

Table 5.4: The estimated effects of patients characteristics and food intakes on elevated Haemoglobin A1C from Model 3 where HbA1C $< 5.5\ \%$ is a reference category

Predictors	RRR	(95% CI)	p-value
HbA1C $\geq 5.5\%$ and $< 6.0\%$			
Alcohol, 1 g	0.987	(0.975, 0.999)	0.040
Veg Raw, 100g	0.53	(0.32, 0.88)	0.014
Age, year	1.05	(1.03, 1.07)	< 0.001
BMI , kg/m2	1.04	(0.97, 1.12)	0.277
Lipid reduction medicine (yes)	1.65	(0.50, 5.52)	0.413
HbA1C $\geq 6.0\ \%$			
Alcohol, 1 g	0.956	(0.932, 0.980)	<0.001
Veg Raw, 100g	0.58	(0.31, 1.09)	0.091
Age, year	1.08	(1.05, 1.12)	< 0.001
BMI , kg/m2	1.20	(1.09, 1.32)	< 0.001
Lipid reduction medicine (yes)	7.67	(2.07, 28.32)	0.002

Model 3: Akaike Information criterion 477.09, Bayesian information criterion 520.18.

5.4 Discussion

This chapter showed how the regression calibration approach coupled with Empirical Bayes estimation and a joint modelling approach can be applied to estimate the effect of multiple occasionally-consumed foods on health outcomes. This extends the method suggested by Kipnis et al. (2009) to take into account the unobserved correlated preferences across various foods and minimise the bias produced by correlated measurement errors. This is the first research, to the best of the author's knowledge, which applies the regression calibration method to estimate the effect of alcohol intake on HbA1C in an NDNS RP sub-population taking into account correlated observed and unobserved preferences across more than one food. For comparative purposes, the alcohol intake was predicted based on a 2-part model, where the correlation between probability and portion size of alcohol intake was taken into account, in a 4-part model, other unobserved food preferences were additionally taken into account. The regression parameters from these models were contrasted with those estimated via the traditional approach based on individual averages of food intake. The 4-part model produced only marginally better regression calibration for alcohol intake with respect to the HbA1C outcome, compared to the two-part model. The results utilising the prediction approach are however, in startling contrast to the results obtained when the observed individual average is fitted as a predictor. This is an important finding as it confirms that the regression calibration approach, which aims to account for measurement error and the correlation between the probability of consumption and portion size, produces much less biased results compared to using an individual average proxy, which has a large measurement error component and is correlated with other predictors measured with error. The results hold even when it is not possible to predict a food intake taking into account all the correlation structure of unobserved preferences. The results also emphasize the importance of gathering extensive behavioural, personal and social information when estimating relationships between food intakes and health outcome as it is highly possible that a minimal observed difference between predictions based on a simpler versus a more complex model is, partially, due to

extensively accounting for social and personal information in the fixed parts of the models.

One of the limitations of the presented research is the low number of unobserved preferences taken into account, to the maximum of four to make the numerical integration feasible, and the choice of random effects for integration was based on eyeballing of the correlation matrix, rather than making the choice based on rigid mathematical criteria. This decision is due to the fact that the task of reducing the dimensions of the correlation matrix of random effects is not trivial. The estimation of the covariance matrix in multivariate cases has been a hot discussion topic in the recent literature, including finance, climate change modelling and genetics, with the main concerns being the reliability of the sample covariance matrix in high dimensions (Ledoit and Wolf, 2015; Bien and Tibshirani, 2011). In our case, the concern is to reduce the dimensions of the matrix when no theoretical ordering of the variable exists and this is an area of future research.

List of abbreviations

Body mass index (BMI)

Composite Likelihood (CL)

Haemoglobin A1C (HbA1C)

National Diet and Nutrition Survey Rolling Programme (NDNS RP)

Primary Sampling Units (PSUs)

Probability density function (pdf)

Postcode Address File (PAF)

Randomised control trial (RCT)

Type 2 diabetes mellitus (T2D)

Chapter 6

Discussion

The thesis provides novel methods of analysis for multivariate data with a complex structure arising from the presence of excess zeroes and correlated measurement errors.

Firstly, we proposed a new numerical approach of estimating the sample distribution of occasionally-consumed food intakes modelled with a two-part model in Chapter 1, which, compared to traditional Monte Carlo simulations reduces the estimation burden and increases precision, especially, on the tails of the distribution. The method was illustrated through the analysis of self-reported alcohol consumption collected during the screening phase of a randomised controlled trial (RCT) investigating the effect of the types of fats and carbohydrates in diet on glucose and insulin metabolism. Chapter 2 also demonstrated how the two-part model can be estimated using freely available software for the case of 2 day food diaries, or two recall records, and one occasionally-consumed food. Chapter 3 further investigated the determinants of several occasionally-consumed food intakes, such as income, qualification and physical activity. This is the first research, to our knowledge, which applies the two-part model for the investigation of the predictors of food intakes in a NDNS RP sub-population. Our findings contribute further to our understanding of people's preferences for certain foods. The richness of the data, the extensive list of potential correlated predictors and the advanced statistical methods make the inference

quite robust and, although the effect of the residual confounding can never be excluded, the conclusions presented in this research add to the current literature of people's food preferences and the effects of various social and personal factors in relation to food intake. Chapter 4 extended the results of Chapter 3 for the estimation of effects of determinants of alcohol intake taking the bigger correlation structure of unobserved preferences into account. It has been carried out by applying composite likelihood and bootstrap methods and this is the first time this novel approach was applied to the investigation of determinants of food intake. The last chapter, Chapter 5, applied the regression calibration and empirical Bayes approaches when occasionally-consumed food intake acts as a predictor of health outcome. The work was built on the estimation models obtained in the previous chapters and extends the previously published work by Kipnis et al. (2009) to take into account the correlation structure of unobserved personal preferences to minimise inferential bias. The method is applied to the sub-sample of male population in NDNS RP (Years 2-4) to estimate the effect of alcohol intake on Haemoglobin A1C, a well recognized precursor of type 2 diabetes. We demonstrated how the method can be applied to minimise inferential bias in nutritional research.

There are other multivariate data analysis methods which are widely used in the nutrition research community. For example, principal component analysis (PCA), which is a technique developed by Hotelling (1933) with an aim to reduce dimensions of correlated multivariate data by finding fewer new uncorrelated directions which describe the biggest part of the overall data variation. It serves well in many applications where reduction in dimensions is of primary importance, for example, classification tasks and where the true variance covariance matrix can be estimated with minimal bias. However, in nutritional research, PCA most often operates on individual averages, which contain measurement error. The application of PCA to individual averages should be interpreted with caution as the presence of correlated measurement errors might lead to biased principal components (Scagliarini, 2011; Raykov et al., 2017) and excess zeros are ignored. There is also an issue with interpretation because it is difficult to replicate the results in another sample

from the same population.

A potential application of PCA within the presented framework is to reduce the estimated correlation matrix of random effects when modelling the intake of several foods. This could include finding the related eigenvectors and their corresponding eigenvalues, and, then, only the directions corresponding to the largest eigenvalues are retained such that a big part of the overall variation is explained by these new remaining fewer directions. Importantly, each of the previous dimensions can be represented as a linear combination of the new uncorrelated dimensions. In our example, if we consider the joint probability distribution function $f(u_1 \dots u_n)$ of random effects $(u_1 \dots u_n)$ then with the newly found dimensions $\tau_1 \dots \tau_k$, $k < n$, each of the random effects $u_1 \dots u_n$ can be represented as a linear combination of newly found independent dimensions $u_j = \sum_{j=1}^{k} \alpha_j \tau_j$. Then, the joint probability density function (pdf) can be factorised as the product of conditional densities $f(u_i \mid \tau_1, \dots \tau_k)$ such that the task of multiple integration will now be reduced to the integrating multiple times over the simpler pdfs of τ_j. The suggestion is not without a drawback as although PCA retains the biggest part of the variation, still some variation will be lost. On the other hand, this approach appears to be more robust compared to just picking out the biggest correlations from the covariance matrix and will be the area of the future research.

In the applications presented in this thesis the regression coefficients, or marginal effects, of explanatory variables of the components of the two-part models and their multivariate extensions were the parameters of inferential interest, whereas the covariance parameters were regarded as nuisance parameters. Estimation of both regression coefficients and variance parameters was based on maximising the likelihood function. The properties of these estimates, for example, describing what aspects of the data and model may lead to estimation issues is an area of future research.

In practice, selection of the appropriate covariance structure may require testing for zero variance components. In this case, the null hypothesis may place some of the variance

components on the boundary of the parameter space. Therefore, the commonly used likelihood ratio test does not have the usual chi-squared distribution with the degrees of freedom equal to the number of independent parameters being tested under the null hypothesis. This is a challenging problem and a current area of research. The problem has been addressed, under certain conditions, by using approximate or exact restricted likelihood ratio tests for the covariance parameters of the linear mixed-effects model. The results from Self and Liang (1987) to obtain the correct null distribution of the likelihood ratio test in the case of generalised linear mixed-effects model or the models developed in this thesis requires the calculation of the Fisher information matrix at the true parameter value under the null hypothesis and to study the topological behaviour of the neighbourhood of the true parameter value. This problem is complex due to the intractable integrals involved in the evaluation of the likelihood function and because the variance components may become numerically unstable when some variance components are small. Further research is required to tackle this complex problem.

The applicability of the 2-part model can be extended by modelling the outcomes through various distributions, for example, generalised gamma to avoid the need of transformation to Normality (Agogo, 2017) and extending the 2-part model to accommodate non-consumers (Keogh, 2011).

The apparent limitations of the presented framework is its reliance on relatively large datasets, cost of meticulous data collection needed to produce robust model estimates and relative complexity of methods. However, in situations where large scale research into nutritional habits of populations is commissioned, the presented framework, if utilised, provides the necessary tools to extract maximum benefit from the project given some additional preliminary planning. The scope of application of the work presented in this thesis in nutritional epidemiology is wide. While RCTs could be thought as the gold standard to assess the effect of dietary exposures, they are often costly, difficult to conduct, of short duration, and they rarely measure the effect of interventions on hard endpoints such as

death, cancer or major cardiovascular events in general populations. For example the PREDIMED trial, a randomised dietary intervention trial of the effects of a Mediterranean diet supplemented with either extra-virgin olive oil or nuts on major cardiovascular events rate in Spanish subjects at high cardiovascular risk, showed significant beneficial effects of a Mediterranean diet with extra-virgin olive oil or nuts supplement compared to a advice to reduce saturated fat intake (Estruch et al., 2013). The findings from the PREDIMED trial were, however, questioned, the original results retracted due to potential biases arising from problems with the randomisation procedure, and the data were re-analised as an observational study (Estruch et al., 2018b,a). Furthermore, in the PREDIMED trial, one of the questions that have been debated is if supplementing any diet (not necessarily a Mediterranean diet) with extra-virgin olive oil or nuts would have a beneficial effect on the major cardiovascular event rate in those at risk (Appel and Van Horn, 2013). Additionally, generalisability of the results might be questionable with respect to larger, more diverse, populations.

Observational studies, on the other hand, can follow a large number of participants from the general population in their natural environment for a long time to be able to observe long-term health outcomes. Moreover, observational studies have the capacity to assess exposures which might be unethical to assess in RCTs with a caveat of them being prone to unobserved confounding, which, ideally, should be thought through at the design stage.

The models developed in this thesis can be applied in a variety of settings, not necessarily restricted to nutrition research. For example, in respiratory therapeutic research, it is plausible that the frequency of asthma exacerbations can be correlated with the intensity of exacerbations. In therapeutic research areas where symptoms are an important indication of disease progression, for example, rheumatoid arthritis, it is plausible that symptoms' frequency can correlate with symptoms' intensity

Overall, this thesis provides a step-by-step tool using advanced statistical techniques in the framework of the estimation of multiple correlated habitually and occasionally-

consumed food intakes with measurement errors and suggests that the effort of taking into account the full correlation structure is worthwhile for both, more robust inference on food intake predictors, and when predicting risk of various nutritional exposures on health outcomes. Hopefully, this work will make the new advanced statistical methods more accessible for future nutrition researchers and will help to minimise inferential bias associated with nutritional research.

List of abbreviations

Haemoglobin A1C (HbA1C)

National Diet and Nutrition Survey Rolling Programme (NDNS RP)

Principal Component Analysis (PCA)

Probability density function (pdf)

Randomised control trial (RCT)

Type 2 diabetes mellitus (T2D)

.

Bibliography

Aerts, M. and Claeskens, G. (1999). Bootstrapping pseudolikelihood models for clustered binary data. *Annals of the Institute of Statistical Mathematics*, 51(3):515–530.

Agogo, G. O. (2017). A zero-augmented generalized gamma regression calibration to adjust for covariate measurement error: A case of an episodically consumed dietary intake. *Biometrics Journal*, 59:94,109.

Albert, P. S. (2005). Letter to the editor. *Biometrics*, 61(3):879–880.

Anyaegbu, G. (2010). Using the OECD equivalence scale in taxes and benefits analysis. *Economic & Labour Market Review*, 4:49–54.

Appel, L. J., Moore, T. J., Obarzanek, E., Vollmer, W. M., Svetkey, L. P., Sacks, F. M., Bray, G. A., Vogt, T. M., Cutler, J. A., Windhauser, M. M., Lin, P.-H., Karanja, N., Simons-Morton, D., McCullough, M., Swain, J., Steele, P., Evans, M. A., Miller, E. R., and Harsha, D. W. (1997). A clinical trial of the effects of dietary patterns on blood pressure. *New England Journal of Medicine*, 336(16):1117–1124.

Appel, L. J. and Van Horn, L. (2013). Did the PREDIMED trial test a mediterranean diet? *New England Journal of Medicine*, 368(14):1353–1354.

Appleton, K. M., Hemingway, A., Saulais, L., Dinnella, C., Monteleone, E., Depezay, L., Morizet, D., Perez-Cueto, A. F. J., Bevan, A., and Hartwell, H. (2016). Increasing vegetable intakes: rationale and systematic review of published interventions. *European Journal of Nutrition*, 55:869–896.

Arab, L., Tseng, C.-H., Ang, A., and Jardack, P. (2011). Validity of a multipass, web-based, 24-hour self-administered recall for assessment of total energy intake in blacks and whites. *American Journal of Epidemiology*, 174(11):1256–1265.

Armstrong, B. (1985). Measurement error in the generalised linear model. *Communications in Statistics - Simulation and Computation*, 14(3):529–544.

Ashfield-Watt, P. A. L., Welch, A. A., Day, N. E., and Bingham, S. A. (2004). Is 'five-a-day' an effective way of increasing fruit and vegetable intakes? *Public Health Nutrition*, 7(02):257–261.

Baliunas, D. O., Taylor, B. J., Irving, H., Roerecke, M., Patra, J., Mohapatra, S., and Rehm, J. (2009). Alcohol as a risk factor for type 2 diabetes: A systematic review and meta-analysis. *Diabetes Care*, 32(11):2123–2132.

Bartoshuk, L. M., Duffy, V. B., and Miller, I. J. (1994). Ptc/prop tasting: anatomy, psychophysics, and sex effects. *Physiol Behav*, 56(6):1165–71.

Basiotis, P. P., Welsh, S. O., Cronin, F. J., Kelsay, J. L., and Mertz, W. (1987). Number of days of food intake records required to estimate individual and group nutrient intakes with defined confidence. *The Journal of Nutrition*, 117(9):1638–1641.

Beaton, G. H., Milner, J., Corey, P., McGuire, V., Cousins, M., Stewart, E., de Ramos, M., Hewitt, D., Grambsch, P. V., Kassim, N., and Little, J. A. (1979). Sources of variance in 24-hour dietary recall data: implications for nutrition study design and interpretation. *The American Journal of Clinical Nutrition*, 32(12):2546–59.

Bellio, R. and Varin, C. (2005). A pairwise likelihood approach to generalized linear models with crossed random effects. *Statistical Modelling*, 5(3):217–227.

Bien, J. and Tibshirani, R. J. (2011). Sparse estimation of a covariance matrix. *Biometrika*, 98:807–820.

Bingham, S. A., Gill, C., Welch, A., Cassidy, A., Runswick, S. A., Oakes, S., Lubin, R.,

Thurnham, D. I., Key, T. J., Roe, L., Khaw, K. T., and Day, N. E. (1997). Validation of dietary assessment methods in the uk arm of epic using weighed records, and 24-hour urinary nitrogen and potassium and serum vitamin c and carotenoids as biomarkers. *International Journal of Epidemiology*, 26(suppl 1):137.

Bingham, S. A., Gill, C., Welch, A., Day, K., Cassidy, A., Khaw, K. T., Sneyd, M. J., Key, T. J. A., Roe, L., and Day, N. E. (1994). Comparison of dietary assessment methods in nutritional epidemiology: weighed records v. 24 h recalls, food-frequency questionnaires and estimated-diet records. *British Journal of Nutrition*, 72(04):619–643.

Black, A. E. (2000a). Critical evaluation of energy intake using the goldberg cut-off for energy intake:basal metabolic rate. a practical guide to its calculation, use and limitations. *International Journal of Obesity*, 24:1119–1130.

Black, A. E. (2000b). The sensitivity and specificity of the goldberg cut-off for ei:bmr for identifying diet reports of poor validity. *European Journal of Clinical Nutrition*, 54.

Bonora, E. and Jaakko, T. (2011). The pros and cons of diagnosing diabetes with a1c. *Diabetes Care*, 34(Suppl 2):S184–90.

Borrelli, R., Simonetti, M. S., and Fidanza, F. (1992). Inter-and intra-individual variability in food intake of elderly people in perugia (italy). *British Journal of Nutrition*, 68(01):3–10.

Braam, L. A. J. L. M., Ocké, M. C., Bueno-de Mesquita, H. B., and Seidell, J. C. (1998). Determinants of obesity-related underreporting of energy intake. *American Journal of Epidemiology*, 147(11):1081–1086.

Bufe, B., Breslin, P. A. S., Kuhn, C., Reed, D. R., Tharp, C. D., Slack, J. P., Kim, U.-K., Drayna, D., and Meyerhof, W. (2005). The molecular basis of individual differences in phenylthiocarbamide and propylthiouracil bitterness perception. *Current biology : CB*, 15(4):322–327.

Buonaccorsi, J. P. (2010). *Measurement error: models, methods and applications*. Interdisciplinary Statistics Series. Chapman and Hall/CRC.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer, 2 edition.

Burrows, T. L., Martin, R. J., and Collins, C. E. (2010). A systematic review of the validity of dietary assessment methods in children when compared with the method of doubly labeled water. *Journal of the American Dietetic Association*, 110(10):1501–1510.

Caragea, P. C. and Smith, R. L. (2007). Asymptotic properties of computationally efficient alternative estimators for a class of multivariate normal models. *Journal of Multivariate Analysis*, 98(7):1417–1440.

Carle, A. C. (2009). Fitting multilevel models in complex survey data with design weights: Recommendations. *BMC Medical Research Methodology*, 9(1):1–13.

Carroll, R. J. (1989). Covariance analysis in generalized linear measurement error models. *Statistics in Medicine*, 8(9):1075–1093.

Carroll, R. J., Gallo, P., and Gleser, L. J. (1985). Comparison of least squares and errors-in-variables regression, with special reference to randomized analysis of covariance. *Journal of the American Statistical Association*, 80(392):929–932.

Carroll, R. J. and Stefanski, L. A. (1990). Approximate quasilikelihood estimation in models with surrogate predictors. *Journal of the American Statistical Association*, 85:652.

Carroll, R. L., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models*. Chapman and Hall/CRC.

Casella, G. (1985). An introduction to empirical bayes data analysis. *The American Statistician*, 39(2):83–87.

Chambers, R. L., Steel, D. G., Wang, S., and Welsh, A. (2012). *Maximum likelihood estimation for sample surveys*. CRC Press Boca Raton, FL.

Chan, T.-C., Luk, J. K. H., Chu, L.-W., and Chan, F. H. W. (2015). Association between

body mass index and cause-specific mortality as well as hospitalization in frail chinese older adults. *Geriatrics & Gerontology International*, 15(1):72–79.

Chen, Y., Hong, C., and Riley, R. D. (2015). An alternative pseudolikelihood method for multivariate random-effects meta-analysis. *Statistics in Medicine*, 34(3):361–380.

Chernova, J. and Solis-Trapala, I. (2016). A simplified approach to estimating the distribution of occasionally-consumed dietary components, applied to alcohol intake. *BMC Medical Research Methodology*, 16:78.

Cloostermans, L., Wendel-Vos, W., Doornbos, G., Howard, B., Craig, C. L., Kivimäki, M., Tabak, A. G., Jefferis, B. J., Ronkainen, K., Brown, W. J., Picavet, S. H. S. J., Ben-Shlomo, Y., Laukkanen, J. A., Kauhanen, J., and Bemelmans, W. J. E. (2015). Independent and combined effects of physical activity and body mass index on the development of type 2 diabetes – a meta-analysis of 9 prospective cohort studies. *The International Journal of Behavioral Nutrition and Physical Activity*, 12:147.

Conklin, A. I., Forouhi, N. G., Suhrcke, M., Surtees, P., Wareham, N. J., and Monsivais, P. (2014). Variety more than quantity of fruit and vegetable intake varies by socioeconomic status and financial hardship. findings from older adults in the epic cohort(). *Appetite*, 83:248–255.

Cooke, L. and Fildes, A. (2011). The impact of flavour exposure in utero and during milk feeding on food acceptance at weaning and beyond. *Appetite*, 57(3):808–811.

Cox, D. R. and Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91(3):729–737.

Darmon, N. and Drewnowski, A. (2008). Does social class predict diet quality? *The American Journal of Clinical Nutrition*, 87(5):1107–1117.

Dashti, H. S., Scheer, F. A. J. L., Jacques, P. F., Lamon-Fava, S., and Ordovas, J. M. (2015). Short sleep duration and dietary intake: Epidemiologic evidence, mechanisms, and health implications. *Advances in Nutrition*, 6(6):648–659.

Day, N. E., McKeown, N., Wong, M. Y., Welch, A., and Bingham, S. A. (2001). Epidemiological assessment of diet: a comparison of a 7-day diary with a food frequency questionnaire using urinary markers of nitrogen, potassium and sodium. *International Journal of Epidemiology*, 30(2):309–317.

Department of Health (2014). A quick guide to the government's healthy eating recommendations. Report, Department of Health.

Department of Health, UK (2016). UK Chief Medical Officers' Low Risk Drinking Guidelines. Technical report, Department of Health, UK.

Diggle, P., Heagerty, P., Liang, K.-Y., and Zeger, S. (2002). *Analysis of Longitudinal Data*. Oxford Statistical Science. OUP Oxford, second edition.

Djoussé, L., Akinkuolie, A. O., Wu, J. H. Y., Ding, E. L., and Gaziano, J. M. (2012). Fish consumption, omega-3 fatty acids and risk of heart failure: a meta-analysis. *Clinical nutrition*, 31(6):846–853.

Dodd, K. W., Guenther, P. M., Freedman, L. S., Subar, A. F., Kipnis, V., Midthune, D., Tooze, J. A., and Krebs-Smith, S. M. (2006). Statistical methods for estimating usual intake of nutrients and foods: A review of the theory. *Journal of the American Dietetic Association*, 106(10):1640–1650.

Estruch, R., Ros, E., Salas-Salvadó, J., Covas, M.-I., Corella, D., Arós, F., Gómez-Gracia, E., Ruiz-Gutiérrez, V., Fiol, M., Lapetra, J., Lamuela-Raventos, R. M., Serra-Majem, L., Pintó, X., Basora, J., Muñoz, M. A., Sorlí, J. V., Martínez, J. A., Fitó, M., Gea, A., Hernán, M. A., and Martínez-González, M. A. (2018a). Primary prevention of cardiovascular disease with a mediterranean diet supplemented with extra-virgin olive oil or nuts. *New England Journal of Medicine*, 378(25):e34. PMID: 29897866.

Estruch, R., Ros, E., Salas-Salvadó, J., Covas, M.-I., Corella, D., Arós, F., Gómez-Gracia, E., Ruiz-Gutiérrez, V., Fiol, M., Lapetra, J., Lamuela-Raventos, R. M., Serra-Majem, L., Pintó, X., Basora, J., Muñoz, M. A., Sorlí, J. V., Martínez, J. A., and Martínez-González,

M. A. (2013). Primary prevention of cardiovascular disease with a mediterranean diet. *New England Journal of Medicine*, 368(14):1279–1290.

Estruch, R., Ros, E., Salas-Salvadó, J., Covas, M.-I., Corella, D., Arós, F., Gómez-Gracia, E., Ruiz-Gutiérrez, V., Fiol, M., Lapetra, J., Lamuela-Raventos, R. M., Serra-Majem, L., Pintó, X., Basora, J., Muñoz, M. A., Sorlí, J. V., Martínez, J. A., and Martínez-González, M. A. (2018b). Retraction and republication: Primary prevention of cardiovascular disease with a mediterranean diet. *New England Journal of Medicine*, 378(25):2441–2442. PMID: 29897867.

Feeney, E., O'Brien, S., Scannell, A., Markey, A., and Gibney, E. R. (2011). Genetic variation in taste perception: does it have a role in healthy eating? *Proceedings of the Nutrition Society*, 70(01):135–143.

Fieuws, S. and Verbeke, G. (2006). Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics*, 62(2):424–431.

Fieuws, S., Verbeke, G., Boen, G., and Delecluse, C. (2006). High dimensional multivariate mixed models for binary questionnaire data. *Applied Statistics*, 55:449–460.

Fitt, E., Mak, T. N., Stephen, A. M., Prynne, C., Roberts, C., Swan, G., and Farron-Wilson, M. (2010). Disaggregating composite food codes in the UK National Diet and Nutrition Survey food composition databank. *Eur J Clin Nutr*, 64(S3):32–36.

Food and Agriculture Organization of the United Nations, United Nations University, World Health Organization (2004). *Human Energy Requirements: Report of a Joint FAO/WHO/ONU Expert Consultation: Rome, 17–24 October 2001*. Food and nutrition technical report series. Food & Agricultural Org.

Fraser, G. E. (2003). A search for truth in dietary epidemiology. *The American Journal of Clinical Nutrition*, 78(3):521–525.

Freedman, L. S., Commins, J. M., Moler, J. E., Arab, L., Baer, D. J., Kipnis, V., Midthune, D., Moshfegh, A. J., Neuhouser, M. L., Prentice, R. L., Schatzkin, A., Spiegelman, D.,

Subar, A. F., Tinker, L. F., and Willett, W. C. (2014). Pooled results from 5 validation studies of dietary self-report instruments using recovery biomarkers for energy and protein intake. *American Journal of Epidemiology*, 180(2):172–188.

Freedman, L. S., Guenther, P. M., Krebs-Smith, S. M., Dodd, K. W., and Midthune, D. (2010). A population's distribution of healthy eating index-2005 component scores can be estimated when more than one 24-hour recall is available. *The Journal of Nutrition*, 140(8):1529–1534.

Frost, C. and Thompson, S. G. (2000). Correcting for regression dilution bias: comparison of methods for a single predictor variable. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 163(2):173–189.

Fuller, W. A. (1980). Properties of some estimators for the errors—in—variables model. *The Annals of Statistics*, 8:407.

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

Gadgil, M. D., Appel, L. J., Yeung, E., Anderson, C. A. M., Sacks, F. M., and Miller, E. R. (2013). The effects of carbohydrate, unsaturated fat, and protein intake on measures of insulin sensitivity: Results from the omniheart trial. *Diabetes Care*, 36(5):1132–1137.

Galobardes, B., Morabia, A., and Bernstein, M. S. (2001). Diet and socioeconomic position: does the use of different indicators matter? *International Journal of Epidemiology*, 30(2):334–340.

Garber, A., Handelsman, Y., Einhorn, D., Bergman, D., Bloomgarden, Z., Fonseca, V., Timothy Garvey, W., Gavin Iii, J., Grunberger, G., Horton, E., Jellinger, P., Jones, K., Lebovitz, H., Levy, P., McGuire, D., Moghissi, E., and Nesto, R. (2008). Diagnosis and management of prediabetes in the continuum of hyperglycemia—when do the risks of diabetes begin? a consensus statement from the american college of endocrinology and the american association of clinical endocrinologists. *Endocrine Practice*, 14(7):933–946.

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, 22(2):153–164.

Geys, H., Molenberghs, G., and Ryan, L. M. (1999). Pseudolikelihood modeling of multivariate outcomes in developmental toxicology. *Journal of the American Statistical Association*, 94(447):734–745.

Gleser, L. J. (1990). *Statistical Analysis of Measurement Error Models and Application*, chapter Improvement of the naive approach to estimation in non-linear errors-in-variables regression models. American Mathematics Society.

Grieger, J. A., Wycherley, T. P., Johnson, B. J., and Golley, R. K. (2016). Discrete strategies to reduce intake of discretionary food choices: a scoping review. *The International Journal of Behavioral Nutrition and Physical Activity*, 13:57.

Groth, M. V., Fagt, S., and Brøndsted, L. (2001). Social determinants of dietary habits in denmark. *Eur J Clin Nutr*, 55(11):959–966.

Guenther, P. M., Casavale, K. O., Kirkpatrick, S. I., Reedy, J., Hiza, H. A. B., Kuczynski, K. J., Kahle, L. L., and Krebs-Smith, S. M. (2013). Update of the healthy eating index: Hei-2010. *Journal of the Academy of Nutrition and Dietetics*, 113(4).

Guenther, P. M., Dodd, K. W., Reedy, J., and Krebs-Smith, S. M. (2006). Most americans eat much less than recommended amounts of fruits and vegetables. *Journal of the American Dietetic Association*, 106(9):1371–1379.

Guenther, P. M., Kott, P. S., and Carriquiry, A. L. (1997). Development of an approach for estimating usual nutrient intake distributions at the population level. *The Journal of Nutrition*, 127(6):1106–1112.

Guolo, A. (2011). Pseudo-likelihood inference for regression models with misclassified and mismeasured variables. *Statistica Sinica*, 21(4):1639–1663.

Hamra, G., MacLehose, R., and Richardson, D. (2013). Markov chain monte carlo: an introduction for epidemiologists. *International Journal of Epidemiology*, 42.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441.

Huang, J., Wang, X., and Zhang, Y. (2017). Specific types of alcoholic beverage consumption and risk of type 2 diabetes: A systematic review and meta-analysis. *Journal of Diabetes Investigation*, 8(1):56–68.

Institute of Medicine, Food and Nutrition Board (2003). Dietary reference intakes: Applications in dietary planning.

Ioannidis, J. (2013). Implausible results in human nutrition research. *BMJ*, 347.

Ivanova, A., Molenberghs, G., and Verbeke, G. (2017). Fast and highly efficient pseudo-likelihood methodology for large and complex ordinal data. *Statistical Methods in Medical Research*, 26(6):2758–2779.

Jebb, S. A., Lovegrove, J. A., Griffin, B. A., Frost, G. S., Moore, C. S., Chatfield, M. D., Bluck, L. J., Williams, C. M., and Sanders, T. A. (2010). Effect of changing the amount and type of fat and carbohydrate on insulin sensitivity and cardiovascular risk: the risck (reading, imperial, surrey, cambridge, and kings) trial. *The American Journal of Clinical Nutrition*, 92(4):748–758.

Jelleyman, C., Yates, T., O'Donovan, G., Gray, L. J., King, J. A., Khunti, K., and Davies, M. J. (2015). The effects of high-intensity interval training on glucose regulation and insulin resistance: a meta-analysis. *Obesity Reviews*, 16(11):942–961.

Jeon, C. Y., Lokken, R. P., Hu, F. B., and van Dam, R. M. (2007). Physical activity of moderate intensity and risk of type 2 diabetes: A systematic review. *Diabetes Care*, 30(3):744–752.

Katzmarzyk, P. T., Janssen, I., and Ardern, C. I. (2003). Physical inactivity, excess adiposity and premature mortality. *Obesity Reviews*, 4(4):257–290.

Kent, S., Fusco, F., Gray, A. R., Jebb, S. A., Cairns, B. J., and Mihaylova, B. (2017). Body mass index and healthcare costs: a systematic literature review of individual participant data studies. *Obesity Reviews*, 18(8):869–879.

Keogh, R. (2011). Allowing for never and episodic consumers when correcting for error in food record measurements of dietary intake. *Biostatistics*, 12(4).

Keogh, R. and White, I. (2014). A toolkit for measurement error correction, with a focus on nutritional epidemiology. *Statistics in Medicine*, 33(12):2137–2155.

Kim, U. K. and Drayna, D. (2005). Genetics of individual differences in bitter taste perception: lessons from the ptc gene. *Clinical Genetics*, 67(4):275–280.

Kipnis, V., Midthune, D., Buckman, D. W., Dodd, K. W., Guenther, P. M., Krebs-Smith, S. M., Subar, A. F., Tooze, J. A., Carroll, R. J., and Freedman, L. S. (2009). Modeling data with excess zeros and measurement error: Application to evaluating relationships between episodically consumed foods and health outcomes. *Biometrics*, 65(4):1003–1010.

Kipnis, V., Midthune, D., Freedman, L., Bingham, S. A., Day, N. E., Riboli, E., Ferrari, P., and Carroll, R. J. (2002). Bias in dietary-report instruments and its implications for nutritional epidemiology. *Public Health Nutrition*, 5(6a):915–923.

Kipnis, V., Subar, A. F., Midthune, D., Freedman, L. S., Ballard-Barbash, R., Troiano, R. P., Bingham, S. A., Schoeller, D. A., Schatzkin, A., and Carroll, R. J. (2003). Structure of dietary measurement error: Results of the open biomarker study. *American Journal of Epidemiology*, 158(1):14–21.

Kloner, R. A. and Rezkalla, S. H. (2007). To drink or not to drink? that is the question. *Circulation*, 116(11):1306–1317.

Knott, C., Bell, S., and Britton, A. (2015). Alcohol consumption and the risk of type 2 diabetes: A systematic review and dose-response meta-analysis of more than 1.9 million individuals from 38 observational studies. *Diabetes Care*, 38(9):1804–1812.

Koloverou, E., Panagiotakos, D. B., Pitsavos, C., Chrysohoou, C., Georgousopoulou, E. N., Metaxa, V., and Stefanadis, C. (2015). Effects of alcohol consumption and the metabolic syndrome on 10-year incidence of diabetes: The attica study. *Diabetes & Metabolism*, 41(2):152–159.

Koppes, L. L. J., Dekker, J. M., Hendriks, H. F. J., Bouter, L. M., and Heine, R. J. (2005). Moderate alcohol consumption lowers the risk of type 2 diabetes. *A meta-analysis of prospective observational studies*, 28(3):719–725.

Kröller, K. and Warschburger, P. (2009). Maternal feeding strategies and child's food intake: considering weight and demographic influences using structural equation modeling. *International Journal of Behavioral Nutrition and Physical Activity*, 6(1):1–9.

Lallukka, T., Laaksonen, M., Rahkonen, O., Roos, E., and Lahelma, E. (2006). Multiple socio-economic circumstances and healthy food habits. *Eur J Clin Nutr*, 61(6):701–710.

Larribe, F. and Fearnhead, P. (2011). On composite likelihoods in statistical genetics. *Statistica Sinica*, 21:43–69.

Ledoit, O. and Wolf, M. (2015). Spectrum estimation: A unified framework for covariance matrix estimation and pca in large dimensions. *Journal of Multivariate Analysis*, 139:360–384.

Lele, S. and L. Taper, M. (2002). A composite likelihood approach to (co)variance components estimation. *Journal of Statistical Planning and Inference*, 103:117–135.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Lindley, D. V. (1953). Estimation of a functional relationship. *Biometrika*, 40:47.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary mathematics*, 80(1):221–239.

Liu, L., Cowen, M. E., Strawderman, R. L., and Shih, Y.-C. T. (2010). A flexible two-part random effects model for correlated medical costs. *Journal of health economics*, 29(1):110–123.

Lobstein, T., Baur, L., and Uauy, R. (2004). Obesity in children and young people: a crisis in public health. *Obesity Reviews*, 5:4–85.

Lusk, J. L., Roosen, J., and Shogren, J. (2013). *The Oxford Handbook of the Economics of Food Consumption and Policy*. OUP Oxford.

Lyles, R. H. and Kupper, L. L. (1997). A detailed evaluation of adjustment methods for multiplicative measurement error in linear regression with applications in occupational epidemiology. *Biometrics*, 53.

Madansky, W. (1959). The fitting of straight lines when both variables are subject to error. *Journal of the American Statistical Association*, 54:173.

Maguire, E. R. and Monsivais, P. (2015). Socio-economic dietary inequalities in uk adults: an updated picture of key food groups and nutrients from national surveillance data. *The British Journal of Nutrition*, 113(1):181–189.

Malik, V. S., Popkin, B. M., Bray, G. A., Després, J.-P., and Hu, F. B. (2010). Sugar-sweetened beverages, obesity, type 2 diabetes mellitus, and cardiovascular disease risk. *Circulation*, 121(11):1356–1364.

Mamdani, M., Sykora, K., Li, P., Normand, S.-L. T., Streiner, D. L., Austin, P. C., Rochon, P. A., and Anderson, G. M. (2005). Reader's guide to critical appraisal of cohort studies: 2. assessing potential for confounding. *BMJ*, 330(7497):960–962.

Mankiw, N. G. and Taylor, M. P. (2007). *Macroeconomics*. Worth Publishers, sixth edition.

McCulloch, C. E. and Neuhaus, J. M. (2011). Misspecifying the shape of a random effects distribution: Why getting it wrong may not matter. 26(3):388–402.

McGowan, L., Croker, H., Wardle, J., and Cooke, L. J. (2012). Environmental and individual determinants of core and non-core food and drink intake in preschool-aged children in the united kingdom. *Eur J Clin Nutr*, 66(3):322–328.

McKeown, N. M., Day, N. E., Welch, A. A., Runswick, S. A., Luben, R. N., Mulligan, A. A., McTaggart, A., and Bingham, S. A. (2001). Use of biological markers to validate self-reported dietary intake in a random sample of the european prospective investigation into cancer united kingdom norfolk cohort. *The American Journal of Clinical Nutrition*, 74(2):188–196.

McKinnon, L., Giskes, K., and Turrell, G. (2014). The contribution of three components of nutrition knowledge to socio-economic differences in food purchasing choices. *Public Health Nutrition*, 17(08):1814–1824.

Mendez, M. A., Popkin, B. M., Buckland, G., Schroder, H., Amiano, P., Barricarte, A., Huerta, J.-M., Quirós, J. R., Sánchez, M.-J., and González, C. A. (2011). Alternative methods of accounting for underreporting and overreporting when measuring dietary intake-obesity relations. *American Journal of Epidemiology*, 173(4):448–458.

Mennella, J. A. (2014). Ontogeny of taste preferences: basic biology and implications for health. *The American Journal of Clinical Nutrition*, 99(3):704–711.

Mennella, J. A., Pepino, M. Y., and Reed, D. R. (2005). Genetic and environmental determinants of bitter perception and sweet preferences. *Pediatrics*, 115(2):216–222.

Mifflin, M. D., St Jeor, S. T., Hill, L. A., Scott, B. J., Daugherty, S. A., and Koh, Y. O. (1990). A new predictive equation for resting energy expenditure in healthy individuals. *The American Journal of Clinical Nutrition*, 51(2):241–247.

Mindell, J. (2014). Appendix V: Measuring physical activity in adults using the Recent Physical Activity Questionnaire (RPAQ), National Diet and Nutrition Survey: results from

years 1 to 4 (combined) of the rolling programme for 2008 and 2009 to 2011 and 2012. Report.

Mozaffarian, D. and Willett, W. C. (2007). Trans fatty acids and cardiovascular risk: a unique cardiometabolic imprint? *Curr Atheroscler Rep*, 9:486.

Muthen, B. O. and Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology*, 25:267–316.

National Cancer Institute (2015). Usual dietary intakes: The NCI method.

Navarro-Allende, A., Khataan, N., and El-Sohemy, A. (2008). Impact of genetic and environmental determinants of taste with food preferences in older adults. *Journal of Nutrition For the Elderly*, 27(3-4):267–276.

Nelson, M., Black, A. E., Morris, J. A., and Cole, T. J. (1989). Between- and within-subject variation in nutrient intake from infancy to old age: estimating the number of days required to rank dietary intakes with desired precision. *The American Journal of Clinical Nutrition*, 50(1):155–67.

Neuhouser, M. L., Tinker, L., Shaw, P. A., Schoeller, D., Bingham, S. A., Horn, L. V., Beresford, S. A. A., Caan, B., Thomson, C., Satterfield, S., Kuller, L., Heiss, G., Smit, E., Sarto, G., Ockene, J., Stefanick, M. L., Assaf, A., Runswick, S. A., and Prentice, R. L. (2008). Use of recovery biomarkers to calibrate nutrient consumption self-reports in the women's health initiative. *American Journal of Epidemiology*, 167(10):1247–1259.

Nicklaus, S. (2011). Children's acceptance of new foods at weaning. role of practices of weaning and of food sensory properties. *Appetite*, 57(3):812–815.

Normand, S.-L. T., Sykora, K., Li, P., Mamdani, M., Rochon, P. A., and Anderson, G. M. (2005). Readers guide to critical appraisal of cohort studies: 3. analytical strategies to reduce confounding. *BMJ*, 330(7498):1021–1023.

Novaković, R., Cavelaars, A., Geelen, A., Nikolić, M., Altaba, I. I., Viñas, B. R., Ngo, J.,

Golsorkhi, M., Medina, M. W., Brzozowska, A., Szczecinska, A., de Cock, D., Vansant, G., Renkema, M., Majem, L. S., Moreno, L. A., Glibetić, M., Gurinović, M., van't Veer, P., and de Groot, L. C. (2014). Review article socio-economic determinants of micronutrient intake and status in europe: a systematic review. *Public Health Nutrition*, 17(5):1031–1045.

Nusser, S. M., Carriquiry, A. L., Dodd, K. W., and Fuller, W. A. (1996). A semiparametric transformation approach to estimating usual daily intake distributions. *Journal of the American Statistical Association*, 91(436):1440–1449.

Nusser, S. M., Fuller, W. A., and Guenther, P. M. (1987). *Estimating usual dietary intake distributions: Adjusting for measurement error and non-normality in 24-hour food intake data. Survey Measurement and Process Quality*. NY: Wiley.

Olsen, M. K. and Schafer, J. L. (2001). A two-part random-effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association*, 96(454):730–745.

Pfeffermann, D. (2011). Modelling of complex survey data: Why model? why is it a problem? how can we approach it. *Survey Methodology*, 37(2):115–136.

Polonsky, K. S. (2012). The past 200 years in diabetes. *New England Journal of Medicine*, 367(14):1332–1340.

Pomerleau, J., Pederson, L. L., Østbye, T., Speechley, M., and Speechley, K. N. (1997). Health behaviours and socio-economic status in ontario, canada. *European Journal of Epidemiology*, 13(6):613–622.

Public Health England (2014). National Diet and Nutrition Survey: results from Years 1 to 4 (combined) of the rolling programme for 2008 and 2009 to 2011 and 2012. Report.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ranjit, N., Wilkinson, A. V., Lytle, L. M., Evans, A. E., Saxton, D., and Hoelscher, D. M. (2015). Socioeconomic inequalities in children's diet: the role of the home food environment. *International Journal of Behavioral Nutrition and Physical Activity*, 12(1):1–9.

Rasmussen, M., Krølner, R., Klepp, K. I., Lytle, L., Brug, J., and Bere, E. (2006). Determinants of fruit and vegetable consumption among children and adolescents: a review of the literature. *Int J Behav Nutr Phys Act*, 3.

Raykov, T., Marcoulides, G. A., and Li, T. (2017). On the fallibility of principal components in research. *Educational and psychological measurement*, 77:165–178.

Rochon, P. A., Gurwitz, J. H., Sykora, K., Mamdani, M., Streiner, D. L., Garfinkel, S., Normand, S.-L. T., and Geoffrey, M. (2005). Reader's guide to critical appraisal of cohort studies: 1. role and design. *BMJ*, 330(7496):895–897.

Roerecke, M. and Rehm, J. (2012). The cardioprotective association of average alcohol consumption and ischaemic heart disease: a systematic review and meta-analysis. *Addiction (Abingdon, England)*, 107(7):1246–1260.

Rosenbaum, M., Ravussin, E., Matthews, D. E., Gilker, C., Ferraro, R., Heymsfield, S. B., Hirsch, J., and Leibel, R. L. (1996). A comparative study of different means of assessing long-term energy expenditure in humans. *American Journal of Physiology*, 270(3):496–504.

Rosner, B., Spiegelman, D., and Willett, W. C. (1990). Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. *American Journal of Epidemiology*, 132(4):734–745.

Rosner, B., Willett, W. C., and Spiegelman, D. (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in Medicine*, 8(9):1051–1069.

Rozin, P., Remick, A. K., and Fischler, C. (2011). Broad themes of difference between

french and americans in attitudes to food and other life domains: Personal versus communal values, quantity versus quality, and comforts versus joys. *Frontiers in Psychology*, 2:177.

Rutishauser, I. and Black, A. (2002). *Measuring food intake*. Oxford: Blackwell Science, first edition edition.

Rydén, P. J. and Hagfors, L. (2011). Diet cost, diet quality and socio-economic position: how are they related and what contributes to differences in diet costs? *Public Health Nutrition*, 14(9):1680–1692.

Scagliarini, M. (2011). Multivariate process capability using principal component analysis in the presence of measurement errors. *AStA Advances in Statistical Analysis*, 95(2).

Schofield, W. N., Schofield, C., and James, W. P. T. (1985). *Basal metabolic rate: review and prediction, together with an annotated bibliography of source material*. Human nutrition. Clinical nutrition, v. 39C, suppl. 1. J. Libbey.

School for Public Health Research (2013). Public Health Practice Evaluation Scheme (PHPES).

Schrieks, I. C., Heil, A. L. J., Hendriks, H. F. J., Mukamal, K. J., and Beulens, J. W. J. (2015). The effect of alcohol consumption on insulin sensitivity and glycemic status: A systematic review and meta-analysis of intervention studies. *Diabetes Care*, 38(4):723–732.

Self, S. G. and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. 82(398):605–610.

Sempos, C. T., Johnson, N. E., Smith, E. L., and Gilligan, C. (1985). Effects of intraindividual and interindividual variation in repeated dietary records. *American Journal of Epidemiology*, 121(1):120–130.

Shepherd, R. and Raats, M. (2013). *The Psychology of Food Choice*. OUP Oxford.

Sinha, R. and Jastreboff, A. M. (2013). Stress as a common risk factor for obesity and addiction. *Biological psychiatry*, 73(9):827–835.

Smeets, P. A. M., Erkner, A., and de Graaf, C. (2010). Cephalic phase responses and appetite. *Nutrition Reviews*, 68(11):643–655.

Smith, V. A., Neelon, B., Preisser, J. S., and Maciejewski, M. L. (2015). A marginalized two-part model for longitudinal semicontinuous data. *Statistical Methods in Medical Research*.

Stampfer, M. J., Colditz, G. A., Willett, W. C., Mamson, J. E., Arky, R. A., Hennekens, C. H., and Speizer, F. E. (1988). A prospective study of moderate alcohol drinking and risk of diabetes in women. *American Journal of Epidemiology*, 128(3):549–558.

Su, L., Tom, B. D. M., and Farewell, V. T. (2009). Bias in 2-part mixed models for longitudinal semicontinuous data. *Biostatistics*, 10(2):374–389.

Su, L., Tom, B. D. M., and Farewell, V. T. (2011). A likelihood-based two-part marginal model for longitudinal semi-continuous data. *Statistical Methods in Medical Research*, 24(2):194–205.

Subar, A. F., Kipnis, V., Troiano, R. P., Midthune, D., Schoeller, D. A., Bingham, S. A., Sharbaugh, C. O., Trabulsi, J., Runswick, S. A., Ballard-Barbash, R., Sunshine, J., and Schatzkin, A. (2003). Using intake biomarkers to evaluate the extent of dietary misreporting in a large sample of adults: The open study. *American Journal of Epidemiology*, 158(1):1–13.

Tepper, B. J., Banni, S., Melis, M., Crnjar, R., and Tomassini, B. I. (2014). Genetic sensitivity to the bitter taste of 6-n-propylthiouracil (prop) and its association with physiological mechanisms controlling body mass index (bmi). *Nutrients*, 6(9):3363–3381.

Tepper, B. J., White, E. A., Koelliker, Y., Lanzara, C., D'Adamo, P., and Gasparini, P. (2009). Genetic variation in taste sensitivity to 6-n-propylthiouracil and its relationship

to taste perception and food selection. *Annals of the New York Academy of Sciences*, 1170(1):126–139.

Tomiyama, A. J., Schamarek, I., Lustig, R. H., Kirschbaum, C., Puterman, E., Havel, P. J., and Epel, E. S. (2012). Leptin concentrations in response to acute stress predict subsequent intake of comfort foods. *Physiology & behavior*, 107(1):34–39.

Tooze, J. A., Grunwald, G. K., and Jones, R. H. (2002). Analysis of repeated measures data with clumping at zero. *Statistical Methods in Medical Research*, 11(4):341–355.

Tooze, J. A., Kipnis, V., Buckman, D. W., Carroll, R. J., Freedman, L. S., Guenther, P. M., Krebs-Smith, S. M., Subar, A. F., and Dodd, K. W. (2010). A mixed-effects model approach for estimating the distribution of usual intake of nutrients: The NCI method. *Statistics in medicine*, 29(27).

Tooze, J. A., Midthune, D., Dodd, K. W., Freedman, L. S., Krebs-Smith, S. M., Subar, A. F., Guenther, P. M., Carroll, R. J., and Kipnis, V. (2006). A new method for estimating the usual intake of episodically-consumed foods with application to their distribution. *Journal of the American Dietetic Association*, 106(10):1575–1587.

Tooze, J. A., Subar, A. F., Thompson, F. E., Troiano, R., Schatzkin, A., and Kipnis, V. (2004). Psychosocial predictors of energy underreporting in a large doubly labeled water study. *The American Journal of Clinical Nutrition*, 79(5):795–804.

Tuomilehto, J., Lindström, J., Eriksson, J. G., Valle, T. T., Hämäläinen, H., Ilanne-Parikka, P., Keinänen-Kiukaanniemi, S., Laakso, M., Louheranta, A., Rastas, M., Salminen, V., Aunola, S., Cepaitis, Z., Moltchanov, V., Hakumäki, M., Mannelin, M., Martikkala, V., Sundvall, J., and Uusitupa, M. (2001). Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. *New England Journal of Medicine*, 344(18):1343–1350.

Turrell, G., Hewitt, B., Patterson, C., and Oldenburg, B. (2003). Measuring socio-economic

position in dietary research: is choice of socio-economic indicator important? *Public Health Nutrition*, 6(02):191–200.

US Department of Health and Human Services (2015). 2015-2020 dietary guidelines for americans. Report.

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21(1):5–42.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25.

Willett, W. C. (2012). Dietary fats and coronary heart disease. *Journal of Internal Medicine*, 272(1):13–24.

World Health Organization (2016). Global report on diabetes.

Xiao, X., White, E. P., Hooten, M. B., and Durham, S. L. (2011). On the use of log-transformation vs. nonlinear regression for analyzing biological power laws. *Ecology*, 92(10):1887–1894.

Yau, Y. H. C. and Potenza, M. N. (2013). Stress and eating behaviors. *Minerva endocrinologica*, 38(3):255–267.

Zeger, S. L. and Karim, M. R. (1991). Generalized linear models with random effects; a gibbs sampling approach. *Journal of the American Statistical Association*, 86(413):79–86.

Zhang, B., Liu, W., Zhang, H., Chen, Q., and Zhang, Z. (2016). Composite likelihood and maximum likelihood methods for joint latent class modeling of disease prevalence and high-dimensional semicontinuous biomarker data. *Computational Statistics*, 31(2):425–449.

Zhang, S., Midthune, D., Guenther, P. M., Krebs-Smith, S. M., Kipnis, V., Dodd, K. W., Buckman, D. W., Tooze, J. A., Freedman, L. S., and Carroll, R. J. (2011). A new

multivariate measurement error model with zero-inflated dietary data, and its application to dietary assessment. *The Annals of Applied Statistics*, pages 1456–1487.