BAYESIAN METHODS IN NUTRITION EPIDEMIOLOGY AND REGRESSION-BASED PREDICTIVE MODELS IN HEALTHCARE

A Dissertation

by

SAIJUAN ZHANG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

December 2010

Major Subject: Statistics

BAYESIAN METHODS IN NUTRITION EPIDEMIOLOGY AND REGRESSION-BASED PREDICTIVE MODELS IN HEALTHCARE

A Dissertation

by

SAIJUAN ZHANG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Co-Chairs of Committee,	Raymond J. Carroll
	Jianhua Huang
Committee Members,	Bani Mallick
	Joanne Lupton
	Susan Krebs-Smith
	Victor Kipnis
Head of Department,	Simon J. Sheather

December 2010

Major Subject: Statistics

ABSTRACT

Bayesian Methods in Nutrition Epidemiology and

Regression-based Predictive Models in Healthcare. (December 2010)

Saijuan Zhang, B.S., Southeast University;

M.A., University of Oklahoma

Co–Chairs of Advisory Committee: Dr. Raymond J. Carroll
Dr. Jianhua Huang

This dissertation has mainly two parts. In the first part, we propose a bivariate nonlinear multivariate measurement error model to understand the distribution of dietary intake and extend it to a multivariate model to capture dietary patterns in nutrition epidemiology. In the second part, we propose regression-based predictive models to accurately predict surgery duration in healthcare.

Understanding the distribution of episodically consumed dietary components is an important problem in public health. Short-term measurements of episodically consumed dietary components are zero-inflated skewed distributions. So-called two-part models have been developed for such data. However, there is much greater public health interest in the usual intake adjusted for caloric intake. Recently a nonlinear mixed effects model has been developed and fit by maximum likelihood using nonlinear mixed effects programs. However, the fitting is slow and unstable. We develop a Monte-Carlo-based fitting method in Chapter II. We demonstrate numerically that our methods lead to increased speed of computation, converge to reasonable solutions, and have the flexibility to be used in either a frequentist or a Bayesian manner. Diet consists of numerous foods, nutrients and other components, each of which have distinctive attributes. Increasingly nutritionists are interested in exploring

them collectively to capture overall dietary patterns. We thus extend the bivariate model described in Chapter III to multivariate level. We use survey-weighted MCMC computations to fit the model, with uncertainty estimation coming from balanced repeated replication. The methodology is illustrated through an application of estimating the population distribution of the Healthy Eating Index-2005 (HEI-2005), a multi-component dietary quality index , among children aged 2-8 in the United States.

The second part of this dissertation is to accurately predict surgery duration. Prior research has identified the current procedural terminology (CPT) codes as the most important factor when predicting surgical case durations but there has been little reporting of a general predictive methodology using it effectively. In Chapter IV, we propose two regression-based predictive models. However, the naively constructed design matrix is singular. We thus devise a systematic procedure to construct a full-ranked design matrix. Using surgical data from a central Texas hospital, we compare the proposed models with a few benchmark methods and demonstrate that our models lead to a remarkable reduction in prediction errors.

To My Family

ACKNOWLEDGMENTS

I would like to thank my committee co-chairs, Dr. Carroll and Dr. Huang, and my committee members, Dr. Mallick, Dr. Lupton, Dr. Krebs-Smith and Dr. Kipnis, for their guidance and support throughout the course of this research.

Thanks also to my friends and colleagues and the department faculty and staff for making my time at Texas A&M University a great experience. I also want to extend my gratitude to the National Cancer Institute which supported my research.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

LIST OF TABLES

TABLE Page

LIST OF FIGURES

CHAPTER I

INTRODUCTION

There has been great public health interest in estimating usual, i.e., long-term average, intake of episodically consumed dietary components that are not consumed daily by everyone, e.g., fish, red meat and whole grains. Short-term measurements of episodically consumed dietary components have zero-inflated skewed distributions. So-called two-part models have been developed for such data, in order to correct for measurement error due to within-person variation and to estimate the distribution of usual intake of the dietary component in the univariate case. However, there is arguably much greater public health interest in the usual intake of an episodically consumed dietary component adjusted for caloric intake, e.g., ounces of whole grains per 1000 kilo-calories, which reflects usual dietary composition and adjusts for different total amounts of caloric intake. Because of this public health interest, it is important to have models to fit such data, and it is important that the model-fitting methods can be applied across the broad range of episodically consumed dietary components. We have recently addressed the first issue by developing a nonlinear mixed effects model (Kipnis et al., 2010a), and have fit it by maximum likelihood using nonlinear mixed effects programs and methodology (the SAS NLMIXED procedure). Maximum likelihood fitting of such a nonlinear mixed model is generally slow because of 3-dimensional adaptive Gaussian quadrature, and there are times when the programs either fail to converge or converge to models with a singular covariance matrix. For these reasons we develop a Monte-Carlo computation of fitting this model, which allows for both frequentist and Bayesian inference. There are technical challenges to

---

This dissertation follows the style of *Biometrics*.

developing this solution because one of the covariance matrices in the model is patterned, having structural zeros. Our main application is to the National Institutes of Health (NIH)-AARP Diet and Health Study, where we illustrate our methods for modeling the energy-adjusted usual intake of fish and whole grains. We demonstrate numerically that our methods lead to increased speed of computation, converge to reasonable solutions, and have the flexibility to be used in either a frequentist or a Bayesian manner.

In the United States the preferred method of obtaining dietary intake data is the 24-hour dietary recall, yet the measure of most interest is usual or long-term average intake, which is impossible to measure. Thus, usual dietary intake is assessed with considerable measurement error. Also, diet represents numerous foods, nutrients and other components, each of which have distinctive attributes. Sometimes, it is useful to examine intake of these components separately, but increasingly nutritionists are interested in exploring them collectively to capture overall dietary patterns. Consumption of these components varies widely: some are consumed daily by almost everyone on every day, while others are episodically consumed so that 24-hour recall data are zero-inflated. In addition, they are often correlated with each other. Finally, it is often preferable to analyze the amount of a dietary component relative to the amount of energy (calories) in a diet because dietary recommendations often vary with energy level. The quest to understand overall dietary patterns of usual intake has to this point reached a standstill. There are no statistical methods or models available to model such complex multivariate data with its measurement error and zero inflation. The second project proposes the first such model, and it proposes the first workable solution to fit such a model. After describing the model, we use survey-weighted MCMC computations to fit the model, with uncertainty estimation coming from balanced repeated replication. The methodology is illustrated through an ap-

plication to estimating the population distribution of the Healthy Eating Index-2005 (HEI-2005), a multi-component dietary quality index involving ratios of interrelated dietary components to energy, among children aged 2-8 in the United States. We pose a number of interesting questions about the HEI-2005 and provide answers that were not previously within the realm of possibility, and we indicate ways that our approach can be used to answer other questions of importance to nutritional science and public health.

Efficient utilization of existing resources is crucial for cost containment in medical institutions. Accurately predicting surgery duration will help improve the utilization of indispensable surgical resources such as surgeons, nurses, and operating rooms. Prior research has identified the current procedural terminology (CPT) codes as the most important factor when predicting surgical case durations. Yet there has been little reporting of a general predictive methodology that can effectively extract information from multiple CPT codes. In the third project, we propose two regression-based predictive models, a linear regression and a log-linear regression. To perform these regression analysis, a full-ranked design matrix based on CPT code inclusions in the surgical cases needs to be constructed. However, some CPT codes only appear in conjunction with others, and as a result, naively constructed design matrix is ill-conditioned (i.e. singular). We devise a systematic procedure to construct a full-ranked design matrix by sifting out the CPT codes without any predictive power while useful information is retained as much as possible. Our proposed models can be applied in general situations where a surgery can have any number of CPT codes and any combination of CPT codes. Using surgical data from a central Texas hospital, we compare the proposed models with a few benchmark methods. The comparison demonstrates that using the proposed predictive models leads to a remarkable reduction in prediction errors.

CHAPTER II

A BIVARIATE MEASUREMENT ERROR MODEL FOR EPISODICALLY CONSUMED DIETARY COMPONENTS

A. Introduction

This project is about the important public health problem of understanding the distribution of episodically consumed dietary component intakes in terms of their energy-adjusted amounts, and relating this to diet-disease relationships. Before commenting in more detail, we first discuss the literature for simpler problems that are also of interest.

In nutritional surveillance and nutritional epidemiology, there is considerable interest in understanding the distribution of usual dietary intake, which is defined as long-term daily average intake. In addition, of interest is the regression of this intake on measured covariates, which is needed to correct diet-disease relationships for measurement error in assessing diet. If the dietary component of interest is ubiquitously consumed, as most nutrients are, the data are continuously distributed and methods are well-established for solving both problems. See for example Nusser et al. (1997) for surveillance and Carroll et al. (2006) for measurement error modeling.

Another class of dietary components is those which are episodically consumed, as is true of most foods, e.g., fish, red meat, dark green vegetables, whole grains. When consumption is measured by a short-term instrument such as a 24 hour recall, hereafter denoted by 24hr, the episodic nature of these dietary components means that their reported intake may either equal zero on a non-consumption day, or is positive on a day the component is consumed. In many studies, non-consumption days predominate for several episodically consumed foods of interest. For exam-

ple, in our data example, for fish and whole grains, 65% and 12% reported no consumption on either day, respectively. Thus, data on episodically consumed dietary components are zero-inflated data with measurement error. Recently, Tooze et al. (2006) for nutritional surveillance and Kipnis et al. (2009) for nutritional epidemiology have reported so-called two-part methods, which are actually nonlinear mixed effects models, for analyzing episodically consumed dietary components in the univariate case. These methods are known commonly as the "NCI method" because many of the co-authors of these papers are members of the National Cancer Institute (NCI), and because SAS routines based upon the NLMIXED procedure are available at http://riskfactor.cancer.gov/diet/usualintakes/, an NCI web site. Other two-part models in different contexts are described for example in Olsen and Schafer (2001), Tooze et al. (2002) and Li et al. (2005).

In this project, we are interested in the more complex public health problem of understanding the usual intake of an episodically consumed dietary component adjusted for energy intake (caloric intake), along with the distribution of usual intake of energy. This is critical because it addresses the issue of dietary component composition, and makes comparable diets of individuals whose usual intakes of energy are very different. As an example, the U.S. Department of Agriculture's Healthy Eating Index-2005 (www.cnpp.usda.gov/HealthyEatingIndex.htm) is a measure of diet quality that assesses conformance to Federal dietary guidance. One component of that index is the number of ounces of whole grains consumed per 1000 kilo-calories: there are other items in the HEI-2005 that deal with episodically consumed dietary components, and all of them are adjusted for energy intake. The data needed to compute such variables are thus the usual intake of the dietary component consumed and the usual amount of calories consumed, and (possibly normalized) ratios of them.

Recently, Kipnis et al. (2010a) have developed a model for an episodically con-

sumed dietary component and energy, see Section B, this chapter. They fit this model using nonlinear mixed effects models with likelihoods computed by adaptive Gaussian quadrature using the SAS procedure NLMIXED. However, as described in Section B and documented in Section D of this chapter, this form of computation can be slow, and can have serious convergence issues. This is extremely problematic, because of the importance of the problem and the fact that solutions will find wide use in the nutrition community, but only if they are numerically stable.

In this project, we take an alternative Markov Chain Monte Carlo (MCMC) approach to computation, which is faster and numerically more stable. There are many good introductory papers reviewing MCMC, such as Casella and George (1992), Chib et al. (1995) and Kass et al. (1998). Effectively, we exploit the well-known fact (Lehmann and Casella, 1998, Chapter 6.8) that in fully parametric regular models of the type we study, Bayesian posterior means of parameters are asymptotically equivalent to their corresponding maximum likelihood estimators. To implement an MCMC approach in our problem, there are technical issues that have to be overcome, including the fact that one of the covariance matrices in the model of Kipnis et al. (2010a) is patterned, with fixed ones and fixed zeros. Besides fitting the model, our main focus in this project is to discuss how to use the parameter estimates to then estimate the distributions of the usual intake of energy and energy-adjusted usual intake of dietary components.

In Section B of this chapter, we describe the model of Kipnis et al. (2010a). In Section B of this chapter, we also briefly outline some of the details of our implementation, although the technical details are given in the Appendix. In Sections C and D of this chapter, we take up the analysis of the NIH-AARP Study of Diet and Health (http://dietandhealth.cancer.gov/) as an illustration of our model and method. Concluding remarks are giving in Chapter V.

B. Data and Model

1. The Data

In practice, the response data often come from repeated 24-hour recalls, hereafter denoted "24hr". Necessarily, due to cost and logistical reasons, the number of recalls is limited, and is rarely greater than 2. In a 24hr, what is observed is whether a dietary component is consumed, and if it is consumed, the reported amount. In addition, the amount of energy reported to be consumed is also available. Thus, for person $i = 1, ..., n$, and for the $k = 1, ..., m_i$ repeats of the 24hr, the data are $\widetilde{\mathbf{Y}}_{\mathbf{ik}} = (Y_{i1k}, ..., Y_{i3k})^{\mathrm{T}}$, where

- Y_{i1k} = Indicator of whether the episodically consumed dietary component is consumed.
- Y_{i2k} = Amount of the dietary component consumed as reported by the 24hr, which equals zero if the dietary component is not consumed.
- Y_{i3k} = Amount of energy consumed as reported by the 24hr.

There are also generally covariates such as age category, ethnic status and in many cases the results of reported intakes from a food frequency questionnaire. We will generically call these covariates $\mathbf{X}$.

2. A Model

Here we describe the latent variable model of Kipnis et al. (2010a). This model is also a nonlinear mixed effects model. As stated above, there are $i = 1, ..., n$ individuals and $k = 1, ..., m_i$ repeats of the 24hr. Also, as stated above, the observed data have three parts, relating to whether the episodically consumed dietary component is consumed,

the amount if it is consumed, and the amount of energy. Also with the observed data, we will have covariates for the individual, generically called $\mathbf{X}$, see below for more precise notation. Finally, Kipnis et al. (2010a) use what are called in nutritional epidemiology "person-specific random effects" which are generically denoted by U, so that individuals actually differ from one another in usual intake when they have the same values of the covariates.

To be more precise, for the i^{th} individual there are covariates $(\mathbf{X_{i1}}, \mathbf{X_{i2}}, \mathbf{X_{i3}})$: $\mathbf{X_{i1}}$ are the covariates for the indicator of consumption, $\mathbf{X_{i2}}$ are the covariates for the consumption amount of the dietary component of interest, and $\mathbf{X_{i3}}$ are the covariates for the consumption of energy. Often, in practice, the covariates for each observed data component are the same, so that $\mathbf{X}_{i1} = \mathbf{X}_{i2} = \mathbf{X}_{i3}$. Along with the covariates, there are corresponding person specific random effects (U_{i1}, U_{i2}, U_{i3}), the role of which is to allow different people who share the same covariates to have different amounts of usual intakes. As we will see shortly, there are also random errors that account for day-to-day variation. Only the covariates, the person-specific random effects, and, because of transformations, the variances of the random errors are relevant to the definitions of usual intake, which are given below at equations (2.6)-(2.7).

Kipnis et al. (2010a) uses a latent variable approach. Let (W_{i1k},W_{i2k},W_{i3k}) be latent variables that are assumed to follow the linear mixed effects model

$$W_{ijk} = \mathbf{X}_{\mathbf{ij}}^{\mathrm{T}}\beta_{\mathbf{j}} + U_{ij} + \epsilon_{ijk} \text{ for } j = 1, 2, 3, \tag{2.1}$$

where $(U_{i1}, U_{i2}, U_{i3}) = \text{Normal}(0, \mathbf{\Sigma_u})$ are the person-specific random effects, while the within-person random errors that account for day-to-day variation $(\epsilon_{i1k}, \epsilon_{i2k}, \epsilon_{i3k}) = \text{Normal}(0, \mathbf{\Sigma}_\epsilon)$. The (U_{i1}, U_{i2}, U_{i3}) and $(\epsilon_{i1k}, \epsilon_{i2k}, \epsilon_{i3k})$ are mutually independent.

The observed data are related to the latent variables as follows:

$$Y_{i1k} = I(W_{i1k} > 0); \tag{2.2}$$

$$Y_{i2k} = Y_{i1k} g^{-1}(W_{i2k}, \lambda_F); \tag{2.3}$$

$$Y_{i3k} = g^{-1}(W_{i3k}, \lambda_E), \tag{2.4}$$

where $I(\cdot)$ is the indicator function and $g^{-1}(x, \lambda)$ is the inverse of the Box-Cox transformation $g(x, \lambda) = (x^\lambda - 1)/\lambda$ for $\lambda \neq 0$ and $= \log(x)$ if $\lambda = 0$. We used the same Box-Cox transformation as those used by Kipnis et al. (2009, 2010a), i.e., the NLMIXED procedure. Under the model defined by (2.1)-(2.4), the probability to consume follows the probit model

$$\text{pr}(Y_{i1k} = 1 | \mathbf{X_{i1}}, U_{i1}, U_{i2}, U_{i3}) = \Phi(\mathbf{X_{i1}^T} \beta_\mathbf{1} + U_{i1}), \tag{2.5}$$

where $\Phi(\cdot)$ is the standard normal distribution function. The probit model is commonly used to model a relationship between a binary dependent variable and one or more independent variables. The probit link was used in Kipnis et al. (2010a) to allow the day-to-day variation in whether a food is consumed to be correlated with the amount of energy consumed, and in such a way that the day-to-day variation random variables $(\epsilon_{i1k}, \epsilon_{i2k}, \epsilon_{i3k})$ are jointly normal, thus facilitating both nonlinear mixed effects software and the MCMC. The Box-Cox transformations in (2.3)-(2.4) allow for skewed distributions typically seen with dietary data. Of course, the notation in (2.5) means that consumption depends on (U_{i1}, U_{i2}, U_{i3}) only through U_{i1}.

In this project, we used the Box-Cox transformation parameters used by Kipnis et al. (2010a), so as to facilitate comparison. It is easy to extend our approach to estimating the transformations.

Under the assumption that the 24hr is unbiased for usual (mean) intake, the usual

intake of the dietary component and energy are given as $T_{Fi} = E(Y_{i2k}|\mathbf{X_{i1}}, \mathbf{X_{i2}}, U_{i1}, U_{i2})$ and $T_{Ei} = E(Y_{i3k}|\mathbf{X_{i3}}, U_{i3})$. Kipnis et al. (2009, 2010a) use a Taylor series approximation to approximate $E\{g^{-1}(v+\epsilon)|v) \approx g^{-1}(v,\lambda) + (1/2)\text{var}(\epsilon)\{\partial^2 g^{-1}(v,\lambda)/\partial v^2\}$. Using this approximation, see equation (19) of Kipnis et al. (2009), and under the covariance matrix restriction described below in the following subsection, they show that the usual intake T_{Fi} of the dietary component and the usual intake T_{Ei} of energy for individual i are given as

$$T_{Fi} \equiv \Phi(\mathbf{X_{i1}}^{\mathrm{T}}\beta_{\mathbf{1}} + U_{i1}) g_*\{\mathbf{X_{i2}}^{\mathrm{T}}\beta_{\mathbf{2}} + U_{i2}, \lambda_F, \mathbf{\Sigma}_\epsilon(2,2)\}, \tag{2.6}$$

$$T_{Ei} \equiv g_*\{\mathbf{X_{i3}}^{\mathrm{T}}\beta_{\mathbf{3}} + U_{i3}, \lambda_E, \mathbf{\Sigma}_\epsilon(3,3)\}, \tag{2.7}$$

where the (j,k) element of $\mathbf{\Sigma}_\epsilon$ is denoted as $\mathbf{\Sigma}_\epsilon(j,k)$ and $g_*(v,\lambda,\sigma_\epsilon^2) = g^{-1}(v,\lambda) + (1/2)\sigma_\epsilon^2\{\partial^2 g^{-1}(v,\lambda)/\partial v^2\}$. The equations (2.6) and (2.7) are relevant to the "long term average intake", the former one is for dietary component and the latter one is for energy. We can combine the usual intakes of dietary component and energy in various ways, e.g., the number of ounces of whole grains per 1000 kilo-calories, i.e., $1000 \times T_{Fi}/T_{Ei}$.

Remark 1 The Taylor series approximation to computing expectations of inverses of the Box-Cox transformation is used here because it was used by Kipnis et al. (2009, 2010a). More precise quadrature formulae can be used, and we have done so, finding almost no numerical changes. The computational convenience of the method makes it attractive.

3. Restriction on the Covariance Matrix

There are two restrictions necessary in the specification of $\mathbf{\Sigma}_\epsilon$. First, following Kipnis et al. (2009, 2010a), we set ϵ_{i1k} and ϵ_{i2k} to be independent. The intuitive way to think

about the independence between the first two is that whether the dietary component is consumed or not and the amount consumed are assumed to be independent. This actually makes sense because a dietary component being consumed cannot indicate how much was consumed. Second, for identifiability of $\beta_{\mathbf{1}}$ and the distribution of U_{i1}, we require that $\text{var}(\epsilon_{i1k}) = 1$, because otherwise the marginal probability of consumption is $\Phi\{(\mathbf{X_{i1}}^{\text{T}}\beta_{\mathbf{1}} + U_{i1})/\text{var}^{1/2}(\epsilon_{i1k})\}$. Without this second restriction, $\beta_{\mathbf{1}}$, $\text{var}(U_{i1})$, $\text{cov}(U_{i1}, U_{i2})$ and $\text{cov}(U_{i1}, U_{i3})$ are identified only up to scale factors. Hence we have that

$$\mathbf{\Sigma}_{\epsilon} = \begin{bmatrix} 1 & 0 & s_{13} \\ 0 & s_{22} & s_{23} \\ s_{13} & s_{23} & s_{33} \end{bmatrix}. \tag{2.8}$$

The difficulty with parameterizations such as (2.8) is that $(s_{13}, s_{23}, s_{22}, s_{33})$ cannot be left unconstrained, or else (2.8) need not be a covariance matrix. Define $s_{13} = \rho_{13} s_{33}^{1/2}$ and $s_{23} = \rho_{23}(s_{22}s_{33})^{1/2}$. Then the determinant $|\mathbf{\Sigma}_{\epsilon}| = s_{22}s_{33}(1 - \rho_{13}^2 - \rho_{23}^2)$. Since $\mathbf{\Sigma}_{\epsilon}$ is a covariance matrix, its determinant must be non-negative, and hence we cannot allow the correlations (ρ_{13}, ρ_{23}) to vary freely. There are many ways to parameterize $\mathbf{\Sigma}_{\epsilon}$ in an unrestricted way that forces it to be positive semi-definite. Here we use a *polar coordinate* representation, $\rho_{13} = \gamma\cos(\theta)$ while $\rho_{23} = \gamma\sin(\theta)$, with $\gamma \in (-1, 1)$ and $\theta \in (-\pi, \pi)$.

The zero entries in (2.8) are not required, although they are implicit in the two part model used in the original papers involving only the episodically consumed dietary component and not energy (Tooze et al., 2006; Kipnis et al., 2009) and they make intuitive sense in our context. We have chosen to use this restriction for these reasons and especially so that the marginal model for the episodically consumed dietary component is the same as that in the literature.

Kipnis et al. (2010a) explore a sample selection model (Heckman, 1976, 1979;

Leung and Yu, 1996; Kyriazidou, 1997; Min and Agresti, 2002) that does not have this restriction. They found that such a sample selection model can be very unstable in our context, with the components of $\mathbf{\Sigma_u}$ and $\mathbf{\Sigma}_\epsilon$ varying wildly. Although it is possible to use MCMC computations to fit the sample selection model, given the acceptance of the restriction in nutritional epidemiology and of the NCI method, we focus on the covariance model (2.8).

Remark 2 It is absolutely vital to allow for $\mathbf{\Sigma}_\epsilon$ being non-diagonal. The term $s_{23} \neq 0$ simply reflects the reality that, within a person and hence conditional on (U_{i1}, U_{i2}, U_{i3}), the amount of food reported consumed and the amount of energy consumed are sometimes highly correlated. The reason we allow $s_{13} \neq 0$ is to account for the very real possibility that, again within a person, the very fact that one consumes a food leads to a higher or lower reported energy (caloric) intake.

4. Model Fitting and Computation

It is possible in principle to fit model (2.1)-(2.7) using nonlinear mixed effects software. Kipnis et al. (2010a) use the SAS procedure PROC NLMIXED. However, we have found that such implementation is slow and not very stable, with many issues of convergence. NLMIXED uses adaptive Gaussian quadrature to integrate the likelihood over the distribution of random effects. NLMIXED can have convergence problems, especially when there are too many, or too few, zeros. What typically happens is that $\text{corr}(U_{i1}, U_{i2})$ tries to go to 1.00 or sometimes even -1.00, or that $\text{var}(U_{i1})$ or $\text{var}(U_{i2})$ tries to go to 0.00. When one of these things happens, the model usually converges, according to the change-in-likelihood criterion, but the Hessian is not positive definite. Occasionally, NLMIXED fails to converge at all. In general, we have found that when NLMIXED does not have such numerical problems, its results

and ours are in reasonable agreement. These issues are described in more detail in Section D of this chapter.

Hence, for both stability and speed, we have turned to Bayesian approaches for fitting the model described by equations (2.1)-(2.8). We emphasize that the Markov Chain Monte Carlo computation can either be thought of as a strictly Bayesian computation with ordinary Bayesian inference, or as a means of developing *frequentist* estimators of the crucial parameters, based on the well-known fact that in parametric models such as ours, the Bayesian posterior mean of the parameters is a consistent and asymptotically normally distributed frequentist estimator, see for example Lehmann and Casella (1998, Chapter 6.8).

Our computational algorithm, described in detail in the appendix, uses Gibbs sampling with some Metropolis-Hastings steps. We have implemented this approach in both Matlab and R, and it is fast enough for practical use. In the NIH-AARP Diet and Health Study described in Section C of this chapter, with a sample size of 899, for a burn-in of $1,000$ steps followed by $10,000$ MCMC iterations, our Matlab and R programs take approximately 2 minutes and 11.7 minutes on an Intel(R) Xeon(TM) CPU with 3.73GHz and 7.8GB of RAM in a Linux system, respectively. For a burn-in of $5,000$ steps followed by $15,000$ MCMC iterations, our Matlab and R programs take approximately 3 minutes and 17.5 minutes, respectively. Both programs are available from the first author.

We have also developed an implementation in WinBUGS with a BUGS model called from R by using the package R2WinBUGS. Details are available from the third author. As to be expected, the WinBUGS code is much slower than the custom programs, taking approximately 5 hours (Pentium computer with 3.5GHz CPU and 1.99GB of RAM in a Windows system) for a burn-in of $1,000$ steps followed by $10,000$ MCMC samples. We are also currently developing a SAS macro for use by

the nutritional community. On various test data sets, the WinBUGS, R, SAS and Matlab code gave very similar answers. In our empirical work, we use the Matlab code.

Remark 3 There are important data conventions that we use. These are described in detail in the Appendix. For example, in Appendix 1, we mention that covariates are always standardized to have sample mean zero and sample variance one. The reason is a matter of scaling: energy intake is in terms of calories, which are typically in the 1,000's, so that the corresponding regression parameters, without standardization, with the FFQ energy as a covariate, would necessarily be tiny, making it hard to develop a plausible prior distribution. As described in Appendix 1, we also standardize the responses for numerical stability and weaken dependence upon the prior distributions, and in Appendix 2 we describe why this standardization makes sense. We have fit our method with various different prior distributions, and there is very little sensitivity to prior specification.

5. Simulation Study

We performed a simulation study that was based upon our empirical study given in Section C of this chapter, in order to ascertain whether the methodology results in reasonably unbiased estimates of $(\beta_{\mathbf{1}}, \beta_{\mathbf{2}}, \beta_{\mathbf{3}}, \mathbf{\Sigma_u}, \mathbf{\Sigma}_\epsilon)$. To test whether our algorithm can produce non-near-zero correlations when the true correlations are actually far from zero, we simulated 200 data sets, each of size $n = 1,000$, roughly the size of the NIH-AARP calibration cohort in Section C of this chapter. In this simulation, we used the same covariates for each of the three outcomes, i.e., we set $\mathbf{X_{i1}} = \mathbf{X_{i2}} = \mathbf{X_{i3}}$. The covariates had three components, the first equal to 1.0 for an intercept, and the other two generated as Normal$(0, 1)$. The parameters $(\beta_{\mathbf{1}}, \beta_{\mathbf{2}}, \beta_{\mathbf{3}})$ were generated as

Uniform$(0,1)$ for each simulated data set. We used

$$\boldsymbol{\Sigma}_{\mathbf{u}} = \begin{bmatrix} 0.50 & 0.24 & 0.24 \\ 0.24 & 0.70 & 0.35 \\ 0.24 & 0.35 & 0.70 \end{bmatrix}; \qquad \boldsymbol{\Sigma}_{\epsilon} = \begin{bmatrix} 1.00 & 0.00 & 0.47 \\ 0.00 & 1.20 & 0.78 \\ 0.47 & 0.78 & 1.40 \end{bmatrix}.$$

The mean of the posterior means of $(\beta_{\mathbf{1}}, \beta_{\mathbf{2}}, \beta_{\mathbf{3}})$ was very nearly unbiased overall and are not reported here. The parameters $(\boldsymbol{\Sigma}_{\mathbf{u}}, \boldsymbol{\Sigma}_{\epsilon})$ are more difficult to estimate, but the mean of their posterior means were

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{u}} = \begin{bmatrix} 0.51 & 0.27 & 0.27 \\ 0.27 & 0.68 & 0.33 \\ 0.27 & 0.33 & 0.67 \end{bmatrix}; \qquad \widehat{\boldsymbol{\Sigma}}_{\epsilon} = \begin{bmatrix} 1.00 & 0.00 & 0.39 \\ 0.00 & 1.23 & 0.80 \\ 0.39 & 0.80 & 1.43 \end{bmatrix}.$$

Crucially, for the main purposes of estimating the distribution of usual intakes, the posterior means were essentially unbiased for estimating $\boldsymbol{\Sigma}_{\mathbf{u}}$. As seen in the Appendix, $\boldsymbol{\Sigma}_{\epsilon}$ also has a role in the definition of usual intake, and it too was essentially unbiased except for a small bias of size 0.08 in estimating $\text{cov}(\epsilon_{i1k}, \epsilon_{i3k})$, a term that does not appear in the definitions of usual intake.

Remark 4 We give here only the results of a single simulation because what we have shown above are representative of other simulations we have done. For example, we have simulated cases where the off-diagonal elements of $\boldsymbol{\Sigma}_{\mathbf{u}}$ were zero and cases where some of them were negative. We have also simulated cases that the diagonal elements of $\boldsymbol{\Sigma}_{\mathbf{u}}$ were smaller and somewhat larger. In none of the cases did we see any significant bias in the estimates.

Remark 5 We have not displayed the simulation results for the Proc NLMIXED procedure because in those cases that it converges, it is very nearly unbiased, just like our method.

C. Empirical Analysis: Methods

1. Introduction to the NIH-AARP Diet and Health Study at the National Cancer Institute

The NIH-AARP Diet and Health Study, see http://dietandhealth.cancer.gov/ and Schatzkin et al. (2001), has two components, the main study with diet assessed by a Food Frequency Questionnaire (FFQ) and a calibration sub-study with additional diet assessment by two 24hr. We considered a part of the main study that consists of $n_p = 142,364$ women, who contributed an FFQ as well as relevant demographic characteristics. The data used were the same as in Sinha et al. (2010). The covariates $\mathbf{X}$ used included an intercept, age, body mass index, the FFQ for energy intake and the FFQ for the dietary component in question. The 24hr was not available for these subjects. Thus, the primary sample represents data on $\mathbf{X_i} = \mathbf{X_{i1}} = \mathbf{X_{i2}} = \mathbf{X_{i3}}$ for $i = 1, ..., n_p$.

In addition to the primary sample, there was a subsample of $n_v = 899$ women in the calibration sub-study who completed an FFQ and demographic characteristics, so that there are $\mathbf{X_i} = \mathbf{X_{i1}} = \mathbf{X_{i2}} = \mathbf{X_{i3}}$ for $i = 1 + n_p, ..., n_v + n_p$. In addition, these women completed two 24hr. Hence we observed $(Y_{i1k}, Y_{i2k}, Y_{i3k})$ for $k = 1, 2$ and for $i = 1 + n_p, ..., n_v + n_p$.

We illustrate our computational algorithm using data from both the two 24hr and the FFQ for whole grains, fish and energy intake, along with covariates. Following Kipnis et al. (2009, 2010a), the FFQ values for fish, whole grain and energy intake were transformed using $\lambda = 0.25$, $\lambda = 0.33$ and $\lambda = 0.00$, respectively. The 24hr used $\lambda = 0.50$, $\lambda = 0.33$ and $\lambda = 0.33$, respectively.

The MCMC output gives samples from the posterior distribution of $\mathbf{\Sigma_u}$, $\mathbf{\Sigma}_\epsilon$, $\mathcal{B}$= $(\beta_{\mathbf{1}}{}^{\mathrm{T}}, \beta_{\mathbf{2}}{}^{\mathrm{T}}, \beta_{\mathbf{3}}{}^{\mathrm{T}})^{\mathrm{T}}$ and (U_{i1}, U_{i2}, U_{i3}), the latter only for $i = 1 + n_p, ..., n_v + n_p$. The

means of the samples for $(\mathcal{B}, \mathbf{\Sigma_u}, \mathbf{\Sigma}_\epsilon)$ can be taken as frequentist point estimates of these quantities, and are denoted here as $(\widehat{\beta}_\mathbf{1}, \widehat{\beta}_\mathbf{2}, \widehat{\beta}_\mathbf{3}, \widehat{\mathbf{\Sigma}}_\mathbf{u}, \widehat{\mathbf{\Sigma}}_\epsilon)$. We will use shorthand notation for usual intake:

Usual dietary component intake is $T_{Fi} = \mathcal{G}_1\{\mathbf{X_{i1}}, \mathbf{X_{i2}}, \beta_\mathbf{1}, \beta_\mathbf{2}, U_{i1}, U_{i2}, \mathbf{\Sigma}_\epsilon(2,2)\}$.

Usual energy intake is $T_{Ei} = \mathcal{G}_2\{\mathbf{X_{i3}}, \beta_\mathbf{3}, U_{i3}, \mathbf{\Sigma}_\epsilon(3,3)\}$.

For both usual dietary component intake and usual energy intake, 24hr samples are available for $i = 1 + n_p, ..., n_v + n_p$.

2. Frequentist Analysis

We are going to write the variable of interest as $\mathcal{H}(T_{Fi}, T_{Ei})$. Thus, (a) the dietary component is $\mathcal{H}(T_{Fi}, T_{Ei}) = T_{Fi}$; (b) energy is $\mathcal{H}(T_{Fi}, T_{Ei}) = T_{Ei}$; and (c) the energy adjusted dietary component is $\mathcal{H}(T_{Fi}, T_{Ei}) = 1000 \times T_{Fi}/T_{Ei}$. In general then, the usual intake variable of interest for person i can be written as

$$\mathcal{Q}_i = \mathcal{H}\left[\mathcal{G}_1\{\mathbf{X_{i1}}, \mathbf{X_{i2}}, \beta_\mathbf{1}, \beta_\mathbf{2}, U_{i1}, U_{i2}, \mathbf{\Sigma}_\epsilon(2,2)\}, \mathcal{G}_2\{\mathbf{X_{i3}}, \beta_\mathbf{3}, U_{i3}, \mathbf{\Sigma}_\epsilon(3,3)\}\right],$$

for $i = 1, ..., n_p + n_v$, where we have that $(U_{i1}, U_{i2}, U_{i3}) = \text{Normal}(0, \mathbf{\Sigma_u})$. Estimation of the distribution of $\mathcal{Q}$ across the population is easily accomplished by a Monte-Carlo computation. Specifically, for a large B, where we took $B = 5,000$, and for $b = 1, ..., B$ generate $(U_{bi1}, U_{bi2}, U_{bi3}) = \text{Normal}(0, \widehat{\mathbf{\Sigma}}_\mathbf{u})$. Here B is not the number of burn-in steps, but simply a large enough number to do numerical integration. Then the distribution of usual intake can be estimated as the empirical distribution of the values

$$\mathcal{Q}_{bi} = \mathcal{H}\left[\mathcal{G}_1\{\mathbf{X_{i1}}, \mathbf{X_{i2}}, \widehat{\beta}_\mathbf{1}, \widehat{\beta}_\mathbf{2}, U_{bi1}, U_{bi2}, \widehat{\mathbf{\Sigma}_\epsilon}(2,2)\}, \mathcal{G}_2\{\mathbf{X_{i3}}, \widehat{\beta}_\mathbf{3}, U_{bi3}, \widehat{\mathbf{\Sigma}_\epsilon}(3,3)\}\right],$$

taken across $i = 1, ..., n_v + n_p$ and $b = 1, ..., B$.

Standard errors and confidence intervals for regression calibration and the distribution of usual intake can be formed easily by bootstrapping. We used 400 bootstrap samples in our numerical work.

Remark 6 For bootstrap confidence intervals, it is often recommended to use at least 399 bootstrap samples, as we have done, see for example Davidson and MacKinnon (1999). We have experimented with using up to $1,000$ bootstrap samples, but this significantly increases computing time without changing the basic results in any material way.

3. Bayesian Analysis

As described below, Bayesian inference on the distribution of usual intake depends on estimating the distribution of the covariates. The distribution of usual intake $\mathcal{H}(T_F, T_E)$ in a population can be described as follows. Let $\mathcal{X} = (\mathbf{X_1}, \mathbf{X_2}, \mathbf{X_3})$ and let $f_X(\mathcal{X}|\theta) = f_X(\mathbf{X_1}, \mathbf{X_2}, \mathbf{X_3}, \theta)$ be the distribution of $\mathcal{X} = (\mathbf{X_1}, \mathbf{X_2}, \mathbf{X_3})$ in the population, based on a parameter θ. Write $\mathcal{U} = (\mathbf{U_1}, \mathbf{U_2}, \mathbf{U_3})$. Use the shorthand notation

$$\mathcal{K}(\mathcal{X}, \mathcal{B}, \mathcal{U}, \mathbf{\Sigma}_\epsilon) = \mathcal{H}\left[\mathcal{G}_1\{\mathbf{X_1}, \mathbf{X_2}, \beta_1, \beta_2, \mathbf{U_1}, \mathbf{U_2}, \mathbf{\Sigma}_\epsilon(2,2)\}, \mathcal{G}_2\{\mathbf{X_3}, \beta_3, \mathbf{U_3}, \mathbf{\Sigma}_\epsilon(3,3)\}\right].$$

Then the distribution of usual intake is

$$\begin{aligned} F(v|\mathcal{B}, \mathbf{\Sigma_u}, \theta, \mathbf{\Sigma}_\epsilon) &= \text{pr}\left\{\mathcal{K}(\mathcal{X}, \mathcal{B}, \mathcal{U}, \mathbf{\Sigma}_\epsilon) \leq v|\mathcal{B}, \mathbf{\Sigma_u}, \mathbf{\Sigma}_\epsilon, \theta\right\} \\ &= \int I\left\{\mathcal{K}(\mathcal{X}, \mathcal{B}, \mathcal{U}, \mathbf{\Sigma}_\epsilon) \leq v\right\} f_{\mathcal{U}}(\mathcal{U}|\mathbf{\Sigma_u}) f_X(\mathcal{X}|\theta) d\mathcal{U} d\mathcal{X}. \end{aligned}$$

We suggest approximating this using Monte-Carlo integration, as follows. Again, let B be large where we took $B = 1,000$, and for $b = 1, ...B$, let $\mathbf{u_b} = \text{Normal}(0, \mathbf{I_3})$. Let

$\boldsymbol{\Sigma_u}^{1/2}$ be the symmetric square root of $\boldsymbol{\Sigma_u}$. Then

$$F(v|\mathcal{B}, \boldsymbol{\Sigma_u}, \theta, \boldsymbol{\Sigma}_\epsilon) \approx B^{-1} \sum_{b=1}^{B} \int I\left\{\mathcal{K}(\mathcal{X}, \mathcal{B}, \boldsymbol{\Sigma_u}^{1/2}\mathbf{u_b}, \boldsymbol{\Sigma}_\epsilon) \leq v\right\} f_X(\mathcal{X}|\theta) d\mathcal{X}.$$

The posterior distribution of $F(v|\mathcal{B}, \boldsymbol{\Sigma_u}, \theta, \boldsymbol{\Sigma}_\epsilon)$ is then calculated from the MCMC samples: our methods in the Appendix are easily generalized to sample from the posterior distribution of θ.

In the NIH-AARP Diet and Health Study, with a sample size of $n_p + n_v > 140,000$, we effectively know the distribution of $\mathcal{X}$. Let the values in the data be $\mathcal{X}_i$ for $i = 1, ..., n_v + n_p$. Then we have

$$F(v|\mathcal{B}, \boldsymbol{\Sigma_u}, \theta, \boldsymbol{\Sigma}_\epsilon) \approx \{(n_v + n_p)B\}^{-1} \sum_{b=1}^{B} \sum_{i=1}^{n_v+n_p} I\left\{\mathcal{K}(\mathcal{X}_i, \mathcal{B}, \boldsymbol{\Sigma_u}^{1/2}\mathbf{u_b}, \boldsymbol{\Sigma}_\epsilon) \leq v\right\}.$$

The posterior distribution of $F(v|\mathcal{B}, \boldsymbol{\Sigma_u}, \theta, \boldsymbol{\Sigma}_\epsilon)$ can then be calculated from the MCMC samples.

D. Results

Along with illustrating the distributions of usual intakes of the dietary components adjusted for energy, we also compared our results with NLMIXED.

1. Basic Analysis

We used a burn-in of 5,000 steps followed by 15,000 MCMC samples. We saved every 10^{th} sample to reduce autocorrelation.

a. Frequentist Analysis

In Table 1 we present summary statistics (mean, standard deviation and selected percentiles) of the usual intakes as well as the usual intakes adjusted for energy. The

Table 1. Estimated distributions of the usual intake for whole grains, fish and energy and the estimated distributions of energy-adjusted usual intake for whole grains and fish, for women.

	Whole Grains		Fish			Energy
	Usual Intake (Unit: cup)	per 1000 kcals	Usual Intake (Unit: oz.)	Freq, per 1000 kcals	Bayes, per 1000 kcals	Usual Intake (Unit:kcal)
Mean	1.013	0.625	0.539	0.338	0.339	1631.77
s.d.	0.631	0.375	0.486	0.309	0.315	369.16
5^{th}	0.181	0.121	0.053	0.033	0.028	1075.70
10^{th}	0.287	0.189	0.089	0.057	0.057	1180.37
25^{th}	0.536	0.345	0.193	0.122	0.122	1370.29
50^{th}	0.911	0.569	0.399	0.249	0.249	1604.04
75^{th}	1.375	0.841	0.736	0.456	0.456	1863.01
90^{th}	1.867	1.127	1.176	0.731	0.731	2118.74
95^{th}	2.195	1.320	1.508	0.945	0.951	2282.50

5^{th} percentile of the distribution is labeled as 5^{th}, etc. For energy-adjusted fish intake, we give the results for both the frequentist ("Freq") and the Bayesian ("Bayes") fits. Estimates were very similar for both Freq and Bayes fits and thus we have only displayed results for fish. Figures 1 and 2 give density estimates for usual intake and energy adjusted intake of fish and whole grains, respectively: a similar plot for usual energy intake was also produced but not displayed here. The solid line is the density estimate for usual intake in the unit of oz. for fish and cups for whole grains. The dashed line is the density estimate for usual intake per 1000 kilo-calories. The evident skewness of the usual intakes of fish and whole grains is expected, as are the somewhat less skew nature of the energy adjusted intakes.

We bootstrapped the validation and primary data sets separately 400 times, see Remark 6, reran the analysis, and formed bootstrap confidence intervals. Since the distribution of the covariates X is essentially known because of the size of the primary study, this bootstrap simply reflects the uncertainty in the parameter estimates as

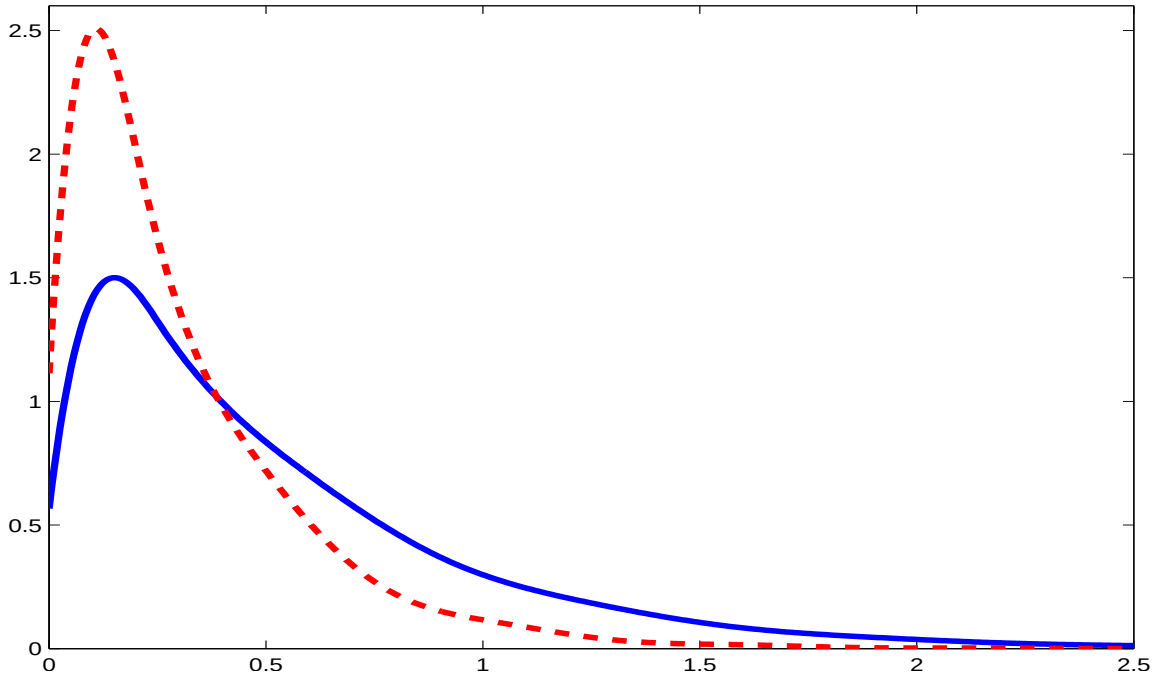

Figure 1. Density estimates for fish.

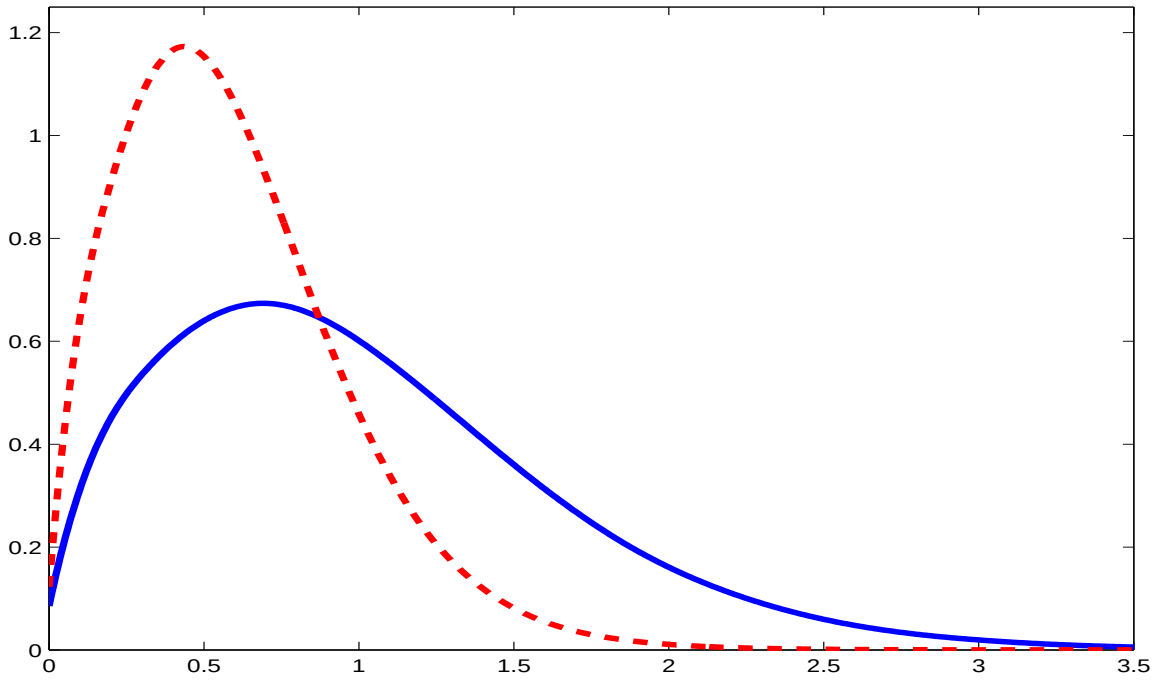

Figure 2. Density estimates for whole grains.

they propagate through to usual intakes. To give a graphical summary including uncertainty, in Figure a we plot the actual estimated percentiles of the distribution of adjusted fish intake against the percentile number, as well as the 95% pointwise bootstrap confidence interval for these percentiles. Horizontal axis is the relative percentile, e.g., the value at 50 is the median. The vertical axis is the estimated percentile (solid line) in the unit of oz./(1000 kcal). Dashed lines are the pointwise 95% bootstrap confidence intervals.

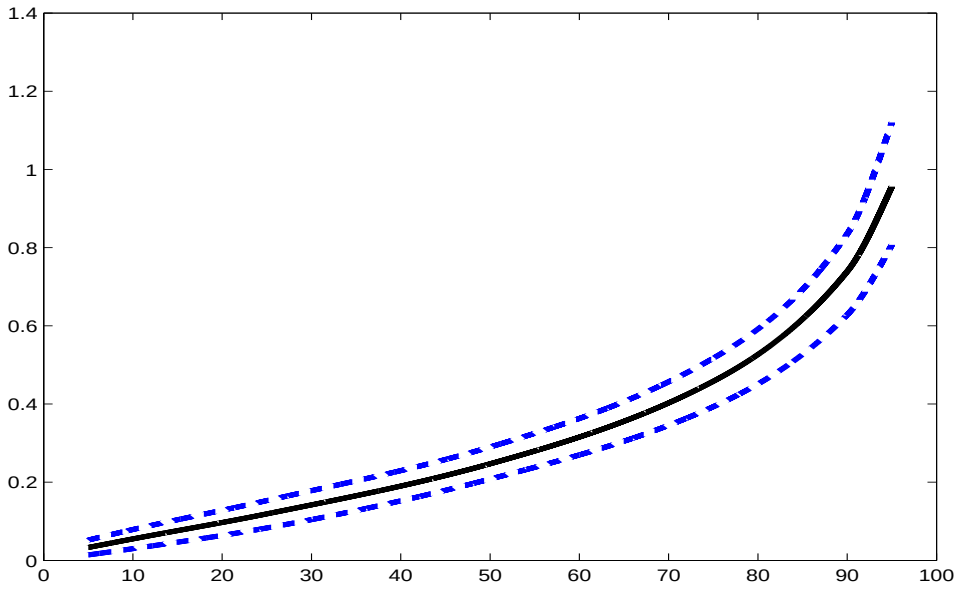

Figure 3. Quantile functions for usual fish intake per 1000 kilo-calories.

b. Bayesian Analysis

In Table 1 we also give the Bayesian analysis for energy-adjusted fish intake. As seen there, the Bayesian analysis posterior means of the distribution of energy-adjusted fish intake is nearly identical to the frequentist analysis. The same thing was found for all the columns in Table 1.

In addition, posterior credible interval lengths were almost equivalent to those of the frequentist method and are not displayed here.

2. Comparison With Proc NLMIXED

We described in Section B of this chapter, some of the motivation for our computational approach. In this section, we show documentation of those claims.

Table 2. Comparison between two approaches, "NLMIXED" and "MCMC".

	Whole Grains		Fish		Dark Green	
	NLMIXED	MCMC	NLMIXED	MCMC	NLMIXED	MCMC
Time in Minutes	20	3	12	3	12	4
% zeros on 24hr	32%		77%		73%	
Correlations						
$\text{corr}(U_{i1}, U_{i2})$	0.65	0.48	-0.39	0.08	1.00	0.48
	(0.17)	(0.09)	(0.44)	(0.07)	(N/A)	(0.06)
$\text{corr}(U_{i1}, U_{i3})$	0.20	0.18	0.28	0.26	0.27	0.24
	(0.08)	(0.07)	(0.14)	(0.07)	(N/A)	(0.06)
$\text{corr}(U_{i2}, U_{i3})$	0.37	0.40	0.02	0.02	0.27	0.28
	(0.10)	(0.07)	(0.16)	(0.09)	(N/A)	(0.06)

First, in Table 2, we describe aspects of the analysis for women of whole grains, fish and dark-green vegetables, using the AARP data set. Table 2 is a comparison between two approaches, "NLMIXED" and "MCMC"., of the nonlinear mixed effects model, for whole grains, fish and dark-green vegetables. Displayed are the estimates of correlations among the components of (U_{i1}, U_{i2}, U_{i3}), the estimates for the MCMC approach being posterior means. The numbers displayed in parentheses are the standard errors from the inverse of the Hessian matrix ("NLMIXED") and from MCMC samples ("MCMC"). Here "Dark Green" refers to Dark-Green vegetables, for which the nonlinear mixed effects analysis converged but to a singular covariance matrix for Σ_u. The phrase "Time in Minutes" refers to computation time to complete the

analysis. The overall % of zeros on the 24hr are also displayed. The first line in the table is the number of minutes of computation for the nonlinear mixed effects program and our MCMC approach. It can be seen that the MCMC approach is considerably faster. While not displayed here, for Milk for men, which had only 12% reported non-consumption on the 24hr, the nonlinear mixed effects program took 200 minutes, while ours took only 4 minutes. This illustrates our claim concerning speed of computation.

A second aspect is that we claimed that sometimes the nonlinear mixed effects analysis of Kipnis et al. (2010a) suffered from convergence to a singular covariance matrix estimate for Σ_u. This occurred for dark-green vegetables, see Table 2, where it was estimated that the correlation between (U_{i1}, U_{i2}), $\text{corr}(U_{i1}, U_{i2})$, was equal to 1.00. This seemingly ridiculous result is in marked contrast to the much more sensible posterior mean of 0.48.

A third aspect of the comparison is that we claimed that when the method of Kipnis et al. (2010a) converged to a reasonable answer, our results were in general agreement with theirs. This is borne out in Table 2, where we have listed the standard errors of the estimates using the Hessian for the nonlinear mixed effects analysis, and using the MCMC samples for our method. The estimates are quite similar with the exception of $\text{corr}(U_{i1}, U_{i2})$ for fish, which can be explained as follows. We performed a separate bootstrap calculation for this correlation with our method and the nonlinear mixed effects analysis, which suggested a standard error as large as the difference between the two. The other standard errors are also different, but this may well reflect imprecision in the former caused by using a Hessian in a nonlinear mixed effects model instead of a bootstrap.

Remark 7 While it may seem obvious, it is useful to clarify what we mean by the

term "*convergence*". We are not meaning asymptotic rates of convergence, because these are the standard $n^{1/2}$-type one sees in parametric models. We are also not talking about theoretical rates of numerical convergence, e.g., how fast is convergence of the Proc NLMIXED procedure in terms of number of iterations. Instead, for us the term convergence has the meaning that Proc NLMIXED announces that it has converged to a solution with a nonsingular Hessian. Of course, our method, being based on proper priors, converges in the usual MCMC sense.

CHAPTER III

A NEW MULTIVARIATE MEASUREMENT ERROR MODEL WITH ZERO-INFLATED DIETARY DATA

A. Introduction

This project presents statistical models and methodology to overcome a major stumbling block in the field of dietary assessment.

More nutritional background is provided in Section B of this chapter: a summary of the key conceptual issues follows.

- Nutritional surveys conducted in the United States typically use 24-hour (24hr) dietary recalls to obtain intake data, i.e., an assessment of what was consumed in the past 24 hours.
- Because dietary recommendations are intended to be met over time, nutritionists are interested in "usual" or long-term average daily intake.
- Dietary intake is thus assessed with considerable measurement error.
- Consumption patterns of dietary components vary widely; some are consumed daily by almost everyone, while others are episodically consumed so that 24-hour recall data are zero-inflated. Further, these components are correlated with one another.
- Nutritionists are interested in dietary components collectively to capture patterns of usual dietary intake, and thus need multivariate models for usual intake.
- These multivariate models for usual intakes, taking into account episodically consumed foods, do not exist, nor do methods exist for fitting them.

One way to capture dietary patterns is by scores, although our work is not limited to scores. The Healthy Eating Index-2005 (HEI-2005), described in detail in Section B of this chapter, is a scoring system based on a priori knowledge of dietary recommendations, and is on a scale of 0 to 100. Ideally, it consists of the usual intake of 6 episodically consumed and thus 24hr-zero inflated foods, 6 daily-consumed dietary components, adjusts these for energy (caloric) intake, and gives a score to each component. The total score is the sum of the individual component scores. Higher scores indicate greater compliance with dietary guidelines and, therefore, a healthier diet. Here are a few questions that nutritionists have not been able to answer, and that our approach can address.

- What is the distribution of the HEI-2005 total score, and what % of Americans are eating a healthier diet defined for example, by a total score exceeding 80?
- What is the correlation between the individual score on each dietary component and the scores of all other dietary components?
- Among those whose total HEI-2005 score is > 50 or ≤ 50, what is the distribution of usual intake of whole grains, whole fruits, dark green and orange vegetables and legumes (DOL) and calories from solid fats, alcoholic beverages and added sugars (SoFAAS)?
- What % of Americans exceed the median score on all 12 HEI-2005 components?

In this project, to answer public health questions such as these that can have policy implications, we build a novel multivariate measurement error model for estimating the distributions of usual intakes, one that accounts for measurement error and zero-inflation, and has a special structure associated with the zero-inflation. Previous attempts to fit even simple versions of this model, using nonlinear mixed effects

software, failed because of the complexity and dimensionality of the model. We use survey-weighted Monte Carlo computations to fit the model with uncertainty estimation coming from balanced repeated replication. The methodology is illustrated using the HEI-2005 to assess the diets of children aged 2-8 in the United States. This work represents the first analysis of joint distributions of usual intakes for multiple food groups and nutrients.

The project is outlined as follows. In Section B of this chapter we give the background for the data we observe. In particular, we provide more information about the HEI-2005. Section C of this chapter describes our model which is a highly nonlinear, zero-inflated, repeated measures model with multiple latent variables. The model also has a patterned covariance matrix with structural zeros and ones. We derive a parameterization that allows estimated covariance matrices to be actual covariance matrices. We also define technically what we mean by usual intake, and illustrate the use of simulation methods used to answer the questions posed above, as well as many others.

Section D of this chapter describes our estimation procedure. Previous attempts using nonlinear mixed effects models to estimate the distribution of episodically consumed food groups (Tooze et al., 2006; Kipnis et al., 2009) do not work here because of the high dimensionality of the problem. We instead develop a Monte Carlo strategy based on the idea of Gibbs sampling; although because of sampling weights, we treat the method as a frequentist (non-Bayesian) one. This section describes some of the basics of the methodology; the full technical details of implementation are given in an appendix.

Section E of this chapter describes the analysis of the HEI-2005 components using the 2001-2004 National Health and Nutrition Examination Survey (NHANES) for children ages 2-8. Important contextual points arise because of the nature of the

data. For example, if whole grains are consumed, then necessarily total grains are consumed with probability one, a restriction that a naive use of our model cannot handle. We develop a simple novel device to uncouple consumption variables that are tightly linked in this way. Finally in this section, we provide the first answers to the four questions we have posed. In Section F of this chapter, we discuss various additional aspects of the problem and the data analysis. Concluding remarks and a policy application are given in Chapter V.

There are a number of general reviews of the measurement error field (Fuller, 1987; Gustafson, 2003; Carroll et al., 2006; Buonaccorsi, 2010). Recent papers that focus on estimating the density function of a univariate continuous random variable subject to measurement error include Delaigle (2008), Delaigle and Hall (2008, 2010), Delaigle and Meister (2008), Delaigle et al. (2008), Staudenmayer et al. (2008) and Wand (1998). The field of measurement error in regression continues to expand rapidly, with some recent contributions including Küchenhoff, et al (2006), Guolo (2008), Liang et al. (2008), Messer and Natarajan (2008) and Natarajan (2009). There is also a large statistical literature on measurement error as it relates to public health nutrition: some recent papers relevant to our work include Carriquiry (1999, 2003), Ferrari et al. (2009), Fraser and Shavlik (2004), Kott et al. (2009), Nusser et al. (1996, 1997), Prentice (1996, 2003), and Tooze et al. (2002, 2006).

B. Data and the HEI-2005 Scores

Here we give more detail about the nutrition context that motivates this work. In surveys conducted in the United States, the preferred method of obtaining intake data is the 24-hour dietary recall because it limits respondent burden and facilitates accurate reporting; yet the measure of greatest interest is "usual" or long-term average

daily intake. Thus dietary intake is assessed with considerable measurement error. Also, diets are comprised of numerous foods, nutrients, and other components, each of which may have distinctive attributes and effects on nutritional health. Sometimes, it is useful to examine intake of these components separately, but increasingly nutritionists are interested in exploring them collectively to capture patterns of dietary intake. Consumption patterns of these components vary widely; some are consumed daily by almost everyone while others are episodically consumed so that 24-hour recall data are zero-inflated. In addition, these various components are often correlated with one other. Finally, it is often preferable to analyze the amount of a dietary component relative to the amount of energy (calories) in a diet because dietary recommendations often vary with energy level, and this approach provides a way of standardizing dietary assessments.

One of the US Department of Agriculture's (USDA's) strategic objectives is "to promote healthy diets" and it has developed an associated performance measure, the Healthy Eating Index-2005 (http://www.cnpp.usda.gov/HealthyEatingIndex.htm, HEI-2005). The HEI-2005 is based on the key recommendations of the 2005 Dietary Guidelines (http://www.health.gov/dietaryguidelines/dga2005/document/default.htm) for Americans. The index includes ratios of interrelated dietary components to energy. The HEI-2005 comprises 12 distinct component scores and a total summary score. See Table 3 for a list of these components and the standards for scoring, and see Guenther et al. (2008a) for details. Except for saturated fat and SoFAAS, density is obtained by multiplying usual intake by 1000 and dividing by usual intake of kilo-calories. In Table 3, for saturated fat, density is 9×100 usual saturated fat (grams) divided by usual calories, i.e., the percentage of usual calories coming from usual saturated fat intake. For SoFAAS, the density is the percentage of usual intake that comes from usual intake of calories, i.e., the division of usual intake of SoFAAS by usual intake

of calories. Here, "DOL" is dark green and orange vegetables and legumes. Also,

Table 3. Description of the HEI-2005 scoring system.

Component	Units	HEI-2005 score calculation
Total Fruit	cups	$\min(5, 5 \times (\text{density}/.8))$
Whole Fruit	cups	$\min(5, 5 \times (\text{density}/.4))$
Total Vegetables	cups	$\min(5, 5 \times (\text{density}/1.1))$
DOL	cups	$\min(5, 5 \times (\text{density}/.4))$
Total Grains	ounces	$\min(5, 5 \times (\text{density}/3))$
Whole Grains	ounces	$\min(5, 5 \times (\text{density}/1.5))$
Milk	cups	$\min(10, 10 \times (\text{density}/1.3))$
Meat and Beans	ounces	$\min(10, 10 \times (\text{density}/2.5))$
Oil	grams	$\min(10, 10 \times (\text{density}/12))$
Saturated Fat	% of energy	if density ≥ 15 score $= 0$ else if density ≤ 7 score $= 10$ else if density > 10 score $= 8 - (8 \times (\text{density} - 10)/5)$ else, score $= 10 - (2 \times (\text{density} - 7)/3)$
Sodium	milligrams	if density ≥ 2000 score=0 else if density ≤ 700 score=10 else if density ≥ 1100 score $= 8 - \{8 \times (\text{density} - 1100)/(2000 - 1100)\}$ else score $= 10 - \{2 \times (\text{density} - 700)/(1100 - 700)\}$
SoFAAS	% of energy	if density ≥ 50 score $= 0$ else if density ≤ 20 score=20 else score $= 20 - \{20 \times (\text{density} - 20)/(50 - 20)\}$

"SoFAAS" is calories from solid fats, alcoholic beverages and added sugars. The total HEI-2005 score is the sum of the individual component scores. Intakes of each food or nutrient, represented by one of the 12 components, are expressed as a ratio to energy intake, assessed, and ascribed a score.

The HEI-2005 is used to evaluate the diets of Americans to assess compliance with the 2005 Dietary Guidelines, yet use of the HEI-2005 is limited by the challenges described above. Until recently, there have been no solutions to these challenges, so published evaluations have been limited to analyses of mean scores for the population and various subgroups. Freedman et al. (2010) have described a method of estimating the population distribution of a single component of HEI-2005, and the prevalence

of high or low scores on that component; but there has been to date no satisfactory way to determine the prevalence of high or low total HEI-2005 scores, considering all of its interrelated components simultaneously. In addition, answers to the complex questions posed in the Introduction remain unavailable. This project aims to provide a means to do these crucial evaluations.

The 12 HEI-2005 components represent 6 episodically consumed food groups (total fruit, whole fruit, total vegetables, dark green and orange vegetables and legumes or DOL, whole grains and milk), 3 daily-consumed food groups (total grains, meat and beans and oils), and 3 other daily-consumed dietary components (saturated fat; sodium; and calories from solid fats, alcoholic beverages and added sugars, or SoFAAS). The classification of food groups as "episodically" and "daily" consumed is based on the number of individuals who report them on 24hr recalls. If there are only a few zeros for a component, we treat that as a daily-consumed food, and replace all zeros with 1/2 the minimum value of the non-zeros for that food. However, the crucial statistical aspect of the data is that six of the food groups are zero-inflated. The percentages of reported non-consumption of total fruit, whole fruit, whole grains, total vegetables, DOL, and milk on any single day are 17%, 40%, 42%, 3%, 50% and 12%, respectively.

We are interested in the usual intake of foods for children aged 2-8. The data available to us, described in more detail in Section E of this chapter, came from the National Health and Nutrition Examination Survey, 2001-2004 (NHANES). The data used here consisted of $n = 2,638$ children, each of whom had a survey weight w_i for $i = 1, ..., n$. In addition, one or two 24hr dietary recalls were available for each individual. Along with the dietary variables, there are covariates such as age, gender, ethnicity, family income and dummy variables that indicate a weekday or a weekend day, and whether the recall was the first or second reported for that individual.

Using the 24hr recall data reported, for each of the episodically consumed food groups, two variables are defined: (a) whether a food from that group was consumed; and (b) the amount of the food that was reported on the 24hr recall. For the 6 daily-consumed food groups and nutrients, only one variable indicating the consumption amount is defined. In addition, the amount of energy that is calculated from the 24hr recall is of interest. The number of dietary variables for each 24hr recall is thus 12+6+1 = 19. The observed data are Y_{ijk} for the i^{th} person, the j^{th} variable and the k^{th} replicate, $j = 1, \ldots, 19$ and $k = 1, \ldots, m_i$. In the data set, at most two 24hr recalls were observed, so that $m_i \leq 2$. Set $\widetilde{Y}_{ik} = (Y_{i1k}, ..., Y_{i,19,k})^{\mathrm{T}}$, where

- $Y_{i,2\ell-1,k}$ = Indicator of whether dietary component # ℓ is consumed, with $\ell = 1, 2, 3, 4, 5, 6$.
- $Y_{i,2\ell,k}$ = Amount of food # ℓ consumed. This equals zero, of course, if none of food #ℓ is consumed, with $\ell = 1, 2, 3, 4, 5, 6$.
- $Y_{i,\ell+6,k}$ = Amount of non-episodically consumed food or nutrient #ℓ, with $\ell = 7, 8, 9, 10, 11, 12$.
- $Y_{i,19,k}$ = Amount of energy consumed as reported by the 24hr recall.

C. Model and Methods

1. Basic Model Description

Our model is a generalization of work by Tooze et al. (2006) and Kipnis et al. (2009) for a single food and Kipnis et al. (2010b) and Zhang et al. (2010) for a single food and nutrient. Observed data will be denoted as Y, and covariates in the model will be denoted as X. As is usual in measurement error problems, there will also be latent variables, which will be denoted by W.

We use a probit threshold model. Each of the 6 episodically consumed foods will have 2 sets of latent variables, one for consumption and one for amount, while the 6 daily-consumed foods and nutrients as well as energy will have 1 set of latent variables, for a total of 19. The latent random variables are ϵ_{ijk} and U_{ij}, where $(U_{i1}, \ldots, U_{i,19}) = \text{Normal}(0, \Sigma_u)$ and $(\epsilon_{i1k}, \ldots, \epsilon_{i,19,k}) = \text{Normal}(0, \Sigma_\epsilon)$ are mutually independent. In this model, food $\ell = 1, ..., 6$ being consumed on day k is equivalent to observing the binary $Y_{i,2\ell-1,k}$, where

$$Y_{i,2\ell-1,k} = 1 \iff W_{i,2\ell-1,k} = X^{\mathrm{T}}_{i,2\ell-1,k}\beta_{2\ell-1} + U_{i,2\ell-1} + \epsilon_{i,2\ell-1,k} > 0. \tag{3.1}$$

If the food is consumed we model the amount reported $Y_{i,2\ell,k}$ as

$$[g_{\mathrm{tr}}(Y_{i,2\ell,k}, \lambda_\ell)|Y_{i,2\ell-1,k} = 1] \;\; = \;\; W_{i,2\ell,k} = X^{\mathrm{T}}_{i,2\ell,k}\beta_{2\ell} + U_{i,2\ell} + \epsilon_{i,2\ell,k}, \tag{3.2}$$

where $g_{\mathrm{tr}}(y, \lambda) = \sqrt{2}\{g(y, \lambda) - \mu(\lambda)\}/\sigma(\lambda)$, $g(y, \lambda)$ is the usual Box-Cox transformation with transformation parameter λ, and $\{\mu(\lambda), \sigma(\lambda)\}$ are the sample mean and standard deviation of $g(y, \lambda)$, computed from the non-zero food data. This standardization is simply a convenient device to improve the numerical performance of our algorithm without affecting the conclusions of our analysis.

The reported consumption of daily consumed foods or nutrients $\ell = 7, \ldots, 12$ are modeled as

$$g_{\mathrm{tr}}(Y_{i,\ell+6,k}, \lambda_\ell) \;\; = \;\; W_{i,\ell+6,k} = X^{\mathrm{T}}_{i,\ell+6,k}\beta_{\ell+6} + U_{i,\ell+6} + \epsilon_{i,\ell+6,k}. \tag{3.3}$$

Finally, energy is modeled as

$$g_{\mathrm{tr}}(Y_{i,19,k}, \lambda_{13}) \;\; = \;\; W_{i,19,k} = X^{\mathrm{T}}_{i,19,k}\beta_{19} + U_{i,19} + \epsilon_{i,19,k}. \tag{3.4}$$

As seen in (3.2)-(3.4), different transformations $(\lambda_1, ..., \lambda_{13})$ are allowed to be used for the different types of dietary components, see Appendix 23.

In summary, there are latent variables $\widetilde{W}_{ik} = (W_{i1k}, ..., W_{i,19,k})^{\mathrm{T}}$, latent random effects $\widetilde{U}_i = (U_{i1}, ..., U_{i,19})^{\mathrm{T}}$, fixed effects $(\beta_1, ..., \beta_{19})$, and design matrices $(X_{i1k}, \ldots, X_{i,19,k})$. Define $\widetilde{\epsilon}_{ik} = (\epsilon_{i1k}, ..., \epsilon_{i,19,k})^{\mathrm{T}}$. The latent variable model is

$$W_{ijk} \quad = \quad X_{ijk}^{\mathrm{T}}\beta_j + U_{ij} + \epsilon_{ijk}, \tag{3.5}$$

where $\widetilde{U}_i = \text{Normal}(0, \Sigma_u)$ and $\widetilde{\epsilon}_{ik} = \text{Normal}(0, \Sigma_\epsilon)$ are mutually independent.

2. Restriction on the Covariance Matrix

Two necessary restrictions are set on Σ_ϵ. First, following Kipnis et al. (2009, 2010b), $\epsilon_{i,2\ell-1,k}$ and $\epsilon_{i,2\ell,k}$, $(\ell = 1, \ldots, 6)$ are set to be independent. Second, in order to technically identify $\beta_{2\ell-1}$ and the distribution of $U_{i,2\ell-1}$ $(\ell = 1, \ldots, 6)$, we require that $\text{var}(\epsilon_{i,2\ell-1,k}) = 1$, because otherwise the marginal probability of consumption of dietary component $\#\ell$ would be $\Phi\{(X_{i,2\ell-1,k}^{\mathrm{T}}\beta_{2\ell-1} + U_{i,2\ell-1})/\text{var}^{1/2}(\epsilon_{i,2\ell-1,k})\}$, and thus components of β and Σ_u would be identified only up to the scale $\text{var}^{1/2}(\epsilon_{i,2\ell-1,k})$.

So that we can handle any number of episodically consumed dietary components and any number of daily consumed components, suppose that there are J episodically consumed dietary components, and K daily consumed dietary components, and in addition there is energy. Then the restrictions defined above lead to the covariance

matrix

$$\Sigma_\epsilon = \begin{bmatrix} 1 & 0 & \cdots & s_{1,2J+1} & \cdots & s_{1,2J+K+1} \\ 0 & s_{22} & \cdots & s_{2,2J+1} & \cdots & s_{2,2J+K+1} \\ s_{13} & s_{23} & \cdots & s_{3,2J+1} & \cdots & s_{3,2J+K+1} \\ s_{14} & s_{24} & \cdots & s_{4,2J+1} & \cdots & s_{4,2J+K+1} \\ \vdots & \vdots & \ddots & \vdots & \cdots & \vdots \\ s_{1,2J+1} & s_{2,2J+1} & \cdots & s_{2J+1,2J+1} & \cdots & s_{2J+1,2J+K+1} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ s_{1,2J+K+1} & s_{2,2J+K+1} & \cdots & s_{2J+1,2J+K+1} & \cdots & s_{2J+K+1,2J+K+1} \end{bmatrix}. \quad (3.6)$$

The difficulty with parameterizations of (3.6) is that the cells that are not constrained to be 0 or 1 cannot be left unconstrained, otherwise (3.6) need not be a covariance matrix, i.e., positive semidefinite.

We have developed an unconstrained parameterization that results in the structure (3.6). Consider an unconstrained lower triangular matrix V and define $\Sigma_\epsilon = VV^{\mathrm{T}}$. This is positive semidefinite and therefore qualifies Σ_ϵ as a proper covariance matrix. The form of V is

$$V = \begin{bmatrix} v_{11} & 0 & \cdots & 0 \\ v_{21} & v_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ v_{2J+K+1,1} & v_{2J+K+1,2} & \cdots & v_{2J+K+1,2J+K+1} \end{bmatrix}.$$

To achieve the desired pattern (3.6), we derive the following four restrictions:

$$\begin{aligned} v_{11} &= 1; \\ v_{21} &= 0; \\ \textstyle\sum_{p=1}^{q} v_{qp}^2 &= 1; \quad q = 3, 5, \ldots, 2J-1; \\ \textstyle\sum_{p=1}^{q} v_{qp} v_{q+1,p} &= 0; \quad q = 3, 5, \ldots, 2J-1. \end{aligned}$$

The third restriction can be ensured by the further parameterization

$$\begin{aligned}
v_{31} &= r_1 \sin(\theta_1); \\
v_{32} &= r_1 \cos(\theta_1); \\
v_{33} &= \sqrt{1 - r_1^2}; \\
v_{2q+1,1} &= r_q \sin(\theta_{1+(q-1)^2}); \\
v_{2q+1,p} &= r_q \cos(\theta_{1+(q-1)^2}) \times \ldots \times \cos(\theta_{p-1+(q-1)^2}) \sin(\theta_{p+(q-1)^2}), \\
&\qquad p = 2, \ldots, 2q-1; \\
v_{2q+1,2q} &= r_q \cos(\theta_{1+(q-1)^2}) \times \ldots \times \cos(\theta_{q^2}); \\
v_{2q+1,2q+1} &= \sqrt{1 - r_q^2},
\end{aligned}$$

where $q = 2, 3, \ldots, J-1$; $|r_t| \leq 1$, $t = 1, \ldots, J-1$, and $|\theta_s| \leq \pi$, $s = 1, \ldots, (J-1)^2$.

Similarly, the fourth restriction can be further expressed by setting

$$v_{q+1,q} = -\sum_{p=1}^{q-1} v_{qp} v_{q+1,p} / v_{qq} = -\sum_{p=1}^{q-1} v_{qp} v_{q+1,p} / \sqrt{1 - r^2_{(q-1)/2}}, \quad q = 3, 5, \ldots, 2J-1.$$

Note that $|\Sigma_\epsilon| = |V|^2 = \prod_{q=1}^{2J+K+1} v_{qq}^2 = \prod_{q=1}^{J} v_{2q,2q}^2 \prod_{q=2J+1}^{2J+K+1} v_{q,q}^2 \prod_{q=1}^{J-1} (1 - r_q^2)$.

3. The Use of Sampling Weights

As described in the Appendix, we used the survey sample weights from NHANES both in the model fitting procedure and, after having fit the model, in estimating the distributions of usual intake.

While not displayed here, we redid the model fitting calculations without weighting, because the covariates we use are major players in determining the sampling weights, hence it is reasonable to believe that the model in Section C of this chapter holds both in the sample and in the population. When we did this, the parameter estimates were essentially unchanged.

Thus, we use the sampling weights only for estimation of the population distributions. We actually did this for the purpose of handling the clustering in the sample design. For such a complex statistical procedure as ours, we knew we could not do theoretical standard errors, so we thought about the bootstrap, and realized that putting together a bootstrap for the complex survey would be nearly impossible. However, we already had developed a set of Balanced Repeated Replication (BRR) weights (Wolter, 1995), see Section E of this chapter for details. These BRR weights have the property that, in the frequentist survey sampling sense, they appropriately reflect the clustering in the standard error calculations.

Of course, the use of sampling weights in the modeling provide unbiased estimates of the (super) population parameters of interest. In addition, the use of sampling weights in the distribution estimation provides an estimated distribution that is representative of the US population, not just the sample.

4. Distribution of Usual Intake and the HEI-2005 Scores

We assume here that estimates of Σ_u, Σ_ϵ and β_j for $j = 1, ..., 19$ have been constructed, see Section D of this chapter. Here we discuss what we mean by usual intake for an individual, how to estimate the distribution of usual intakes, how to convert usual intakes into HEI-2005 scores, and how to assess uncertainty.

Consider the first episodically consumed dietary component, a food group, with reporting being done on a weekend. Set $X_{i1,\text{wkend}}$ and $X_{i2,\text{wkend}}$ to be the versions of X_{i1k} and X_{i2k} where the dummy variable has the indicator of the weekend and that the recall is the first one. Following Kipnis et al. (2009), we define the usual intake for an individual on the weekend to be the expectation of the reported intake conditional on the person's random effects $\widetilde{U}_i$. Let the (q, p) element of Σ_ϵ be denoted

as $\Sigma_{\epsilon,q,p}$. As in Kipnis et al. (2009) define

$$g^*_{\rm tr}\{v,\lambda,\Sigma_{\epsilon,q,p}\} \;=\; g^{-1}_{\rm tr}(v,\lambda)+(1/2)\Sigma_{\epsilon,q,p}\frac{\partial^2 g^{-1}_{\rm tr}(v,\lambda)}{\partial v^2}. \tag{3.7}$$

Detailed formulas for this are given in Appendix 22. Then, following the convention of Kipnis et al. (2009), the person's usual intake of the first episodically consumed dietary component on the weekend is defined as

$$T_{i1,\rm wkend}=\Phi(X^{\rm T}_{i1,\rm wkend}\beta_1+U_{i1})g^*_{\rm tr}\left(X^{\rm T}_{i2,\rm wkend}\beta_2+U_{i2},\lambda_1,\Sigma_{\epsilon,2,2}\right).$$

Similarly, let $X_{i1,\rm wkday}$ and $X_{i2,\rm wkday}$ be as above but the dummy variable is appropriate for a weekday. Then the person's usual intake of the first episodically consumed food group on weekdays is defined as

$$T_{i1,\rm wkday}=\Phi(X^{\rm T}_{i1,\rm wkday}\beta_1+U_{i1})g^*_{\rm tr}\left(X^{\rm T}_{i2,\rm wkday}\beta_2+U_{i2},\lambda_1,\Sigma_{\epsilon,2,2}\right).$$

Finally, the usual intake of the first episodically consumed food for the individual is

$$T_{i1}=(4T_{i1,\rm wkday}+3T_{i1,\rm wkend})/7,$$

since Fridays, Saturdays and Sundays are considered to be weekend days. Usual intake for the other episodically consumed food groups is defined similarly.

A person's usual intake of a daily-consumed food group/nutrient and energy on the original scale is defined similarly. Consider, for example, energy, which is the 13^{th} dietary component and the 19^{th} set of terms in the model. Let $X_{i,19,\rm wkend}$ and $X_{i,19,\rm wkday}$ be the versions of $X_{i,19,k}$ where the dummy variable has the indicator of the weekend or weekday, respectively, and that the recall is the first one. Then

$$\begin{aligned}
T_{i,13,\rm wkend} &= g^*_{\rm tr}\left(X^{\rm T}_{i,19,\rm wkend}\beta_{19}+U_{i,19},\lambda_{13},\Sigma_{\epsilon,19,19}\right);\\
T_{i,13,\rm wkday} &= g^*_{\rm tr}\left(X^{\rm T}_{i,19,\rm wkday}\beta_{19}+U_{i,19},\lambda_{13},\Sigma_{\epsilon,19,19}\right);
\end{aligned}$$

$$T_{i,13} \quad = \quad (4T_{i,13,\text{wkday}} + 3T_{i,13,\text{wkend}})/7.$$

Similar formulae are used for the other daily-consumed foods and nutrients.

Finally, the energy-adjusted usual intakes and the HEI-2005 scores are then obtained as in Table 3, using the estimated usual intakes of the dietary components.

To find the joint distribution of usual intakes of the HEI-2005 scores, it is convenient to use Monte-Carlo methods. Recall that w_i is the sampling weight for individual i. Let B be a large number: we set $B = 5,000$. Generate $b = 1, ..., B$ observations $\widetilde{U}_{bi} = \text{Normal}(0, \Sigma_u)$ and then obtain $\widetilde{T}_{bi} = (T_{bi\ell})_{\ell=1}^{13}$ by replacing U_{ij} in their formulae by U_{bij}. With appropriate sample weighting, the $\widetilde{T}_{bi}$ can be used to estimate joint and marginal distributions. Thus, for example, consider the total HEI-2005 score, which is a deterministic function of the usual intakes, say $G(\widetilde{T}_i)$. Its cumulative distribution function is estimated as

$$\widehat{F}(x) \quad = \quad \frac{\sum_{i=1}^{n}\sum_{b=1}^{B} I\{G(\widetilde{T}_{bi}) \leq x\} w_i}{\sum_{i=1}^{n}\sum_{b=1}^{B} w_i}. \tag{3.8}$$

Frequentist standard errors of derived quantities such a mean, median and quantiles can be estimated using the Balanced Repeated Replication (BRR) method (Wolter, 1995), see Section E of this chapter for details.

D. Comments on the Approach to Estimation

Our model (3.2)-(3.4) is a highly nonlinear, mixed effects model with many latent variables and nonlinear restrictions on the covariance matrix Σ_ϵ. As seen in Section C of this chapter, we can estimate relevant distributions of usual intake in the population if we can estimate Σ_u, Σ_ϵ and β_j for $j = 1, ..., 19$. We have found that working within a pseudo-likelihood Bayesian paradigm is a convenient way to do this computation. We emphasize, however, that we are doing this only to get frequentist

parameter estimates based on the well-known asymptotic equivalence of frequentist likelihood estimators and Bayesian posterior means, and especially the consistency of both (Lehmann and Casella, 1998). We are specifically not doing Bayesian posterior inference, since valid Bayesian inference in a complex survey such as NHANES is an immensely challenging task, and because frequentist estimation and inference are the standard in the nutrition community.

Kipnis et al. (2009) were able to get estimates of parameters separately for each food group using the nonlinear mixed effects program NLMIXED in SAS with sampling weights. While this gives estimates of β_j for $j = 1, ..., 19$, it only gives us parts of the covariance matrices Σ_u and Σ_ϵ, and not all the entries. Using the 2001-2004 NHANES data, we have verified that our estimates and the subset of the parameters that can be estimated by one food group at a time using NLMIXED are in close agreement, and that estimates of the distributions of usual intake and HEI-2005 component scores are also in close agreement. We expect this because of the rather large sample size in our data set. Zhang et al. (2010) have shown that even considering a single food group plus energy is a challenge for the NLMIXED procedure, both in time and in convergence, and using this method for the entire HEI-2005 constellation of dietary components is impossible.

Full technical details of the model fitting procedure are given in Appendices 12 - 21.

E. Empirical Work

1. Basic Analysis

We analyzed data from the 2001-2004 National Health and Nutrition Examination Survey (NHANES) for children age 2-8. The study sample consisted of $2,638$ children,

among whom $1,103$ children have two 24hr recalls and the rest have only one. We used the dietary intake data to calculate the 12 HEI-2005 components plus energy. In addition, besides age, gender, race and interaction terms, two covariates were employed, along with an intercept. The first was a dummy variable indicating whether or not the recall was for a weekend day (Friday, Saturday, or Sunday) because food intakes are known to differ systematically on weekends and weekdays. The second was a dummy variable indicating whether the 24hr recall was the first or second such recall, the idea being that there may be systematic differences attributable to the repeated administration of the instrument.

2. Contextual Information

When we ran our program based on the variables in Table 3, the results were disastrous. Mixing of the MCMC sampler was very poor, with long sojourns in different regions.

The reason for this failure to converge depends on the context of the dietary variables. For example, whole grains are a subset of total grains. Thus, if someone consumes any whole grains, then necessarily, with probability 1.0, that person also consumes total grains. Such a restriction cannot be handled by our model, because it would force one of the random effects U to equal infinity. A similar thing happens for energy. Calories coming from saturated fat are a subset of total calories as are calories from SoFAAS, so there is a restriction that total calories must be greater than calories from saturated fat and also greater than calories from SoFAAS. Since the latter sum makes up a significant portion of calories, this restriction is not something that our model can handle well.

Luckily, there is an easy and natural context-based solution. Instead of using total grains in the model, we used grains that are not whole grains, i.e., refined grains,

thus decoupling whole grains and total grains, and removing the restriction mentioned above. Similarly, instead of using total fruit, we use fruit that is not whole fruits, i.e., fruit juices. Additionally, instead of using total vegetables, we use total vegetables excluding dark green and orange vegetables and legumes. Finally, instead of total energy, we use total energy minus the sum of energy from saturated fat (11% of mean energy) and from SoFAAS (35% of mean energy). We recognize that there is overlap of energy from saturated fat and energy from solid fat, but this has no impact on our analysis since total energy has sources other than these two. An alternative of course, would have been to simply use total energy minus energy from SoFAAS,

This is sufficient to estimate the distributions of interest. If, for example, in the new data set T_{i1} represents usual intake of non-whole fruits, and T_{i2} is usual intake of whole fruits, then the usual intake of total fruits is $T_{i1} + T_{i2}$. Similar remarks apply for total grains and total vegetables.

With these new variables, our model mixed well and gave reasonable looking answers that, as mentioned in Section D of this chapter, give similar results to other methods employed with smaller parts of the data set.

3. Estimation of the HEI-2005 Scores

In the introduction, we posed 4 questions to which answers had not been possible previously. The first open question concerned the distribution of the HEI total score. Along the way towards this, Table 4 presents the energy-adjusted distributions of the dietary components used in the HEI-2005. For each dietary component, the first line = estimate from our model, while the second line is its BRR-estimated standard error. Total Fruit, Whole Fruit, Total Vegetables, DOL and Milk are in cups. Total Grains, Whole Grains and Meat and Beans are in ounces. Oil and Sodium are in grams. Saturated Fat and SoFAAS are in % of energy. Further discussion of the size

Table 4. Estimated distributions of energy-adjusted usual intakes for children aged 2-8; NHANES, 2001-2004.

			Percentile						
Component	Units	Mean	5^{th}	10^{th}	25^{th}	50^{th}	75^{th}	90^{th}	95^{th}
Total Fruit	cups/(1000 kcal)	0.70	0.14	0.21	0.37	0.62	0.95	1.30	1.54
		0.02	0.02	0.02	0.02	0.02	0.03	0.05	0.07
Whole Fruit	cups/(1000 kcal)	0.31	0.04	0.07	0.14	0.26	0.42	0.61	0.73
		0.02	0.01	0.01	0.02	0.02	0.03	0.04	0.06
Total Vegetables	cups/(1000 kcal)	0.47	0.23	0.27	0.36	0.46	0.58	0.69	0.77
		0.01	0.02	0.02	0.02	0.01	0.02	0.03	0.03
DOL	cups/(1000 kcal)	0.05	0.00	0.01	0.02	0.03	0.07	0.11	0.15
		0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01
Total Grains	ounces/(1000 kcal)	3.32	2.35	2.54	2.87	3.28	3.72	4.16	4.45
		0.05	0.08	0.07	0.06	0.05	0.06	0.08	0.10
Whole Grains	ounces/(1000 kcal)	0.27	0.05	0.07	0.13	0.23	0.36	0.52	0.64
		0.01	0.01	0.01	0.02	0.01	0.02	0.03	0.04
Milk	cups/(1000 kcal)	0.97	0.28	0.38	0.60	0.90	1.26	1.64	1.90
		0.02	0.03	0.03	0.02	0.02	0.03	0.05	0.07
Meat and Beans	ounces/(1000 kcal)	1.84	1.06	1.21	1.48	1.80	2.16	2.51	2.73
		0.04	0.09	0.08	0.06	0.04	0.04	0.05	0.07
Oil	grams/(1000 kcal)	7.13	4.05	4.60	5.63	6.93	8.41	9.90	10.89
		0.23	0.24	0.21	0.17	0.20	0.35	0.54	0.68
Saturated Fat	% of Energy	11.71	8.56	9.20	10.33	11.64	13.01	14.32	15.13
		0.15	0.25	0.20	0.15	0.15	0.22	0.32	0.38
Sodium	grams/(1000 kcal)	1.49	1.16	1.23	1.34	1.48	1.63	1.77	1.86
		0.01	0.02	0.02	0.01	0.01	0.02	0.03	0.03
SoFAAS	% of Energy	36.93	27.19	29.28	32.87	36.90	40.96	44.61	46.77
		0.48	0.93	0.81	0.63	0.48	0.49	0.64	0.75

of the BRR-estimated standard errors is given in Section F of this chapter. Table 5 presents the distributions of the HEI-2005 individual component scores and the total score, with a graphical view given in Figure 4. In Table 5, for each component score, the first line = estimate from our model, while the second line is its BRR-estimated standard error. The total score is the sum of the individual scores. Further discussion of the size of the BRR-estimated standard errors is given in Section F of this chapter and in the supplementary material. In Figure 4, the horizontal axis is the percentile of interest, e.g., 0.5 refers to the median, while the vertical axis gives percentile of the HEI-2005 scores. Standard error estimates are given in Table 4.

Table 5 presents the first estimates of the distribution of HEI-2005 scores for a vulnerable subgroup of the population, namely children aged 2-8 years. A previous analysis of 2003-04 NHANES data, looking separately at 2-5 year olds and 6-11 year olds, was limited to estimates of mean usual HEI-2005 scores (59.6 and 54.7, respectively, see Fungwe et al., 2009). The mean scores noted here are comparable to those

Table 5. Estimated distributions of the usual intake HEI-2005 scores.

Component	Mean	Percentile 5^{th}	10^{th}	25^{th}	50^{th}	75^{th}	90^{th}	95^{th}
Total Fruit	3.55	0.87	1.31	2.33	3.90	5.00	5.00	5.00
	0.09	0.13	0.14	0.15	0.15	0.00	0.00	0.00
Whole Fruit	3.14	0.49	0.82	1.71	3.24	5.00	5.00	5.00
	0.14	0.12	0.16	0.21	0.26	0.03	0.00	0.00
Total Vegetables	2.16	1.02	1.24	1.63	2.10	2.62	3.15	3.48
	0.06	0.10	0.10	0.07	0.06	0.07	0.12	0.16
DOL	0.62	0.05	0.09	0.21	0.45	0.86	1.38	1.76
	0.04	0.02	0.03	0.04	0.05	0.06	0.08	0.13
Total Grains	4.81	3.92	4.23	4.79	5.00	5.00	5.00	5.00
	0.03	0.13	0.12	0.09	0.00	0.00	0.00	0.00
Whole Grains	0.90	0.16	0.24	0.43	0.75	1.21	1.74	2.13
	0.04	0.04	0.05	0.05	0.05	0.05	0.10	0.14
Milk	6.77	2.15	2.96	4.62	6.91	9.67	10.00	10.00
	0.12	0.23	0.22	0.18	0.17	0.25	0.00	0.00
Meat and Beans	7.22	4.23	4.83	5.91	7.21	8.64	10.00	10.00
	0.16	0.34	0.30	0.23	0.17	0.15	0.11	0.00
Oil	5.92	3.37	3.83	4.69	5.77	7.01	8.25	9.07
	0.18	0.20	0.18	0.14	0.17	0.29	0.45	0.57
Saturated Fat	5.16	0.00	1.09	3.18	5.38	7.48	8.53	8.96
	0.21	0.35	0.51	0.35	0.24	0.23	0.13	0.16
Sodium	4.52	1.25	2.05	3.31	4.62	5.83	6.85	7.44
	0.09	0.30	0.24	0.15	0.09	0.11	0.16	0.19
SoFAAS	8.73	2.15	3.60	6.02	8.73	11.42	13.81	15.21
	0.32	0.50	0.42	0.33	0.32	0.42	0.54	0.62
Total Score	53.50	37.42	40.74	46.73	53.68	60.36	65.87	68.96
	0.81	1.45	1.34	1.09	0.83	0.82	0.96	1.08

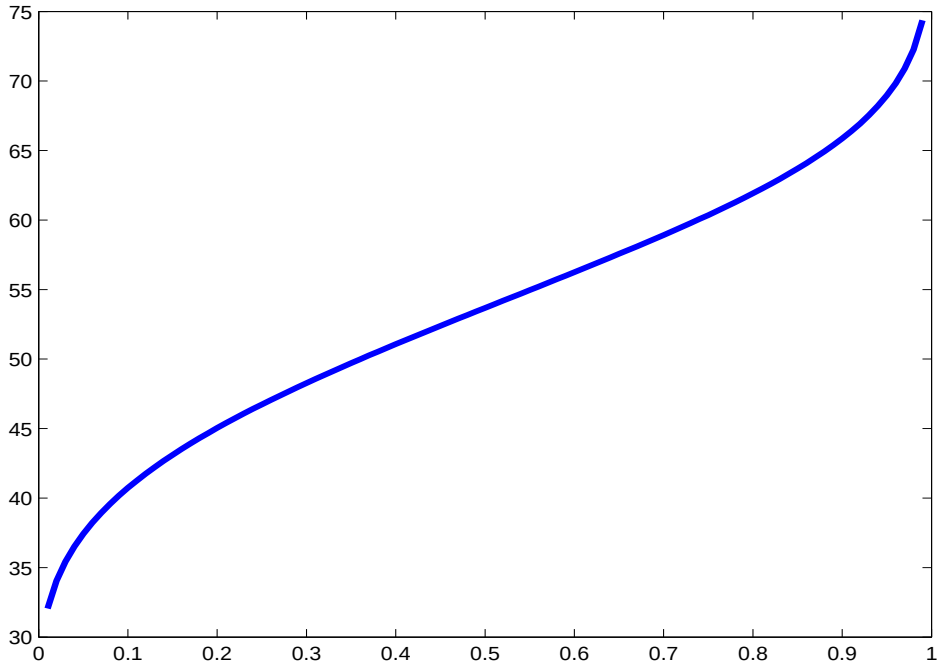

Figure 4. The estimated percentiles of the HEI-2005 total score.

and reinforce the notion that children's diets, on average, are far from ideal. However, this analysis provides a more complete picture of the state of US children's diets. By including the scores at various percentiles, we estimate that only 5% of children have a score of 69 or greater and another 10% have scores of 41 or lower. While not in the table, we also estimate that the 99^{th} percentile is 74. This analysis suggests that virtually all children in the US have suboptimal diets and that a sizeable fraction (10%) have alarmingly low scores (41 or lower.)

We have also considered whether our multivariate model fitting procedure gives reasonable marginal answers. To check this, we note that it is possible to use the SAS procedure NLMIXED *separately for each component* to fit a model with one episodically consumed food group or daily consumed dietary component together with energy. The marginal distributions of each such component done separately are quite close to what we have reported in Table 5, as is our mean, which is 53.50 compared to the mean of 53.25 based on analyzing one HEI-2005 component at a time with the NLMIXED procedure. The only case where there is a mild discrepancy is in the estimated variability of the energy-adjusted usual intake of oils, likely caused by the NLMIXED procedure itself, which has an estimated variance 9 times greater than our estimated variance.

Of course, it is the distribution of the HEI-2005 total score that cannot be estimated by analysis of one component at a time.

There are other things that have not been computed previously that are simple by-products of our analysis. For example, the correlations among energy-adjusted usual intakes involving episodically consumed foods have not been estimated previously, but this is easy for us, see Table 6. Here TF = Total Fruits, WF = Whole Fruits, TV = Total Vegetables, WG = Whole Grains, TG = Total Grains, SatFat = Saturated Fat. The estimated correlation of -0.64 between energy-adjusted to-

Table 6. Estimated correlation matrix for energy-adjusted usual intakes.

Component	TF	WF	TV	DOL	TG	WG	Milk	Meat	Oil	SatFat	Sodium	SoFAAS
TF	1	0.76	0.07	0.41	-0.10	0.33	0.16	0.08	-0.35	-0.38	-0.25	-0.64
WF		1	0.14	0.49	0.03	0.35	0.10	0.05	-0.17	-0.30	-0.20	-0.51
TV			1	0.51	-0.25	-0.23	-0.09	0.51	-0.08	0.08	0.42	-0.16
DOL				1	-0.08	0.11	0.14	0.25	-0.06	-0.23	0.01	-0.47
TG					1	0.30	-0.30	-0.13	0.44	-0.36	0.17	-0.22
WG						1	0.18	-0.18	-0.11	-0.29	-0.17	-0.46
Milk							1	-0.37	-0.21	0.21	-0.27	-0.21
Meat & Beans								1	-0.06	-0.08	0.39	-0.19
Oil									1	-0.06	0.11	0.05
SatFat										1	0.09	0.46
Sodium											1	0.04
SoFAAS												1

tal fruit and energy-adjusted SoFAAS, and the -0.47 correlation between DOL and SoFAAS are surprisingly high.

4. Component Scores and Other Scores

As described in the introduction, an open problem has been to estimate the correlation between the individual score on each dietary component and the scores of all other dietary components. In their Table 3, Guenther et al. (2008b) consider this problem, but of course they did not have a model for usual energy adjusted intakes, and instead they used a single 24hr recall. In Table 7, we show the resulting correlations using (a) a single 24hr recall; (b) the mean of two 24hr recalls for those who have two 24hr recalls; and (c) our model for usual intake. The column labeled "Two 24hr" is the naive analysis that uses the mean of the two 24hr recalls, while the column labeled "First 24hr" is the naive analysis that uses the first 24hr recall. The column labeled "Model" is our analysis, and the column labeled "BRR s.e." is the estimated standard error of our estimates. The numbers for the former differ from that of Guenther et al. (2008b) because we are considering here a different population than do they. A striking and not unexpected aspect of this table is that for those components with non-trivial correlations, the correlations all increase as one moves from a single 24hr

Table 7. Estimated correlations between each individual HEI-2005 component score and the sum of the other HEI component scores, i.e., the difference of the total score and each individual component.

	First 24hr	Two 24hr	Model	BRR s.e.
Total Fruit	0.38	0.44	0.62	0.05
Whole Fruit	0.31	0.37	0.59	0.10
Total Vegetables	0.09	0.11	0.10	0.11
DOL	0.18	0.24	0.41	0.07
Total Grains	0.00	0.00	0.06	0.11
Whole Grains	0.12	0.16	0.53	0.08
Milk	-0.07	-0.01	0.01	0.08
Mean and Beans	-0.03	-0.01	-0.03	0.15
Oil	0.08	0.05	-0.17	0.08
Saturated Fat	0.21	0.23	0.36	0.06
Sodium	-0.03	0.05	0.07	0.12
SoFAAS	0.52	0.59	0.72	0.04

recall to the mean of two 24hr recalls and then finally to estimated usual intake. Thus, for example, the correlation between the HEI-2005 score for total fruit and its difference with the total score is 0.38 for a single 24hr recall, 0.44 for the mean of two 24hr recalls and then finally 0.62 for usual intake.

5. Distributions of Intakes for HEI Total Scores

A third open question is: among those whose total HEI-2005 score is > 50 or ≤ 50, what is the distribution of energy-adjusted usual intake of whole grains, whole fruits, dark green and orange vegetables and legumes (DOL) and calories from solid fats, alcoholic beverages and added sugars (SoFAAS)? This follows naturally from our method. Following (3.8), let $G_1(\widetilde{T}_{bi})$ be energy adjusted usual intake and let $G_2(\widetilde{T}_{bi})$ be the HEI total score. Then the distributions in question for when the total HEI-2005 score is > 50 can be estimated as $\widehat{F}(x) = \sum_{i=1}^{n}\sum_{b=1}^{B} w_i I\{G_1(\widetilde{T}_{bi}) \leq x\} I\{G_2(\widetilde{T}_{bi}) > 50\} / \sum_{i=1}^{n}\sum_{b=1}^{B} w_i I\{G_2(\widetilde{T}_{bi}) > 50\}$. The results are provided in Table 8, with a graphical view in Figure 5. Units of measurement in Table 8 are given in Table 4. Figure 5 gives the estimated percentiles of the energy-adjusted usual intakes for whole fruits (Top left) in cups/(1000 kcal), whole grains (Top right) in ounces/(1000 kcal),

Table 8. Estimated distributions of energy-adjusted usual intake for those whose total HEI-2005 total scores are ≤ 50 and > 50.

Component	Mean	s.d	Percentile 5^{th}	10^{th}	25^{th}	50^{th}	75^{th}	90^{th}	95^{th}
Whole Fruit									
Total Score ≤ 50	0.15	0.12	0.02	0.03	0.07	0.12	0.21	0.30	0.38
Total Score > 50	0.39	0.22	0.11	0.15	0.23	0.35	0.51	0.68	0.80
Whole Grains									
Total Score ≤ 50	0.18	0.13	0.03	0.05	0.09	0.15	0.25	0.36	0.44
Total Score > 50	0.32	0.20	0.07	0.10	0.17	0.28	0.42	0.59	0.70
DOL									
Total Score ≤ 50	0.02	0.02	0.00	0.00	0.01	0.02	0.03	0.05	0.07
Total Score > 50	0.06	0.05	0.01	0.01	0.03	0.05	0.09	0.13	0.17
SoFAAS									
Total Score ≤ 50	42.43	3.97	36.40	37.59	39.66	42.16	44.92	47.67	49.42
Total Score > 50	33.83	4.44	26.01	27.89	30.97	34.15	36.98	39.28	40.57
Total Score	53.50	9.58	37.42	40.74	46.73	53.68	60.36	65.87	68.96

DOL (bottom left) in cups/(1000 kcal) and calories from SoFAAS (bottom right) in % of Energy. The solid lines are for those whose usual HEI-2005 total score is ≤ 50, i.e., poorer diets, while the dashed lines are for those whose usual HEI-2005 total score is > 50, i.e., better diets. The results show that those who have poorer diets with usual HEI-2005 total score ≤ 50 are consistently eating poorer diets, i.e., less whole fruits, less whole grains and less DOL, but higher SoFAAS.

6. Dietary Consistency

We stated in the introduction that it is interesting to understand the percentage of children whose usual intake HEI score exceeds the median HEI score on all 12 HEI components. Those median scores, say $(\kappa_1, ..., \kappa_{12})$, are estimated in Table 5. If $G_j(\widetilde{T}_{bi})$ is the HEI component score for episodically consumed food j, then following (3.8) the quantity in question can be estimated as $\sum_{i=1}^{n}\sum_{b=1}^{B} w_i \prod_{j=1}^{6} I\{G_j(\widetilde{T}_{bi}) \geq \kappa_j\}/\sum_{i=1}^{n}\sum_{b=1}^{B} w_i$. We estimate that the percentage is 6%, woefully small. The percentage of children whose usual intake HEI score exceeds the median HEI score on all 12 HEI components is 0.24%. Figure 6 gives the estimated probabilities (Y-axis)

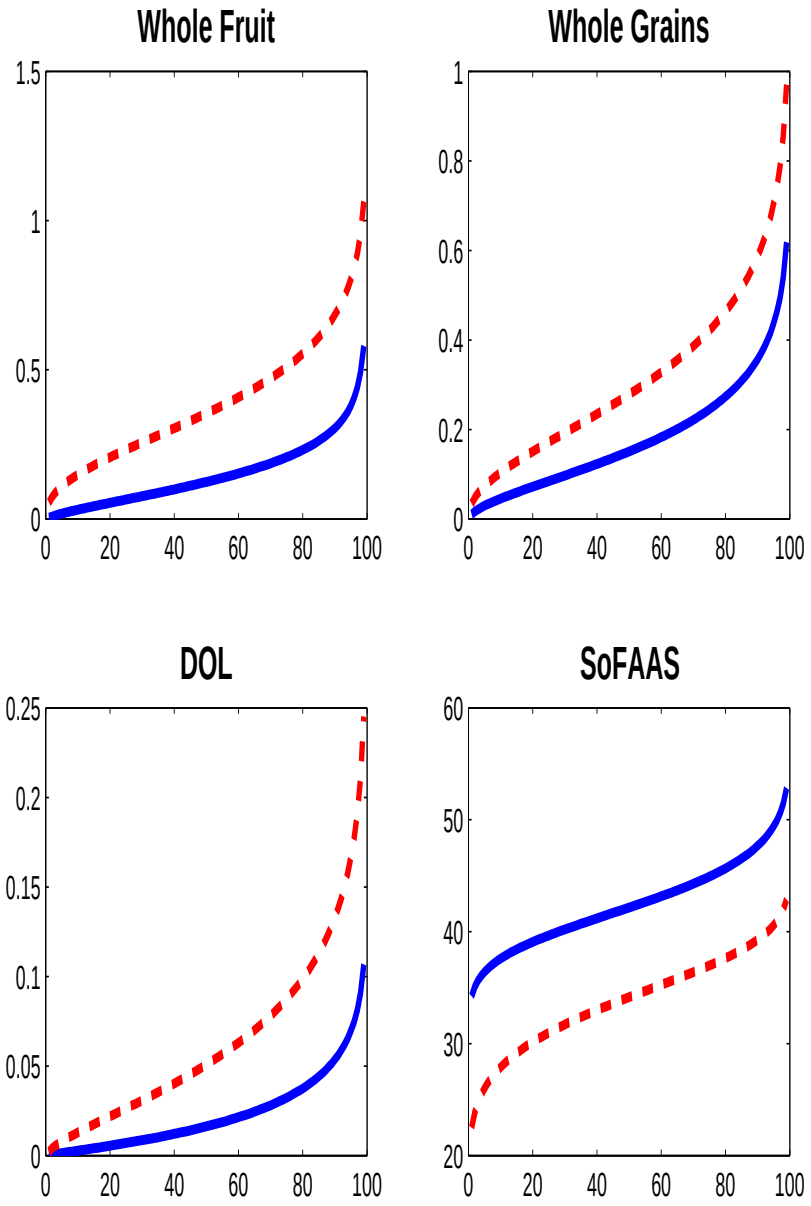

Figure 5. The estimated percentiles of the energy-adjusted usual intakes for whole fruits, whole grains, DOL and calories from SoFAAS.

of exceeding the κ (X-axis) percentile on all 12 HEI components simultaneously, for $\kappa = 1, 2, ..., 99$.

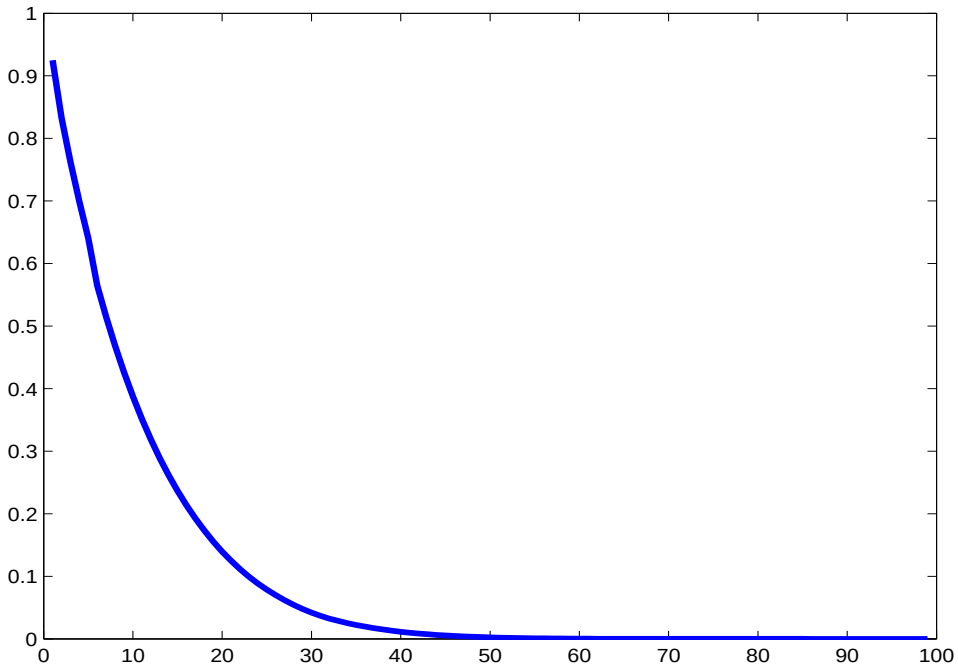

Figure 6. Dietary consistency.

7. Uncertainty Quantification

The BRR standard errors of HEI-2005 components' adjusted usual intakes and scores are shown in Tables 4 and 5. The BRR weights are only used in variance calculations. Once we have estimated some quantity, say $\widehat{\theta}$, from the sample using sample weight, we will need to compute the same quantity using, in succession, the 32 BRR weights. This will give us 32 estimates $\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_{32}$. The BRR estimate for the variance of $\widehat{\theta}$ is $(32 \times 0.49)^{-1} \sum_{p=1}^{32} (\widehat{\theta}_p - \widehat{\theta})^2$. The 32 in the denominator is for the 32 different estimates from the 32 different sets of weights, and the 0.49 is the square of the

perturbation factor used to construct the BRR weight sets (Wolter, 1995).

F. Further Discussion of the Analysis

1. Never Consumers

An aspect of the modeling that we have not discussed is the possibility that some people never, ever consume an episodically consumed dietary component. Our model does not allow for this, for general reasons and for reasons that are specific to our data analysis.

It is in principle possible to add an additional modeling step for non-consumers, via fixed effects probit regression, but we do not think this is a practical issue in our case, for two reasons.

- The first is that the HEI-2005 is based on 6 episodically consumed dietary components, namely total fruit, whole fruit, whole grains, total vegetables, DOL, and milk, the latter of which includes cheese, yogurt and soy beverages. None of these are "lifestyle adverse", unlike say alcohol. While 40% of the responses for whole fruits, for example, equal zero, the percentage of children who never eat any whole fruits at all is likely to be minuscule.

- Even if one disputes whether there are very few individuals who never consume one of the dietary components, then it necessarily follows that we have overestimated the HEI-2005 total scores, and hence the estimates of the proportion of individuals with alarmingly low HEI scores are deflated, and not inflated. The reason is that our model suggests everyone has a positive usual intake of the 6 episodically consumed dietary components. Since the HEI-2005 score components are nondecreasing functions of usual intake of the episodi-

Table 9. BRR estimated standard errors of HEI-2005 component energy-adjusted usual intakes for 250 randomly selected children with replicate 24hr.

Component	se(Mean)	se(Q5)	se(Q10)	se(Q25)	se(Q50)	se(Q75)	se(Q90)	se(Q95)
Total Fruit	0.06	0.06	0.07	0.07	0.07	0.09	0.12	0.14
Whole Fruit	0.04	0.01	0.02	0.02	0.03	0.05	0.07	0.09
Total Vegetables	0.04	0.06	0.06	0.05	0.04	0.04	0.06	0.07
DOL	0.02	0.00	0.00	0.01	0.01	0.02	0.03	0.04
Total Grains	0.17	0.14	0.14	0.15	0.17	0.20	0.25	0.29
Whole Grains	0.03	0.02	0.02	0.03	0.04	0.04	0.06	0.07
Milk	0.08	0.07	0.06	0.05	0.06	0.12	0.19	0.25
Meat and Beans	0.10	0.17	0.15	0.13	0.10	0.09	0.10	0.11
Oil	0.50	0.42	0.41	0.41	0.46	0.61	0.87	1.09
Saturated Fat	0.51	0.47	0.43	0.44	0.52	0.65	0.77	0.85
Sodium	0.03	0.05	0.04	0.03	0.03	0.04	0.07	0.09
SoFAAS	0.99	1.63	1.48	1.24	1.04	1.05	1.28	1.50

cally consumed dietary components, this would mean that we overestimate the HEI-2005 total score.

2. Complexity of the Data and Sample Size

The complexity of the modeling may make it seem like a miracle that we have been able to get results. Actually, because we have $1,103$ children with replicate 24hr measurements, and the highest amount of reported zeros is $< 50\%$, we have a great deal of data for estimating Σ_ϵ and for estimating Σ_u. To show that smaller sample sizes result in significantly larger variability, we reran the analysis by using only a randomly selected 250 of the children with replicate 24hr, and the BRR estimated variances go up more than a factor of 7, on average, see Tables 9 and 10. The point of these tables is to show that in smaller sample sizes, standard errors do increase substantially. The phrase "se(Mean)" is the standard error estimate for the estimated mean, and "se(Q)" is the estimated standard error of the particular quantile.

3. Comparisons When Measurement Error is Ignored

It is interesting to understand how some of the results change if the 24hr recalls were used directly as if they were usual intake. We redo Tables 4 and 5 in two ways:

Table 10. BRR estimated standard errors of HEI-2005 component scores for 250 randomly selected children with replicate 24hr.

Component	se(Mean)	se(Q5)	se(Q10)	se(Q25)	se(Q50)	se(Q75)	se(Q90)	se(Q95)
Total Fruit	0.24	0.40	0.46	0.45	0.41	0.00	0.00	0.00
Whole Fruit	0.31	0.19	0.21	0.28	0.42	0.62	0.09	0.00
Total Vegetables	0.16	0.28	0.27	0.23	0.18	0.17	0.25	0.33
DOL	0.20	0.03	0.04	0.08	0.16	0.29	0.44	0.55
Total Grains	0.15	0.24	0.24	0.25	0.24	0.00	0.00	0.00
Whole Grains	0.11	0.06	0.08	0.10	0.12	0.14	0.19	0.24
Milk	0.35	0.54	0.50	0.42	0.49	0.83	0.09	0.01
Meat and Beans	0.38	0.67	0.62	0.51	0.40	0.35	0.06	0.01
Oil	0.38	0.35	0.34	0.34	0.38	0.51	0.72	0.68
Saturated Fat	0.72	0.88	1.17	1.03	0.84	0.70	0.48	0.31
Sodium	0.26	0.79	0.63	0.40	0.25	0.28	0.39	0.44
SoFAAS	0.66	1.00	0.86	0.70	0.69	0.82	0.99	1.09
Total Score	2.14	3.48	3.09	2.57	2.30	2.21	2.26	2.36

Table 11. Estimated distributions of a single energy-adjusted 24-hour recall for children ages 2-8; NHANES, 2001-2004.

			Percentile						
Component	Units	Mean	5^{th}	10^{th}	25^{th}	50^{th}	75^{th}	90^{th}	95^{th}
Total Fruit	cups/(1000 kcal)	0.72	0.00	0.00	0.11	0.55	1.09	1.67	2.12
Whole Fruit	cups/(1000 kcal)	0.31	0.00	0.00	0.00	0.06	0.48	0.90	1.26
Total Vegetables	cups/(1000 kcal)	0.48	0.01	0.07	0.19	0.39	0.66	1.02	1.29
DOL	cups/(1000 kcal)	0.05	0.00	0.00	0.00	0.00	0.02	0.19	0.32
Total Grains	ounces/(1000 kcal)	3.31	1.35	1.73	2.38	3.19	4.10	5.13	5.64
Whole Grains	ounces/(1000 kcal)	0.27	0.00	0.00	0.00	0.13	0.39	0.79	1.07
Milk	cups/(1000 kcal)	0.99	0.00	0.08	0.44	0.87	1.40	1.94	2.38
Meat and Beans	ounces/(1000 kcal)	1.86	0.17	0.42	0.98	1.75	2.50	3.26	3.89
Oil	grams/(1000 kcal)	6.88	0.15	0.72	2.36	5.60	10.09	14.34	19.31
Saturated Fat	% of energy	11.67	6.27	7.41	9.26	11.41	13.89	16.05	17.73
Sodium	grams/(1000 kcal)	1.50	0.90	1.03	1.22	1.46	1.71	2.02	2.23
SoFAAS	% of energy	36.72	19.41	23.99	29.95	36.75	43.60	49.87	53.22

- Using only the first 24 hour recall for everyone as a measure of their usual intake.

- Using the mean of the two 24 hour recalls as a measure of their usual intake for those who have two 24 hour recalls.

Because of the measurement error, the naive methods give distributions with too large a variance. This is seen in Tables 11-14. The point of this table is to compare our results, which correct for the measurement errors in the 24hr, with the naive results that ignore measurement error. The total score in Tables 12 and 14 is the sum of the individual scores.

Table 12. Estimated distributions of the HEI-2005 scores for a single energy-adjusted 24-hour recall for children ages 2-8; NHANES, 2001-2004.

Component	Mean	Percentile 5^{th}	10^{th}	25^{th}	50^{th}	75^{th}	90^{th}	95^{th}
Total Fruit	2.94	0.00	0.00	0.66	3.46	5.00	5.00	5.00
Whole Fruit	2.14	0.00	0.00	0.00	0.72	5.00	5.00	5.00
Total Vegetables	2.06	0.02	0.30	0.86	1.79	2.99	4.64	5.00
DOL	0.56	0.00	0.00	0.00	0.00	0.31	2.31	3.96
Total Grains	4.39	2.25	2.88	3.96	5.00	5.00	5.00	5.00
Whole Grains	0.88	0.00	0.00	0.00	0.43	1.32	2.62	3.57
Milk	6.22	0.02	0.61	3.38	6.66	10.00	10.00	10.00
Meat and Beans	6.52	0.66	1.69	3.91	7.01	10.00	10.00	10.00
Oil	5.06	0.12	0.60	1.97	4.67	8.41	10.00	10.00
Saturated Fat	5.23	0.00	0.00	1.78	5.75	8.49	9.72	10.00
Sodium	4.65	0.00	0.00	2.56	4.78	6.92	8.35	9.00
SoFAAS	9.01	0.00	0.09	4.27	8.83	13.36	17.34	20.00
Total Score	49.66	28.30	31.66	40.20	50.01	58.93	67.28	71.72

Table 13. Estimated distributions of the energy-adjusted 2-day mean 24-hour recall for children ages 2-8, NHANES, 2001-2004.

Component	Units	Mean	Percentile 5^{th}	10^{th}	25^{th}	50^{th}	75^{th}	90^{th}	95^{th}
Total Fruit	cups/(1000 kcal)	0.76	0.01	0.10	0.31	0.62	1.11	1.59	1.86
Whole Fruit	cups/(1000 kcal)	0.34	0.00	0.00	0.02	0.24	0.53	0.87	1.09
Total Vegetables	cups/(1000 kcal)	0.50	0.08	0.15	0.28	0.46	0.68	0.90	1.07
DOL	cups/(1000 kcal)	0.05	0.00	0.00	0.00	0.00	0.07	0.15	0.22
Total Grains	ounces/(1000 kcal)	3.17	1.68	1.95	2.51	3.10	3.79	4.44	4.81
Whole Grains	ounces/(1000 kcal)	0.30	0.00	0.00	0.05	0.20	0.43	0.72	0.90
Milk	cups/(1000 kcal)	0.99	0.11	0.25	0.50	0.87	1.36	1.88	2.24
Meat and Beans	ounces/(1000 kcal)	1.88	0.52	0.84	1.27	1.83	2.45	3.04	3.44
Oil	grams/(1000 kcal)	7.10	1.15	1.89	4.01	6.66	9.62	12.76	14.36
Saturated Fat	% of energy	11.82	7.58	8.48	10.00	11.73	13.41	15.06	16.43
Sodium	grams/(1000 kcal)	1.50	1.05	1.14	1.26	1.47	1.68	1.90	2.01
SoFAAS	% of energy	35.32	21.80	25.20	29.87	34.63	40.73	46.52	49.65

Table 14. Estimated distributions of the HEI-2005 scores for the 2-day mean 24-hour recall for children ages 2-8, NHANES, 2001-2004.

Component	Mean	Percentile 5^{th}	10^{th}	25^{th}	50^{th}	75^{th}	90^{th}	95^{th}
Total Fruit	3.37	0.03	0.60	1.92	3.89	5.00	5.00	5.00
Whole Fruit	2.73	0.00	0.00	0.27	3.06	5.00	5.00	5.00
Total Vegetables	2.24	0.37	0.68	1.28	2.08	3.10	4.08	4.88
DOL	0.60	0.00	0.00	0.00	0.03	0.88	1.92	2.72
Total Grains	4.49	2.80	3.26	4.18	5.00	5.00	5.00	5.00
Whole Grains	0.97	0.00	0.00	0.18	0.65	1.44	2.41	2.99
Milk	6.47	0.88	1.96	3.83	6.68	10.00	10.00	10.00
Meat and Beans	6.98	2.09	3.37	5.07	7.32	9.80	10.00	10.00
Oil	5.62	0.96	1.58	3.34	5.55	8.02	10.00	10.00
Saturated Fat	5.03	0.00	0.00	2.54	5.24	8.00	9.02	9.62
Sodium	4.58	0.00	0.87	2.83	4.70	6.55	7.69	8.24
SoFAAS	9.82	0.23	2.32	6.18	10.24	13.42	16.54	18.80
Total Score	52.90	32.86	35.97	44.47	53.52	61.24	68.71	72.22

4. Sizes of Standard Errors

The standard errors of the BRR-estimated standard deviations may appear too small relative to the percentiles themselves. To check whether this is the case, we took the mean and standard deviation reported in Table 4, and used the method of moments to fit a Gamma distribution as a rough approximation. We then computed model-based 95^{th} percentiles over $1,000$ simulated data sets of size 100 and 500. We used model-based percentiles because except for the weights, ours is a model-based estimator.

The results are displayed in Table 15. "se(Q95) Paper" is the standard error of the 95^{th} percentile as reported in the paper, while "se(Q95) Model" is the standard error of the 95^{th} model-based percentile, based on samples of effective size $n = 100$ and $n = 500$. The main point of this table is to show that the BRR-estimated standard errors of the 95^{th} percentiles in our data analysis are consistent with the standard errors that would have been obtained if the Gamma distribution were correct. They show that the BRR-estimated standard errors of the 95^{th} percentiles are consistent with the standard errors that would have been obtained if the Gamma distribution were correct, even if the effective sample size is 100. If the effective sample size is 500, then our estimated standard deviations of the 95^{th} percentile are far too large. The key point is that the standard errors are not minute compared to what they should reasonably be.

One might also notice that the standard errors of the 5^{th} percentile can be smaller than the standard errors of the 25^{th} percentile. This actually makes sense, because the data are all positive, not normally distributed, so at the left tail the 5^{th} percentile might actually be rather well-determined. Using the same Gamma model as above, with a sample size of $n = 100$, we compared the ratio of the standard deviations of the model-based 95^{th} percentile to the sample standard deviation, and the ratio of

Table 15. Comparison of standard errors when the data are Gamma distributed with method of moments parameter estimates.

Dietary Component	se(Q95) Paper	se(Q95) Model $n = 100$	se(Q95) Model $n = 500$
Total Fruit	0.07	0.13	0.06
Whole Fruit	0.06	0.06	0.03
Total Vegetables	0.03	0.12	0.05
DOL	0.01	0.05	0.02
Total Grains	0.10	0.04	0.02
Whole Grains	0.04	0.02	0.01
Milk	0.07	0.12	0.06
Meat and Beans	0.07	0.10	0.05
Oil	0.68	0.41	0.20
Saturated Fat	0.38	0.37	0.17
Sodium	0.03	0.04	0.02
SoFAAS	0.75	1.09	0.51

the standard deviations of the 5^{th} and 25^{th} percentiles. The point of Table 16 is to show that, with various Gamma distributions, the standard error of the model-based estimated 5^{th} percentile is not necessarily larger than that of the 25^{th} percentile. As seen in Table 16, for most of the episodically consumed dietary components, the standard deviation of the 5^{th} percentile was smaller than the standard deviation of the 25^{th} percentile, as observed with the BRR-estimated standard errors. "5 vs 25" is the ratio of the standard error of the 5^{th} model-based percentile to the 25^{th} model-based percentile, with a sample of size $n = 100$.

Table 16. Comparison of standard errors when the data are Gamma distributed with method of moments parameter estimates.

Dietary Component	5 vs 25
Total Fruit	0.80
Whole Fruit	0.71
Total Vegetables	1.21
DOL	0.72
Total Grains	1.09
Whole Grains	0.41
Milk	0.93
Meat and Beans	1.17
Oil	1.13
Saturated Fat	1.24
Sodium	1.25
SoFAAS	1.21

One might also notice in Table 5 that the expected pattern of higher standard errors for higher percentiles does not always obtain. The reason for this has to do

with the nature of the HEI-2005 scores, which have maximum values. Looking at Total Fruit, for example, we see that the 75^{th} percentile is already at the maximum HEI-2005 component score, namely 5.0, and so there is no real variability in the estimate of the 95^{th} percentile, which necessarily also equals 5.0.

5. Computing and Data

Our program was written in Matlab. It is available in the *Annals of Applied Statistics* online archive, and also on the last author's web site. In addition, we have created data that mimic the NHANES data, and put it in the online archive. Although a much smaller amount of computing effort yields similar results, using $70,000$ MCMC steps with a burn-in of $20,000$ takes approximately 10 hours on a Linux server.

We also estimated the Monte Carlo standard error which is defined by Flegal et al. (2008) as $\widehat{\sigma}_g/\sqrt{n}$, where n is the total of iterations, and $n = ab$, where a is the number of blocks and b is the block size, and where

$$\bar{Y_j} = b^{-1} \sum_{i=(j-1)b+1}^{jb} g(X_i) \quad for \quad j = 1, \ldots, a.$$

The batch means estimate of σ_g^2 is

$$\widehat{\sigma}_g^2 = \frac{b}{a-1} \sum_{j=1}^{a} (\bar{Y_j} - \bar{g}_n)^2.$$

The ratio of the Monte Carlo standard error to the estimated standard deviation of the estimated parameters averages 3.4% for Σ_u and 1.7% for β.

Because of the public health importance of the problem, the National Cancer Institute has contracted for the creation of a SAS program that performs our analysis. It will allow any number of episodically and daily consumed dietary components. The first draft of this program, written independently in a different programming language,

gives almost identical results to what we have obtained, at least suggesting that our results are not the product of a programming error.

CHAPTER IV

REGRESSION-BASED PREDICTIVE MODELS IN HEALTHCARE

A. Introduction

The pressure of cost containment makes efficient utilization of existing resources a top priority for medical institutions. Surgeons, nurses and operating rooms are indispensable for a surgery to be performed, and are important resources for a medical service provider. The schedule of these resources should be based upon surgical case durations (Weiss, 1990; Olivares et al., 2008), which, however, can be highly variable. Such variability poses a serious challenge to surgical scheduling and resource utilization (Litvak and Long, 2000; McManus et al., 2003). Accurately predicting surgical case duration is a pressing need in hopital management.

Surgery, by nature, involves a series of physical activities. Each surgery is characterized by one or multiple current procedural terminology (CPT) codes. A CPT code is a five-digit number that represents a set of medical, surgical or diagnostic services. The CPT code or the combination of CPT codes that is prescribed for a surgery dictates the core actions taken during the surgery. CPT codes are maintained by the American Medical Association for uniformity. Naturally, CPT codes are a key factor that determines the duration of a surgery. Having surveyed the articles in the area of general thoracic surgery, Dexter et al. (2008) find that the precise procedure types, which are represented by CPT codes, were the *most important* factor when predicting surgical case durations. Ignoring the critical information conveyed in CPT codes will often lead to unsatisfactory predictions of surgical case durations. For instance, Combes et al. (2008) use data mining tools to predict the duration of surgeries. Their results, while shedding lights on the benefit of applying data warehousing

models, are reportedly not satisfactory. The authors believe that their grouping of surgeries based on diagnoses rather than procedure types is the main reason for inaccuracy. Motivated by the need to incorporate surgery type information, we present in this project predictive methods for surgical case durations based directly on the CPT codes included in each surgery.

Modeling surgical case durations has been a topic of interest for operations management and medical communities. Given the role CPT codes play in a surgery, the majority of the existing literature involve CPT codes directly or indirectly. There are two major lines of approaches among existent work utilizing CPT related information. One relies on linear regressions for estimating surgical case durations or identifying the crucial factors that affect variability in surgeries. In a multi-phase study, Wright et al. (1996) find that surgeons provide better time estimates than the scheduling software adopted in their institution. Due to this finding, Wright et al. (1996) develop regression models for predicting surgical case durations by including, as the explanatory variables (or independent variables), the surgeons' own estimates, the estimates from the scheduling software, and several other characteristics of a surgery. CPT codes are not directly included as part of the explanatory variables in their regression models. But surgeons are aware of the CPT codes prescribed for a surgery when making their estimates. As such, CPT codes are implicitly utilized. The regression models are shown to outperform both surgeons and the scheduling software. This study supports the inclusion of CPT codes as the explanatory variables for predicting surgical case durations. Strum et al. (2000a) investigate factors associated with variability in surgery durations. They select surgeries with a single CPT code that were repeatedly performed by one or multiple surgeons. A five-factor main-effects linear model is established for each CPT code under consideration. This study identified surgeon as the second most important source of variability after the CPT codes.

The other line of work studies the fitness of known distributions, notably the normal distribution and the lognormal distribution, for the purpose of predicting surgery case durations. Strum et al. (2000b) examine a large set of real surgery data and test how well the lognormal and normal distributions fit the data set. In their study, only surgeries with a single CPT code are considered, and the surgeries are categorized based on its CPT-anesthesia combination. Goodness-of-fit tests are conducted for each of those CPT-anesthesia combinations. Strum et al. (2000b) conclude that lognormal distributions fit the surgery data better than normal distributions. They also note that the Shapiro-Wilk goodness-of-fit test can sometimes reject lognormal distributions that seemingly fit the surgery data, and thereby suggest using normal probability plots together with goodness-of-fit tests. The lognormal distributions investigated by Strum et al. (2000b) have two parameters, namely the mean and variance associated with the normal distribution after a logarithm transformation. Their work is extended in May et al. (2000) and Spangler et al. (2004), where a third parameter (called *location parameter*) is added to a lognormal distribution. Both papers compare various strategies that estimate the location parameter. Using the same data set as the one in Strum et al. (2000b), May et al. (2000) show that the skewness of data is an effective indicator that identifies the best estimation strategy. They also observe that when the skewness of data is small, the two-parameter lognormal models outperform the three-parameter lognormal models (i.e., the one with a location parameter). Spangler et al. (2004) suggest using a properly chosen order statistics for estimating the location parameter in a lognormal distribution. Both simulated and real surgical data (again, with single CPT codes) are used to test different estimation strategies in Spangler et al. (2004).

Surgeries consisting of exactly two CPT codes are the focus of Strum et al. (2003). Treating permutations of the same CPT codes as different combinations of

CPT codes, Strum et al. (2003) perform Shapiro-Wilk goodness-of-fit tests to examine the fitness of the lognormal and normal distributions for each combination of CPT codes. They conclude that lognormal distributions provide a better fit. Building on this result, Strum et al. (2003) apply logarithm transformations to normalize surgical case durations prior to conducting hypothesis testings with linear models. Their hypothesis tests show that permutations of CPT codes do not affect the accuracy of predictions of surgery case durations. Their results confirm that CPT code is the most important factor when predicting surgical case durations. Anesthesia types, emergency status, patient ages and surgery departments are also found to be relevant factors.

Although the importance of CPT codes in predicting surgical case durations has been noted for at least a decade, Strum et al. (2003) present the only work in the existing literature that uses CPT codes as explanatory variables for surgeries containing more than one CPT code. The limitation of their approach can be understood as follows. Their method provides a prediction only for surgeries consisting of exactly two CPT codes. When applying their method, a sufficient number of surgeries with the *same combination* of CPT codes must exist in the data samples. This requirement limits the application of their approach even for surgeries consisting of exactly two CPT codes. Moreover, it is difficult to extend Strum et al. (2003)'s distribution-fitting approach to surgeries with three or more CPT codes, due to lack of historical data. — Although there will be plenty of surgical cases with three or more CPT codes, there are not that many cases with exactly the same combination of three or more CPT codes.

In this project, we propose two regression models to predict surgical case durations. Different from previous regression-based approaches, our models explicitly include CPT codes as the explanatory variables. The proposed models can be applied

in general situations where a surgery can have any number of CPT codes and any combination of CPT codes. To the best of our knowledge, this project is the first that develops a systematic approach to predict surgery case durations based on multiple CPT codes.

Utilizing CPT code information is not a trivial matter because (i) many CPT codes only appear in conjunction with others and thus do not have any predictive power on their own; (ii) combination of CPT codes varys from surgery to surgery. Incorporating CPT information in our regression models hinges upon constructing a suitable design matrix of existing CPT codes, which frees us from relying on the occurrence of the same combination of CPT codes in historical data. The main challenge of constructing such a suitable design matrix is that naively constructed design matrix is usually ill-conditioned (i.e. singular). We devise a construction procedure to obtain a nonsingular, well-conditioned design matrix for our regression models. Our procedure carefully sifts out those CPT codes without any predictive power while retaining useful information as much as possible.

We compare our two regression-based models with three benchmark methods, one uses a lognormal distribution for prediction and the other two involve making predictions based on sample means. We measure the models' performances in terms of both mean squared errors (MSE) and mean relative absolute errors (MRAE). These performance measures show that our proposed models make more accurate predictions of surgical case durations than the benchmark methods although the magnitude of improvement varies for different service departments.

The rest of the project unfolds as follows. In Section B of this chapter, we describe the surgical data set with which we establish our prediction models and validate our approaches. Details of our predictive models are presented in Section C of this chapter. We compare our predictions to that of three benchmark approaches

in Section D of this chapter. Section E of this chapter concludes the project.

B. Data Set

Our surgical data set is from a large teaching hospital in central Texas. The data set consists of 48,714 surgical cases from 10/1/2004 to 3/31/2008. It involves 25 operating rooms (OR), 115 surgeons and 19 service departments. Variables collected include surgery date, operating room number, surgeon's name initial, the date and time at which a patient was admitted into an OR (*pt_in*), surgery preparation began (*prep_pos*), surgery began (*incision*), surgery ended (*closure*), dressing ended (*dress_end*) and the patient left OR (*pt_out*), as well as the CPT code(s), which accurately describe the surgical procedures performed. Here is an example of a surgery. The surgery was performed on a weekday in March 2008. According to the records, the patient entered the operating room at 11:34am. Preparation for surgery started at 11:58am. The surgeon made the first incision at 12:03pm and closed the patient up at 13:10pm. By 13:20pm the patient was completed dressed. He/she was transported out of the operating room at 13:28pm. Three CPT codes were performed during the period from 12:03pm (incision) to 13:10pm (closure). They are 25607, 64415, and 76942. In our data set, a surgical case could include as many as eight CPT codes.

Among the various segments of time included in a surgery, we are most interested in the *surgical time* (the duration from *incision* to *closure*). This is the time during which surgical actions take place. The CPT codes prescribed for a surgery are carried out during this time, and hence have a direct impact on the length of this time. The surgical case durations we study in this project refer to such surgical times.

Remark. We recognize that other durations, for example, the *total time* (the duration from *pt_in* to *pt_out*), can also be of interest to practitioners and researchers.

In addition to surgical times, Strum et al. (2003) also examine total times in their work. Although our focus is on surgical times, our proposed methodology can be easily adapted to the modeling of total times. The adaptation, however, adds little insight. In order to avoid repetitions, we present our methods in the context of surgical times.

Before we establish our predictive models using the historical data, a data cleaning action is performed to eliminate data records that are deemed "invalid." The following considerations are used to identify invalid records: (i) A record should have a starting time entry and an ending time entry to calculate a time duration. When a data record lacks any one of the two entries, it is incomplete and thus not used. (ii) The starting time should be earlier than the ending time. (iii) The duration should not be unreasonably long; for instance, 36 hours would be considered anomalous for this hospital given the nature of their operations. If these conditions are not met, a record is invalid and removed from the data set. Tables 17 and 18 provide summary information regarding our surgical data after data cleaning actions. From Table 17, one observes that data cleaning only eliminates a tiny portion of the data records (about 0.7% of the original data with a total of $48,714$ cases). One also observes that a large portion of the surgical cases include no more than two CPT codes. However, Strum et al. (2003)'s approach can not be readily applied here because those cases with two CPT codes do not necessarily share two common CPT codes. In our data set, there were $11,771$ combinations of CPT codes among $48,373$ valid cases. That is, there were, on average, about 4.1 cases with the same combination of CPT codes. Furthermore, even though the cases with more than two CPT codes are in minority among the total of forty-eight thousand plus cases, the absolute number of those cases (totaling $8,754$) is remarkable. The importance of making accurate duration predictions of these cases with at least three CPT codes should not be understated.

Table 17. Number of valid cases with exactly k ($k = 1, \cdots, 8$) CPT codes after data cleaning.

Number of CPT codes included in a surgery (k)	Number of valid cases
$k = 1$	29,039
$k = 2$	10,580
$k = 3$	4,065
$k = 4$	2,172
$k = 5$	1,182
$k = 6$	574
$k = 7$	366
$k = 8$	395
SUM	48,373

Table 18 provides departmental statistics of the surgical cases. The surgical cases in our data set were performed by 19 service departments, each in charge of a specialty area, for example, orthopedics, oncology, etc. Each department is represented by an acronym (consisting of two or three letters) commonly used and readily recognizable in medical profession; hence we skip the explanation of these acronyms. Although CPT codes associated with surgeries performed by different service departments often differ, we find that some CPT codes are shared across various departments. This is not surprising since two surgical cases serving different purposes could have a common set of surgical actions, which is represented by a common CPT code.

For each department, Table 18 lists the number of valid cases, the number of CPT codes, and the number of CPT combinations performed by that department. Note that a CPT combination is a set of CPT codes that appear together in a surgical case. Permutations of the same set of CPT codes are treated as the same CPT combination because permutations do not have any significant impact on surgical case durations (see Strum et al., 2003). The goal of this research is to predict surgical case durations based on the CPT codes included in a surgery. Table 18 roughly outlines the size of the problem we are dealing with. It is also clear from Table 18 that the surgical case

Table 18. Number of valid cases, CPT codes and CPT combinations performed by each service department after data cleaning.

Dept.	# of valid cases performed by a dept.	# of CPT codes performed by a dept.	# of CPT combinations performed by a dept.
NS	1,060	207	312
ORT	8,606	928	2,308
TPT	1,500	100	199
URO	5,223	489	1,108
CT	1,825	194	606
THO	721	241	502
UMC	1,390	280	338
GEN	8,386	656	1,507
ONC	4,755	579	1,493
GYN	4,726	366	906
ORA	555	181	206
PLA	3,118	736	1,486
EYE	89	54	54
VAS	1,549	309	818
PDS	2,489	397	636
ENT	1,960	363	544
OTH	65	23	20
POD	354	85	113
RAD	2	3	2

load distributions across various departments are uneven; most departments have performed over one thousand surgeries over the three-and-half-year period, whereas a few departments have performed fewer than one hundred cases.

C. Solution Approaches

Surgical case durations are predicted for each service department separately. The reason is threefold. Firstly, each service department handles their own surgery schedules. Secondly, the CPT codes that describe the surgical procedures, despite certain degree of sharing, are by and large different across the service departments. Thirdly, service departments are found to be a relevant factor that affects the prediction of surgical case durations (see Strum et al., 2003).

The need of department-specific predictions further renders the existing lognormal distribution based approach less effective. If we are to estimate the lognormal

distribution for a given CPT combination, we hardly have enough data points in the sample. To see this, one can simply compare the number of valid cases and the number of CPT combinations performed by each department in Table 18. The average number of cases per CPT combination ranges from 1 to 7.54 among the 19 service departments; apparently 7 cases per CPT combination are not enough data points for distribution estimation.

Figure 7 shows the histograms of data for three departments (CT, UMC and ORA), where the data deviate significantly from lognormal distributions. The unit for the horizontal axis is in hours. In particular, the case durations in service department CT has a bi-modal distribution, which cannot be well approximated by either the lognormal distribution or the normal distribution.

What we propose here is a regression-based approach, which made no assumption on normality or normality after logarithm transformation. Our models explicitly use the CPT codes describing surgical procedures as explanatory variables. This allows the specific knowledge regarding a surgical procedure to be incorporated. We would like to note that in the current research we consider only the CPT codes, while ignoring other possible covariates, since the CPT codes are recognized as the most important factor in representing surgical case durations in the literature. Our later numerical results indeed demonstrate enough benefit of our research undertaking. We do acknowledge that considering both the CPT codes and other important covariates (such as surgeons or anesthesia types) could potentially further improve the prediction of surgical case durations. But doing so will require a different model, more data collection, and is thus out of the scope of this project.

In the sequel, we will first present two regression models that predict the surgical time (the duration from *incision* to *closure*). We then describe a singularity problem encountered in applying these regression models. In the rest of Section C of this

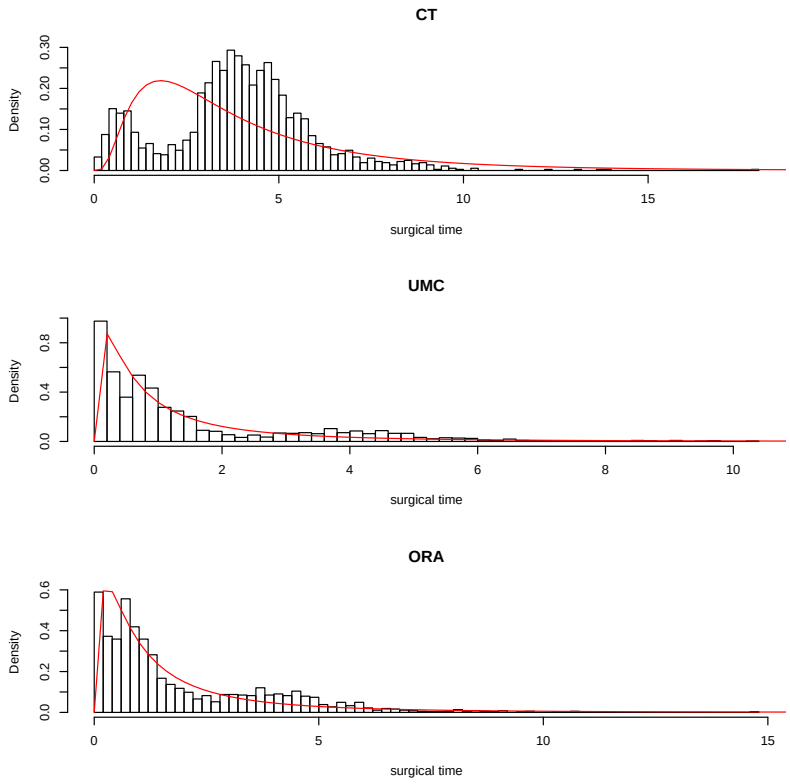

Figure 7. Histograms and best-fit lognormal densities of the surgical time for three service departments.

chapter we propose systematic procedures that address the problem.

1. Regression Models

Since CPT codes describe specific surgical actions undertaken in a surgery, the surgical time can naturally be considered as the summation of all the component surgical actions. Suppose there are n surgical cases performed by a given service department involving a total of m CPT codes. Denote by y_i the surgical time of case i. We introduce here an indicator variable, x_{ij}, of the inclusion of the j_{th} CPT code in the i_{th} surgical procedure. In other words, when CPT code j shows up in surgical case i,

$x_{ij} = 1$; otherwise $x_{ij} = 0$. Denote by β_j the expected time of performing the surgical action specified by CPT code j. We have the following model to describe the surgical time:

$$y_i = \sum_{j=1}^{m} x_{ij}\beta_j + e_i, \quad i = 1, \ldots, n$$

where y_i is the summation of the expected times associated with all the CPT codes involved in case i, plus e_i, which is the residual error of case i that cannot be modeled by the expected times. Residual error e_i is assumed to be a zero-mean random variable.

The above model can be expressed in a matrix form as:

$$Y = X\beta + e \tag{4.1}$$

where $Y = (y_1, \ldots, y_n)'$ is the $n \times 1$ vector of surgical times, $\beta = (\beta_1, \ldots, \beta_m)'$ is the $m \times 1$ vector of the expected times associated with m CPT codes, $X = (x_{ij})$ is the $n \times m$ design matrix, representing the inclusion of CPT codes in surgical cases.

Equation (4.1) represents a typical linear regression model. Once surgeries are performed, Y and X are known, and β is the one to be estimated from historical data. Denote by $\hat{\beta}$ the estimate of β, and $\hat{\beta}$ will be used in future predictions. Since we do not restrict the sign of our estimates, it is possible that we obtain negative estimates of the expected times for certain CPT codes. The predicted surgical time could still be positive because it is determined by the combination of the comprising CPT codes. The negativity can be completely avoided by adding a non-negativity constraint on the β_j's. We did not impose this constraint in our implementation of the regression models because negative values rarely appear in our analysis. More details and explanations are reported at the end of Section C of this chapter.

In order to predict the surgical time of a new case z, one needs to look at the CPT

codes to be performed in the surgery and create a design vector x_z by assigning "1" or "0" to the corresponding x_{zj} for $j = 1 \dots m$. Then, calculating the inner product of this design vector $x_z = (x_{z1}, ..., x_{zj}, ..., x_{zm})$ with the estimates $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_m)'$ gives the predicted surgical time of the new case. Precisely, let Y_z^{new} denote the surgical time of the new case, then the linear regression model predicts that $\widehat{Y_z^{new}} = x_z\hat{\beta}$.

The above model is flexible and easy to use in predicting surgical times composed of any number of CPT codes. Suppose one extra set of surgical actions is added to a series of existing actions, then the surgical time will be simply increased by the length of the corresponding actions described by the additional CPT code(s). The model sets no restrictions on how many CPT codes a surgical procedure can include or what CPT combinations should appear.

Next we present an alternative model, motivated by the arguments in existing literature that surgery data are better fit by a lognormal distribution (Strum et al., 2000a; Strum et al., 2003; May et al., 2000 and Spangler et al., 2004, among others). These arguments help legitimize the use of logarithm transformation to normalize surgical procedure times. In light of this, we take logarithm transformations of surgical times before fitting them to a linear regression model. Consequently, our second model reads as

$$\log(Y) = X\beta + e \tag{4.2}$$

where $\log(Y) \equiv (\log(y_1), \dots, \log(y_n))'$, an $n \times 1$ vector, and e is a vector of zero-mean residuals. We refer to equation (4.2) as a log-regression model. Again, Y and X are known from historical data, and β is to be estimated. The estimate $\hat{\beta}$ will be used for future predictions, in a similar fashion as in the linear regression model explained above. For a new case z with corresponding design vector x_z, the prediction of the surgical time Y_z^{new} is given by $\widehat{Y_z^{new}} = \exp(x_z\hat{\beta})$.

Compared to the linear regression model (4.1), the log-regression model (4.2) is less intuitive in terms of its practical interpretation. Note that a surgical case comprises a series of procedures (each of which is represented by a CPT code). The linear regression model implies that the surgical time is the summation of the times associated with the component procedures, while the log-regression model suggests that the surgical time is the product of the exponentials of the times associated with the component procedures. The advantage of the log-regression model is that its prediction is always positive. Our numerical results (presented in Section D of this chapter) show that both models perform well.

A point worth noting is that we do not make distribution assumptions in (4.1) and (4.2). By using the method of least squares to fit model (4.1), we find the best linear predictor of the surgical time. Although model (4.2) is motivated by log-normal distribution arguments, least squares fitting of the model can be considered as finding the best linear prediction of the log surgical time. The logarithmic transformation is simply used as a device for ensuring a positive prediction.

Remark. If the duration of interest is *total time*, namely the duration a patient spends in an OR, the models in (4.1) and (4.2) only need to be slightly modified. Noticing that the total time is the addition of the surgical time and the pre- and post-surgery processing times, we can add an intercept term β_0 to both models in (4.1) and (4.2). As such, the models for the total time read:

$$Y = \beta_0 \cdot 1_n + X\beta + e \tag{4.3}$$

and

$$\log(Y) = \beta_0 \cdot 1_n + X\beta + e \tag{4.4}$$

where Y now represents the total time, $\log(Y)$ follows the same notation as in model

(4.2), and 1_n is an $n \times 1$ vector whose elements are all 1's. In model (4.3), β_0 represents the expected time consumed collectively by all the pre- and post-surgery actions. Similar meaning applies to β_0 in model (4.4) which is after a logarithm transformation. The inclusion of extra durations bring in additional variability, which is absorbed into the residual error e in the above models. After models (4.3) and (4.4) are fit using the training data, predictions of the total time can be easily computed. Let x_z denote the design vector for a new case, the corresponding total time Y_z^{new} can be predicted using $\widehat{Y_z^{new}} = \hat{\beta}_0 + x_z\hat{\beta}$ for model (4.3) and $\widehat{Y_z^{new}} = \exp(\hat{\beta}_0 + x_z\hat{\beta})$ for model (4.4). In the interest of conciseness, we report our proposed procedures in the context of surgical times.

2. Singularity of Design Matrix X

After establishing the regression models (4.1) and (4.2), if the design matrix X is of full rank, we can estimate β and β_0 in the linear regression models through a standard least-squares estimation (Weisberg, 2005). Specifically, for model (4.1)

$$\hat{\beta} = (X'X)^{-1}X'Y; \tag{4.5}$$

and for model (4.2), one needs to simply replace y_i with $\log(y_i)$ and Y with $\log(Y)$. The reason that a fully ranked X is required is because of the inversion on $X'X$.

Whether the design matrix X is of full rank, however, depends on how it is constructed. If we list all the CPT codes performed by a service department and naively use this list to construct X, we will obtain an ill-conditioned X. As a result, $X'X$ is not invertible and β_i's cannot be estimated. Consequently, the expected times associated with the corresponding CPT codes cannot be estimated.

Consider for example three CPT codes, A, B, and C. (Note that we use capital letters to denote CPT codes for the sake of simplicity although an actual CPT code

is a five-digit number.) Assume that the expected times for the three CPT codes are β_1, β_2, and β_3, respectively. Suppose that there are only three surgical cases: the first case uses all three CPT codes, the second case uses CPT codes B and C, and the third case uses only CPT code A. Then, the design matrix X will be

$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

which is singular.

The singularity in the above illustration is caused by the co-appearence of CPT codes B and C. From the surgical cases performed and times measured, we will not be able to tell the times associated with individual CPT codes B or C but only the combined times of the two CPT codes. In general, to check whether a certain CPT code B always appears in conjunction with another code C, we can perform a simple test as follows: count the number of appearances of CPT codes B and C in the surgical cases within a given service department; suppose both appear, for instance, h times. Then, count the number of CPT codes B and C appearing together (we call this CPT combination BC). If BC also appears h times, then it implies that CPT codes B and C always appear together in conjunction with each other. This appearance pattern will result in a singular design matrix. Furthermore, a CPT code can appear in conjunction with different CPT codes in different cases. Therefore, one needs to exercise extra care when constructing a design matrix. Recall that we have to deal with a large number of CPT codes for each department. Next, we propose a systematic procedure that thoroughly and efficiently sifts out CPT codes/combinations that cause singularity, without losing information.

3. Grouping CPT Combinations

In order to avoid singularity, CPT codes that always appear together should be treated as a whole as if they formed a new CPT code. For instance, in the singularity example above, instead of attempting to estimate individually the times associated with A, B, and C, one should only try to estimate the times associated with a single CPT code A and a CPT combination BC.

In light of this, we need to group CPT codes and combinations in our data set appropriately. The purpose of grouping is to establish the set of single CPT codes whose execution times can be estimated, the set of two-code CPT combinations whose combined time can be estimated, the set of three-code CPT combinations whose combined time can be estimated, and so on. A full-rank design matrix can then be constructed based on the grouping results. We will explain the construction of a design matrix in the next subsection.

Before we present our detailed grouping procedure, we introduce the concept of *code length*, which is defined as the number of component CPT codes in a CPT combination. Denote by k the code length of a CPT combination. In our data set, the largest k is eight. In the sequel we only illustrate implementation details for k up to eight, although the general procedure applies to any value of k. Given the fact that our data set covers nearly 50 thousand cases over three and half years, the scenario in which k could be greater than eight should rarely happen in reality. The grouping procedure is as follows.

- First, construct $k = 8$ empty sets $S_1, \ldots, S_k, \ldots, S_8$, where S_k will hold the grouping results for CPT combinations of length k.
- Repeat the following for $k = 1, 2, ...8$

- Identify all the surgical cases with exactly k CPT codes. Put them in S_k. If there are no such cases, we have finished selecting the CPT combinations of length k, so go to the next value of k.
- For each CPT combination of length k in S_k, determine whether it is "distinctive." We now describe how the distinctiveness of a CPT combination is determined. A CPT combination of length k can be decomposed into a number of CPT codes or code combinations of length 1 to length $k-1$. For instance, a CPT combination ABC of length 3 can be decomposed into three single CPT codes of length 1, A, B, C, or a CPT combination of length 2 plus a single code; there are three possibilities, i.e., AB and C, or AC and B, or BC and A. For any given scheme of decomposition, if all the decomposed component codes or code combinations can be found in sets S_1 to S_{k-1}, then the CPT combination of length k is not distinctive; otherwise it is.
- Remove all the non-distinctive CPT combinations of length k from S_k.

In the above procedure, the step of determining the distinctiveness of a CPT combination is relatively involved. For $k = 1$, it is straightforward since there is no set S_0, all single CPT codes automatically satisfy the distinctiveness condition. For $k = 2, \ldots 8$, we have to go through all possible decomposition schemes of a CPT combination of length k. The larger the k, the more complicated a decomposition process becomes.

Table 19 helps sort out the decomposition schemes for $k = 2, ..., 8$. To understand the notation in the table, take $k = 4$ as an example. The entry of $4 = 3 + 1$ means that a CPT combination of length 4 can be decomposed into a CPT combination of length 3 and one of length 1 (a single code); the next lines of $4 = 2+2$, $4 = 2+1+1$,

Table 19. Decomposition schemes of a CPT combination of length k.

k=2	k=3	k=4	k = 5	k = 6	k = 7	k = 8
2 = 1+1	3 = 2+1	4=3+1	5=4+1	6=5+1	7=6+1	8=7+1
	3= 1+1+1	4=2+2	5=3+2	6=4+2	7=5+2	8=6+2
		4=2+1+1	5=3+1+1	6=4+1+1	7=5+1+1	8=6+1+1
		4=1+1+1+1	5=2+2+1	6=3+3	7=4+3	8=5+3
			5=2+1+1+1	6=3+2+1	7=4+2+1	8=5+2+1
			5=1+1+1+1+1	6=3+1+1+1	7=4+1+1+1	8=5+1+1+1
				6=2+2+2	7=3+3+1	8=4+4
				6=2+2+1+1	7=3+2+2	8=4+3+1
				6=2+1+1+1+1	7=3+2+1+1	8=4+2+2
				6=1+1+1+1+1+1	7=3+1+1+1+1	8=4+2+1+1
					7= 2+2+2+1	8=4+1+1+1+1
					7= 2+2+1+1+1	8=3+3+2
					7= 2+1+1+1+1+1	8=3+3+1+1
					7= 1+1+1+1+1+1+1	8=3+2+2+1
						8=3+2+1+1+1
						8=3+1+1+1+1+1
						8=2+2+2+2
						8=2+2+2+1+1
						8=2+2+1+1+1+1
						8=2+1+1+1+1+1+1
						8=1+1+1+1+1+1+1+1

and $4 = 1+1+1+1$ mean that the same CPT combination can also be decomposed into two CPT combinations of length 2, or a CPT combination of length 2 plus two single codes, or four single codes, respectively. Collectively, those are all the possible decomposition schemes for a CPT combination of length 4.

Recall that Strum et al. (2003) found that permutations of component CPT codes did not significantly affect surgical case durations. For this reason, we do not consider permutations of a CPT combination any different than the original CPT combination. Our definition of the distinctiveness of a CPT combination is based on the decomposition of the CPT combination, not permutations. Another note is that the above grouping procedure can be applied to any surgical data set but in this research we apply them to the data of individual service departments due to the department-specific approach we undertake in predicting the surgical case durations.

As an illustration, we present Table 20, which summarizes the the number of valid cases, the number of CPT combinations, and the number of distinctive CPT combinations for the service department of ENT. There are a total of 1,960 valid cases and 544 CPT combinations (including single CPT codes). Among the 544 CPT

Table 20. Summary of CPT combinations in Department ENT.

	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	Total
# of cases with k codes in ENT	1297	482	111	37	19	7	3	4	1960
# of CPT combinations of length k	182	205	92	32	19	7	3	4	544
# of distinctive CPT combinations of length k	182	119	45	14	12	7	3	4	386

combinations, 386 of them are distinctive according to the aforementioned definition, and others can be decomposed into components found in the sets of shorter code length. For instance, there are 205 CPT combinations of length 2, but 86 of them can be decomposed into two single CPT codes that are both present in S_1. That leaves 119 (= 205 - 86) distinctive CPT combinations of length 2 in S_2. Therefore, the size of S_2 becomes 119 after our grouping procedure is applied.

4. Constructing a Design Matrix

Constructing a design matrix is to assign "1" or "0" to each element x_{ij} in matrix X. Recall that when previously introduced, the first index i is the case index, ranging from 1 to n, and the second index j is the CPT code index, ranging from 1 to m. After applying the grouping procedure described in Section C of this chapter, we will estimate the expected times β_j not only for single CPT codes but also for distinctive CPT combinations of length $k \geq 2$. So the value of m depends on the number of distinctive CPT combinations (including single CPT codes) in a given data set. For example, let's take the data in Table 20 for illustration. If surgeries with code length up to 8 are to be included in the regression model, then $m = 386$, But if only surgeries with the single CPT codes and those with code length of 2 are to be included, then $m = 301$ (= 182 + 119). Suppose that we include all the surgeries with code length up to 8. We should, then, aggregate all the distinctive CPT combinations (sets S_1 to S_8) into a set $S \equiv \bigcup_{k=1}^{8} S_k$. If there are a total of m elements in S, a row vector

of matrix X, $x = (x_{i1}, \ldots, x_{ij}, \ldots, x_{im})'$, has a one-to-one correspondence to the m elements in S.

For convenience, we order the surgical cases based on the number of CPT codes they have, namely that first comes the surgical cases with a single CPT code, followed by the surgical cases with exactly two CPT codes, and then followed by the cases with exactly three CPT codes, and so on. So eventually, surgical cases 1 to i_1 has a single CPT code, cases $i_1 + 1$ to i_2 has two CPT codes, ..., and cases $i_7 + 1$ to i_8 has eight CPT codes, where $1 \leq i_1 \leq ... \leq i_8 = n$.

Before the construction of design matrix X, we set all x_{ij}'s to zero. The basic idea of constructing a design matrix is that for each $i = 1, \ldots, i_8$, take the corresponding surgical case and match the CPT codes it has with the distinctive CPT combinations in S. If a match is found, then the corresponding x_{ij} will be set to "1"; otherwise x_{ij} will be left as "0". We here assume that the set S is well maintained and timely updated using our grouping procedure. So the finding of a match is guaranteed.

Despite the simplicity of this idea, certain complexities have to be dealt with. For surgical cases with a single CPT code (cases 1 to i_1), the procedure is just like what the basic idea describes, except that one need not search the set of S but only S_1. For surgical cases with two CPT codes (cases $i_1 + 1$ to i_2), the two codes could appear as a CPT combination of length 2 or they may have appeared as two single CPT codes. What one needs to do is to search first the set of S_2 in order to check if there is a match for a CPT combination of length 2, and if not, then search S_1 for the matches of the two single CPT codes. Depending on the outcome of the search, the appropriate x_{ij} can be set to 1. For surgical cases with $k \geq 3$ CPT codes (cases from $i_2 + 1$ onward), one needs to search for matches in different sets from S_k to S_1, similarly as one does for surgical cases with two CPT codes. Because there are many different ways of decomposing a CPT combination when k gets large, Table 19 is a

good reference that can guide the search process.

In addition to searching for matches coming from all possible schemes of decomposition, one more complexity arises for surgical cases having three or more CPT codes. To understand this, take a surgery case with three CPT codes as an example. Suppose that the CPT codes prescribed for the surgery are A, B and C. Also suppose when searching the set of S_3, we do not find any matches; and when searching the set of S_1, not all three of the single codes found their matches, either. Then, we need to search S_2 for possible matches of a CPT combination of length 2. Doing so could give us multiple matches: for example, we could have AB in S_2 while C in S_1, this is one match; or AC in S_2 while B in S_1, this is another match. If both matches are found, the surgical case in question can be used to estimate both the expected times of AB, C, and the expected times of AC, B, unless one has profound prior knowledge suggesting otherwise. As a matter of fact, in order to extract the most information from this surgical case, its duration should be taken into account when we estimate both the expected times of AB, C, and the expected times of AC, B. To do so properly, we should include this surgical case twice in our design matrix. One inclusion represents the decomposition AB+C, and the other AC+B. In order to account for duplicate use of the same data, caused by multiple inclusions of a single surgical case, we use the weighted least-squares approach by applying weights that are inversely proportional to the number of inclusions.

To illustrate, consider the following example. Suppose a data set contains only eight surgical cases. The CPT codes prescribed for each case are listed in Table 21.

Apparently there are six CPT codes (A, B, C, D, E and F) ever performed. It is straightforward to verify that a naively constructed design matrix that assigns "0" and "1" to each case based on their inclusion of each of the six CPT codes is singular. Therefore, the design matrix should be constructed differently. Applying

Table 21. CPT codes relating to the cases in the design matrix example.

Case #	CPT codes performed in a case
1	A
2	C
3	AB
4	BC
5	ABC
6	ABD
7	ABCD
8	ABCDEF

the grouping procedure from Section C of this chapter, we obtain the following sets $S_1 = \{A, C\}$, $S_2 = \{AB, BC\}$, $S_3 = \{ABD\}$, and $S_6 = \{ABCDEF\}$. Aggregating S_1, S_2, S_3 and S_6 generates $S = \{A, C, AB, BC, ABD, ABCDEF\}$, implying that $m = 6$. That is, there are six columns in the design matrix. Next we construct each row of the design matrix by including each of the surgical cases in the data set.

For Case 1, apparently, $x_{11} = 1$ and all other entries in the first row of the design matrix X are zeros since the CPT code that Case 1 uses matches the first element in S. Following the same reasoning, for Case 2, $x_{22} = 1$; for Case 3, $x_{33} = 1$; and for Case 4, $x_{44} = 1$. For Case 5, ABC can have two different decompositions, $C + AB$ or $A + BC$. Both decompositions are possible in S. Therefore we include Case 5 twice (occupying two rows) in the design matrix: for the fifth row, $x_{51} = 1$ and $x_{54} = 1$ (corresponding to $A+BC$); and for the sixth row, $x_{62} = 1$ and $x_{63} = 1$ (corresponding to $C + AB$). Because Case 5 is included twice, Cases 6, 7 and 8 will then correspond to rows 7, 8 and 9 (instead of rows 6, 7 and 8) of the design matrix X, respectively. For Case 6, $x_{75} = 1$; for Case 7, since $ABCD$ can be decomposed into $C + ABD$, $x_{82} = 1$ and $x_{85} = 1$; for Case 8, $x_{96} = 1$. Ultimately, the design matrix X, which is of full rank, looks like the matrix presented in Figure 8. The code or code combination on the top indicate the columns corresponding to $A, C, AB, BC, ABD, ABCDEF$, respectively, and the texts on the right side of the matrix identify the corresponding

$$X = \begin{array}{c} \begin{matrix} A & C & AB & BC & ABD & ABCDEF \end{matrix} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array} \begin{array}{l} \\ \text{Case 1} \\ \text{Case 2} \\ \text{Case 3} \\ \text{Case 4} \\ \text{Case 5 (first inclusion)} \\ \text{Case 5 (second inclusion)} \\ \text{Case 6} \\ \text{Case 7} \\ \text{Case 8} \end{array}$$

Figure 8. Design matrix of the $m = 6$ example.

cases. Note that there are 8 surgical cases in the example while the resulting design matrix has 9 rows.

Below we outline a general procedure for the construction of a design matrix assuming that the set S has already been obtained. Let m be the total number of elements in the set S, and n the total number of surgical cases in the data set.

(1) Set the number of columns in the design matrix to m. Order the elements in the set S from 1 to m, and use them to label the columns of the design matrix.

(2) Order all the surgical cases from 1 to n.

(3) Set $i = 1$ and $r = 1$.

(4) Decompose the i_{th} surgical case using the elements in the set S. Let d_i be the number of possible decompositions. Order the possible decompositions from 1 to d_i. Set $c = 1$.

(5) For $j = 1, ..., m$, use x_{rj} to denote the value of the entry in the r_{th} row and the j_{th} column. Set $x_{rj} = 1$ if the c_{th} decomposition of the i_{th} case uses the

j_{th} element of the set S. Otherwise, $x_{rj} = 0$.

(6) If $c = d_i$ then set $r = \sum_{t=1}^{i} d_t + 1$ and go to Step (7). Otherwise set $c = c+1$, $r = r + 1$ and go to Step (5).

(7) If $i = n$ then the design matrix is completed. Otherwise set $i = i + 1$ and go to Step (4).

Next we illustrate the application of our design matrix using data from the service department ENT. We set $S = (S_1, S_2, S_3, S_4)$, our $m = 360$ (=182+119+45+14), and $n = 1,927$ (=1,297+482+111+37). The corresponding design matrix X is of dimension 1,950 $\times$ 360 rather than 1,927 $\times$ 360 because of repeated inclusions of certain cases for the reasons explained earlier. We then proceed to estimate the β's associated with the 360 distinctive CPT combinations (using equation (4.5)); they are the expected times for performing the corresponding combination of CPT codes. Figure 9 presents the histogram for the 360 values of $\hat{\beta}$'s. Unit for the horizontal axis is in hours. Most of these CPT combinations have an estimated time in the range of (0, 10) hours. The vertical line is the mean of the times of the 360 distinctive CPT combinations, which is about 1.532 hours. Almost half of the distinctive CPT combinations used in department ENT takes fewer than one hour to complete. We also observe from Figure 9 that a very small portion of distinctive CPT combinations have a negative time estimate. In fact, 4 out of these 360 (1.11%) CPT combinations have a negative time estimate. To explain why we may have negative estimates, consider the following example of two surgical cases: Case 1 includes CPT codes A and B, while Case 2 only uses CPT code A. When surgeons actually perform these surgical cases, the surgical time of Case 1 could be shorter than that of Case 2. When this happens, the estimated time of CPT code B becomes negative. This negativity rarely happens, as evident from the ENT departmental data (data from

other departments generate the same conclusion). Moreover, the code B is likely to appear together with another CPT code and thus still gives a positive prediction of the surgical time. We are confident that the rare appearance of negativity does not cause our prediction of surgical case durations to go off the marks.

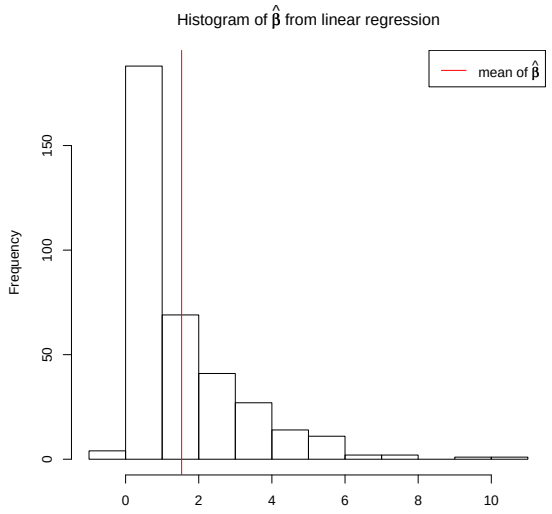

Figure 9. Histogram of $\hat{\beta}$'s from the linear regression model for department ENT.

D. Prediction and Comparison

1. Construction of Training and Test Data Sets

According to the statistical literature(Witten and Frank, 2005; Mitchell, 1997), the typical protocol for validating empirical models is to split the original data set into a training set and a test set. Suppose the n surgical cases in the original dataset are divided as the n_t cases in the training set and the n_s cases in the test set, where $n = n_t + n_s$. The training data set is used to obtain $\hat{\beta}$ based on equation (4.5); this is known as *model fitting*. After the model is fit, i.e., all β's are estimated, one can use

the model to make predictions of surgical case durations on the data records in the test set, which are never used in the model fitting process. Then, the predictions are compared with the real measurements of surgical case durations in the test set. The differences between the predictions and the actual surgical case durations are good indications of how well a model works.

Suppose we would like to predict the duration for a surgical case i in the test set. A design vector x_i for the case can be generated based on the set S, which is obtained after our grouping procedure is applied. Use the random variable Y_i^{new} to denote the length of the duration of interest for the case i. The predicted value is denoted as $\widehat{Y_i^{new}}$. The difference between the predictions and the actual surgical case durations is measured by two metrics: the mean squared errors (MSE) and the mean relative absolute errors (MRAE). They are defined as:

$$MSE = \frac{1}{n_s}\sum_{i=1}^{n_s}\{y_i - \hat{Y_i^{new}}\}^2 \quad and \quad MRAE = \frac{1}{n_s}\sum_{i=1}^{n_s}\frac{|\hat{Y_i^{new}} - y_i|}{y_i}$$

where y_i is the recorded duration of the i_{th} surgical case in the test set and $\widehat{Y_i^{new}}$ is the predicted duration for the same case. MRAE characterizes how well a model makes prediction.

In this research, we assign two thirds of the historical cases to the training set and one third to the test set, i.e., $n_t \approx \frac{2}{3}n$ and $n_s \approx \frac{1}{3}n$, where "$\approx$" is used because n_t and n_s need to be rounded off to the closest integer number. To avoid any systematic bias, the assignment of a case to one of the sets is randomly decided. Moreover, we repeat the assignment process 1,000 times, meaning that we randomly split the original data set 1,000 times, and consequently, we obtain 1,000 pairs of training/test sets. The performance measures MSE/MRAE, are then calculated 1,000 times using the 1,000 pairs of training/test sets. The MSE/MRAE values reported later in this

section are the average of the 1,000 individual MSE/MRAE values.

2. Three Benchmark Methods

We compare our predictive models with three benchmark models below.

- Lognormal Model

  This model assumes that the surgical time or the total time Y follows the lognormal distribution. It means that $Y \sim$ lognormal (μ, σ), or equivalently, $\log(Y) \sim$ normal (μ, σ), where μ and σ are the parameters to be estimated. One can estimate them by using the data in the training set, such as:

  $$\hat{\mu} = \frac{1}{n_t}\sum_{i=1}^{n_t}\log(y_i); \quad \hat{\sigma}^2 = \frac{1}{n_t}\sum_{i=1}^{n_t}(\log(y_i) - \hat{\mu})^2$$

  Then, the surgical time for surgical case z in the test set is predicted using

  $$\widehat{Y_z^{new}} = \exp(\hat{\mu} + \frac{\hat{\sigma}^2}{2})$$

  because that is the expectation of a lognormal model with parameters μ and σ (Casella and Berger, 2001).

  When using the lognormal benchmark model, we compute an estimated surgery length for all the surgeries in a department based on all the historical data in the training set for the same department. Ideally, we would like to find benchmark predictions for each surgery of a specific CPT combination. However, one would run into the insufficient-sample-size problem frequently when implementing this ideal approach for the lognormal Model. As aforementioned, the average number of cases per CPT combination ranges from 1 to 7.54 among the various departments.

- Departmental Sample-mean Model

This model takes the sample mean of the case durations within a service department in the training set, and treat it as the prediction for the cases within the same department in the test set, namely

$$\widehat{Y_z^{new}} = \frac{1}{n_t}\sum_{i=1}^{n_t} y_i.$$

- Hybrid Sample-mean Model

  When departmental sample means are used, the individuality of each surgery, which is manifested by various CPT codes, is lost. Meanwhile, such individuality often results in lack of historical data. Taking into account both concerns, our third benchmark model calculates sample means differently for different surgeries based on the existence of historical data. For a surgery in the test set, if its CPT combination can be found in the training set, the mean of all the surgeries with the same CPT combination is the predicted duration; otherwise, the departmental sample mean serves as the predicted value.

Inclusion of the lognormal model in our comparison is easily understood since previous research has argued for its use. Sample means are intuitive benchmarks because they are common practices when surgery schedules are determined in hospitals. One can certainly find drawbacks in these benchmark models or their implementations. But, the lack of a better benchmark model also validates the necessity of our work. We believe our proposed prediction models in this project set a reasonable benchmark for future research.

3. Comparison

Three departments, “EYE”, “OTH”, and “RAD”, have too few surgical cases, which are 88, 65, and 2, respectively, and will be omitted in this section for prediction

Table 22. Mean squared errors of out-of-sample prediction of surgical times for several competing methods.

Dept	Mean Squared Error				
	Reg	LogReg	Lognormal	Dept-mean	Hybrid-mean
NS	**0.828**(0.003)	0.851(0.004)	1.865(0.005)	1.848(0.006)	1.086(0.004)
ORT	**0.417**(0.001)	0.551(0.002)	1.026(0.001)	1.024(0.001)	0.576(0.001)
TPT	**0.711**(0.003)	0.891(0.003)	1.704(0.005)	1.700(0.005)	0.787(0.004)
URO	**0.403**(0.001)	0.487(0.001)	1.272(0.002)	1.272(0.002)	0.530(0.001)
CT	0.899(0.003)	**0.898**(0.003)	2.769(0.004)	2.682(0.005)	1.539(0.005)
THO	0.730(0.004)	**0.708**(0.004)	0.913(0.004)	0.912(0.004)	0.734(0.004)
UMC	**0.095**(0.000)	0.186(0.001)	0.377(0.001)	0.375(0.001)	0.107(0.000)
GEN	**0.532**(0.001)	0.569(0.001)	1.157(0.001)	1.156(0.001)	0.571(0.001)
ONC	**0.559**(0.001)	0.605(0.001)	1.875(0.002)	1.872(0.002)	1.009(0.002)
GYN	**0.371**(0.001)	0.424(0.001)	0.885(0.001)	0.884(0.001)	0.490(0.001)
ORA	**0.497**(0.002)	0.552(0.003)	2.714(0.007)	2.705(0.007)	0.550(0.005)
PLA	**0.707**(0.003)	2.281(0.060)	2.940(0.005)	2.926(0.005)	1.414(0.004)
VAS	**0.540**(0.002)	0.588(0.002)	1.349(0.004)	1.348(0.004)	0.799(0.003)
PDS	0.190(0.001)	0.252(0.002)	0.418(0.001)	0.416(0.001)	**0.176**(0.001)
ENT	**0.282**(0.002)	0.493(0.003)	1.126(0.002)	1.118(0.003)	0.406(0.002)
POD	0.109(0.001)	0.131(0.001)	0.148(0.001)	0.148(0.001)	**0.095**(0.001)

and comparison. We apply two proposed regression-based methods and three benchmark methods to the remaining 16 departments. When reporting our results, we use "Reg", "LogReg", "Lognormal", "Dept-mean", and "Hybrid-mean" to represent the linear regression model (4.1), the log-regression model (4.2), the lognormal model, the departmental sample-mean model, and the hybrid sample-mean model, respectively. For each department, all the five models are employed to make predictions over the test data sets coming from the 1,000 random splitting of the surgical data (with $S = \bigcup_{k=1}^{4} S_k$). Both MSE and MRAE are calculated. Results of the comparison are summarized in Tables 22 and 23. Numbers shown are means and corresponding standard derivations, based on 1000 random splits of the data into training and test sets. The unit is in hour.

The highlighted numbers in the two tables represent the smallest MSE or MRAE of prediction, or the best performance, in each respective department. From Tables 22 and 23, we observe the following:

- When MSE/MRAE is used as performance measure, the highlighted numbers

Table 23. Mean relative absolute errors of out-of-sample prediction of surgical times for several competing methods.

Dept	Mean Relative Absolute Error				
	Reg	LogReg	Lognormal	Dept-mean	Hybrid-mean
NS	0.440(0.001)	**0.402**(0.001)	1.313(0.003)	1.238(0.003)	0.608(0.001)
ORT	**0.377**(0.000)	0.403(0.000)	0.865(0.001)	0.844(0.001)	0.489(0.001)
TPT	**0.356**(0.001)	0.461(0.001)	0.995(0.002)	0.959(0.002)	0.390(0.001)
URO	**0.474**(0.001)	0.568(0.001)	1.277(0.001)	1.266(0.001)	0.561(0.001)
CT	0.275(0.001)	**0.253**(0.001)	1.221(0.003)	1.113(0.003)	0.570(0.001)
THO	**0.406**(0.001)	0.452(0.003)	0.680(0.004)	0.667(0.004)	0.463(0.002)
UMC	**0.332**(0.000)	0.523(0.001)	1.789(0.002)	1.684(0.002)	0.408(0.001)
GEN	**0.387**(0.000)	0.406(0.000)	0.943(0.001)	0.922(0.001)	0.461(0.000)
ONC	**0.354**(0.000)	0.396(0.000)	1.130(0.001)	1.091(0.001)	0.585(0.001)
GYN	**0.339**(0.000)	0.436(0.001)	0.962(0.001)	0.930(0.001)	0.515(0.001)
ORA	0.502(0.002)	**0.477**(0.002)	1.445(0.005)	1.395(0.005)	0.529(0.003)
PLA	**0.401**(0.001)	0.453(0.001)	1.881(0.002)	1.745(0.002)	0.802(0.003)
VAS	0.305(0.001)	**0.304**(0.001)	0.608(0.001)	0.602(0.002)	0.378(0.001)
PDS	**0.514**(0.002)	0.717(0.002)	1.592(0.003)	1.506(0.003)	0.536(0.001)
ENT	**0.491**(0.001)	0.831(0.002)	2.975(0.005)	2.747(0.004)	0.956(0.003)
POD	0.420(0.001)	0.466(0.002)	0.549(0.002)	0.546(0.002)	**0.386**(0.001)

occur in the first two columns for 14/15 out of 16 departments. This demonstrates the superior performance of our proposed regression-based methods, as compared to the three benchmark models.

- Between the two of our proposed methods, the linear regression model claims the best performance more often than the log-regression model. This observation suggests that when using CPT codes as explanatory variables to make predictions, the benefit of applying a logarithm transformation to the data is no longer obvious.

- Our proposed regression-based methods significantly improve the two benchmark methods that make predictions based on departmental data. This is consistent to the observation in the literature that CPT code information plays an important role in predicting surgical case durations.

- The hybrid sample-mean method, which utilizes the CPT code information, does perform better than the other two benchmark methods, which do not

use the CPT code information. The hybrid sample-mean method cannot outperform our regression-based methods for most departments, because it uses the CPT code information only when there is an exact matching CPT code combination in the training set. Understandably, the hybrid sample-mean method performs well only when there are "sufficient" number of historical cases with the same CPT combination. In practice, it is not always easy to decide how many are many enough, and there do exist circumstances when there is only a handful of historical cases or there is no such case at all. Looking at the comparison result tables, there are several departments (e.g., CT, PLA, ENT) where the hybrid-mean predictions have MRAEs almost as twice large as those using our regression methods. Similar large differences in MSE can also be found. This observation suggests that the hybrid sample-mean method is not a suitable tool for predicting surgical durations when numerous and complex CPT code combinations are used.

- For 8 of the 16 departments, the reduction of MRAE by using the linear regression model instead of the hybrid sample-mean method (the best performer of the three benchmark methods) is bigger than 0.10, which corresponds to a 30 minutes reduction of prediction error for a 5-hour long surgery.

- The two benchmark models, the lognormal model and the departmental sample-mean model, have similar performances. The lognormal model does not exhibit any noticeable edge in terms of prediction quality over the simple sample-mean model. To some extent, this result "validates" the use of the sample-mean model in practice. We believe that the lack of difference between these two benchmark models is due to the fact that the surgical data within a department do not always follow a lognormal distribution (see Section C of this chapter and

Figure 7).

CHAPTER V

CONCLUSIONS AND DISCUSSION

A. Application of Our Bivariate Model

The public health importance of understanding the distribution of energy-adjusted usual intake of episodically consumed dietary components is very great, having implications for basic understanding of both dietary component composition and policy. Being able to correct for measurement error due to within-person variation in short-term assessment of intake, when investigating diet-disease relationships in cohort studies, is equally important. Because of the importance of these problems, models and fitting methods for addressing them will find wide use in the nutrition community. Thus, it is not only vital that the models are reasonable, but that the fitting methods be reasonably fast, that they converge, and that the answers from the fitting methods usually make sense. The main point of this project has been to show that an MCMC approach satisfies these criteria, and has the potential to be used widely in the nutrition community. The fact that the MCMC approach can be used in a frequentist sense is a new insight for nutritional epidemiology, which is decidedly frequentist in orientation, although of course the MCMC model fitting can also allow Bayesian inference.

Our methods are not limited to estimating the distributions of usual intake. Indeed, they can also be applied to the problem of analyzing the relationship between energy-adjusted usual intake and disease. The typical method for such analysis as applied to studies such as the NIH-AARP Diet and Health Study is regression calibration (Carroll et al., 2006). In this methodology, the unobserved usual intake is replaced by its regression of usual intake on covariates and the FFQ. While our main

focus has been on distributions of usual intake, it is trivial to extend the methods in Section C of Chapter II to regression calibration.

There is of course an enormous literature on measurement error models, both parametric and nonparametric, for estimating distributions (e.g., Fan, 1991; Wand, 1998; Johnson et al., 2007; Staudenmeyer et al., 2008; Delaigle et al., 2008, among many others) and in regression (Ferrari et al., 2004; Liang and Wang, 2005, among many others). Many more references are given in Carroll et al. (2006). However, none of these papers deal with our topic of episodically consumed and hence zero-inflated dietary components along with continuous components that involve skewness, a structured covariance matrix, correlations of random effects, and usual intakes on the original data scale.

An issue of practically much less importance is that the model of Kipnis et al. (2010b) in equation (2.6) assumes that each food is consumed by all individuals. Kipnis et al. (2009) address this issue, by adding a fixed effect regression so as to model never-consumers. They show that even without energy in the model, and with only two 24hr as is standard for such data, their method was numerically very unstable. Our method easily handles such an extension, but its practical implications are not particularly clear when, for example, in other studies, less than 0.5% of subjects claimed on the FFQ never to eat fish or whole grains.

B. Extension of Our Multivariate Model

1. Transformations

In Appendix 23, we describe how we estimated the transformation parameters as a separate component-wise calculation. We have done some analyses where we simultaneously transform each component, and found very little difference with our results.

However, the computing time to implement this is extremely high, because of the fact that different transformations make data on different scales, so we have to compute the usual intakes at each step in the MCMC, and not just at the end.

2. What Have We Learned That Is New

There are many important questions in dietary assessment that have not been able to be answered because of a lack of multivariate models for complex, zero-inflated data with measurement errors and a lack of ability to fit such multivariate models. Nutrients and foods are not consumed in isolation, but rather as part of a broader pattern of eating. There is reason to believe that these various dietary components interact with one another in their effect on health, sometimes working synergistically and sometimes in opposition. Nonetheless, simply characterizing various patterns of eating has presented enormous statistical challenge. Until now, descriptive statistics on the HEI-2005 have been limited to examination of either the total scores or only a single energy-adjusted component at a time. This has precluded characterization of various patterns of dietary quality as well as any subsequent analyses of how such patterns might relate to health.

This methodology presented in the second project presents a workable solution to these problems which has already proven valuable. In May 2010, just as we were submitting the paper, a White House Task Force on Childhood Obesity created a report. They had wanted to set a goal of all children having a total HEI score of 80 or more by 2030, but when they learned we estimated only 10% of the children ages 2-8 had a score of 66 or higher, they decided to set a more realistic target. The facility to estimate distributions of the multiple component scores simultaneously will be important in tracking progress toward that goal.

3. In What Other Arenas Will Our Work Have Impact

There are many other important problems where multivariate models such as ours will be important. One such problem arises when studying the relationship between multiple dietary components or dietary patterns and health outcomes. Traditionally, for cost reasons, large cohort studies have used a food frequency questionnaire (FFQ) to measure dietary intake, sometimes with a small calibration study including short-term measures such as 24hr recalls. However, there is a new web-based instrument called the Automated Self-administered 24-hour Dietary Recall (ASA24$^{\text{TM}}$), see http://riskfactor.cancer.gov/tools/instruments/asa24, which has been proposed to replace or at least supplement the FFQ and which is currently undergoing extensive testing. The dietary data we will see then is what we have called Y_{ijk}, i.e., 24hr recall data. In order to correct relative risk estimates for the measurement error inherent in the ASA24$^{\text{TM}}$, regression calibration (Carroll et al., 2006) will almost certainly be the method of choice, as it is in most of nutritional epidemiology. This method attempts to produce an estimate of the regression of usual intake on the observed intakes, and then to use these estimates in Cox and logistic regression for the health outcome. In order to perform this regression, a multivariate measurement error model will be required, since the regression is on <u>all</u> the observed dietary intake components in the regression model measured by the ASA24$^{\text{TM}}$, and not on each individual component. Our methodology is easily extended to address this problem.

C. Significance of Our Predictive Model

The third project presents regression-based methodologies that take multiple CPT codes as explanatory variables when predicting surgical case durations. Our research is motivated by the fact that CPT codes describe how a surgical case should be

performed, and thus provide specific knowledge and information relevant to individual surgical cases. The importance of CPT codes in predicting surgical case durations has been noted in health care literature for years. Our research demonstrates the benefit of utilizing the CPT codes by a prediction comparision using real data from a large central Texas hospital. The reduction of prediction errors due to utilization of the CPT codes will certainly boost certainty in the scheduling process and help cut "white spaces" between surgeries (i.e., buffer time inserted between surgeries to accommodate variability) or overruns so that more surgeries can be scheduled with a higher start time reliability. Our proposed methodology could help a great deal with the issues related to the operating room scheduling and resource utilization, and consequently, will bring considerable economic benefits to the bottom line of a hospital and lead to greater patient satisfaction.

To the best of our knowledge, our project is the first that predicts surgery case durations based on multiple CPT codes that a surgical case performs. In our opinion, one of the reasons that such a predictive model was not available prior to our research is perhaps caused by the complexity involved in devising a proper design matrix. If naively constructing a design matrix according to the appearances of CPT codes in surgical cases, one will likely end up with an ill-conditioned matrix that is not solvable. In our research, we develop general procedures to overcome this difficulty by systematically grouping CPT combinations and treat not only single CPT codes but also distinctive CPT combinations with multiple CPT codes as separate explanatory variables. Our algorithm guarantees a fully ranked design matrix, and consequently, the solvability of the least-squares estimation. Although the implementation details are provided for surgical cases using up to eight CPT codes (which in itself has already represented very complicated surgeries), our models and algorithms can be applied to surgical cases using any number of CPT codes or any combination of CPT codes.

One possible extension of our research is to consider other important covariates together with CPT codes. In addition to CPT codes, which is arguably the most important factor relevant to the prediction of surgical case durations, prior research also identified a number of other factors influencing surgical case durations (such as surgeon experience, anesthesia type, patient's status). The inclusion of those factors is methodologically straightforward — we can extend our predictive models by simply adding the relevant covariates. Although we believe that an extended model that incorporates both CPT codes and other relevant covariates as explanatory variables has the potential to further reduce prediction uncertainty, testing the extended model using real data would require a different data set than the one we have, and another round of (possibly very lengthy) data collection efforts.

REFERENCES

Buonaccorsi, J. (2010). *Measurement Error: Models, Methods and Applications*. New York: Chapman & Hall/CRC Interdisciplinary Statistics Series.

Carriquiry, A. L. (1999). Assessing the prevalence of nutrient inadequacy. *Public Health Nutrition* **2**, 23-33.

Carriquiry, A. L. (2003). Estimation of usual intake distributions of nutrients and foods. *Journal of Nutrition* **133**, 601-608.

Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*, Second Edition. New York: Chapman and Hall CRC Press.

Casella, G. and Berger, R. L. (2001). *Statistical Inference*, 2nd Edition. New York: Duxbury Press.

Casella, G. and George, E. (1992). Explaining the Gibbs Sampler. *The American Statistician* **46**, 167-174.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings Algorithm. *The American Statistician* **49**, 327-335.

Combes, C., Meskens, N., Rivat, C. and Vandamme, J. P. (2008). Using a KDD process to forecast the duration of surgery. *International Journal of Production Economics* **112**, 279-293.

Davidson, R. and MacKinnon, J. G. (1999). Bootstrap testing in nonlinear models. *International Economic Review* **40**, 487-508.

Delaigle, A. (2008). An alternative view of the deconvolution problem. *Statistica Sinica* **18**, 1025-1045.

Delaigle, A. and Hall, P. (2008). Using SIMEX for smoothing-parameter choice in errors-in-variables problems. *Journal of the American Statistical Association* **103**, 280-287.

Delaigle, A. and Hall, P. (2010). Estimation of observation-error variance in errors-in-variables regression. *Statistica Sinica*, to appear.

Delaigle, A., Hall, P. and Meister, A. (2008). On deconvolution with repeated measurements. *Annals of Statistics* **36**, 665-685.

Delaigle, A. and Meister, A. (2008). Density estimation with heteroscedastic error. *Bernoulli* **14**, 562-579.

Dexter, F., Dexter, E. U., Masursky, D. and Nussmeier, N. A. (2008). Systematic review of general thoracic surgery articles to identify predictors of operating room case durations. *Anesthesia and Analgesia* **106**, 1232-1241.

Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Annals of Statistics* **19**, 1257-1272.

Ferrari, P., Kaaks, R., Fahey, M. T., Slimani, N., Day, N. E., Pera, G., Boshuizen, H. C., Roddam, A., Boeing, H., Nagel, G., Thiebaut, A., Orfanos, P., Krogh, P., Braaten, T., and Riboli, E. (2004). Within- and between-cohort variation in measured macronutrient intakes, taking account of measurement errors, in the European Prospective Investigation Into Cancer and Nutrition Study. *American Journal of Epidemiology* **160**, 814-822.

Ferrari, P., Roddam, A., Fahey, M. T., Jenab, M., Bamia, C., Ock, M., Amiano, P., Hjartker, A., Biessy, C., Rinaldi, S., Huybrechts, I., Tjnneland, A., Dethlefsen, C., Niravong, M., Clavel-Chapelon, F., Linseisen, J., Boeing, H., Oikonomou, E., Orfanos, P., Palli, D., Santucci de Magistris, M., Bueno-de-Mesquita, H.

B., Peeters, P. H., Parr, C. L., Braaten, T., Dorronsoro, M., Berenguer, T., Gullberg, B., Johansson, I., Welch, A. A., Riboli, E., Bingham, S. and Slimani, N. (2009). A bivariate measurement error model for nitrogen and potassium intakes to evaluate the performance of regression calibration in the European Prospective Investigation into Cancer and Nutrition study. *European Journal of Clinical Nutrition* **63**, Supplement 4, 179-187.

Flegal, J. M., Haran, M. and Jones, G. L. (2008). Markov Chain Monte Carlo: Can we trust the third significant figure? *Statistical Science* **23**, 250-260.

Fraser, G. E. and Shavlik, D. J. (2004). Correlations between estimated and true dietary intakes. *Annals of Epidemiology* **14**, 287-95.

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. *Journal of Nutrition* **140**, 1529-1534.

Fuller, W. A. (1987). *Measurement Error Models*. New York: Wiley.

Fungwe, T., Guenther, P. M., Juan, W. Y., Hiza, H, and Lino, M. (2009). The quality of children's diets in 2003-04 as measured by the Healthy Eating Index-2005. *Nutrition Insight* **43**, USDA Center for Nutrition Policy and Promotion.

Guenther, P. M., Reedy, J. and Krebs-Smith, S. M. (2008a). Development of the Healthy Eating Index-2005. *Journal of the American Dietetic Association* **108**, 1896-1901.

Guenther, P. M., Reedy, J., Krebs-Smith, S. M. and Reeve, B. B. (2008b). Evaluation of the Healthy Eating Index-2005. *Journal of the American Dietetic Association* **108**, 1854-1864.

Guolo, A. (2008). A flexible approach to measurement error correction in casecontrol studies. *Biometrics* **64**, 1207-1214.

Gustafson, P. (2003). *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments.* New York: Chapman and Hall/CRC Press.

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection, and limited dependent variables and a simple estimator for such models. *Annals of Economics and Social Management* **5**, 475-592.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica* **47**, 153-161.

Johnson, B. A., Herring, A. H., Ibrahim, J. G., and Siega-Riz, A. M. (2007). Structured measurement error in nutritional epidemiology: Applications in the pregnancy, infection, and nutrition (PIN) study. *Journal of the American Statistical Association* **102**, 856-866.

Kass, R. E., Carlin, B. P., Gelman, A., and Neal, R. M. (1998). Markov Chain Monte Carlo in Practice: A Roundtable Discussion. *Statistical Science* **52**(2), 93-100.

Kipnis, V., Freedman, L. S., Carroll, R. J. and Midthune, D. (2010a). A bivariate measurement error model for an episodically consumed dietary component and energy: Application to epidemiology. Preprint.

Kipnis, V., Freedman, L. S., Carroll, R. J. and Midthune, D. (2010b). A measurement error model for episodically consumed foods and energy. Preprint.

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**, 1003-1010.

Kott, P. S., Guenther, P. M., Wagstaff, D. A., Juan W. Y. and Kranz, S. (2009). Fitting a linear model to survey data when the long-term average daily intake of a dietary component is an explanatory variable. *Survey Research Methods* **3**, No 3, 157-165.

Küchenhoff, H., Mwalili, S. M. and Lesaffre, E. (2006). A general method for dealing with misclassification in regression: The misclassification SIMEX. *Biometrics* **62**, 85-96.

Kyriazidou, E. (1997). Estimation of a panel data sample selection model. *Econometrica* **65**, 1335-1364.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*. New York: Springer.

Leung, S. F. and Yu, S. (1996). On the choice between sample selection and two-part models. *Journal of Econometrics* **72**, 197-229.

Li, L., Shao, J., and Palta, M. (2005). A longitudinal measurement error model with a semicontinuous covariate. *Biometrics* **61**, 824-830.

Liang, H., Thurston, S., Ruppert, D., Apanasovich, T. and Hauser, R. (2008). Additive partial linear models with measurement errors. *Biometrika* **95**, 667-678.

Liang, H. and Wang, N. S. (2005). Partially linear single-index measurement error models. *Statistica Sinica* **15**, 99-116.

Litvak, E. and Long, M. C. (2000). Cost and quality under managed care: Irreconcilable differences? *American Journal of Managed Care* **6**, 305-312.

May, J. H. , Strum, D. P. and Vargas, L. G. (2000). Fitting the lognormal distribution to surgical procedure times. *Decision Sciences* **31**, 129-148.

McManus, M. L., Long, M. C. , Cooper, A., Mandell, J., Berwick, D. M., Pagano, M. and Litvak, E. (2003). Variability in surgical caseload and access to intensive care services. *Anesthesiology* **98**, 1491-1496.

Messer, K. and Natarajan, L. (2008). Maximum likelihood, multiple imputation and regression calibration for measurement error adjustment. *Statistics in Medicine* **27**, 6332-6350.

Min, Y. and Agresti, A. (2002). Modeling nonnegative data with clumping at zero: A survey. *Journal of the Iranian Statistical Society* **1**, 7-33.

Mitchell, T. M.(1997). *Machine Learning*. New York: McGraw-Hill.

Natarajan, L. (2009). Regression calibration for dichotomized mismeasured predictors. *International Journal of Biostatistics* **5**(1), Article 12.

Nusser, S. M., Carriquiry, A. L., Dodd, K. W., and Fuller, W. A. (1996). A semiparametric approach to estimating usual intake distributions. *Journal of the American Statistical Association* **91**, 1440-1449.

Nusser, S. M., Fuller, W. A., and Guenther, P. M. (1997). Estimating usual dietary intake distributions: Adjusting for measurement error and nonnormality in 24-hour food intake data. In Lyberg, L, Biemer, P, Collins, M, Deleeuw, E, Dippo, C, Schwartz, N, and Trewin, D (editors). *Survey Measurement and Process Quality*, pp.670-689, New York: Wiley.

Olivares, M. and Terwiesch, C. and Cassorla, L. (2008). Structural estimation of the newsvendor model: An application to reserving operating room time. *Management Science* **54**, 41-55.

Olsen, M. K. and Schafer, J. L. (2001). A two-part random-effects model for semi-continuous longitudinal data. *Journal of the American Statistical Association* **96**, 730-745.

Prentice, R. L. (1996). Measurement error and results from analytic epidemiology: Dietary fat and breast cancer. *Journal of the National Cancer Institute* **88**, 1738-1747.

Prentice, R. L. (2003). Dietary assessment and the reliability of nutritional epidemiology reports. *Lancet* **362**, 182-183.

Robert, C. P. (1995). Simulation of truncated normal variables. *Statistics and Computing* **5**, 121-125.

Schatzkin, A., Subar, A. F., Thompson, F. E., Harlan, L. C., Tangrea, J., Hollenbeck, A. R., Hurwitz, P. E., Coyle, L., Schussler, N., Michaud, D. S., Freedman, L. S., Brown, C. C., Midthune, D. and Kipnis, V. (2001). Design and serendipity in establishing a large cohort with wide dietary intake distributions: The National Institutes of Health-AARP Diet and Health Study. *American Journal of Epidemiology* **154**, 1119-1125.

Sinha, S., Mallick, B. K., Kipnis, V. and Carroll, R. J. (2010). Semiparametric Bayesian analysis of nutritional epidemiology data in the presence of measurement error. *Biometrics*, to appear.

Spangler, W. E., Strum, D. P. , Vargas, L. G. and May, J. H. (2004). Estimating procedure times for surgeries by determining location parameters for the lognormal model. *Health Care Management Sciences* **7**, 97-104.

Staudenmayer, J., Ruppert, D., and Buonaccorsi, J. (2008). Density estimation in the presence of heteroskedastic measurement error. *Journal of the American*

Statistical Association **103**, 726-736.

Strum, D. P., May, J. H. and Vargas, L. G. (2000a). Modeling the uncertainty of surgical procedure times: Comparison of the log-normal and normal models. *Anesthesiology* **92**, 1160-1167.

Strum, D. P., May, J. H., Sampson, A. R., Vargas, L. G. and Spangler, W. E. (2003). Estimating times of surgeries with two component procedures. *Anesthesiology* **98**, 232-240.

Strum, D. P., Sampson, A. R. , May, J. H. and Vargas, L. G. (2000b). Surgeon and type of anesthesia predict variability in surgical procedure times. *Anesthesiology* **92**, 1454-1466.

Subar, A. F., Dodd, K. W., Guenther, P. M., Kipnis, V., Midthune, D., McDowell, M., Tooze, J. A., Freedman, L. S. and Krebs-Smith, S. M. (2006). The food propensity questionnaire: Concept, development, and validation for use as a covariate in a model to estimate usual food intake. *Journal of the American Dietetic Association* **106**, 1556-1563.

Tooze, J. A., Grunwald, G. K., and Jones, R. H. (2002). Analysis of repeated measures data clumping at zero. *Statistical Methods in Medical Research* **11**, 341-355.

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 statistical method for estimating the usual intake of episodically consumed foods with application to their distribution. *Journal of the American Dietetic Association* **106**, 1575-1587.

Wand, M. P. (1998). Finite sample performance of deconvolving kernel density estimators. *Statistics and Probability Letters* **37**, 131-139.

Weisberg, S. (2005). *Applied Linear Regression*, 3rd Edition. New York: Wiley/Interscience.

Weiss, E. N. (1990). Models for determining estimated start times and case ordering in hospital operating rooms. *IIE Transactions* **22**, 143-150.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd Edition. San Francisco: Morgan Kaufmann.

Wolter, K. M. (1995). *Introduction to Variance Estimation*. New York: Springer-Verlag.

Wright, I. H., Kooperberg, C., Bonar, B. and Bashein, G. (1996). Statistical modeling to predict elective surgery time: Comparison with a computer scheduling system and surgeon-provided estimates. *Anesthesiology* **85**, 1235-1245.

Zhang, S., Midthune, D., Pérez, A, Buckman, D. W., Kipnis, V., Freedman, L. S., Dodd, K. W., Krebs-Smith, S. M. and Carroll, R. J. (2010). A bivariate measurement error model for episodically consumed dietary components. Preprint.

APPENDIX A

APPENDIX OF CHAPTER II

1. Notational Convention

Standardization is important in MCMC applications both for numerical stability and to allow fairly off-the-shelf prior distributions to make sense. Prior to analysis, we standardized the covariates to have mean 0.0 and variance 1.0. The observed, transformed non-zero 24hr were standardized to have mean 0.0 and variance 2.0. More precisely, we first transformed the non-zero dietary component data as $Z_{i2k} = g(Y_{i2k}, \lambda_F)$, and then we standardized these data as $Q_{i2k} = \sqrt{2}(Z_{i2k} - a_F)/s_F$. Similarly, for energy we transformed to $Z_{i3k} = g(Y_{i3k}, \lambda_E)$ and then standardized to $Q_{i3k} = \sqrt{2}(Z_{i3k} - a_E)/s_E$. Of course, whether the dietary component is consumed or not is $Q_{i1k} = Y_{i1k}$. Collected, the data are $\widetilde{\mathbf{Q}}_{\mathbf{ik}} = (Q_{i1k}, Q_{i2k}, Q_{i3k})^{\mathrm{T}}$. The terms (a_F, s_F, a_E, s_E) are not random variables but are merely constants used for standardization, and we need not consider inference for them.

We will first describe the algorithm used in terms of the Q_{ijk}, and then in Appendix 11, we describe the back-transformation method that we used to obtain estimation and inference for usual intake.

Remark 8 Having the total variability of the non-zero transformed responses equal to 2.0 is extraordinarily convenient. Effectively, this means that $\mathrm{var}(U_{ij}) + \mathrm{var}(\epsilon_{ij}) \approx 2.0$ for $j = 1, 2$. Thus, neither component of this sum is at all likely to be large. Hence, a prior mean for the diagonal elements of Σ_u all equalling 1.0, while too large in our examples, is at least within nodding distance of a reasonable answer. Having priors for $\mathrm{var}(\epsilon_{ij})$ for $j = 1, 2$ that are Uniform$[0, 3]$ is flexible and does not allow ridiculous answers.

2. Prior Distributions

Because the data were standardized, following the discussion of Remark 8, we used the following conventions.

- The priors for all $\beta_{\mathbf{j}}$ were normal with mean zero and variance 100.
- The prior for $\mathbf{\Sigma_u}$ was exchangeable with diagonal entries all equal to 1.0 and correlations 0.50. There was 5 degrees of freedom in the inverse Wishart prior, i.e., $m_u = 5$. Thus, the prior is IW$\{(m_u - 3 - 1)\mathbf{\Omega_u}, m_u\}$.
- The priors for s_{22} and s_{33} were Uniform[0,3]. This range is reasonable because of the standardization.
- The priors for (γ, θ) were uniform on their range.

We experimented with different priors for $\mathbf{\Sigma_u}$, e.g., setting the correlations equal to 0.0, setting the diagonal elements equal to 0.5, etc. The results were essentially unchanged when these were done.

3. Generating Starting Values for the Latent Variables

While we observe $\widetilde{\mathbf{Q}}_{\mathbf{ik}}$, in the MCMC we need to generate the latent variables $\widetilde{\mathbf{W}}_{\mathbf{ik}}$ to initiate the MCMC.

- For energy, $Q_{i3k} = W_{i3k}$, no data need to be generated.
- For the amounts, Q_{i2k}, we just simply set $W_{i2k} = Q_{i2k}$.
- For consumption, we generate $\widetilde{\mathbf{U}}_{\mathbf{i}}$ as normally distribution with mean zero and covariance matrix given as the prior covariance matrix for $\mathbf{\Sigma_u}$. We then also compute $z_{ik} = |\mathbf{X_{i1}}^{\mathrm{T}}\beta_{\mathbf{1},\mathrm{prior}} + U_{i1} + \mathcal{Z}_{ik}|$, where $\mathcal{Z}_{ik} = \mathrm{Normal}(0, 1)$ are generated independently. We then set $W_{i1k} = z_{ik}Q_{i1k} - z_{ik}(1 - Q_{i1k})$.

- We then updated $\widetilde{\mathbf{W}}_{\mathbf{ik}}$ by a single application of the updates given in Appendix 9.

4. Complete Data Loglikelihood

The loglikelihood of the complete data is

$$\begin{aligned}
&\textstyle\sum_{i=1}^{n}\sum_{k=1}^{2}\log\{Q_{i1k}I(W_{i1k}>0)+(1-Q_{i1k})I(W_{i1k}<0)\}\\
&+(n/2)\log(|\boldsymbol{\Sigma_u}^{-1}|)-(1/2)\textstyle\sum_{i=1}^{n}\widetilde{\mathbf{U}}_{\mathbf{i}}^{\mathrm{T}}\boldsymbol{\Sigma_u}^{-1}\widetilde{\mathbf{U}}_{\mathbf{i}}\\
&-(1/2)\textstyle\sum_{j=1}^{3}(\beta_{\mathbf{j}}-\beta_{\mathbf{j},\mathrm{prior}})^{\mathrm{T}}\boldsymbol{\Omega}_{\beta,\mathbf{j}}{}^{-1}(\beta_{\mathbf{j}}-\beta_{\mathbf{j},\mathrm{prior}})\\
&+\{(m_u+3+1)/2\}\log(|\boldsymbol{\Sigma_u}^{-1}|)-(1/2)\mathrm{trace}(\boldsymbol{\Omega_u}\boldsymbol{\Sigma_u}^{-1})\\
&-(1/2)(2n)\log\{s_{22}s_{33}(1-\gamma^2)\}\\
&-(1/2)\textstyle\sum_{i=1}^{n}\sum_{k=1}^{2}\{\widetilde{\mathbf{W}}_{\mathbf{ik}}-(\mathbf{X_{i1}}^{\mathrm{T}}\beta_{\mathbf{1}},...,\mathbf{X_{i3}}^{\mathrm{T}}\beta_{\mathbf{3}})^{\mathrm{T}}-\widetilde{\mathbf{U}}_{\mathbf{i}}\}^{\mathrm{T}}\boldsymbol{\Sigma}_{\epsilon}{}^{-1}\{\bullet\},
\end{aligned}$$

where $\{\bullet\}$ means that the term before $f(\cdot)$ is transposed and substituted.

Remark 9 In the NIH-AARP Study, only the calibration sub-study has any information about the parameters $(\beta_{\mathbf{1}},\beta_{\mathbf{2}},\beta_{\mathbf{3}},\boldsymbol{\Sigma_u},\boldsymbol{\Sigma}_{\epsilon})$. Consequently, our methodology is run only on the calibration sub-study.

5. Complete Conditionals for $(\gamma,\theta,s_{22},s_{33})$

The complete conditionals for $(\gamma,\theta,s_{22},s_{33})$ do not have an explicit form, so we use a Metropolis-Hastings within Gibbs sampler to generate them in turn. Since $\boldsymbol{\Sigma}_{\epsilon}$ is determined by γ, θ, s_{22} and s_{33}, we write it as $\boldsymbol{\Sigma}_{\epsilon}{}^{-1}\equiv f(\gamma,\theta,s_{22},s_{33})$. Also, current values are γ_t, θ_t, $s_{22,t}$ and $s_{33,t}$.

Generation of γ. For convenience, we set γ to be discrete with 41 equally-spaced values on its range. Let γ_t be the current value. The candidate value y is selected

randomly from γ_t and its two nearest neighbors. The candidate value y is accepted with probability $\alpha(\gamma_t, y)$, $\alpha(\gamma_t, y) = \min\{1, g(y)/g(\gamma_t)\}$, where

$$\begin{aligned} g(y) &\propto (1-y^2)^{-n} \\ &\quad \times \exp\Big[-\frac{1}{2}\textstyle\sum_{i=1}^{n}\sum_{k=1}^{2}\{\widetilde{\mathbf{W}}_{\mathbf{ik}} - (\mathbf{X_{i1}}^{\mathrm{T}}\beta_{\mathbf{1}}, ..., \mathbf{X_{i3}}^{\mathrm{T}}\beta_{\mathbf{3}})^{\mathrm{T}} - \widetilde{\mathbf{U}}_{\mathbf{i}}\}^{\mathrm{T}} f(y, \theta_t, s_{22,t}, s_{33,t})\{\bullet\}\Big], \end{aligned}$$

If the candidate y is accepted, then $\gamma_{t+1} = y$. Otherwise, $\gamma_{t+1} = \gamma_t$.

Generation of θ. This is done exactly as for γ, except now

$$g(y) \propto \exp\Big[-\frac{1}{2}\textstyle\sum_{i=1}^{n}\sum_{k=1}^{2}\{\widetilde{\mathbf{W}}_{\mathbf{ik}} - (\mathbf{X_{i1}}^{\mathrm{T}}\beta_{\mathbf{1}}, ..., \mathbf{X_{i3}}^{\mathrm{T}}\beta_{\mathbf{3}})^{\mathrm{T}} - \widetilde{\mathbf{U}}_{\mathbf{i}}\}^{\mathrm{T}} f(\gamma_{t+1}, y, s_{22,t}, s_{33,t})\{\bullet\}\Big].$$

If the candidate y is accepted, then $\theta_{t+1} = y$. Otherwise, $\theta_{t+1} = \theta_t$.

Generation of s_{22}. Suppose the current value of s_{22} is $s_{22,t}$. A candidate value y is generated from the Uniform distribution of length 0.4 with mean $s_{22,t}$: y $\sim$ Uniform[$s_{22,t}$ - 0.2, $s_{22,t}$ + 0.2]. The candidate value y is accepted with probability $\alpha(s_{22,t}, y)$, where

$$\begin{aligned} \alpha(s_{22,t}, y) &= \min\left\{(1, g(y)I_{[0,3]}(y)/g(s_{22,t})\right\}; \\ g(y) &\propto y^{-n}\exp\Big[-\frac{1}{2}\textstyle\sum_{i=1}^{n}\sum_{k=1}^{2}\{\widetilde{\mathbf{W}}_{\mathbf{ik}} - (\mathbf{X_{i1}}^{\mathrm{T}}\beta_{\mathbf{1}}, ..., \mathbf{X_{i3}}^{\mathrm{T}}\beta_{\mathbf{3}})^{\mathrm{T}} - \widetilde{\mathbf{U}}_{\mathbf{i}}\}^{\mathrm{T}} \\ &\qquad \times f(\gamma_{t+1}, \theta_{t+1}, y, s_{33,t})\{\bullet\}\Big] \end{aligned}$$

If the candidate is accepted, then $s_{22,t+1} = y$. Otherwise, $s_{22,t+1} = s_{22,t}$.

Generation of s_{33}. This is the same as that for s_{22}, except now

$$\begin{aligned} \alpha(s_{33,t}, y) &= \min\left\{1, g(y)I_{[0,3]}(y)/g(s_{33,t})\right\}; \\ g(y) &\propto y^{-n}\exp\Big[-\frac{1}{2}\textstyle\sum_{i=1}^{n}\sum_{k=1}^{2}\{\widetilde{\mathbf{W}}_{\mathbf{ik}} - (\mathbf{X_{i1}}^{\mathrm{T}}\beta_{\mathbf{1}}, ..., \mathbf{X_{i3}}^{\mathrm{T}}\beta_{\mathbf{3}})^{\mathrm{T}} - \widetilde{\mathbf{U}}_{\mathbf{i}}\}^{\mathrm{T}} \\ &\qquad \times f(\gamma_{t+1}, \theta_{t+1}, s_{22,t+1}, y)\{\bullet\}\Big]. \end{aligned}$$

If the candidate is accepted, then $s_{33,t+1} = y$. Otherwise, $s_{33,t+1} = s_{33,t}$.

6. Complete Conditional for $\mathbf{\Sigma_u}$

By "rest", we mean all the observable data, latent variables and parameters other than the one in question. By inspection, the complete conditional for $\mathbf{\Sigma_u}$ is

$$[\mathbf{\Sigma_u}|\text{rest}] \;=\; \text{IW}\{(m_u - K - 1)\mathbf{\Omega_u} + \textstyle\sum_{i=1}^{n} \widetilde{\mathbf{U}}_\mathbf{i}\widetilde{\mathbf{U}}_\mathbf{i}^{\mathrm{T}}, n + m_u\}.$$

7. Complete Conditionals for β

Let the elements of $\mathbf{\Sigma_\epsilon^{-1}}$ be $\sigma_\epsilon^{j\ell}$. For any j, except for irrelevant constants,

$$\begin{aligned}
\log[\beta_\mathbf{j}|\text{rest}] &= -(1/2)(\beta_\mathbf{j} - \beta_{\mathbf{j},\text{prior}})^{\mathrm{T}}\mathbf{\Omega}_{\beta,\mathbf{j}}{}^{-1}(\beta_\mathbf{j} - \beta_{\mathbf{j},\text{prior}}) \\
&\quad -(1/2)\textstyle\sum_{i=1}^{n}\sum_{k=1}^{2}(W_{ijk} - \mathbf{X_{ij}}{}^{\mathrm{T}}\beta_\mathbf{j} - U_{ij})^2\sigma_\epsilon^{jj} \\
&\quad -\textstyle\sum_{i=1}^{n}\sum_{k=1}^{2}\sum_{\ell\neq j}\sigma_\epsilon^{j\ell}(W_{ijk} - \mathbf{X_{ij}}{}^{\mathrm{T}}\beta_\mathbf{j} - U_{ij})(W_{i\ell k} - \mathbf{X_{i\ell}}{}^{\mathrm{T}}\beta_\ell - U_{i\ell}) \\
&= \mathcal{C}_\mathbf{1}{}^{\mathrm{T}}\beta_\mathbf{j} - (1/2)\beta_\mathbf{j}{}^{\mathrm{T}}\mathcal{C}_\mathbf{2}{}^{-1}\beta_\mathbf{j}
\end{aligned}$$

which implies $[\beta_\mathbf{j}|\text{rest}] \sim \text{Normal}(\mathcal{C}_\mathbf{2}\mathcal{C}_\mathbf{1}, \mathcal{C}_\mathbf{2})$, where

$$\begin{aligned}
\mathcal{C}_\mathbf{2} &= (\mathbf{\Omega}_{\beta,\mathbf{j}}{}^{-1} + 2\textstyle\sum_{i=1}^{n}\sigma_\epsilon^{jj}\mathbf{X_{ij}}\mathbf{X_{ij}}{}^{\mathrm{T}})^{-1}; \\
\mathcal{C}_\mathbf{1} &= \mathbf{\Omega}_{\beta,\mathbf{j}}{}^{-1}\beta_{\mathbf{j},\text{prior}} + \textstyle\sum_{i=1}^{n}\sum_{k=1}^{2}\sigma_\epsilon^{jj}\mathbf{X_{ij}}(W_{ijk} - U_{ij}) \\
&\quad +\textstyle\sum_{i=1}^{n}\sum_{k=1}^{2}\sum_{\ell\neq j}\sigma_\epsilon^{j\ell}(W_{i\ell k} - \mathbf{X_{i\ell}}{}^{\mathrm{T}}\beta_\ell - U_{i\ell})\mathbf{X_{ij}}.
\end{aligned}$$

8. Complete Conditionals for $\widetilde{\mathbf{U}}_\mathbf{i}$

Except for irrelevant constants, and remembering that $j = 1, ..., 3$,

$$\begin{aligned}
\log\left[\widetilde{\mathbf{U}}_\mathbf{i}|\text{rest}\right] &= -(1/2)\widetilde{\mathbf{U}}_\mathbf{i}^{\mathrm{T}}\mathbf{\Sigma_u^{-1}}\widetilde{\mathbf{U}}_\mathbf{i} \\
&\quad -(1/2)\textstyle\sum_{k=1}^{2}\{\widetilde{\mathbf{W}}_\mathbf{ik} - (\mathbf{X_{i1}}{}^{\mathrm{T}}\beta_\mathbf{1}, ..., \mathbf{X_{i3}}{}^{\mathrm{T}}\beta_\mathbf{3})^{\mathrm{T}} - \widetilde{\mathbf{U}}_\mathbf{i}\}^{\mathrm{T}}\mathbf{\Sigma}_\epsilon{}^{-1} \\
&\qquad\qquad \times\{\widetilde{\mathbf{W}}_\mathbf{ik} - (\mathbf{X_{i1}}{}^{\mathrm{T}}\beta_\mathbf{1}, ..., \mathbf{X_{i3}}{}^{\mathrm{T}}\beta_\mathbf{3})^{\mathrm{T}} - \widetilde{\mathbf{U}}_\mathbf{i}\} \\
&= \mathcal{C}_\mathbf{1}{}^{\mathrm{T}}\widetilde{\mathbf{U}}_\mathbf{i} - (1/2)\widetilde{\mathbf{U}}_\mathbf{i}^{\mathrm{T}}\mathcal{C}_\mathbf{2}{}^{-1}\widetilde{\mathbf{U}}_\mathbf{i}
\end{aligned}$$

which implies $\left[\widetilde{\mathbf{U}}_\mathbf{i}|\text{rest}\right] \sim \text{Normal}(\mathcal{C}_\mathbf{2}\mathcal{C}_\mathbf{1}, \mathcal{C}_\mathbf{2})$, where

$$\begin{aligned}\mathcal{C}_\mathbf{2} &= (\boldsymbol{\Sigma}_\mathbf{u}{}^{-1} + 2\boldsymbol{\Sigma}_\epsilon{}^{-1})^{-1};\\ \mathcal{C}_\mathbf{1} &= \textstyle\sum_{k=1}^{2} \boldsymbol{\Sigma}_\epsilon{}^{-1}\{\widetilde{\mathbf{W}}_\mathbf{ik} - (\mathbf{X_{i1}}{}^{\mathrm{T}}\beta_\mathbf{1}, ..., \mathbf{X_{i3}}{}^{\mathrm{T}}\beta_\mathbf{3})^{\mathrm{T}}\}.\end{aligned}$$

9. Complete Conditionals for W_{i1k}

Here we do the complete conditional for $W_{i\ell k}$ with $\ell = 1$. Except for irrelevant constants,

$$\begin{aligned}\log\left[W_{i\ell k}|\text{rest}\right] &= \log\{Q_{i\ell k}I(W_{i\ell k} > 0) + (1 - Q_{i\ell k})I(W_{i\ell k} < 0)\}\\ &\quad -(1/2)(W_{i1k} - \mathbf{X_{i1}}{}^{\mathrm{T}}\beta_\mathbf{1} - U_{i1}, ..., W_{i3k} - \mathbf{X_{i3}}{}^{\mathrm{T}}\beta_\mathbf{3} - U_{i3})\boldsymbol{\Sigma}_\epsilon{}^{-1}(\bullet)\\ &= \log\{Q_{i\ell k}I(W_{i\ell k} > 0) + (1 - Q_{i\ell k})I(W_{i\ell k} < 0)\}\\ &\quad -(1/2)\sigma_\epsilon^{\ell\ell}(W_{i\ell k} - \mathbf{X_{i\ell}}{}^{\mathrm{T}}\beta_\ell - U_{i\ell})^2\\ &\quad -\textstyle\sum_{j\neq\ell}\sigma_\epsilon^{\ell j}(W_{i\ell k} - \mathbf{X_{i\ell}}{}^{\mathrm{T}}\beta_\ell - U_{i\ell})(W_{ijk} - \mathbf{X_{ij}}{}^{\mathrm{T}}\beta_\mathbf{j} - U_{ij})\\ &= \log\{Q_{i\ell k}I(W_{i\ell k} > 0) + (1 - Q_{i\ell k})I(W_{i\ell k} < 0)\}\\ &\quad +\mathcal{C}_1 W_{i\ell k} - (1/2)W_{i\ell k}^2\mathcal{C}_2^{-1},\end{aligned}$$

where

$$\begin{aligned}\mathcal{C}_2 &= 1/(\sigma_\epsilon^{\ell\ell})\\ \mathcal{C}_1 &= \sigma_\epsilon^{\ell\ell}(\mathbf{X_{i\ell}}{}^{\mathrm{T}}\beta_\ell + U_{i\ell}) - \textstyle\sum_{j\neq\ell}\sigma_\epsilon^{\ell j}(W_{ijk} - \mathbf{X_{ij}}{}^{\mathrm{T}}\beta_\mathbf{j} - U_{ij}).\end{aligned}$$

If we use the notation $\text{TN}_+(\mu, \sigma, c)$ for a normal random variable with mean μ, standard deviation σ is truncated from the left at c, and $\text{TN}_-(\mu, \sigma, c)$ is truncated from the right at c, then it follows that with $\mu = \mathcal{C}_2\mathcal{C}_1$ and $\sigma = \mathcal{C}_2^{1/2}$,

$$\begin{aligned}[W_{i\ell k}|\text{rest}] &= Q_{i\ell k}\text{TN}_+(\mu, \sigma, 0) + (1 - Q_{i\ell k})\text{TN}_-(\mu, \sigma, 0)\\ &= \mu + Q_{i\ell k}\text{TN}_+(0, \sigma, -\mu) + (1 - Q_{i\ell k})\text{TN}_-(0, \sigma, -\mu)\end{aligned}$$

$$
\begin{aligned}
&= \mu + Q_{i\ell k}\mathrm{TN}_+(0,\sigma,-\mu) - (1-Q_{i\ell k})\mathrm{TN}_+(0,\sigma,\mu) \\
&= \mu + \sigma\{Q_{i\ell k}\mathrm{TN}_+(0,1,-\mu/\sigma) - (1-Q_{i\ell k})\mathrm{TN}_+(0,1,\mu/\sigma)\}.
\end{aligned}
$$

Generating $\mathrm{TN}_+(0,1,c)$ is easy: if $c<0$, simply do rejection sampling of a Normal$(0,1)$ until you get one that is $> c$. If $c>0$, there is an adaptive rejection scheme (Robert, 1995). The "truncated normal" was used because the latent variable W_{i1k} is associated with Y_{i1k} which indicates whether the dietary component is consumed or not. If the dietary component is indeed consumed, then based on our model (2.2), W_{i1k} should have a positive value. Similarly, if the dietary component is actually not consumed, then W_{i1k} should have a negative value. In order to achieve these, we need a truncated distribution. Besides, the conditional distribution of W_{i1k} proportional to a normal distribution, thus we chose truncated normal.

10. Complete Conditionals for W_{i2k} When it is Not Observed

For $p=2$, the variable W_{ipk} is not observed when $Q_{i,p-1,k}=0$, or, equivalently, when $W_{i,p-1,k}<0$. Except for irrelevant constants,

$$
\begin{aligned}
\log\left[W_{ipk}|\mathrm{rest}\right] &= -(1/2)\sum_j\sum_\ell \sigma_\epsilon^{j\ell}(W_{ijk}-\mathbf{X_{ij}}^{\mathrm{T}}\beta_{\mathbf{j}}-U_{ij})(W_{i\ell k}-\mathbf{X_{i\ell}}^{\mathrm{T}}\beta_\ell-U_{i\ell}) \\
&= -(1/2)W_{ipk}^2\mathcal{C}_2^{-1}+\mathcal{C}_1W_{ipk}
\end{aligned}
$$

where

$$
\begin{aligned}
\mathcal{C}_2 &= 1/(\sigma_\epsilon^{pp}); \\
\mathcal{C}_1 &= \sigma_\epsilon^{pp}(\mathbf{X_{ip}}^{\mathrm{T}}\beta_{\mathbf{p}}+U_{ip})-\sum_{\ell\neq p}\sigma_\epsilon^{p\ell}(W_{i\ell k}-\mathbf{X_{i\ell}}^{\mathrm{T}}\beta_\ell-U_{i\ell}).
\end{aligned}
$$

Therefore,

$$
[W_{ipk}|\mathrm{rest}] = Q_{ipk}Q_{i,p-1,k}+(1-Q_{i,p-1,k})\mathrm{Normal}(\mathcal{C}_2\mathcal{C}_1,\mathcal{C}_2).
$$

11. Usual Intake, Standardization and Transformation

Here we show how to go from the transformed and standardized data to usual intakes. We first consider energy, where we used the transformation

$$Q_{i3k} = \sqrt{2}\{g(Y_{i3k}, \lambda_E) - a_E\}/s_E = g_{\rm tr}(Y_{i3k}, \lambda_E, a_E, s_E) = \mathbf{X_{i3}}^{\rm T}\beta_{\mathbf{3}} + U_{i3} + \epsilon_{i3k}.$$

When $\lambda_E = 0$, the back-transformation is

$$\begin{aligned} g_{\rm tr}^{-1}(z, 0, a_E, s_E) &= \exp\left\{a_E + s_E z/\sqrt{2}\right\}; \\ \partial^2 g_{\rm tr}^{-1}(z, 0, a_E, s_E)/\partial z^2 &= \frac{s_E^2}{2} g_{\rm tr}^{-1}(z, 0). \end{aligned}$$

When $\lambda_E \neq 0$, the back-transformation is

$$g_{\rm tr}^{-1}(z, \lambda_E, a_E, s_E) = \left[1 + \lambda_E\left\{a_E + s_E z/\sqrt{2}\right\}\right]^{1/\lambda_E}; \quad \text{(A.1)}$$

$$\partial^2 g_{\rm tr}^{-1}(z, \lambda_E, a_E, s_E)/\partial z^2 = \frac{s_E^2}{2}(1 - \lambda_E)\left[1 + \lambda_E\left\{a_E + s_E z/\sqrt{2}\right\}\right]^{-2+1/\lambda_E} \quad \text{(A.2)}$$

Define

$$g_{\rm tr}^{*}\{v, \lambda_E, a_E, s_E, \mathbf{\Sigma}_\epsilon(3,3)\} = g_{\rm tr}^{-1}(v, \lambda_E, a_E, s_E) + (1/2)\mathbf{\Sigma}_\epsilon(3,3)\frac{\partial^2 g_{\rm tr}^{-1}(v, \lambda_E, a_E, s_E)}{\partial v^2}.$$

As in Kipnis et al. (2009), the usual intake of energy for person i is

$$\begin{aligned} T_{Ei} &= E\left\{g_{\rm tr}^{-1}(\mathbf{X_{i3}}^{\rm T}\beta_{\mathbf{3}} + U_{i3} + \epsilon_{i3}, \lambda_E, a_E, s_E)|\mathbf{X_{i3}}, U_{i3}\right\} \\ &\approx g_{\rm tr}^{*}\left\{\mathbf{X_{i3}}^{\rm T}\beta_{\mathbf{3}} + U_{i3}, \lambda_E, a_E, s_E, \mathbf{\Sigma}_\epsilon(3,3)\right\}. \end{aligned}$$

Similarly, a person's usual intake of the dietary component on the original scale is defined as

$$T_{Fi} = \Phi(\mathbf{X_{i1}}^{\rm T}\beta_{\mathbf{1}} + U_{i1}) g_{\rm tr}^{*}\left\{\mathbf{X_{i2}}^{\rm T}\beta_{\mathbf{2}} + U_{i2}, \lambda_F, a_F, s_F, \mathbf{\Sigma}_\epsilon(2,2)\right\}.$$

APPENDIX B

APPENDIX OF CHAPTER III

In this Appendix we give the full details of the model fitting procedure.

12. Notational Convention

In our example, age was standardized to have mean 0.0 and variance 1.0, to improve numerical stability.

As described in Appendix 12, the observed, transformed non-zero 24hr recalls were standardized to have mean 0.0 and variance 2.0. More precisely, for $\ell = 1, 2, ..., 6$, we first transformed the non-zero food group data as $Z_{i,2\ell,k} = g(Y_{i,2\ell,k}, \lambda_\ell)$, and then we standardized these data as $Q_{i,2\ell,k} = \sqrt{2}\{Z_{i,2\ell,k} - \mu(\lambda_\ell)\}/\sigma(\lambda_\ell)$, where $\{\mu(\lambda_\ell), \sigma(\lambda_\ell)\}$ are the mean and standard deviation of the non-zero food intakes $Z_{i,2\ell,k}$. Similarly, for non-episodically consumed dietary components and energy we transformed to $Z_{i,6+\ell,k} = g(Y_{i,6+\ell,k}, \lambda_\ell)$ for $\ell = 7, ..., 13$, and then standardized to $Q_{i,6+\ell,k} = \sqrt{2}\{Z_{i,6+\ell,k} - \mu(\lambda_\ell)\}/\sigma(\lambda_\ell)$. Of course, whether the food group is consumed or not is $Q_{i,2\ell-1,k} = Y_{i,2\ell-1,k}$ for $\ell = 1, ..., 6$. Collected, the data are $\widetilde{Q}_{ik} = (Q_{ijk})_{j=1}^{19}$. The terms $\{\mu(\lambda_\ell), \sigma(\lambda_\ell)\}$ are not random variables but are merely constants used for standardization, and we need not consider inference for them. Back-transformation is discussed in Appendix 22.

13. Prior Distributions

Because the data were standardized, we used the following conventions.

- The prior for all β_j were normal with mean zero and variance 100.

- The prior for Σ_u was exchangeable with diagonal entries all equal to 1.0 and correlations all equal to 0.50. There were 21 degrees of freedom in the inverse Wishart prior, i.e., $m_u = 21$. Thus, the prior is IW$\{(m_u - 19 - 1)\Sigma_{u,\text{prior}}, m_u\}$. We experimented with this prior by using zero correlation, and the results were essentially unchanged.

- The prior for r_k is Uniform[-1, 1]. Set the initial value: $r_k = 0$, $k = 1, \ldots, 5$.

- The prior for θ_k is Uniform$[-\pi, \pi]$. Set the initial value: $\theta_k = 0$, $k = 1, \ldots, 25$.

- The priors for $v_{22}, v_{44}, \ldots, v_{12,12}$ and $v_{13,13}, \ldots, v_{19,19}$ were Uniform[-3,3]. Set the initial values: $v_{22} = v_{44} = \ldots = v_{12,12} = v_{13,13} = \ldots = v_{19,19} = 1$.

- For the rest of the non-diagonal v_{ij}'s which could not be determined by the restrictions, we used Uniform[-3,3] priors. Set the initial values to be 0.

The constraints on Σ_ϵ are nonlinear, and our parameterization enforces them easily without having to have prior distributions for the original parameterization that satisfy the nonlinear constraints.

The key thing that makes things work well with the other components of the matrix V with $\Sigma_\epsilon = VV^{\mathrm{T}}$ is that we have standardized the data as described in Section C. With this standardization, things become much nicer. For example, the variance of the ϵ's for energy is $\sum_{j=1}^{19} v_{19,j}^2$. However, since the sample variance for energy is standardized to equal 2.0, we simply just need to make priors for $v_{19,j}$ be uniform on a modest range to have real flexibility.

14. Generating Starting Values for the Latent Variables

While we observe $\widetilde{Q}_{ik}$, in the MCMC we need to generate starting values for the latent variables $\widetilde{W}_{ik} = (W_{ijk})_{j=1}^{19}$ to initiate the MCMC.

- For nutrients and energy, $Q_{ijk} = W_{ijk}$, no data need be generated, $j = 13, \ldots, 19$.
- For the amounts, Q_{i2k}, Q_{i4k}, Q_{i6k}, Q_{i8k}, $Q_{i,10,k}$ and $Q_{i,12,k}$, we set $W_{i2k} = Q_{i2k}$, $W_{i4k} = Q_{i4k}$, $W_{i6k} = Q_{i6k}$, $W_{i8k} = Q_{i8k}$, $W_{i,10,k} = Q_{i,10,k}$ and $W_{i,12,k} = Q_{i,12,k}$.
- For consumption, we generate $\widetilde{U}_i$ as normally distributed with mean zero and covariance matrix given as the prior covariance matrix for Σ_u. For $\ell = 1, \ldots, 6$, we also compute $z_{ik} = |X^{\mathrm{T}}_{i,2\ell-1,k}\beta_{2\ell-1,\mathrm{prior}} + U_{i,2\ell-1} + \mathcal{Z}_{ik}|$, where $\mathcal{Z}_{ik} = \mathrm{Normal}(0,1)$ are generated independently. We then set $W_{i,2\ell-1,k} = z_{ik}Q_{i,2\ell-1,k} - z_{ik}(1 - Q_{i,2\ell-1,k})$.
- Finally, we then updated $\widetilde{W}_{ik}$ by a single application of the updates given in Appendix 20.

15. Complete Data Loglikelihood

Let $J = 19$. The complete data include the indicators of whether a food was consumed, the W variables, and the random effect U variables. The loglikelihood of the complete data is

$$
\begin{aligned}
&\textstyle\sum_{\ell=1}^{6}\sum_{i=1}^{n}\sum_{k=1}^{2}\log\{Q_{i,2\ell-1,k}I(W_{i,2\ell-1,k} > 0) + (1 - Q_{i,2\ell-1,k})I(W_{i,2\ell-1,k} < 0)\} \\
&+(\textstyle\sum_{i=1}^{n} w_i/2)\log(|\Sigma_u^{-1}|) - (1/2)\sum_{i=1}^{n} w_i \widetilde{U}_i^{\mathrm{T}}\Sigma_u^{-1}\widetilde{U}_i \\
&-(1/2)\textstyle\sum_{j=1}^{J}(\beta_j - \beta_{j,\mathrm{prior}})^{\mathrm{T}}\Omega_{\beta,j}^{-1}(\beta_j - \beta_{j,\mathrm{prior}}) \\
&+\{(m_u + J + 1)/2\}\log(|\Sigma_u^{-1}|) - \{(m_u - J - 1)/2\}\mathrm{trace}(\Sigma_{u,\mathrm{prior}}\Sigma_u^{-1}) \\
&-(1/2)\textstyle\sum_{i=1}^{n} w_i m_i \log\{(v_{22}^2 v_{44}^2 v_{66}^2 v_{88}^2 v_{10,10}^2 v_{12,12}^2 v_{13,13}^2 \ldots v_{JJ}^2)\prod_{q=1}^{5}(1 - r_q^2)\} \\
&-(1/2)\textstyle\sum_{i=1}^{n} w_i \sum_{k=1}^{2}\{\widetilde{W}_{ik} - (X_{i1k}^{\mathrm{T}}\beta_1, ..., X_{iJk}^{\mathrm{T}}\beta_J)^{\mathrm{T}} - \widetilde{U}_i\}^{\mathrm{T}}\Sigma_\epsilon^{-1} \\
&\qquad\qquad \times\{\widetilde{W}_{ik} - (X_{i1k}^{\mathrm{T}}\beta_1, ..., X_{iJk}^{\mathrm{T}}\beta_J)^{\mathrm{T}} - \widetilde{U}_i\}.
\end{aligned}
$$

We used Gibbs sampling to update this complete data loglikelihood, the details for

which are given in subsequent appendices. The weights w_i are integers and are used here in a pseudo-likelihood fashion. One can also think of this as expanding each individual into w_i individuals, each with the same observed data but different latent variables. For computational convenience, since we are only asking for a frequentist estimator and not doing full Bayesian inference, the latent variables in the process are generated once for each individual. Estimates of Σ_u, Σ_ϵ and β_j for $j = 1, ..., J$ were computed as the means from the Gibbs samples. Once again, we emphasize that we are not doing a proper Bayesian analysis, but only using MCMC techniques to obtain a frequentist estimate, with uncertainty assessed using the frequentist BRR method.

16. Complete Conditionals for r_q, θ_q and v_{pq}

Except for irrelevant constants, the complete conditional for r_q $(q = 1, \ldots, 5)$ is

$$\begin{aligned}\log[r_q|\text{rest}] &= -(1/2)\textstyle\sum_{i=1}^{n} w_i m_i \log(1-r_q^2) \\ &\quad -(1/2)\textstyle\sum_{i=1}^{n} w_i \sum_{k=1}^{2} \{\widetilde{W}_{ik} - (X_{i1k}^{\mathrm{T}}\beta_1, ..., X_{i,19,k}^{\mathrm{T}}\beta_{19})^{\mathrm{T}} - \widetilde{U}_i\}^{\mathrm{T}} \Sigma_\epsilon^{-1} \\ &\quad \times \{\widetilde{W}_{ik} - (X_{i1k}^{\mathrm{T}}\beta_1, ..., X_{i,19,k}^{\mathrm{T}}\beta_{19})^{\mathrm{T}} - \widetilde{U}_i\}.\end{aligned}$$

Except for irrelevant constants, the complete conditionals for v_{qq} (q= 2, 4, 6, 8, 10, 12, 13, ..., 19) are

$$\begin{aligned}\log[v_{qq}|\text{rest}] &= -(1/2)\textstyle\sum_{i=1}^{n} w_i m_i \log(v_{qq}^2) \\ &\quad -(1/2)\textstyle\sum_{i=1}^{n} w_i \sum_{k=1}^{2} \{\widetilde{W}_{ik} - (X_{i1k}^{\mathrm{T}}\beta_1, ..., X_{i,19,k}^{\mathrm{T}}\beta_{19})^{\mathrm{T}} - \widetilde{U}_i\}^{\mathrm{T}} \Sigma_\epsilon^{-1} \\ &\quad \times \{\widetilde{W}_{ik} - (X_{i1k}^{\mathrm{T}}\beta_1, ..., X_{i,19,k}^{\mathrm{T}}\beta_{19})^{\mathrm{T}} - \widetilde{U}_i\}.\end{aligned}$$

Except for irrelevant constants, the compete conditionals for θ_q, $(q = 1, \ldots, 25)$ and non-diagonal free parameters v_{pq} are

$$\begin{aligned}\log[x|\text{rest}] &= -(1/2)\textstyle\sum_{i=1}^{n} w_i \sum_{k=1}^{2} \{\widetilde{W}_{ik} - (X_{i1k}^{\mathrm{T}}\beta_1, ..., X_{i,19,k}^{\mathrm{T}}\beta_{19})^{\mathrm{T}} - \widetilde{U}_i\}^{\mathrm{T}} \Sigma_\epsilon^{-1}\end{aligned}$$

$$\times\{\widetilde{W}_{ik} - (X_{i1k}^{\mathrm{T}}\beta_1, ..., X_{i,19,k}^{\mathrm{T}}\beta_{19})^{\mathrm{T}} - \widetilde{U}_i\}.$$

The full conditionals do not have an explicit form, so we use a Metropolis-Hastings within a Gibbs sampler to generate it.

- r_q $(q = 1, \ldots, 5)$

  We discretize the values of r_q to the set $\{-0.99 + 2 \times 0.99(j-1)/(M-1)\}$, where $j = 1, ..., M$ and we choose $M = 41$.

  Proposal: The current value is $r_{q,t}$. The proposed value of $r_{q,t+1}$ is selected randomly from the current value and the two nearest neighbors of $r_{q,t}$. Then $r_{q,t+1}$ is accepted with probability $\min\{1, g(r_{q,t+1})/g(r_{q,t})\}$, where

  $$\begin{aligned} g(y) \;&\propto\; (1-y^2)^{-(1/2)\sum_{i=1}^{n} w_i m_i} \\ &\quad \times \exp\left[-(1/2)\textstyle\sum_{i=1}^{n} w_i \sum_{k=1}^{2} \{\widetilde{W}_{ik} - (X_{i1k}^{\mathrm{T}}\beta_1, ..., X_{i,19,k}^{\mathrm{T}}\beta_{19})^{\mathrm{T}} - \widetilde{U}_i\}^{\mathrm{T}} \Sigma_\epsilon^{-1}(\bullet)\right], \end{aligned}$$

  where here and in what follows, for any A, $A^{\mathrm{T}}\Sigma_\epsilon^{-1}(\bullet) = A^{\mathrm{T}}\Sigma_\epsilon^{-1}A$.

- θ_q $(q = 1, \ldots, 25)$

  We discretize similarly as above.

  Proposal: The current value is $\theta_{q,t}$. The proposed value $\theta_{q,t+1}$ is selected randomly from the current value and the two nearest neighbors of $\theta_{q,t}$. Then $\theta_{q,t+1}$ is accepted with probability $\min\{1, g(\theta_{q,t+1})/g(\theta_{q,t})\}$, where

  $$g(y) \;\propto\; \exp\left[-(1/2)\textstyle\sum_{i=1}^{n} w_i \sum_{k=1}^{2} \{\widetilde{W}_{ik} - (X_{i1k}^{\mathrm{T}}\beta_1, ..., X_{i,19,k}^{\mathrm{T}}\beta_{19})^{\mathrm{T}} - \widetilde{U}_i\}^{\mathrm{T}} \Sigma_\epsilon^{-1}(\bullet)\right].$$

- v_{qq} $(q = 2, 4, 6, 8, 10, 12, 13, \ldots, 19)$

  Proposal: The current value is $v_{qq,t}$. A candidate $v_{qq,t+1}$ is generated from the Uniform distribution of length 0.4 with mean $v_{qq,t}$. The candidate value $v_{qq,t+1}$

is accepted with probability $\min\{1, g(v_{qq,t+1})/g(v_{qq,t})\}$, where

$$\begin{aligned} g(y) &\propto y^{-\sum_{i=1}^{n} w_i m_i} \\ &\quad \times \exp\left[-(1/2)\textstyle\sum_{i=1}^{n} w_i \sum_{k=1}^{2} \{\widetilde{W}_{ik} - (X_{i1k}^{\mathrm{T}}\beta_1, ..., X_{i,19,k}^{\mathrm{T}}\beta_{19})^{\mathrm{T}} - \widetilde{U}_i\}^{\mathrm{T}} \Sigma_\epsilon^{-1} (\bullet)\right]. \end{aligned}$$

- non-diagonal free parameters v_{pq}

  Proposal: The current value is $v_{pq,t}$. The candidate value $v_{pq,t+1}$ is generated from the Uniform distribution of length 0.4 with mean $v_{pq,t}$. The candidate value is accepted with probability $\min\{1, g(v_{pq,t+1})/g(v_{pq,t})\}$, where

$$g(y) \propto \exp\left[-(1/2)\textstyle\sum_{i=1}^{n} w_i \sum_{k=1}^{2} \{\widetilde{W}_{ik} - (X_{i1k}^{\mathrm{T}}\beta_1, ..., X_{i,19,k}^{\mathrm{T}}\beta_{19})^{\mathrm{T}} - \widetilde{U}_i\}^{\mathrm{T}} \Sigma_\epsilon^{-1} (\bullet)\right].$$

17. Complete Conditionals for Σ_u

The dimension of the covariance matrices is $J = 19$. By inspection, the complete conditional for Σ_u is

$$[\Sigma_u|\text{rest}] = \text{IW}\{(m_u - J - 1)\Sigma_{u,\text{prior}} + \textstyle\sum_{i=1}^{n} w_i \widetilde{U}_i \widetilde{U}_i^{\mathrm{T}}, n + m_u\}$$

where here IW = the Inverse-Wishart distribution. The density of $\text{IW}(\Omega, m)$ for a $J \times J$ random variable is

$$\text{IW}(\Omega, m) = f(Q|\Omega, m) \propto |Q|^{-(m+J+1)/2} \exp\{-(1/2)\text{trace}(\Omega Q^{-1})\}.$$

This has expectation $\Omega/(m - J - 1)$.

18. Complete Conditionals for β

Let the elements of Σ_ϵ^{-1} be $\sigma_\epsilon^{j\ell}$. For any j, except for irrelevant constants,

$$\log[\beta_j|\text{rest}] = -(1/2)(\beta_j - \beta_{j,\text{prior}})^{\mathrm{T}} \Omega_{\beta,j}^{-1} (\beta_j - \beta_{j,\text{prior}})$$

$$
\begin{aligned}
& -(1/2)\textstyle\sum_{i=1}^{n} w_i \sum_{k=1}^{2} (W_{ijk} - X_{ijk}^{\mathrm{T}}\beta_j - U_{ij})^2 \sigma_\epsilon^{jj} \\
& -\textstyle\sum_{i=1}^{n} w_i \sum_{k=1}^{2} \sum_{\ell \neq j} \sigma_\epsilon^{j\ell} (W_{ijk} - X_{ijk}^{\mathrm{T}}\beta_j - U_{ij})(W_{i\ell k} - X_{i\ell k}^{\mathrm{T}}\beta_\ell - U_{i\ell}) \\
= \; & \mathcal{C}_1^{\mathrm{T}}\beta_j - (1/2)\beta_j^{\mathrm{T}}\mathcal{C}_2^{-1}\beta_j,
\end{aligned}
$$

which implies $[\beta_j|\text{rest}] = \text{Normal}(\mathcal{C}_2\mathcal{C}_1, \mathcal{C}_2)$, where

$$
\begin{aligned}
\mathcal{C}_2 &= \left(\Omega_{\beta,j}^{-1} + \textstyle\sum_{i=1}^{n} w_i \sigma_\epsilon^{jj} \sum_{k=1}^{2} X_{ijk} X_{ijk}^{\mathrm{T}}\right)^{-1}; \\
\mathcal{C}_1 &= \Omega_{\beta,j}^{-1}\beta_{j,\text{prior}} + \textstyle\sum_{i=1}^{n} w_i \sum_{k=1}^{2} \sigma_\epsilon^{jj} X_{ijk}(W_{ijk} - U_{ij}) \\
&\quad + \textstyle\sum_{i=1}^{n} w_i \sum_{k=1}^{2} \sum_{\ell \neq j} \sigma_\epsilon^{j\ell}(W_{i\ell k} - X_{i\ell k}^{\mathrm{T}}\beta_\ell - U_{i\ell}) X_{ijk}.
\end{aligned}
$$

19. Complete Conditionals for $\widetilde{U}_i$

The NHANES 2001-2004 weights are integers, representing the number of children that each sampled child represents. Thus, as described therein, the loglikelihood in Appendix 15 could also be rewritten equivalently by developing w_i pseudo-children, each with the same observed data values. It thus does not make sense to use the weights to generate an individual $\widetilde{U}_i$. Instead, as described in Appendix 15, for computational convenience for generating a $\widetilde{U}_i$ to represent w_i children, we set the weight for that child temporarily = 1.0. Then, except for irrelevant constants,

$$
\begin{aligned}
\log[\widetilde{U}_i|\text{rest}] &= -(1/2) w_i \widetilde{U}_i^{\mathrm{T}} \Sigma_u^{-1} \widetilde{U}_i \\
&\quad -(1/2) w_i \textstyle\sum_{k=1}^{2} \{\widetilde{W}_{ik} - (X_{i1k}^{\mathrm{T}}\beta_1, ..., X_{i,19,k}^{\mathrm{T}}\beta_{19})^{\mathrm{T}} - \widetilde{U}_i\}^{\mathrm{T}} \Sigma_\epsilon^{-1} \\
&\qquad\qquad \times \{\widetilde{W}_{ik} - (X_{i1k}^{\mathrm{T}}\beta_1, ..., X_{i,19,k}^{\mathrm{T}}\beta_{19})^{\mathrm{T}} - \widetilde{U}_i\} \\
&= \mathcal{C}_1^{\mathrm{T}}\widetilde{U}_i - (1/2)\widetilde{U}_i^{\mathrm{T}}\mathcal{C}_2^{-1}\widetilde{U}_i.
\end{aligned}
$$

Remembering that for purposes of this section we are setting $w_i = 1.0$, this implies that $[\widetilde{U}_i|\text{rest}] = \text{Normal}(\mathcal{C}_2\mathcal{C}_1, \mathcal{C}_2)$, where

$$
\mathcal{C}_2 = (\Sigma_u^{-1} + m_i \Sigma_\epsilon^{-1})^{-1};
$$

$$\mathcal{C}_1 \;=\; \textstyle\sum_{k=1}^{2}\Sigma_\epsilon^{-1}\{\widetilde{W}_{ik} - (X_{i1k}^{\mathrm{T}}\beta_1, ..., X_{i,19,k}^{\mathrm{T}}\beta_{19})^{\mathrm{T}}\}.$$

20. Complete Conditional for $W_{i\ell k}$, $\ell = 1, 3, 5, 7, 9, 11$

Here we do the complete conditional for $W_{i\ell k}$ with $\ell = 1, 3, 5, 7, 9, 11$. Except for irrelevant constants,

$$\begin{aligned}
\log\left[W_{i\ell k}|\text{rest}\right] &= \log\{Q_{i\ell k}I(W_{i\ell k} > 0) + (1 - Q_{i\ell k})I(W_{i\ell k} < 0)\} \\
&\quad -(1/2)w_i(W_{i1k} - X_{i1k}^{\mathrm{T}}\beta_1 - U_{i1}, ..., W_{i,19,k} - X_{i,19,k}^{\mathrm{T}}\beta_{19} - U_{i,19})\Sigma_\epsilon^{-1}(\bullet)^{\mathrm{T}} \\
&= \log\{Q_{i\ell k}I(W_{i\ell k} > 0) + (1 - Q_{i\ell k})I(W_{i\ell k} < 0)\} \\
&\quad -(1/2)w_i\sigma_\epsilon^{\ell\ell}(W_{i\ell k} - X_{i\ell k}^{\mathrm{T}}\beta_\ell - U_{i\ell})^2 \\
&\quad -w_i\textstyle\sum_{j\neq\ell}\sigma_\epsilon^{\ell j}(W_{i\ell k} - X_{i\ell k}^{\mathrm{T}}\beta_\ell - U_{i\ell})(W_{ijk} - X_{ijk}^{\mathrm{T}}\beta_j - U_{ij}) \\
&= \log\{Q_{i\ell k}I(W_{i\ell k} > 0) + (1 - Q_{i\ell k})I(W_{i\ell k} < 0)\} + \mathcal{C}_1 W_{i\ell k} - (1/2)W_{i\ell k}^2\mathcal{C}_2^{-1},
\end{aligned}$$

where, using the convention of Appendix 19,

$$\begin{aligned}
\mathcal{C}_2 &= 1/(\sigma_\epsilon^{\ell\ell}) \\
\mathcal{C}_1 &= \sigma_\epsilon^{\ell\ell}(X_{i\ell k}^{\mathrm{T}}\beta_\ell + U_{i\ell}) - \textstyle\sum_{j\neq\ell}\sigma_\epsilon^{\ell j}(W_{ijk} - X_{ijk}^{\mathrm{T}}\beta_j - U_{ij}).
\end{aligned}$$

If we use the notation $\mathrm{TN}_+(\mu, \sigma, c)$ for a normal random variable with mean μ and standard deviation σ that is truncated from the left at c, and similarly use $\mathrm{TN}_-(\mu, \sigma, c)$ when truncation is from the right at c, then it follows that with $\mu = \mathcal{C}_2\mathcal{C}_1$ and $\sigma = \mathcal{C}_2^{1/2}$,

$$\begin{aligned}
\left[W_{i\ell k}|\text{rest}\right] &= Q_{i\ell k}\mathrm{TN}_+(\mu, \sigma, 0) + (1 - Q_{i\ell k})\mathrm{TN}_-(\mu, \sigma, 0) \\
&= \mu + Q_{i\ell k}\mathrm{TN}_+(0, \sigma, -\mu) + (1 - Q_{i\ell k})\mathrm{TN}_-(0, \sigma, -\mu) \\
&= \mu + Q_{i\ell k}\mathrm{TN}_+(0, \sigma, -\mu) - (1 - Q_{i\ell k})\mathrm{TN}_+(0, \sigma, \mu) \\
&= \mu + \sigma\{Q_{i\ell k}\mathrm{TN}_+(0, 1, -\mu/\sigma) - (1 - Q_{i\ell k})\mathrm{TN}_+(0, 1, \mu/\sigma)\}.
\end{aligned}$$

Generating $\mathrm{TN}_+(0, 1, c)$ is easy: if $c < 0$, simply do rejection sampling of a Normal$(0, 1)$

until you get one that is $> c$. If $c > 0$, there is an adaptive rejection scheme (Robert, 1995).

21. Complete Conditionals for W_{i2k}, W_{i4k}, W_{i6k}, W_{i8k}, $W_{i,10,k}$ and $W_{i,12,k}$ When Not Observed

For $p = 2, 4, 6, 8, 10, 12$, the variable W_{ipk} is not observed when $Q_{i,p-1,k} = 0$, or, equivalently, when $W_{i,p-1,k} < 0$. Except for irrelevant constants,

$$\begin{aligned} \log\left[W_{ipk}|\text{rest}\right] &= -(1/2)w_i \textstyle\sum_j \sum_\ell \sigma_\epsilon^{j\ell}(W_{ijk} - X_{ijk}^{\mathrm{T}}\beta_j - U_{ij})(W_{i\ell k} - X_{i\ell k}^{\mathrm{T}}\beta_\ell - U_{i\ell}) \\ &= -(1/2)W_{ipk}^2\mathcal{C}_2^{-1} + \mathcal{C}_1 W_{ipk}, \end{aligned}$$

where, using the convention of Appendix 19,

$$\begin{aligned} \mathcal{C}_2 &= 1/(\sigma_\epsilon^{pp}); \\ \mathcal{C}_1 &= \sigma_\epsilon^{pp}(X_{ipk}^{\mathrm{T}}\beta_p + U_{ip}) - \textstyle\sum_{\ell \neq p} \sigma_\epsilon^{p\ell}(W_{i\ell k} - X_{i\ell k}^{\mathrm{T}}\beta_\ell - U_{i\ell}). \end{aligned}$$

Therefore,

$$[W_{ipk}|\text{rest}] = Q_{ipk}Q_{i,p-1,k} + (1 - Q_{i,p-1,k})\text{Normal}(\mathcal{C}_2\mathcal{C}_1, \mathcal{C}_2).$$

22. Usual Intake, Standardization and Transformation

Here we present detailed formulas for functions defined in Appendix 4. When $\lambda = 0$, the back-transformation is

$$\begin{aligned} g_{\text{tr}}^{-1}(z, 0) &= \exp\left\{\mu(0) + \sigma(0)z/\sqrt{2}\right\}; \\ \partial^2 g_{\text{tr}}^{-1}(z, 0)/\partial z^2 &= \frac{\sigma^2(0)}{2} g_{\text{tr}}^{-1}(z, 0). \end{aligned}$$

When $\lambda \neq 0$, the back-transformation is

$$g_{\text{tr}}^{-1}(z, \lambda) = \left[1 + \lambda\left\{\mu(\lambda) + \sigma(\lambda)z/\sqrt{2}\right\}\right]^{1/\lambda};$$

$$\partial^2 g_{\text{tr}}^{-1}(z,\lambda)/\partial z^2 \;=\; \frac{\sigma^2(\lambda)}{2}(1-\lambda)\left[1+\lambda\left\{\mu(\lambda)+\sigma(\lambda)z/\sqrt{2}\right\}\right]^{-2+1/\lambda}.$$

22. Transformation Estimation

As part of an earlier project (Freedman et al., 2009), we estimated the transformations for one food/nutrient at a time using the method of Kipnis et al. (2009), both for the data and also for each BRR weighted data set. To facilitate comparison with the one food/nutrient at a time analysis, in our analysis of all HEI-2005 components, we used these transformations as well. Of course, our methods can be generalized to allow for estimation of the transformations as well. By allowing a different transformation for each BRR weighted data set, we have captured the variation due to estimation of the transformations.

VITA

Saijuan Zhang received her B.S. in Mathematics and Applied Mathematics from Southeast University, China in June 2003 and M.A. in Mathematics from the University of Oklahoma in Dec 2005. She entered the Department of Statistics at Texas A&M University in August 2006 and received her Ph.D. in Dec 2010. Her research interests include Measurement Error Models, Bayesian Methods, Multivariate Data Analysis, Bioinformatics, Data Mining/Machine Learning, Semi-parametric Methods. She is a student member of Institute of Mathematical Statistics (IMS) and International Society of Bayesian Analysis (ISBA).

Ms. Zhang may be reached at Department of Statistics at Texas A&M University, 3143 TAMU, College Station, TX 77843. Her email address is sjzhang@stat.tamu.edu

The typist for this dissertation was Saijuan Zhang.