NTNU
Norwegian University of
Science and Technology

# Statistical Modeling and Analysis of Repeated Measures, using the Linear Mixed Effects Model.

**Eirin Tangen Østgård**

Master of Science in Physics and Mathematics
Submission date: June 2011
Supervisor: Mette Langaas, MATH

Norwegian University of Science and Technology
Department of Mathematical Sciences

# Problem Description

The aim of this master thesis is to present and discuss the linear mixed effects model for analyzing repeated measures data.

We will use a complex repeated measures data set from a diet intervention study to illustrate model choice for the linear mixed effects model. We will also show in detail how the model is fitted and how to interpret the results from the data analysis. These analysis are preformed in the frequentist approach.

Several statistical aspects of the linear mixed effects model will be studied, including estimation of contrasts, the structure of the implied marginal variance-covariance matrix, the intraclass correlation and using integrated nested Laplace approximations (INLA) to fit a linear mixed effects model.

Assignment given: 14. January 2011
Supervisor: Mette Langaas, MATH

# Preface

This thesis concludes my Master of Science in Applied Physics and Mathematics, at the Norwegian University of Science and Technology (NTNU). The work was carried out at the Department of Mathematical Sciences.

I would like to thank my supervisor Associate Professor Mette Langaas for motivation and excellent guidance.

Trondheim, June 2011

# Abstract

Our main objective for this thesis is to present and discuss the linear mixed effects model and, in particular, the different possible covariance structures for the random effects and the residuals. The linear mixed effects model is widely used in biology and medical research.

We use data from the diet intervention study, Arbo, Brattbakk, Langaas, Kuiper, Lindberg, Kulseng and Johansen (2010), where the aim was to investigate the difference between a diet rich in carbohydrates and a diet rich in fat and protein. Data from 32 participants were available. A series of biomarkers were measured before and after both diets, giving repeated measurements from each participant across time and diet.

We have studied different linear mixed effects models varying in covariance structure for the random effects and the residuals. Further, we have focused on a thorough treatment of statistical contrasts. The contrasts of interest in this study are estimates of the effect of the two diets and the difference in effect between the two diets, and is especially relevant to biologists and medical researchers. Statistically, there is no common agreement on how degrees of freedom should be calculated when testing contrasts. We will show that using different parameter coding for a between-subject factor in the same model, yield different results.

The linear mixed effects model allows complex structures in correlated data to be modeled, and so it is important to look at the implied marginal variance-covariance matrix to understand the structure. We have calculated the empirical variance-covariance matrix of the data, and compared it to the estimated implied marginal variance-covariance matrix, in an attempt to get a more thorough understanding of the covariance structures for the random effects and the residuals.

The estimated implied marginal variance-covariance matrix have also been used to estimate the intraclass correlations.

Finally, we have fitted the linear mixed effects model using the Bayesian approach, integrated nested Laplace approximations (INLA), and compared the results to the results of the frequentist approach.

# Contents

# Chapter 1

# Introduction

In biology and medical analysis, we often meet data sets in which the response variable is measured more than once for each subject across levels of one or more factors, referred to as repeated measures. We can not analyze these data with linear regression, because the residuals from each subject are correlated. A popular regression model for analyzing repeated measures is the linear mixed effects model, which combine both fixed and random effects on a linear scale. Since we expect the subjects to vary independently in the random effect, we allow the observations within a subject to be correlated.

In Chapter 2 we will define the statistical methods used when fitting a linear mixed effects model using the frequentist approach. In Chapter 3 we will present the diet intervention study, Arbo et al. (2010), and use the frequentist approach to fit linear mixed effects models. In particular we will take a closer look at four different forms of the linear mixed effects model, varying in covariance structures for both random effects and residuals. In Chapter 4 we will calculate and discuss contrasts. In Chapter 5 we will study the implied marginal variance-covariance matrix associated with the four fitted linear mixed effects models, in order to get a deeper understanding of how the covariance in the data are structured in the random effects and the residuals. We will also estimate the empirical variance-covariance matrix directly from data and compare it to the implied marginal variance-covariance matrix associated with the four fitted linear mixed effects models. In Chapter 6 we will define a version of the intraclass correlation for LMEs, calculated by using the estimated implied marginal variance-covariance matrix. In Chapter 7 we will use the Bayesian approach, integrated nested Laplace approximations (INLA), to fit the four models and compare the results to the results from the frequentist approach. Finally, in Chapter 8 we will discuss our findings and conclude.

All statistical analysis and variable construction in this thesis were done using the statistical software *R*, R Development Core Team (2010). The packages used were *nlme* by Pinheiro, Bates, DebRoy, Sarkar and R Development Core Team (2010) for model construction, *gmodels* by Warnes (2011) for contrast estimation, and *inla* by Rue and Martino (2009) for Bayesian model construction.

# Chapter 2

# Method

The presentation of the methods used in this chapter is based on Chapter 2 of West, Welch and Galecki (2007) and Chapter 2 of Pinheiro and Bates (2000).

## 2.1 Types, structures and levels of data

The linear mixed effects model, LME, combine both fixed and random effects on a linear scale. Fixed effects are parameters associated with an entire population or with certain levels of factors. Random effects are associated with subjects, or clusters, drawn at random from the population. Because we expect the subjects, or clusters, to vary independently, we will have correlated observations within a subject, or cluster.

Linear mixed effects models are primarily used to describe relationships between a response variable and one or more explanatory variables or factors, for the following types of data:

- **Clustered data**
  Data in which the response variable is measured once for each subject, and the subjects are nested within clusters. An example is measures for students, which are nested within different school classes.

- **Longitudinal data**
  Data in which the response variable is measured repeatedly through time for each subject.

- **Clustered longitudinal data**
  Data in which the response variable is measured repeatedly through time for each subject, and the subjects are nested within clusters.

- **Repeated measures**
  Data in which the response variable is measured more than once for each subject across levels of one or more factors.

According to West et al. (2007) we can also think of these data as multilevel data sets. The concept of "levels" of data is based on ideas from the hierarchical linear modeling (HLM). Level 1 denotes the observations at the most detailed level of the data, level 2 represent the next level of the hierarchy and so forth. The multiple levels of the four data types can be seen in Table 2.1. We stop at level 3, but clustered data may have additional levels.

| Type of data: | Level 1 | Level 2 | Level 3 |
|---|---|---|---|
| Clustered<br>(Two-level) | Subject<br>*Student* | Cluster of units<br>*Class* | |
| Clustered<br>(Three-level) | Subject<br>*Student* | Cluster of units<br>*Class* | Cluster of cluster<br>*School* |
| Longitudinal | Longitudinal measure<br>*Height*<br>*(At age 1,2,3 and 4)* | Subject<br>*Child* | |
| Clustered<br>longitudinal | Longitudinal measure<br>*Height*<br>*(At age 1,2,3 and 4)* | Subject<br>*Child* | Cluster of unit<br>*Family* |
| Repeated<br>measures | Repeated measure<br>*Insulin*<br>*(Measured for two diets)* | Subject<br>*Person* | |

Table 2.1: Multilevel data sets, with *examples*.

In this thesis we will only consider two-level repeated measures data.

## 2.2 Nested vs. crossed factors

There are two types of both fixed and random effects, called nested and crossed factors. A nested factor is a factor in which one level only can be measured within a single level of another factor and not across multiple levels. Then the level of the first factor are said to be nested within levels of the second factor. A crossed factor is a factor in which one level can be measured across multiple levels of another factor.

In this thesis we will only consider data with crossed fixed and random effects.

## 2.3 Specification of the linear mixed effects model

We will consider a two-level repeated measures data set with crossed fixed and random effects, where level 1 represents the repeated measurements and level 2 represents the subjects. The two-level linear mixed effects model is defined as

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\varepsilon}_i, \text{ for } i = 1, ..., m, \tag{2.1}$$

where $\mathbf{Y}_i$ is a vector of continuous responses for the $i$th subject defined by

$$\mathbf{Y}_i = \begin{bmatrix} Y_{1i} \\ Y_{2i} \\ \vdots \\ Y_{n_i i} \end{bmatrix}.$$

Note that $n_i$ is dependent on $i$, hence the number of observations for each subject may differ. We have $m$ subject, in total $n = \sum_i^m n_i$ observations.

The fixed effect design matrix, $\mathbf{X}_i$, is a $n_i \times p$ matrix, which represents $p$ covariates corresponding to the fixed effects for each observation of the $i$th subject. The fixed effect design matrix is defined as

$$\mathbf{X}_i = \begin{bmatrix} x_{1i}^{(1)} & x_{1i}^{(2)} & \cdots & x_{1i}^{(p)} \\ x_{2i}^{(1)} & x_{2i}^{(2)} & \cdots & x_{2i}^{(p)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n_i i}^{(1)} & x_{n_i i}^{(2)} & \cdots & x_{n_i i}^{(p)} \end{bmatrix}.$$

The first column of the design matrix is often equal to 1 for all observations to include an intercept term in the model.

The fixed effects vector, $\boldsymbol{\beta}$, is a vector consisting of $p$ unknown regression coefficients associated with the covariates from the design matrix $\mathbf{X}_i$, and is defined as

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}.$$

The random effect design matrix, $\mathbf{Z}_i$, is a $n_i \times q$ matrix, which represents $q$ covariates corresponding to the random effects for each observation of the $i$th subject. The random effect design matrix is defined as

$$\mathbf{Z}_i = \begin{bmatrix} z_{1i}^{(1)} & z_{1i}^{(2)} & \cdots & z_{1i}^{(q)} \\ z_{2i}^{(1)} & z_{2i}^{(2)} & \cdots & z_{2i}^{(q)} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n_i i}^{(1)} & z_{n_i i}^{(2)} & \cdots & z_{n_i i}^{(q)} \end{bmatrix}.$$

The random effects are effects that vary randomly across subjects. Hence, it includes the individual differences for the subjects. Covariates with random effect are often represented both in the $\mathbf{X}_i$ matrix and the $\mathbf{Z}_i$ matrix. In the simplest example of the linear mixed effects model, only the intercepts are assumed to vary randomly from subject to subject. Hence, in this case the $\mathbf{Z}_i$ matrix is simply reduced to a vector of $n_i$ 1's.

The random effect vector, $\mathbf{u}_i$, is a vector consisting of $q$ random effects associated with the covariates from the design matrix $\mathbf{Z}_i$, and is defined by

$$\mathbf{u}_i = \begin{bmatrix} u_{1i} \\ u_{2i} \\ \vdots \\ u_{qi} \end{bmatrix}.$$

We assume that the random effect vector, $\mathbf{u}_i$, follows a multivariate normal distribution,

$$\mathbf{u}_i \sim N_q(\mathbf{0}, \mathbf{D}),$$

where the positive definite symmetric covariance matrix $\mathbf{D}$ is defined as

$$\mathbf{D} = \mathrm{Var}(\mathbf{u}_i) = \begin{bmatrix} \mathrm{Var}(u_{1i}) & \mathrm{Cov}(u_{1i}, u_{2i}) & \cdots & \mathrm{Cov}(u_{1i}, u_{qi}) \\ \mathrm{Cov}(u_{1i}, u_{2i}) & \mathrm{Var}(u_{2i}) & \cdots & \mathrm{Cov}(u_{2i}, u_{qi}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(u_{1i}, u_{qi}) & \mathrm{Cov}(u_{2i}, u_{qi}) & \cdots & \mathrm{Var}(u_{qi}) \end{bmatrix}. \tag{2.2}$$

Finally, the residual $\boldsymbol{\varepsilon}_i$ vector is defined by

$$\boldsymbol{\varepsilon}_i = \begin{bmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \vdots \\ \varepsilon_{n_i i} \end{bmatrix},$$

where each element represents the residual associated with each response for the $i$th subject. Unlike the residuals in standard linear models, the residuals associated with repeated observations on the same subject in a linear mixed effects model can be correlated. We assume that the $n_i$ residuals in the $\boldsymbol{\varepsilon}_i$ vector follow a multivariate normal distribution,

$$\boldsymbol{\varepsilon}_i \sim N_{n_i}(\mathbf{0}, \mathbf{R}_i),$$

where the positive definite symmetric covariance matrix $\mathbf{R}_i$ is defined as

$$\mathbf{R}_i = \mathrm{Var}(\boldsymbol{\varepsilon}_i) = \begin{bmatrix} \mathrm{Var}(\varepsilon_{1i}) & \mathrm{Cov}(\varepsilon_{1i}, \varepsilon_{2i}) & \cdots & \mathrm{Cov}(\varepsilon_{1i}, \varepsilon_{n_i i}) \\ \mathrm{Cov}(\varepsilon_{1i}, \varepsilon_{2i}) & \mathrm{Var}(\varepsilon_{2i}) & \cdots & \mathrm{Cov}(\varepsilon_{2i}, \varepsilon_{n_i i}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(\varepsilon_{1i}, \varepsilon_{n_i i}) & \mathrm{Cov}(\varepsilon_{2i}, \varepsilon_{n_i i}) & \cdots & \mathrm{Var}(\varepsilon_{n_i i}) \end{bmatrix}. \tag{2.3}$$

We assume that the vectors of residuals, $\boldsymbol{\varepsilon}_1, ..., \boldsymbol{\varepsilon}_m$, and the random effects, $\mathbf{u}_1, ..., \mathbf{u}_m$, are independent of each other.

## 2.4 Covariance parameters, $\theta$

We now want to introduce the vector of covariance parameters,

$$\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\theta_D} \\ \boldsymbol{\theta_R} \end{bmatrix}, \tag{2.4}$$

which combines all parameters from the covariance matrices $\mathbf{D}$ and $\mathbf{R}_i$, respectively contained in the vectors $\boldsymbol{\theta_D}$ and $\boldsymbol{\theta_R}$. Hence, in order to find the vector, $\boldsymbol{\theta}$, we need to know the covariance structure of the matrices $\mathbf{D}$ and $\mathbf{R}_i$. We will now take a closer look at the most commonly used covariance structures for these matrices.

The two most commonly used structures for the positive definite symmetric covariance matrix $\mathbf{D}$ in Equation (2.2), are the unstructured and the variance components structure. The unstructured $\mathbf{D}$ matrix has no other constraints than being positive definite and symmetric. If the linear mixed effects model have two random effects associated with the $i$th subject, the unstructured $\mathbf{D}$ matrix is given as

$$\mathbf{D} = \mathrm{Var}(\mathbf{u}_i) = \begin{bmatrix} \sigma^2_{u1} & \sigma_{u1,u2} \\ \sigma_{u1,u2} & \sigma^2_{u2} \end{bmatrix}. \tag{2.5}$$

In this case, the vector $\boldsymbol{\theta_D}$ contains three covariance parameters,

$$\boldsymbol{\theta_D} = \begin{bmatrix} \sigma^2_{u1} \\ \sigma_{u1,u2} \\ \sigma^2_{u2} \end{bmatrix}. \tag{2.6}$$

The variance components structure of the covariance matrix $\mathbf{D}$ is also called the diagonal structure. If the linear mixed effects model have two random effects associated with the $i$th subject, the variance components structured $\mathbf{D}$ matrix is given as

$$\mathbf{D} = \mathrm{Var}(\mathbf{u}_i) = \begin{bmatrix} \sigma^2_{u1} & 0 \\ 0 & \sigma^2_{u2} \end{bmatrix}. \tag{2.7}$$

In this case, the vector $\boldsymbol{\theta_D}$ contains two covariance parameters,

$$\boldsymbol{\theta_D} = \begin{bmatrix} \sigma^2_{u1} \\ \sigma^2_{u2} \end{bmatrix}. \tag{2.8}$$

The two most commonly used structures for the positive definite symmetric covariance matrix $\mathbf{R}_i$ in Equation (2.3), are the diagonal and the compound symmetry structure. The diagonal structure of the covariance matrix $\mathbf{R}_i$ is the simplest structure, in which the residuals within one subject are assumed to be uncorrelated and have equal variances. Hence, the diagonal structure of the covariance matrix $\mathbf{R}_i$ is given as

$$\mathbf{R}_i = \mathrm{Var}(\boldsymbol{\varepsilon}_i) = \sigma^2\mathbf{I} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}. \tag{2.9}$$

In this case, the vector $\boldsymbol{\theta_R}$ only contain one covariance parameter,

$$\boldsymbol{\theta_R} = \sigma^2. \tag{2.10}$$

The compound symmetry structure of the covariance matrix $\mathbf{R}_i$, assumes that the residuals within one subject have a constant covariance, $\sigma_1$, and a constant variance, $\sigma^2 + \sigma_1$. Hence, the compound symmetry structure of the covariance matrix $\mathbf{R}_i$, is given as

$$\mathbf{R}_i = \mathrm{Var}(\boldsymbol{\varepsilon}_i) = \begin{bmatrix} \sigma^2 + \sigma_1 & \sigma_1 & \cdots & \sigma_1 \\ \sigma_1 & \sigma^2 + \sigma_1 & \cdots & \sigma_1 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_1 & \sigma_1 & \cdots & \sigma^2 + \sigma_1 \end{bmatrix}. \tag{2.11}$$

In this case, the vector $\boldsymbol{\theta_R}$ contains two covariance parameters,

$$\boldsymbol{\theta_R} = \begin{bmatrix} \sigma^2 \\ \sigma_1 \end{bmatrix}. \tag{2.12}$$

Both the covariance matrices $\mathbf{D}$ and $\mathbf{R}_i$ can also be specified to allow heterogeneous variances for different levels of a specific factor.

## 2.5 The implied marginal model

The linear mixed effects model (2.1) implies the marginal linear model

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i^\star, \tag{2.13}$$

where

$$\boldsymbol{\varepsilon}_i^\star = \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\varepsilon}_i.$$

Hence the $\boldsymbol{\varepsilon}_i^\star$ is normally distributed with expected value

$$\begin{aligned} \mathrm{E}(\boldsymbol{\varepsilon}_i^\star) &= \mathrm{E}(\mathbf{Z}_i\mathbf{u}_i) + \mathrm{E}(\boldsymbol{\varepsilon}_i) \\ &= \mathbf{Z}_i\mathrm{E}(\mathbf{u}_i) + \mathrm{E}(\boldsymbol{\varepsilon}_i) \\ &= \mathbf{Z}_i\mathbf{0} + \mathbf{0} \\ &= \mathbf{0} \end{aligned}$$

and covariance matrix

$$\begin{aligned} \mathrm{Cov}(\boldsymbol{\varepsilon}_i^\star) &= \mathrm{Cov}(\mathbf{Z}_i\mathbf{u}_i) + \mathrm{Cov}(\boldsymbol{\varepsilon}_i) \\ &= \mathbf{Z}_i\mathrm{Cov}(\mathbf{u}_i)\mathbf{Z}_i^T + \mathrm{Cov}(\boldsymbol{\varepsilon}_i) \\ &= \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T + \mathbf{R}_i. \end{aligned}$$

By defining the marginal variance-covariance matrix as

$$\mathbf{V}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T + \mathbf{R}_i, \tag{2.14}$$

we get

$$\boldsymbol{\varepsilon}_i^\star \sim N_{n_i}(\mathbf{0}, \mathbf{V}_i)$$

Hence, the marginal distribution of $\mathbf{Y}_i$ is defined as

$$\mathbf{Y}_i \sim N_{n_i}(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{V}_i). \tag{2.15}$$

## 2.6 Maximum likelihood estimation

The marginal distribution of $\mathbf{Y}_i$, (2.15), is the multivariate normal probability density function

$$f(\mathbf{Y}_i|\boldsymbol{\beta},\boldsymbol{\theta}) = (2\pi)^{-\frac{n_i}{2}} det(\mathbf{V}_i)^{-\frac{1}{2}} \exp(-\frac{1}{2}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})^T \mathbf{V}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})),$$

where $det$ is the determinant and $\mathbf{V}_i$ is given by (2.14).

Hence, given the observed data $\mathbf{Y}_i = \mathbf{y}_i$, the likelihood function contribution for the $i$th subject is

$$L_i(\boldsymbol{\beta},\boldsymbol{\theta}) = (2\pi)^{-\frac{n_i}{2}} det(\mathbf{V}_i)^{-\frac{1}{2}} \exp(-\frac{1}{2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^T \mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})).$$

We have observed $m$ independent subjects and the product of these $m$ likelihood functions, gives us the joint likelihood function

$$\begin{aligned} L(\boldsymbol{\beta},\boldsymbol{\theta}) &= \prod_{i=1}^{m} L_i(\boldsymbol{\beta},\boldsymbol{\theta}) \\ &= \prod_{i=1}^{m} (2\pi)^{-\frac{n_i}{2}} det(\mathbf{V}_i)^{-\frac{1}{2}} \exp(-\frac{1}{2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^T \mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})). \end{aligned}$$

Hence, the log-likelihood function is

$$l(\boldsymbol{\beta},\boldsymbol{\theta}) = -\frac{1}{2} n \ln(2\pi) - \frac{1}{2}\sum_{i=1}^{m} \ln(det(\mathbf{V}_i)) - \frac{1}{2}\sum_{i=1}^{m}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^T \mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}). \quad (2.16)$$

By assuming that $\boldsymbol{\theta}$ is known, the log-likelihood function becomes a function of $\boldsymbol{\beta}$ only. This leads to the maximization of the log-likelihood function (2.16) being equivalent to the minimization of its last term

$$q(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{m}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^T \mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}).$$

By using the method of generalized least squares, we minimize $q(\beta)$ to find $\hat{\boldsymbol{\beta}}$.

$$\frac{\partial q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}} \frac{1}{2} \sum_{i=1}^{m} \mathbf{y}_i^T \mathbf{V}_i^{-1}\mathbf{y}_i - \mathbf{y}_i^T \mathbf{V}_i^{-1}\mathbf{X}_i\boldsymbol{\beta} - \boldsymbol{\beta}^T\mathbf{X}_i^T \mathbf{V}_i^{-1}\mathbf{y}_i + \boldsymbol{\beta}^T\mathbf{X}_i^T \mathbf{V}_i^{-1}\mathbf{X}_i\boldsymbol{\beta} = 0$$

$$\Rightarrow \sum_{i=1}^{m} -\mathbf{X}_i^T \mathbf{V}_i^{-1}\mathbf{y}_i + \mathbf{X}_i^T \mathbf{V}_i^{-1}\mathbf{X}_i\boldsymbol{\beta} = 0$$

$$\Rightarrow \hat{\boldsymbol{\beta}} = (\sum_{i=1}^{m} \mathbf{X}_i^T \mathbf{V}_i^{-1}\mathbf{X}_i)^{-1} \sum_{i=1}^{m} \mathbf{X}_i^T \mathbf{V}_i^{-1}\mathbf{y}_i. \quad (2.17)$$

Since $\hat{\boldsymbol{\beta}}$ can be written as $b^T Y$, it is the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$. Which means that $E[b^T Y] = \boldsymbol{\beta}$ and that it has the smallest variance among all unbiased linear estimators.

Further, to obtain the maximum likelihood estimates of the covariance parameters, $\boldsymbol{\theta}$, we construct a profile log-likelihood function, $l_{ML}(\boldsymbol{\theta})$. This is done by replacing the fixed effects $\boldsymbol{\beta}$ with the best linear unbiased estimator of $\boldsymbol{\beta}$, (2.17). Hence,

$$l_{ML}(\boldsymbol{\theta}) = -\frac{1}{2}n\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{m}\ln(det(\mathbf{V}_i)) - \frac{1}{2}\sum_{i=1}^{m}(\mathbf{r}_i^T\mathbf{V}_i^{-1}\mathbf{r}_i), \tag{2.18}$$

where

$$\mathbf{r}_i = \mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}} = \mathbf{y}_i - \mathbf{X}_i((\sum_{i=1}^{m}\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i)^{-1}\sum_{i=1}^{m}\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{y}_i). \tag{2.19}$$

The maximum likelihood estimates of the covariance parameters $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$, can not be given in closed form. $\hat{\boldsymbol{\theta}}$ can be found by numerical optimization. The *R* function *nlme* by Pinheiro et al. (2010) uses a hybrid approach, where an initial $\boldsymbol{\theta}_0$ is computed, then 25 expectation-maximization iterations in performed to refine the estimate and finally, Newton-Raphson iterations is performed until convergence is obtained.

The expectation-maximization algorithm, EM, is an iterative algorithm for likelihood estimation in models with incomplete data. The iterations are based on regarding the random effects as unobserved data. At iteration $w$, where $w = 1, ..., 25$, we use the current covariance parameter vector $\boldsymbol{\theta}^{(w)}$ to evaluate the distribution of $\boldsymbol{\beta}|\mathbf{y}$ and derive the expectation of the log-likelihood for a new value of $\boldsymbol{\theta}$ given this conditional distribution. Then we maximize this expectation with respect to $\boldsymbol{\theta}$, to produce $\boldsymbol{\theta}^{(w+1)}$.

The Newton-Raphson algorithm is a optimization algorithm which uses a first-order expansion of the score function, which is the gradient of the log-likelihood function, around the current estimate of the covariance parameter vector, $\boldsymbol{\theta}^{(w)}$, to produce the next estimate $\boldsymbol{\theta}^{(w+1)}$.

Now we are ready to calculate $\hat{\boldsymbol{\beta}}$. We insert the estimated $\hat{\boldsymbol{\theta}}$ into $\mathbf{D}$ and $\mathbf{R}_i$ to obtain $\hat{\mathbf{D}}$ and $\hat{\mathbf{R}}_i$. And by inserting these estimates into $\mathbf{V}_i$ in Equation (2.14), we get an estimate of $\mathbf{V}_i$,

$$\hat{\mathbf{V}}_i = \mathbf{Z}_i\hat{\mathbf{D}}\mathbf{Z}_i^T + \hat{\mathbf{R}}_i. \tag{2.20}$$

Further, by replacing $\mathbf{V}_i$ by $\hat{\mathbf{V}}_i$ in the log-likelihood function (2.16), $\boldsymbol{\theta}$ is known (implicit assumed), the maximization of the log-likelihood function is equivalent to the minimization of its last term. Hence, by using the method of weighted least squares we obtain the empirical best linear unbiased estimator (EBLUE) of $\boldsymbol{\beta}$,

$$\hat{\boldsymbol{\beta}} = (\sum_{i=1}^{m}\mathbf{X}_i^T\hat{\mathbf{V}}_i^{-1}\mathbf{X}_i)^{-1}\sum_{i=1}^{m}\mathbf{X}_i^T\hat{\mathbf{V}}_i^{-1}\mathbf{y}_i. \tag{2.21}$$

## 2.7 Restricted maximum likelihood estimation

The REML estimation is an alternative way of estimating the covariance parameters in $\boldsymbol{\theta}$, which is often preferred to ML estimation due to the fact that it produces unbiased estimates of covariance parameters by taking into account the loss of degrees of freedom that results from estimating the linear fixed effects in $\boldsymbol{\beta}$.

The REML log-likelihood function is given by

$$\begin{aligned} l_{REML}(\boldsymbol{\theta}) = &-\frac{1}{2}(n-p)\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{m}\ln(det(\mathbf{V}_i)) \qquad (2.22) \\ &-\frac{1}{2}\sum_{i=1}^{m}(\mathbf{r}_i^T\mathbf{V}_i^{-1}\mathbf{r}_i) - \frac{1}{2}\sum_{i=1}^{m}\ln(det(\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i)), \qquad (2.23) \end{aligned}$$

where $\mathbf{r}_i$ is given by Equation (2.19). Here we observe that the difference between the ML- and the REML log-likelihood function is that the REML subtracts less in the first term, $n-p$ instead of $n$, and an extra term $\frac{1}{2}\sum_{i=1}^{m}\ln(det(\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i))$. The general motivation for using REML is to obtain unbiased estimates of the covariance parameters.

By optimization of this REML log-likelihood function we obtain the REML estimates of the covariance parameters in $\boldsymbol{\theta}$. Once an estimate of the variance-covariance matrix $\mathbf{V}_i$, $\hat{\mathbf{V}}_i$, has been obtained, the REML-based estimates of the fixed effect parameters, $\hat{\boldsymbol{\beta}}$, can be computed by using Equation (5.2) and (2.21) from the ML estimation. Hence, the ML-based and the REML-based estimates of the fixed effect parameters, $\hat{\boldsymbol{\beta}}$, differ due to the fact that the estimate of the variance-covariance matrix, $\hat{\mathbf{V}}_i$, is different.

## 2.8 Likelihood ratio test

The likelihood ratio tests are a class of tests based on comparing the values of likelihood functions for two nested models defining a hypothesis being tested. Such hypotheses can be about both fixed effect parameters or covariance parameters in a linear mixed effects model. In general, the likelihood ratio test requires that both the nested model and the reference model corresponding to a specific hypothesis are fitted to the same subset of data. The likelihood ratio test statistic, is according to West et al. (2007) defined as

$$-2\log(\frac{L_{nested}}{L_{reference}}) = -2\log(L_{nested}) - (-2\log(L_{reference})) \sim \chi^2_{df},$$

where $L_{nested}$ refers to the value of the likelihood function evaluated at the ML or REML estimates of the parameters in the nested model, and $L_{reference}$ refers to the value of the likelihood function in the reference model. Likelihood theory states that under mild regularity conditions the likelihood ratio test statistic asymptotically follows a $\chi^2$ distribution, in which the degrees of freedom, $df$, is obtained by subtracting the number of parameters estimated in the nested model from the number of parameters estimated in the reference model. Hence, if the test statistic is sufficiently large there is evidence against the null hypothesis, which is that the nested model is a better fit for the data than the reference model. Similarly, if the test statistic is sufficiently small there is evidence of the null hypothesis, and the nested model is the best fit for the data.

### 2.8.1 Likelihood ratio test for fixed effect parameters

The likelihood ratio test for fixed effect parameters should be used when the estimated fixed effects have been obtained by ML estimation. In this case the nested model and the reference model have the same set of covariance parameters, but different sets of fixed effect parameters. The likelihood ratio test statistic is in this case defined as

$$-2l_{nested} - (-2l_{reference}) \sim \chi^2_{df}, \tag{2.24}$$

where $l$ is the log-likelihood function, $l_{ML}$, given by Equation (2.18). Hence the test statistic has a $\chi^2$ asymptotic null distribution, with degrees of freedom, $df$, equal to the difference in fixed effect parameters between the two models. Pinheiro and Bates (2000) do not recommend this method for testing fixed effect parameters, due to the fact that p-values calculated might be greater than they should be, referred to as "anticonservative".

### 2.8.2 Likelihood ratio test for covariance parameters

The likelihood ratio test for covariance parameters should be used when the estimated covariance parameters have been obtained by REML estimation. We assume that the nested model and the reference model have the same set of fixed effect parameters, but different sets of covariance parameters. The likelihood ratio test statistic is in this case defined as

$$-2l_{nested} - (-2l_{reference}), \tag{2.25}$$

where $l$ is the log-likelihood function, $l_{REML}$, given by Equation (2.23). The null distribution of the test statistic depends on whether the null hypothesis values for the covariance parameters lie on the boundary of the parameter space for the covariance parameters or not.

The first case is that the covariance parameters satisfying the null hypothesis do not lie on the boundary of the parameter space. In this case the test statistic has a $\chi^2$ asymptotic null distribution, with degrees of freedom, $df$, equal to the difference in number of covariance parameters between the nested model and the reference model.

The second case is when the covariance parameters satisfying the null hypothesis lie on the boundary of the parameter space. This case often arises when we test whether a given random effect should be kept in a model or not. Which is tested by whether the variances and covariances corresponding to the given random effect, are equal to zero or not.

In the case where a model has a single random effect, we might wish to test whether that random effect can be omitted. That is,

$$H_0 : D = 0 \text{ versus } H_1 : D = \sigma^2. \tag{2.26}$$

It has been shown by Verbeke and Molenberghs (2000) (page 69-70) that the likelihood ratio test statistic, (2.25), has a asymptotic null distribution that is a mixture of $\chi_1^2$ and $\chi_0^2$ with equal weights 0.5.

In the case where a model contains two random effects, we might wish to test whether one of them can be omitted. That is,

$$H_0 : D = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & 0 \end{bmatrix} \text{ versus } H_1 : D = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{1,2} & \sigma_2^2 \end{bmatrix}. \tag{2.27}$$

It has been shown by Verbeke and Molenberghs (2000) (page 70) that the the likelihood ratio test statistic, (2.25), has a asymptotic null distribution that is a mixture of $\chi_1^2$ and $\chi_2^2$ with equal weights 0.5.

Hence, we can calculate a p-value for the test statistic as follows:

$$p = 0.5(1 - \chi^2_{test-statistic,1}) + 0.5(1 - \chi^2_{test-statistic,2}), \tag{2.28}$$

where the test statistic is defined in Equation (2.25). If the test statistic is significant on a $\alpha$-level, that is if $p < \alpha$, we reject the null hypothesis and retain the random effect tested in the model.

Since most statistical software procedures capable of fitting linear mixed effects models provide the option of using either ML estimation or REML estimation, we can choose to use the estimation method suitable for the hypothesis we want to test.

## 2.9 Conditional tests for fixed effect parameters

The conditional tests for fixed effect parameters include the t-test and the F-test, which are both conditioned on the estimates of the covariance parameters, $\widehat{\boldsymbol{\theta}}$.

### 2.9.1 The conditional t-test

The conditional t-test for fixed effect parameters, tests the hypothesis given as

$$H_0\colon\ \beta = 0 \text{ vs. } H_1\colon\ \beta \neq 0.$$

The corresponding $t$-statistic, or t value, is defined by

$$t = \frac{\hat{\beta}}{se(\hat{\beta})}. \tag{2.29}$$

The $t$-statistic follows an approximate $t$ distribution, with degrees of freedom determined by the grouping level at which the term is estimated.

Using the *lme* function of Pinheiro et al. (2010), the conditional t-tests are implemented in the *summary* method. Here the significance of each fixed effect parameter are conditional on all other fixed effects in the model.

### 2.9.2 The conditional F-test

The conditional F-test for fixed effect parameters, tests the hypothesis given as

$$H_0\text{: } \mathbf{C}\boldsymbol{\beta} = 0 \text{ vs. } H_1\text{: } \mathbf{C}\boldsymbol{\beta} \neq 0,$$

where $\mathbf{C}$ is a known matrix. The $F$-statistic is defined by

$$F = \frac{\hat{\boldsymbol{\beta}}\mathbf{C}^T(\mathbf{C}(\sum_i^m \mathbf{X}_i^T \hat{\mathbf{V}}_i^{-1}\mathbf{X}_i)^{-1}\mathbf{C}^T)^{-1}\mathbf{C}\hat{\boldsymbol{\beta}}}{rank(\mathbf{C})}. \tag{2.30}$$

The $F$-statistic follows an approximate $F$ distribution, with numerator degrees of freedom equal to the rank of the matrix $\mathbf{C}$ and denominator degrees of freedom determined by the grouping level at which the term is estimated.

The conditional F-tests and t-tests is according to Pinheiro and Bates (2000) preferred for assessing the significance of fixed effect parameters, due to the fact that p-values are more realistic than the p-values from the likelihood ratio test, (2.24).

The conditional F-tests are implemented in the ANOVA method of R Development Core Team (2010). It is a Type I F-test, which means that the fixed effects are tested sequentially. That is, the significance of each fixed effect is conditional on the fixed effects listed prior in the model.

### 2.9.3 Denominator degrees of freedom

The conditional t-test and F-test for fixed effect parameters both require denominator degrees of freedom, given by

$$denDF_i = m_i - (m_{i-1} + p_i), \tag{2.31}$$

where $i$ is the level at which the term is estimated. A term is estimated at level $i$ if it is *inner* to the $(i-1)$th grouping factor and *outer* to the $i$th grouping factor. If a term is *inner* to all $Q$ grouping factors it is at $(Q+1)$st level. If a term is *inner*, its value can change within a given level of the grouping factor and if a term is *outer* its value can not change within a given level of the grouping factor.

More specifically, $m_i$ is the total number of groups in the $i$th grouping factor, where $m_0 = 1$ if intercept is included in the model, $m_0 = 0$ otherwise and $m_{Q+1} = n$ which is the total number of observations. Finally, the sum of numerator degrees of freedom for terms estimated at level $i$, is given as $p_i$.

Observe that the denominator degrees of freedom is not influenced by the structure of the covariance matrices, $\mathbf{D}$ and $\mathbf{R}_i$.

It is important for us to specify that the term "level", is used in two different ways. In Table 2.1, we use the term to denote levels of data. When calculating denominator degrees of freedom however, the term is used to denote the level at which a fixed effect term is estimated.

In this thesis we have considered a two-level repeated measures data set, where level 1 represents the repeated measures made on the same subject and level 2 represents the subjects. Hence, we have one grouping factor, the subject, and so $Q = 1$. We have chosen to include intercept in the model and so $m_0 = 1$. The total number of groups in the 1st grouping factor, $m_1$, is the number of subjects and $m_2 = n$ is the total number of observations. The sum of numerator degrees of freedom for terms estimated at level 1, $p_1$, is the sum of numerator degrees of freedom for the between-subject factors. Finally, the sum of numerator degrees of freedom for terms estimated at level 2, $p_2$, is the sum of numerator degrees of freedom for the within-subject factors.

## 2.10 The top-down strategy

There are several ways of fitting a linear mixed effects model. The aim of model selection is to find the simplest model with the best fit for the data. We will in this thesis use the top-down strategy, as performed in Chapter 5 of West et al. (2007).

The top-down strategy starts with a model which includes the maximum number of fixed effects, called the model with the loaded mean structure. We select a structure for the random effects in the model by performing REML-based likelihood ratio test for the associated covariance parameters. Further, we select a covariance structure for the residuals in the model by performing REML-based likelihood ratio test, using the ANOVA method of R Development Core Team (2010).

Finally, we reduce the model by performing an type I F-test using the ANOVA method of R Development Core Team (2010), to determine whether each of the fixed effect parameters should be included in the model. Since the type I F-test tests the fixed effect sequentially, we iteratively test the fixed effects and remove the first term in the sequence of fixed effects which is not significant on a $\alpha = 0.05$ significance level. We do not allow interaction terms to be included unless all factors in that interaction term is present in the model. This is done until we are left with only significant fixed effects. Notice that the denominator degrees of freedom, given in Equation (2.31), changes as the number of fixed effect factors are excluded from the model.

## 2.11 Information criteria

There are two types of information criteria often used to choose the best fitted model for the data, the Akaike information criteria and the Bayes information criteria.

The Akaike information criteria, $AIC$, is defined by

$$AIC = -2l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}) + 2p \tag{2.32}$$

where $l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$ can be either the ML- or REML log-likelihood function and $p$ represents the total number of parameters, both the fixed and random effects, being estimated in the model. The model with the lowest $AIC$ value is assumed to be the best fit for the data.

The Bayes information criteria, $BIC$, is defined by

$$BIC = -2l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}) + p \ln(n) \tag{2.33}$$

where $l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$ is the ML log-likelihood function, $p$ represents the total number of parameters, both the fixed and random effects, being estimated in the model and $n$ is the total number of observations used in estimation of the model. That is, $n = \sum_{i=1}^{m} n_i$.

According to Pinheiro and Bates (2000) we can calculate the REML version of the BIC by simply using the REML log-likelihood function and replacing $ln(n)$ by $ln(n - p_{fixed})$, where $p_{fixed}$ is the number of estimated fixed effect parameters in the model, in Equation (2.33).

In other words, the $BIC$ applies a greater penalty for models with more parameters than the $AIC$. And as for the $AIC$, the model with the lowest $BIC$ value is assumed to be the best fit for the data.

According to West et al. (2007) there is no information criterion which stands apart as the best criterion to be used when selecting linear mixed effects models.

## 2.12 Diagnostics

After a linear mixed effects model is fitted it is important to check whether the underlying distributional assumptions for the random effects and the residuals appear valid for the data. Diagnostic methods for linear models are well established, but diagnostics for linear mixed effects models are however more difficult to perform and interpret due to the complexity of the model. The most useful method for diagnostics, are according to Pinheiro and Bates (2000), based on plots of the residuals, the fitted values and the estimated random effects.

In this thesis we will do all diagnostic by using the functions *qqnorm.lme* and *plot.lme* in Pinheiro et al. (2010). Here the standardized, or Pearson residuals, defined as the raw residuals divided by the estimated corresponding standard deviation, are used.

# Chapter 3

# The diet intervention study

In this chapter we will illustrate the methods in Chapter 2, using the *nlme* package in *R* by Pinheiro et al. (2010).

In the diet intervention study, Arbo et al. (2010), the aim was to examine differences between a diet rich in carbohydrates and a diet rich in fat and protein. The participants volunteered to join the study, but only those who met certain criteria where asked to participate. The participants had to be between $18-30$ years of age, with BMI between $24,5-27,5$ and they had to pass a health check which checked if their biomarkers where inside reference areas.

Out of the 56 persons who met the requirements, 32 completed the study. The participants where given two fluid diets with different macronutrient composition. The high-carbohydrate diet is referred to as diet A, and the moderate-carbohydrate diet is referred to as diet B. The different nutrition compositions of diet A and diet B can be seen in Table 3.1, where $E\%$ is the percentage of the individual total energy intake.

| | Carbohydrates | Fat | Protein |
|---|---:|---:|---:|
| Diet A | $65E\%$ | $20E\%$ | $15E\%$ |
| Diet B | $27E\%$ | $43\ E\%$ | $30E\%$ |

Table 3.1: Composition of diet A and diet B in the diet intervention study.

All 32 participants were assigned to start on either diet A or diet B by randomization, controlling for gender, age and waist circumference. Both diets were given for six days, with a wash-out period of eight days between the two diets. Blood samples were taken before and after each diet, hence at day zero and day six for both diet A and diet B, yielding four blood samples for each individual. From these blood samples Arbo et al. (2010) investigated 32 biomarkers.

Hence, each participant in the diet intervention study have following four measurements for each biomarker:

- $A0$: Measurement of diet $A$ at day 0.
- $AB$: Measurement of diet $A$ at day 6.

- $B0$: Measurement of diet $B$ at day 0.
- $B6$: Measurement of diet $B$ at day 6.

## 3.1 Fitting linear mixed effects models

In analyzing these biomarkers we will exemplify different linear mixed effect models, with varying complexity. In order to choose which biomarkers to have a closer look at, we have to find the model with the best fit for all 32 biomarkers. Since the varying complexity of interest lies in the covariance matrix of the random effects, $D$ and in the covariance matrix of the residuals $R_i$, we are only interested in the first steps of the top-down strategy, fitting a model with a loaded mean structure, selecting a structure for the random effects and selecting a covariance structure for the residuals in the model.

In order to decide if the responses should be analyzed on a original or logarithmic scale, we first modeled all biomarkers by the simplest model and examined $QQ$-plots.

### Step 1

The linear mixed effect model with a loaded mean is defined by

$$\mathbf{Y}_i = \boldsymbol{\beta}\mathbf{X}_i + \mathbf{u}_i\mathbf{Z}_i + \boldsymbol{\varepsilon}_i, \tag{3.1}$$

where the loaded fixed effect vector is given by

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{intercept} \\ \beta_{sex} \\ \beta_{time} \\ \beta_{diet} \\ \beta_{sex:time} \\ \beta_{sex:diet} \\ \beta_{time:diet} \\ \beta_{sex:time:diet} \end{bmatrix}, \tag{3.2}$$

$\mathbf{X}_i$ is the design matrix for subject $i$, $\mathbf{Z}_i$ is a vector of $n_i$ ones, $\mathbf{u}_i = u_{int,i} \sim N(\mathbf{0}, \sigma^2_{int})$, $\boldsymbol{\varepsilon}_i \sim N_{n_i}(\mathbf{0}, \boldsymbol{R_i})$ and $R_i$ is a $n_i \times n_i$ covariance matrix, with $\sigma^2$ on the diagonal.

### Step 2

We then include a second subject-specific random effect, $u_{2,i}$, to the model with the loaded mean structure. Hence, the new model is similar to the loaded model, in Equation (3.1), but where $\mathbf{Z}_i$ is a $n_i \times 2$ matrix with ones i the first column and the design vector corresponding to the random effect, $u_{2,i}$, in the second column. That is, $\mathbf{u}_i = [u_{1,i}u_{2,i}]^T \sim N_2(\mathbf{0}, \mathbf{D})$ and

$$\mathbf{D} = \begin{bmatrix} \sigma^2_{u_1} & \sigma_{u_1,u_2} \\ \sigma_{u_1,u_2} & \sigma^2_{u_2} \end{bmatrix},$$

where $u_{1,i}$ is the random effect associated with the intercept, $u_{int,i}$, for the $i$th subject.

In Hypothesis (3.3), we test if the random effect associated with time for each subject can be omitted from the model.

$$H_0\colon\ \mathbf{D} = \begin{bmatrix} \sigma^2_{int} & 0 \\ 0 & 0 \end{bmatrix} \text{ vs. } H_1\colon\ \mathbf{D} = \begin{bmatrix} \sigma^2_{int} & \sigma_{int,time} \\ \sigma_{int,time} & \sigma^2_{time} \end{bmatrix}. \tag{3.3}$$

We use REML-based likelihood ratio test, defined in Equation (2.25), and calculate a p-value for the test statistic according to Equation (2.28).

Similarly, in Hypothesis (3.4) we test if the random effect associated with diet for each subject can be omitted from the model.

$$H_0\colon\ \mathbf{D} = \begin{bmatrix} \sigma^2_{int} & 0 \\ 0 & 0 \end{bmatrix} \text{ vs. } H_1\colon\ \mathbf{D} = \begin{bmatrix} \sigma^2_{int} & \sigma_{int,diet} \\ \sigma_{int,diet} & \sigma^2_{diet} \end{bmatrix}. \tag{3.4}$$

If none of the hypotheses have significant test statistics, the original model, given in Equation (3.1), is the preferred model. But if one of the the hypothesis have significant test statistics, then the corresponding model is the preferred model. However, if both hypothesis have significant test statistics, we decide which random effect to include by using the Akaike information criteria, (2.32). Then the model with the lowest $AIC$ is the preferred model at this stage of the analysis.

Notice that we have chosen to follow the top-down strategy as preformed in Chapter 5 of West et al. (2007) and therefore we do not test whether the diagonal structure, given in Equation (2.7), is the best fit for the covariance matrix $\mathbf{D}$ associated with the random effects. There is a series of other structures which could be tested. For example whether the best fit is a covariance matrix $\mathbf{D}$ which allows heterogeneous variances for different levels of a specific factor.

## Step 3

When selecting a covariance structure for the residuals in the model, we start by investigating if the residual variances differ for the two levels of time. Hence, we replace $R_i$ in the preferred model at this stage of the analysis with

$$\begin{aligned} \mathbf{R}_i &= \begin{bmatrix} \sigma^2_{time=0} & 0 & 0 & 0 \\ 0 & \sigma^2_{time=1} & 0 & 0 \\ 0 & 0 & \sigma^2_{time=0} & 0 \\ 0 & 0 & 0 & \sigma^2_{time=1} \end{bmatrix} \\ \text{or } \mathbf{R}_i &= \begin{bmatrix} \sigma^2_{diet=A} & 0 & 0 & 0 \\ 0 & \sigma^2_{diet=A} & 0 & 0 \\ 0 & 0 & \sigma^2_{diet=B} & 0 \\ 0 & 0 & 0 & \sigma^2_{diet=B} \end{bmatrix}, \end{aligned}$$

where the order of the diagonal elements are dependent on the order of the measurements of subject $i$.

In Hypothesis (3.5), we test whether we should retain the heterogeneous residual variance structure, associated with time, for the $R_i$ matrix or not.

$$H_0\colon\ \sigma^2_{time=0} = \sigma^2_{time=1} \text{ vs. } H_1\colon\ \sigma^2_{time=0} \neq \sigma^2_{time=1}. \tag{3.5}$$

We use the likelihood ratio test, defined in Equation (2.8), where the reference model is the new model with heterogeneous residual variance structure and the nested model is the preferred model at this stage of the analysis. In order to calculate the p-value for the test statistic we simply use the ANOVA method of R Development Core Team (2010).

Similarly, in Hypothesis (3.6) we test if the residual variances differ for the two levels of diet.

$$H_0\colon\ \sigma^2_{diet=A} = \sigma^2_{diet=B} \text{ vs. } H_1\colon\ \sigma^2_{diet=A} \neq \sigma^2_{diet=B}. \tag{3.6}$$

If none of the hypotheses have significant test statistics, the preferred model from step 2 is still the preferred model. But if one of the the hypothesis have significant test statistics, then the corresponding model is the preferred model. However, if both hypothesis have significant test statistics, we decide the covariance structure for the residuals in the model by using the Akaike information criteria, (2.32). Then the model with the lowest $AIC$ is the preferred model.

Notice that we have chosen to follow the top-down strategy as preformed in Chapter 5 of West et al. (2007) and therefore we do not test whether the compound symmetry structure, given in Equation (2.11), is the best fit for the covariance matrix $\mathbf{R}_i$ associated with the residuals. There is a series of other structures which could be tested. For example whether the best fit is a covariance matrix $\mathbf{R}_i$ which allows heterogeneous variances for all four measures.

## Summary

We end up with five potential models for each of the biomarkers, with four corresponding hypotheses. A summary of what these five different models contain, can be seen in Table 3.2. Here entries marked with ✓ means that the term in the given row is present in the model of the given column. Entries marked with $\star$ depends on the result of hypothesis 1 and 2.

| | Variable | Notation | Model | | | | |
|---|---|---|---|---|---|---|---|
| | | | 1 | 2 | 3 | 4 | 5 |
| Random effects | Intercept | $u_{int,i}$ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Time | $u_{time,i}$ | | ✓ | | ⋆ | ⋆ |
| | Diet | $u_{diet,i}$ | | | ✓ | ⋆ | ⋆ |
| Residuals | | $\boldsymbol{\varepsilon}_i$ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Covariance parameters for **D** | Variance of intercepts | $\sigma^2_{int}$ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Variance of time effects | $\sigma^2_{time}$ | | ✓ | | ⋆ | ⋆ |
| | Covariance of intercepts and time effects | $\sigma_{int,time}$ | | ✓ | | ⋆ | ⋆ |
| | Variance of diet effects | $\sigma^2_{diet}$ | | | ✓ | ⋆ | ⋆ |
| | Covariance of intercepts and diet effects | $\sigma_{int,diet}$ | | | ✓ | ⋆ | ⋆ |
| Covariance parameters for $\mathbf{R}_i$ | Variance of residuals | $\sigma^2$ | ✓ | ✓ | ✓ | $\sigma^2_{t=0} \neq \sigma^2_{t=1}$ | $\sigma^2_{d=A} \neq \sigma^2_{d=B}$ |
| | Structure | $\mathbf{R}_i$ | $\sigma^2 I_{n_i}$ | $\sigma^2 I_{n_i}$ | $\sigma^2 I_{n_i}$ | *Het* | *Het* |

Table 3.2: The five different possible models for the biomarkers in the diet intervention study, where *Het* is a heterogeneous residual variance across time or diet.

## Results

In Table 3.3 we report the p-values from the four hypothesis tests for all 32 biomarkers. In the second column, we see whether the response should be transformed on a natural logarithmic scale or not, according to the *QQ*-plots when the data are fitted by the simplest LME model, given by Equation (3.1). And finally in the last column, Model, we have concluded based on all the hypothesis tests and given the preferred model for the specific biomarker.

From Table 3.3 we choose to take a closer look at four biomarkers:

- Resistin, which has the simplest form of the LME as its best fit.
- Uric acid, which has an LME where the random effect is associated with time for each subject.
- Triglycerides, which has an LME where the residual variances differ for the two levels of time.
- Visfatin, which has an LME where the random effect is associated with diet for each subject and the residual variances differ for the two levels of diet.

| Biomarker | log | Hyp (3.3) | Hyp (3.4) | Hyp (3.5) | Hyp (3.6) | Model |
|---:|:---:|---:|---:|---:|---:|---:|
| Glucose |  | 0.5088 | 1.0000 | 0.4155 | 0.3153 | 1 |
| Insulin | ✓ | 0.1412 | 0.0275 | 0.6434 | 0.2149 | 3 |
| C-peptide | ✓ | 0.2401 | 0.0914 | 0.5322 | 0.6839 | 1 |
| hsCRP | ✓ | 0.8916 | 0.4710 | 0.1807 | 0.0072 | 5 |
| Adiponectin | ✓ | 0.0028 | 0.0286 | <0.0001 | <0.0001 | 4 |
| PAI-1 | ✓ | 0.0002 | 0.1584 | 0.9426 | 0.3242 | 2 |
| Glucagon | ✓ | 0.7945 | 0.0021 | 0.1189 | 0.8694 | 3 |
| GLP-1 | ✓ | 0.5082 | 0.0635 | 0.3398 | 0.4026 | 1 |
| HOMA2 IR | ✓ | 0.3920 | 0.1401 | 0.4837 | 0.5889 | 1 |
| HOMA2 B | ✓ | 0.6118 | 0.0963 | 0.8686 | 0.1808 | 1 |
| HOMA2 S | ✓ | 0.3741 | 0.0988 | 0.6348 | 0.5866 | 1 |
| Triglycerides | ✓ | 0.8018 | 0.4492 | 0.0366 | 0.7946 | 4 |
| Total cholesterol | ✓ | 0.2932 | 0.0233 | 0.7992 | 0.8062 | 3 |
| LDL-cholesterol | ✓ | 0.0753 | 0.6295 | 0.1380 | 0.9981 | 1 |
| HDL-cholesterol | ✓ | 0.0127 | 0.1468 | 0.0001 | No conv. | 4 |
| Tri-HDL ratio | ✓ | 0.2953 | 0.0193 | 0.1760 | 0.9723 | 3 |
| TNF-alpha | ✓ | 0.0296 | 0.0319 | 0.8580 | 0.7574 | 2 |
| IL-6 | ✓ | 0.0028 | 0.0147 | 0.6448 | 0.7028 | 2 |
| Serum amyloid A | ✓ | 0.4120 | 0.7471 | 0.3941 | 0.7128 | 1 |
| GIP | ✓ | 0.0700 | 0.0291 | 0.2789 | 0.9391 | 3 |
| Ghrelin | ✓ | 0.7186 | 0.0350 | 0.1643 | 0.7679 | 3 |
| Leptin | ✓ | 0.0283 | 0.3872 | 0.4423 | 0.4072 | 2 |
| Visfatin | ✓ | 0.2447 | 0.0079 | 0.6653 | 0.0002 | 5 |
| Resistin |  | 0.1624 | 0.3853 | 0.1411 | 0.0709 | 1 |
| Uric acid | ✓ | 0.0006 | 0.9644 | 0.0627 | 0.6196 | 2 |
| Leukocytes | ✓ | 0.5598 | 0.2864 | 0.3193 | 0.3260 | 1 |
| Monocytes | ✓ | 0.8175 | 0.2859 | 0.2743 | 0.5732 | 1 |
| Eosinophils |  | 0.3555 | 0.3696 | 0.1242 | 0.2103 | 1 |
| Neutrophiles | ✓ | 0.1932 | 0.6935 | 0.2518 | 0.1728 | 1 |
| Lymphocytes |  | 0.2758 | 0.5887 | 0.5020 | 0.3886 | 1 |
| Basophiles |  | 0.9296 | 0.7709 | 0.0298 | 0.8805 | 4 |
| Platlets | ✓ | 0.2939 | 0.4128 | 0.8460 | 0.9381 | 1 |

Table 3.3: P-values of all hypothesis for the 32 biomarkers from the diet intervention study, where "No conv." means that the hypothesis could not be tested due to lack of convergence.

## 3.2 Simplest form of a LME: Resistin

Resistin is according to Berger (2001) a protein in the human body, which links obesity to type 2 diabetes. The name resistin comes from "resistance to insulin".

### Descriptive statistics

The descriptive statistics for the resistin measurements from the diet intervention study, can be seen in Table 3.4. Here we observe that the resistin measurements are higher in diet A, than in diet B. Further, we observe that the stating values for the two different diets are quite different. Hence, the wash-out period of the diet intervention study was perhaps too short. We also notice that the female participants have higher values of resistin than the male participants in the study.

| | Mean | N | Std.Deviation | Minimum | Maximum |
|---:|---|---|---:|---:|---:|
| A0 | 607.50 | 32 | 272.98 | 246.00 | 1198.00 |
| A6 | 810.44 | 32 | 386.00 | 294.00 | 1584.00 |
| A-total | 708.97 | 64 | 347.05 | 246.00 | 1584.00 |
| B0 | 566.66 | 32 | 294.31 | 91.00 | 1533.00 |
| B6 | 706.44 | 32 | 286.63 | 281.00 | 1375.00 |
| B-total | 636.55 | 64 | 296.66 | 91.00 | 1533.00 |
| T0 | 587.08 | 64 | 282.34 | 91.00 | 1533.00 |
| T6 | 758.44 | 64 | 341.30 | 281.00 | 1584.00 |
| T-total | 672.76 | 128 | 323.62 | 91.00 | 1584.00 |
| Female | 738.69 | 52 | 395.62 | 91.00 | 1584.00 |
| Male | 627.64 | 76 | 256.58 | 176.00 | 1434.00 |

Table 3.4: Descriptive statistics for the resistin measurements in the diet intervention study.

In Figure 3.1, we can see line graphs of resistin measurements for each individual by diet, from day zero to day six, where the different colors represent each individual. Here we observe that both diets seem to increase the resistin measurements, probably a little more in diet A than in diet B. We also observe that the between-participant variation is large in both diets.

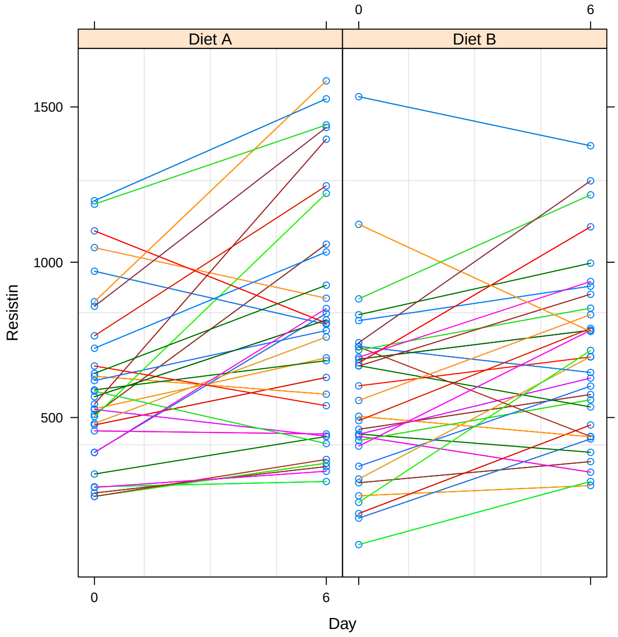

Figure 3.1: Line graphs of resistin for each individual (marked by separate colors) by time within levels of diet.

## Fitting the linear mixed effects model

Resistin was not significant for any of the Hypotheses, (3.3), (3.4), (3.5) or (3.6). Hence, the best fit for the resistin data is given by

$$Y_{ti} = \mathbf{X_i}\boldsymbol{\beta} + u_{int,i} + \varepsilon_{ti}, \tag{3.7}$$

where $Y_{ti}$ is the resistin measurement number $t$ ($t = 1, 2, 3, 4$) for the $i$-th subject ($i = 1, ..., 32$),

$$u_{int,i} \sim N(0, \sigma^2_{int}) \text{ and}$$

$$\boldsymbol{\varepsilon}_i = \begin{bmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \varepsilon_{3i} \\ \varepsilon_{4i} \end{bmatrix} \sim N_4(\mathbf{0}, \sigma^2 \mathbf{I}_4).$$

Following the top-down strategy, we now want to reduce the loaded model by preforming type I F-tests iteratively, using the ANOVA method of R Development Core Team (2010).

In Table 3.5 we report the results of the final F-test. Here we observe that only time and diet should be included as fixed effects in the model. Hence, the fixed effect vector is given by

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{intercept} \\ \beta_{time} \\ \beta_{diet} \end{bmatrix}. \tag{3.8}$$

| Fixed effect | numDF | denDF | F-value | p-value |
|---|---|---|---|---|
| Intercept | 1 | 94 | 200.64 | 0.00e+00 |
| Time | 1 | 94 | 27.25 | 1.07e-06 |
| Diet | 1 | 94 | 4.87 | 2.98e-02 |

Table 3.5: The final F-test results for the fitted resistin model.

The denominator degrees of freedom in the fitted resistin model (3.7), with the fixed effect vector given by Equation (3.8), is calculated according to Equation (2.31). Since the fixed effect vector only contains within-subject factors, which are estimated at level 2, the denominator degrees of freedom is given by

$$denDF_2 = m_2 - (m_1 + p_2) = 128 - (32 + 2) = 94.$$

## Results

The results of the estimation of the fitted resistin model (3.7), with the fixed effect vector given by Equation (3.8), using the *lme* function in Pinheiro et al. (2010), can be seen in Table 3.6.

| Notation | Estimate | Standard error | 95% Confidence Interval |
|---|---|---|---|
| **Fixed effects** | | | |
| Intercept, $\beta_0$ | 623.29 | 52.86 | (518.32, 728.25) |
| Time, $\beta_1$ | 171.36 | 32.83 | (106.18, 236.54) |
| Diet, $\beta_2$ | -72.42 | 32.83 | (-137.60, -7.24) |
| **Random effects** | | | |
| Intercept, $\sigma_{int}$ | 252.12 | | (189.91, 334.71) |
| **Residuals** | | | |
| Intercept, $\sigma$ | 185.70 | | (160.97, 214.24) |

Table 3.6: Results for the resistin model (3.7), using the *lme* function in *R*.

## Diagnostics

We obtain diagnostic plots for assessing the normality of residuals and random effects in the linear mixed effects model, by using the functions *qqnorm.lme* and *plot.lme* in Pinheiro et al. (2010).

### Residual diagnostics

The normal plot of the residuals, conditioned on diet, for the fitted resistin model (3.7), with the fixed effect vector given by Equation (3.8), can be seen in Figure 3.2. Here we observe that the normality assumption for the residuals seems plausible for both diets.

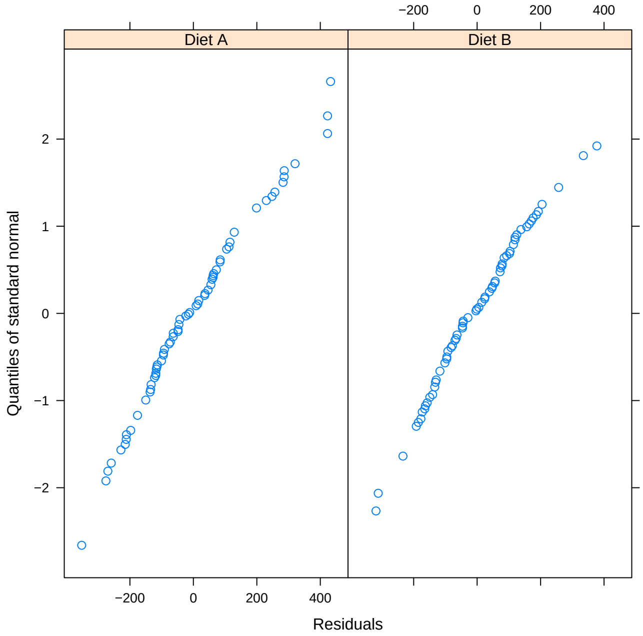

Figure 3.2: Normal plot of the residuals, conditioned on diet, from the resistin model (3.7).

In Figure 3.3 we have plotted the observed versus the fitted resistin measures, for the fitted resistin model (3.7), with the fixed effect vector given by Equation (3.8). This plot strengthens our belief that the that the normality assumption for the residuals seems plausible for both diets.

### Random effect diagnostics

The normal plot of estimated random effects, for the fitted resistin model (3.7), with the fixed effect vector given by Equation (3.8), can be seen in Figure 3.4. Here we notice two outliers, however for the rest of the observations the normality assumption seems reasonable.

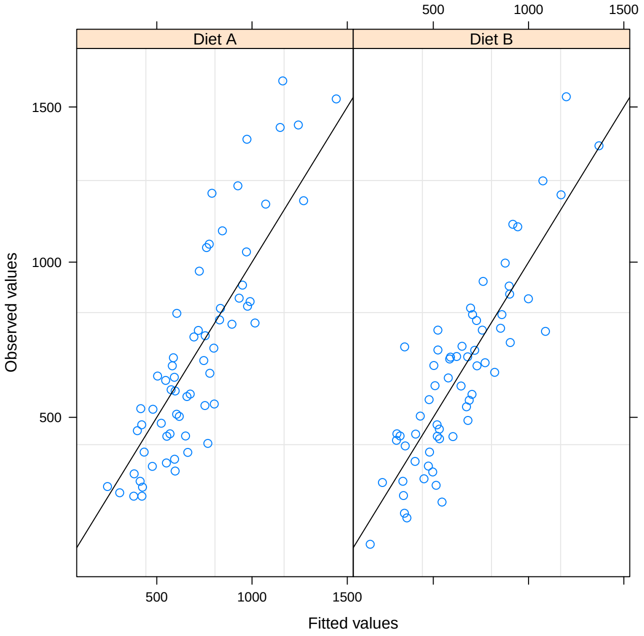

Figure 3.3: Observed versus fitted values plot for the resistin data.

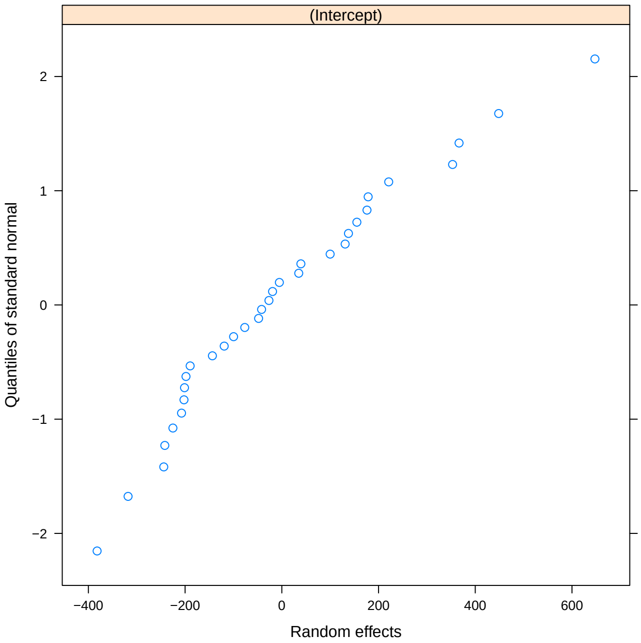

Figure 3.4: Normal plot of the estimated random effects from the resistin model (3.7).

## 3.3 Time associated random effect: Uric acid

Uric acid is according to Dugdale (2009) created when the body breaks down purines. Purines are found in various foods, such as dried beans and beer. Most uric acid dissolves in blood and passes out in urine, but high concentrations of uric acid in the blood may be harmful.

### Descriptive statistics

The descriptive statistics for the uric acid measurements on a natural logarithmic scale, can be seen in Table 3.7. Here we observe that the uric acid measurements are higher in diet B, than in diet A. Notice that there is a missing value at the starting time of diet B. Further, we observe that the female participants have lower values of uric acid than the male participants in the study, but with a higher standard deviation.

| | Mean | N | Std.Deviation | Minimum | Maximum |
|---:|---:|---:|---:|---:|---:|
| A0 | 5.67 | 32 | 0.18 | 5.21 | 6.14 |
| A6 | 5.58 | 32 | 0.20 | 5.19 | 5.89 |
| A-total | 5.62 | 64 | 0.19 | 5.19 | 6.14 |
| B0 | 5.73 | 31 | 0.18 | 5.38 | 6.15 |
| B6 | 5.62 | 32 | 0.19 | 5.29 | 6.12 |
| B-total | 5.67 | 63 | 0.19 | 5.29 | 6.15 |
| T0 | 5.70 | 63 | 0.18 | 5.21 | 6.15 |
| T6 | 5.60 | 64 | 0.19 | 5.19 | 6.12 |
| T-total | 5.65 | 127 | 0.19 | 5.19 | 6.15 |
| Female | 5.54 | 51 | 0.18 | 5.19 | 6.12 |
| Male | 5.72 | 76 | 0.16 | 5.35 | 6.15 |

Table 3.7: Descriptive statistics for the uric acid measurements in the diet intervention study, on a natural logarithmic scale.

In Figure 3.5, we can see line graphs of uric acid measurements on a natural logarithmic scale for each individual by diet, from day zero to day six. The different colors represent each individual. Here we observe that the uric acid measurements for both diets seems to both decrease and increase for different individuals. We also observe that the between-participant variation is large in both diets.

### Fitting the linear mixed effects model

Uric acid was only significant for Hypothesis (3.3). Hence, the best fit for the uric acid data is given by

$$Y_{ti} = \mathbf{X_i}\boldsymbol{\beta} + \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix} \mathbf{u}_i + \varepsilon_{ti}, \tag{3.9}$$

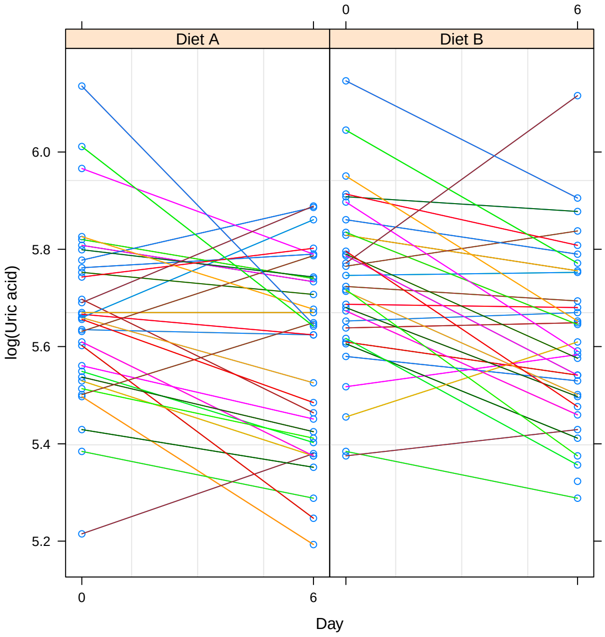

Figure 3.5: Line graphs of uric acid for each individual (marked by separate colors) by time within levels of diet.

where $Y_{ti}$ is the uric acid measurement number $t$ ($t = 1, 2, 3, 4$) for the $i$-th subject ($i = 1, ..., 32$),

$$\mathbf{u}_i = \begin{bmatrix} u_{int,i} \\ u_{time,i} \end{bmatrix} \sim N_2(\mathbf{0}, \mathbf{D}),$$

$$\mathbf{D} = \begin{bmatrix} \sigma^2_{int} & \sigma_{int,time} \\ \sigma_{int,time} & \sigma^2_{time} \end{bmatrix} \text{ and}$$

$$\boldsymbol{\varepsilon}_i = \begin{bmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \varepsilon_{3i} \\ \varepsilon_{4i} \end{bmatrix} \sim N_4(\mathbf{0}, \sigma^2 \mathbf{I}_4).$$

In Table 3.8 we report the results of the final F-test. Here we observe that sex, time and diet, but no interaction terms, should be included as fixed effects in the model. Hence, the fixed effect vector is given by

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{intercept} \\ \beta_{sex} \\ \beta_{time} \\ \beta_{diet} \end{bmatrix}. \tag{3.10}$$

The denominator degrees of freedom in the fitted uric acid model (3.9), with the fixed effect vector given by Equation (3.10), is calculated according to Equation (2.31). As we saw in Table 3.7, the uric acid data have one missing value. Hence

| Fixed effect | numDF | denDF | F-value | p-value |
|---|---|---|---|---|
| Intercept | 1 | 93 | 54188.02 | 0.00e+00 |
| Sex | 1 | 30 | 13.64 | 8.81e-04 |
| Time | 1 | 93 | 14.31 | 2.74e-04 |
| Diet | 1 | 93 | 10.10 | 2.02e-03 |

Table 3.8: The final F-test results for the fitted uric acid model.

the denominator degrees of freedom for the two types of factors in the fitted uric acid model are given by

$$denDF_1 = m_1 - (m_0 + p_1) = 32 - (1 + 1) = 30$$

and

$$denDF_2 = m_2 - (m_1 + p_2) = 127 - (32 + 2) = 93,$$

where the within-subject factors, time and diet, and the intercept are estimated at level 2 and the between-subject factor, sex, is estimated at level 1.

## Results

The results of the estimation of Model (3.9), with the fixed effect vector given by Equation (3.10), using the *lme* function in Pinheiro et al. (2010), can be seen in Table 3.9.

| Notation | Estimate | Standard error | 95% Confidence Interval |
|---|---|---|---|
| **Fixed effects** | | | |
| Intercept, $\beta_0$ | 5.5643 | 0.0399 | (5.4851, 5.6436) |
| Sex, $\beta_1$ | 0.1814 | 0.0495 | (0.0802, 0.2825) |
| Time, $\beta_2$ | -0.0962 | 0.0253 | (-0.1464, -0.0460) |
| Diet, $\beta_3$ | 0.0463 | 0.0146 | (0.0173, 0.0752) |
| **Random effects** | | | |
| Intercept, $\sigma_{int}$ | 0.1348 | | (0.0998, 0.1822) |
| Time, $\sigma_{time}$ | 0.1169 | | (0.0798, 0.1715) |
| Intercept:Time, $\rho_{int,time}$ | -0.253 | | (-0.6187, 0.2029) |
| **Residuals** | | | |
| Intercept, $\sigma$ | 0.0818 | | (0.0686, 0.0975) |

Table 3.9: Results for the uric acid model (3.9), using the *lme* function in *R*.

## Diagnostics

### Residual diagnostics

The normal plot of the residuals, conditioned on diet, for the fitted uric acid model (3.9), with the fixed effect vector given by Equation (3.10), can be seen in Figure 3.6. Here we notice one outlier for diet A, however for the rest of the observations the normality assumption for the residuals seems plausible for both diets.

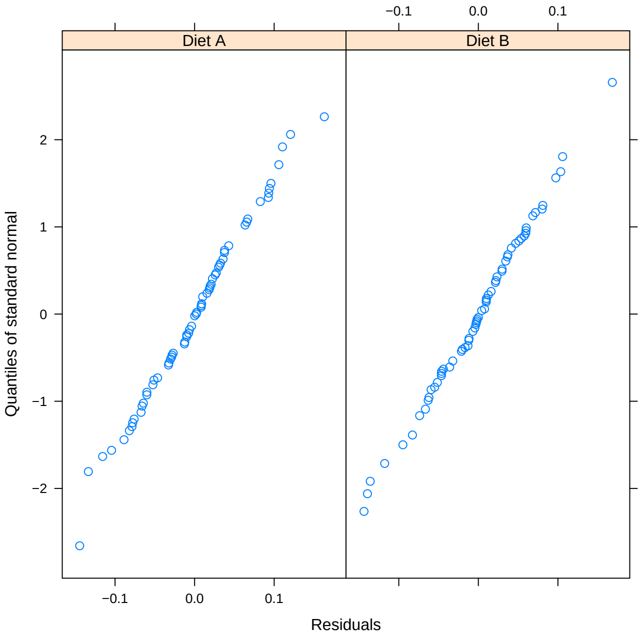

Figure 3.6: Normal plot of the residuals, conditioned on diet, from the uric acid model (3.9).

In Figure 3.7 we have plotted the observed versus the fitted uric acid measures, for the fitted uric acid model (3.9), with the fixed effect vector given by Equation (3.10). This plot strengthens our belief that the normality assumption for the residuals seems plausible for both diets.

### Random effect diagnostics

The normal plot of estimated random effects, for the fitted uric acid model (3.9), with the fixed effect vector given by Equation (3.10), can be seen in Figure 3.8. Here the assumption of normality seems reasonable for both random effects.

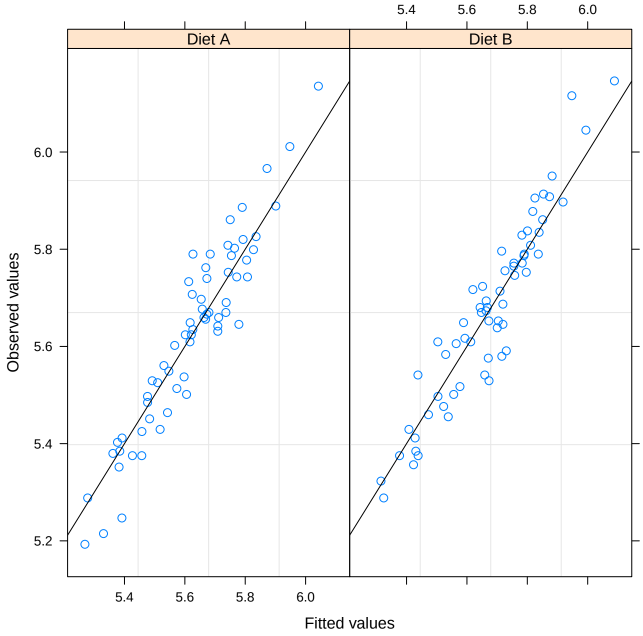

Figure 3.7: Observed versus fitted values plot for the uric acid data.

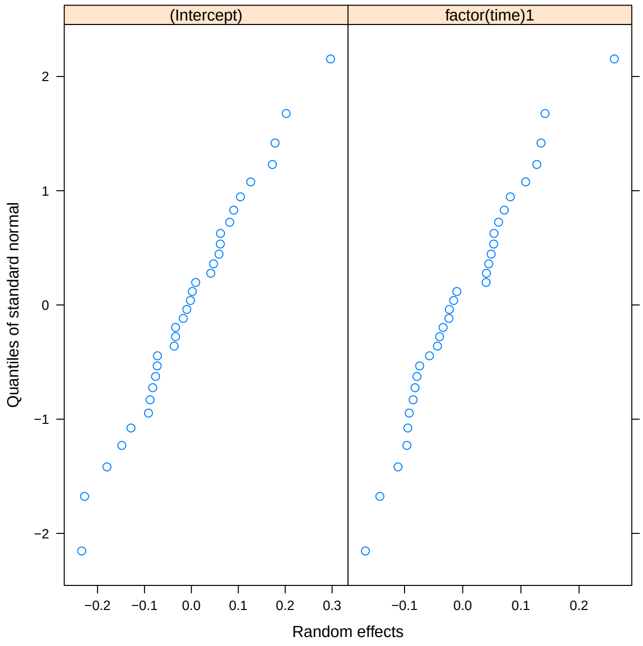

Figure 3.8: Normal plot of the estimated random effects from the uric acid model (3.9).

## 3.4 Time associated residual variances: Triglyceride

Triglyceride is according to Dugdale (2010) a type of fat in the human body, which in high levels may lead to atherosclerosis. Atherosclerosis increases the risk of heart attack and stroke. High triglyceride levels may also cause inflammation of the pancreas.

### Descriptive statistics

In Table 3.10 we can see the descriptive statistics for the triglyceride measurements on a natural logarithmic scale, from the diet intervention study. Here we observe that the triglyceride measurements are highly decreasing over time for both diets, however more so for diet B than for diet A. We also notice that the triglyceride measurements are very similar for male and female participants, though the female participants have a larger standard deviation.

In Figure 3.9, we can see line graphs of triglyceride measurements on a natural logarithmic scale, for each individual by diet, from day zero to day six. The different colors represent each individual. Here we observe that the triglyceride measurements for both diets seems to decrease for the most part. There are however participant, whom triglyceride levels increase. We notice that the between-participant variation is large in both diets.

| | Mean | N | Std.Deviation | Minimum | Maximum |
|---:|---:|---:|---:|---:|---:|
| A0 | -0.11 | 32 | 0.31 | -0.69 | 0.53 |
| A6 | -0.26 | 32 | 0.32 | -0.92 | 0.34 |
| A-total | -0.19 | 64 | 0.32 | -0.92 | 0.53 |
| B0 | -0.14 | 32 | 0.39 | -0.92 | 0.53 |
| B6 | -0.44 | 32 | 0.30 | -0.92 | 0.10 |
| B-total | -0.29 | 64 | 0.37 | -0.92 | 0.53 |
| T0 | -0.13 | 64 | 0.35 | -0.92 | 0.53 |
| T6 | -0.35 | 64 | 0.32 | -0.92 | 0.34 |
| T-total | -0.24 | 128 | 0.35 | -0.92 | 0.53 |
| Female | -0.24 | 52 | 0.32 | -0.92 | 0.53 |
| Male | -0.23 | 76 | 0.37 | -0.92 | 0.53 |

Table 3.10: Descriptive statistics for the triglyceride measurements in the diet intervention study, on the natural logarithmic scale.

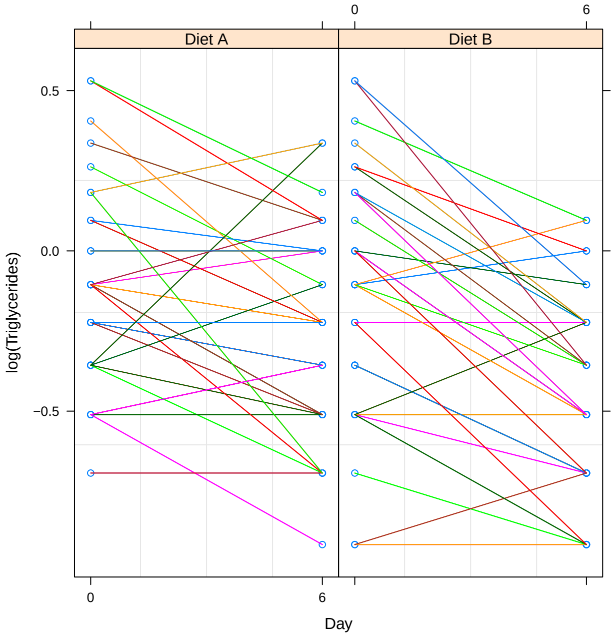

Figure 3.9: Line graphs of triglyceride for each individual (marked by separate colors) by time within levels of diet.

## Fitting the linear mixed effects model

Triglyceride was significant for Hypothesis 3.5. Hence, the best fit for the triglyceride data is given by

$$Y_{ti} = \mathbf{X_i}\boldsymbol{\beta} + u_{int,i} + \varepsilon_{ti}, \tag{3.11}$$

where $Y_{ti}$ is the triglyceride measurement number $t$ ($t = 1, 2, 3, 4$) for the $i$-th subject ($i = 1, ..., 32$),

$$u_{int,i} \sim N(0, \sigma^2_{int}),$$

$$\boldsymbol{\varepsilon}_i = \begin{bmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \varepsilon_{3i} \\ \varepsilon_{4i} \end{bmatrix} \sim N_4(\mathbf{0}, \mathbf{R}_i) \text{ and}$$

$$\mathbf{R}_i = \begin{bmatrix} \sigma^2_{Day0} & 0 & 0 & 0 \\ 0 & \sigma^2_{Day6} & 0 & 0 \\ 0 & 0 & \sigma^2_{Day0} & 0 \\ 0 & 0 & 0 & \sigma^2_{Day6} \end{bmatrix}.$$

In Table 3.11 we report the results of the final F-test. Here we observe that time and diet are the only terms included as fixed effects in the model. Hence, the fixed effect vector is given by

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{intercept} \\ \beta_{time} \\ \beta_{diet} \end{bmatrix}. \tag{3.12}$$

| Fixed effect | numDF | denDF | F-value | p-value |
|---|---|---|---|---|
| Intercept | 1 | 94 | 32.61 | 1.31e-07 |
| Time | 1 | 94 | 32.70 | 1.27e-07 |
| Diet | 1 | 94 | 12.11 | 7.64e-04 |

Table 3.11: The final F-test results for the fitted triglyceride model.

The denominator degrees of freedom in the fitted triglyceride model (3.11), with the fixed effect vector given by Equation (3.12), is calculated according to Equation (2.31). Since the fixed effect vector only contains within-subject factors, which are estimated at level 2, the denominator degrees of freedom is given by

$$denDF_2 = m_2 - (m_1 + p_2) = 128 - (32 + 2) = 93.$$

## Results

The results of the estimation of Model 3.11, with the fixed effect vector given by Equation 3.12, using the *lme* function in Pinheiro et al. (2010), can be seen in Table 3.12.

| Notation | Estimate | Standard error | 95% Confidence Interval |
|---|---|---|---|
| **Fixed effects** | | | |
| Intercept, $\beta_0$ | -0.0623 | 0.0577 | (-0.1769, 0.0523) |
| Time, $\beta_1$ | -0.2228 | 0.0390 | (-0.3001, -0.1454) |
| Diet, $\beta_2$ | -0.1277 | 0.0367 | (-0.2005, -0.0548) |
| **Random effects** | | | |
| Intercept, $\sigma_{int}$ | 0.2518 | | (0.1879, 0.3374) |
| **Residuals** | | | |
| Time, $\sigma_{Day0}$ | 0.2547 | | (0.2086, 0.3110) |
| Time, $\sigma_{Day6}$ | 0.1796 | | (0.1070, 0.3014) |

Table 3.12: Results for the triglyceride model (3.11), using the *lme* function in *R*.

# Diagnostics

## Residual diagnostics

The normal plot of the residuals, conditioned on diet, for the fitted triglyceride model (3.11), with the fixed effect vector given by Equation (3.12), can be seen in Figure 3.10. Here the normality assumption for the residuals seems reasonable for both diets.

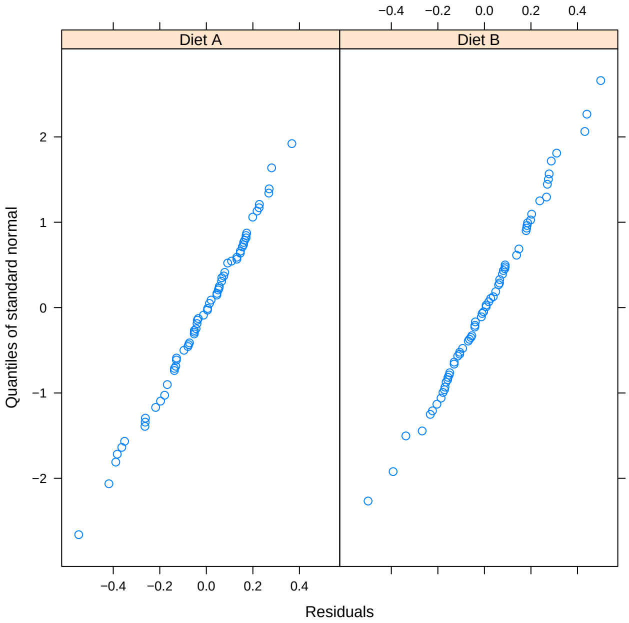

Figure 3.10: Normal plot of the residuals, conditioned on diet, from the triglyceride model (3.11).

In Figure 3.11 we have plotted the observed versus the fitted triglyceride measures, for the fitted triglyceride model (3.11), with the fixed effect vector given by Equation (3.12). This plot strengthens our belief that the normality assumption for the residuals seems plausible for both diets.

## Random effect diagnostics

The normal plot of estimated random effects, for the fitted triglyceride model (3.11), with the fixed effect vector given by Equation (3.12), can be seen in Figure 3.12. Here the assumption of normality seems highly reasonable for the random effects.

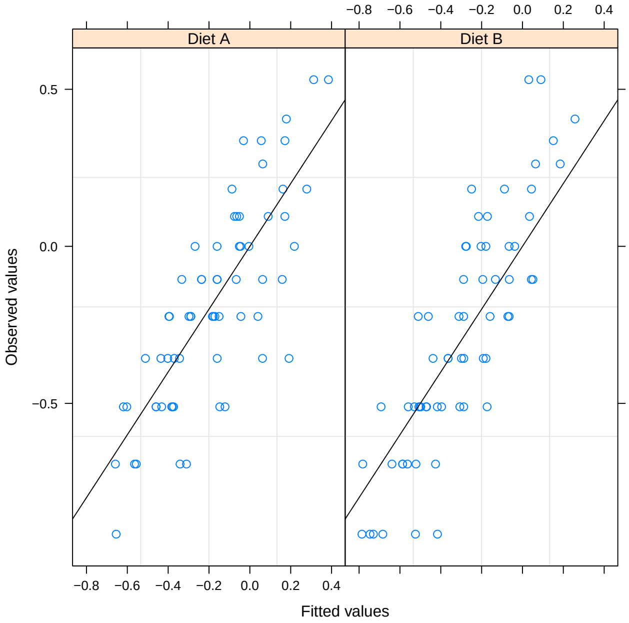

Figure 3.11: Observed versus fitted values plot for the triglyceride data.

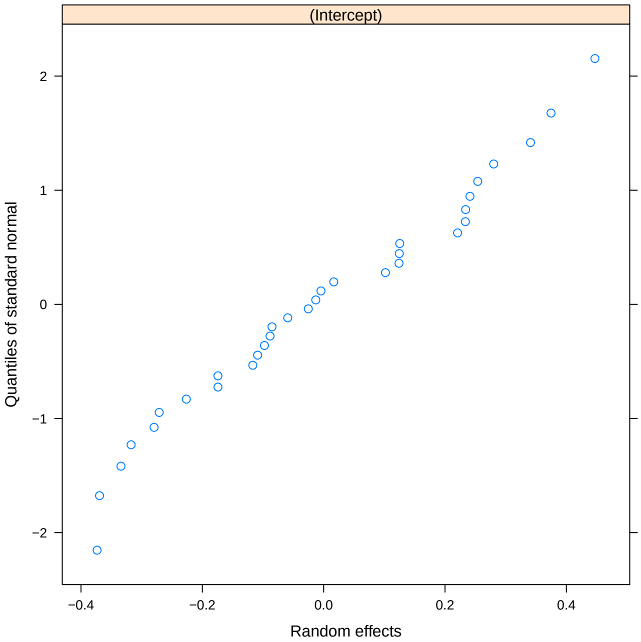

Figure 3.12: Normal plot of the estimated random effects from the triglyceride model (3.11).

## 3.5 Diet associated random effect and residual variances: Visfatin

Visfatin is according to Chi (2007) a protein in fat cells. Its physiological role has been subject of much controversy regarding its insulin-mimetic properties and potential of being a drug target for type 2 diabetes.

### Descriptive statistics

In Table 3.13 we can see the descriptive statistics for the visfatin measurements on a natural logarithmic scale. Here we observe that the measures from diet A both have higher mean values and standard deviation, than diet B. Hence, the wash-out period of the diet intervention study was perhaps too short. We also notice that the visfatin data have 2 missing values at day zero and 5 missing values at day six of diet A.

In Figure 3.13, we can see line graphs of visfatin measurements on the natural logarithmic scale for each individual by diet, from day zero to day six. The different colors represent each individual. Here we observe that there seem to be differences between in the measurements from diet A and diet B. In diet A participants seem to be both increasing and decreasing in visfatin measurements over time. In diet B however, the participants seems to mainly have decreasing visfatin measurements over time. We also observe that the between-participant variation is large in both diets.

| | Mean | N | Std.Deviation | Minimum | Maximum |
|---:|---:|---:|---:|---:|---:|
| A0 | 7.97 | 30 | 1.08 | 5.18 | 9.73 |
| A6 | 7.46 | 27 | 1.11 | 4.80 | 9.99 |
| A-total | 7.73 | 57 | 1.11 | 4.80 | 9.99 |
| B0 | 7.16 | 32 | 0.92 | 4.14 | 8.48 |
| B6 | 6.93 | 32 | 0.73 | 5.48 | 8.20 |
| B-total | 7.05 | 64 | 0.83 | 4.14 | 8.48 |
| T0 | 7.55 | 62 | 1.07 | 4.14 | 9.73 |
| T6 | 7.17 | 59 | 0.95 | 4.80 | 9.99 |
| T-total | 7.37 | 121 | 1.03 | 4.14 | 9.99 |
| Female | 7.40 | 49 | 1.04 | 4.80 | 9.99 |
| Male | 7.34 | 72 | 1.03 | 4.14 | 9.73 |

Table 3.13: Descriptive statistics for the visfatin measurements in the diet intervention study, on a natural logarithmic scale.

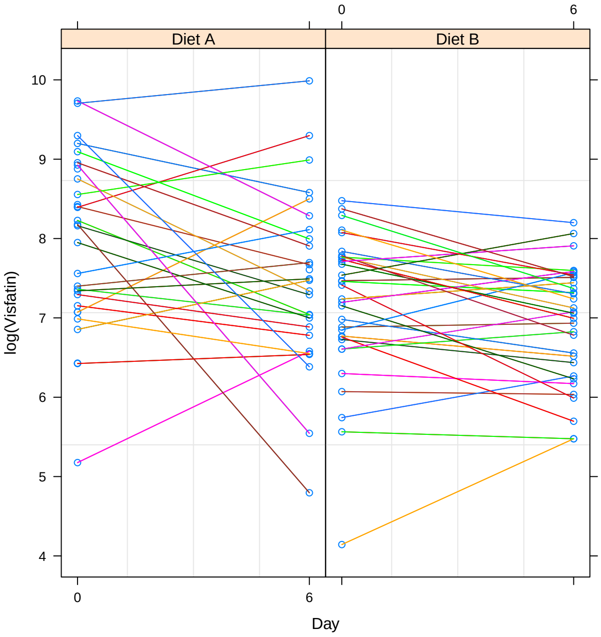

Figure 3.13: Line graphs of visfatin for each individual (marked by separate colors) by time within levels of diet.

## Fitting the linear mixed effects model

Visfatin was significant for both Hypothesis 3.4 and Hypothesis 3.6. Hence, the best fit for the visfatin data is given by

$$Y_{ti} = \mathbf{X_i}\boldsymbol{\beta} + \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \mathbf{u}_i + \varepsilon_{ti}, \tag{3.13}$$

where $Y_{ti}$ is the visfatin measurement number $t$ ($t = 1, 2, 3, 4$) for the $i$-th subject ($i = 1, ..., 32$),

$$\mathbf{u}_i = \begin{bmatrix} u_{int,i} \\ u_{diet,i} \end{bmatrix} \sim N_2(\mathbf{0}, \mathbf{D}),$$

$$\mathbf{D} = \begin{bmatrix} \sigma^2_{int} & \sigma_{int,diet} \\ \sigma_{int,diet} & \sigma^2_{diet} \end{bmatrix},$$

$$\boldsymbol{\varepsilon}_i = \begin{bmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \varepsilon_{3i} \\ \varepsilon_{4i} \end{bmatrix} \sim N_4(\mathbf{0}, \mathbf{R}_i) \text{ and}$$

$$\mathbf{R}_i = \begin{bmatrix} \sigma^2_{DietA} & 0 & 0 & 0 \\ 0 & \sigma^2_{DietA} & 0 & 0 \\ 0 & 0 & \sigma^2_{DietB} & 0 \\ 0 & 0 & 0 & \sigma^2_{DietB} \end{bmatrix}.$$

In Table 3.14 we report the results of the final F-test. Here we observe that only time and diet should be included as fixed effects in the model. Hence, the fixed effect vector is given by

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{intercept} \\ \beta_{time} \\ \beta_{diet} \end{bmatrix}. \tag{3.14}$$

| | numDF | denDF | F-value | p-value |
|---|---|---|---|---|
| (Intercept) | 1 | 87 | 4102.24 | 0.00e+00 |
| factor(time) | 1 | 87 | 8.16 | 5.35e-03 |
| factor(diet) | 1 | 87 | 11.89 | 8.74e-04 |

Table 3.14: The final F-test results for the fitted visfatin model.

The denominator degrees of freedom in the fitted visfatin model (3.13), with the fixed effect vector given by Equation (3.14), is calculated according to Equation (2.31). As we saw in Table 3.13, the visfatin data have seven missing values and since the fixed effect vector only contains within-subject factors, which are estimated at level 2, the denominator degrees of freedom is given by

$$denDF_2 = m_2 - (m_1 + p_2) = 121 - (32 + 2) = 87.$$

## Results

The results of the estimation of Model 3.13, with the fixed effect vector given by Equation 3.14, using the *lme* function in Pinheiro et al. (2010), can be seen in Table 3.15.

| Notation | Estimate | Standard error | 95% Confidence Interval |
|---|---|---|---|
| **Fixed effects** | | | |
| Intercept, $\beta_0$ | 7.8503 | 0.1710 | (7.5103, 8.1902) |
| Time, $\beta_1$ | -0.2744 | 0.0981 | (-0.4693, -0.0794) |
| Diet, $\beta_2$ | -0.6680 | 0.1938 | (-1.0532, -0.2829) |
| **Random effects** | | | |
| Intercept, $\sigma_{int}$ | 0.6315 | | (0.3495, 1.1409) |
| Diet, $\sigma_{diet}$ | 0.7991 | | (0.4874, 1.3101) |
| Intercept:Time, $\rho_{int,time}$ | -0.529 | | (-0.8537,0.0914) |
| **Residuals** | | | |
| Diet, $\sigma_{DietA}$ | 0.8936 | | (0.6852, 1.1653) |
| Diet, $\sigma_{DietB}$ | 0.4297 | | (0.2287, 0.8074) |

Table 3.15: Results for the visfatin model (3.13), using the *lme* function in *R*.

## Diagnostics

### Residual diagnostics

The normal plot of the residuals, conditioned on diet, for the fitted visfatin model (3.13), with the fixed effect vector given by Equation (3.14), can be seen in Figure 3.14. This plot shows that the distribution of the residuals for diet A have heavier tails than expected under normality. According to Pinheiro and Bates (2000), this suggests that a mixture of normal distributions or a t-distribution with a moderate number of degrees of freedom perhaps would be a better distribution of the residuals for diet A. But because the tails seem to be symmetric around zero, the estimates of the fixed effects should not change substantially under either a mixture model or a t-model. However, the assumption of normality for the residuals seems highly reasonable for diet B, but with one outlier.

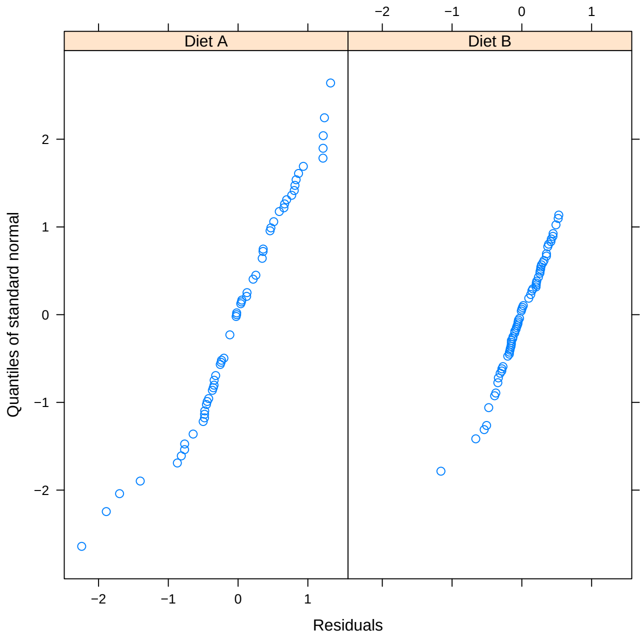

Figure 3.14: Normal plot of the residuals, conditioned on diet, from the visfatin model (3.13).

In Figure 3.15 we have plotted the observed versus the fitted visfatin measures, for the fitted visfatin model (3.13), with the fixed effect vector given by Equation (3.14). This plot strengthens our belief that the normality assumption for the residuals of diet A is not a good assumption. However, the assumption of normality seems highly reasonable for diet B.

### Random effect diagnostics

The normal plot of estimated random effects, for the fitted visfatin model (3.13), with the fixed effect vector given by Equation (3.14), can be seen in Figure 3.16. Here the assumption of normality seems reasonable for both random effects, though there is some asymmetry in the distribution of the random effect associated with diet.

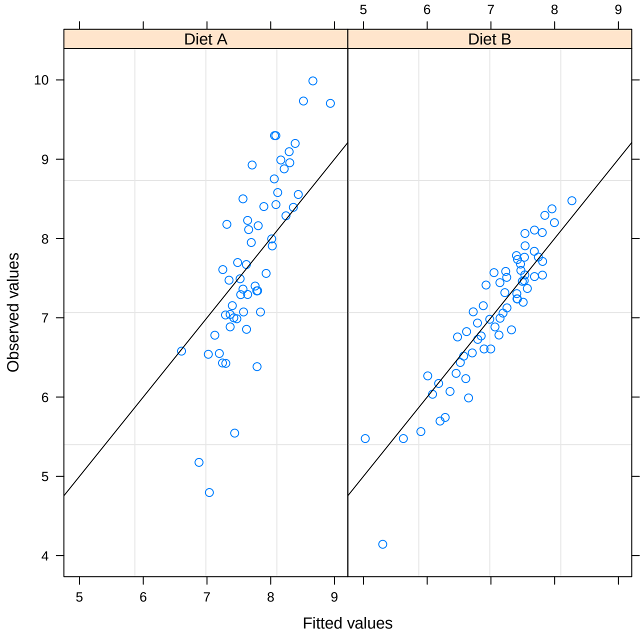

Figure 3.15: Observed versus fitted values plot for visfatin data.

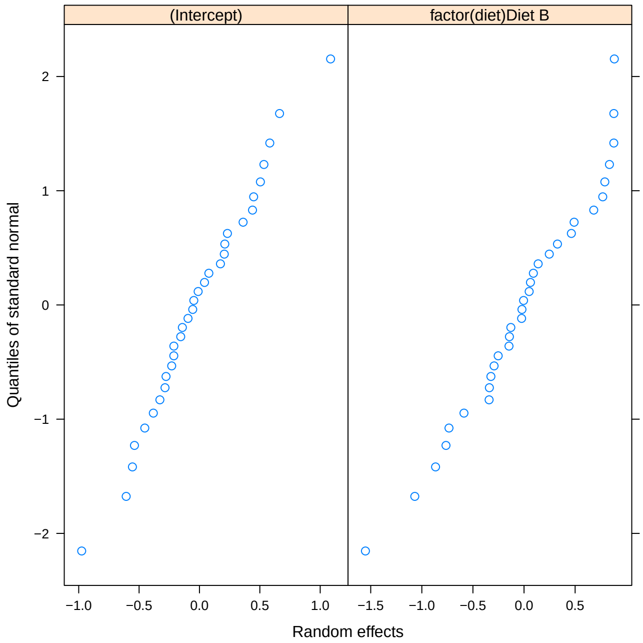

Figure 3.16: Normal plot of the estimated random effects from the visfatin model (3.13).

# Chapter 4

# Contrasts

When linear mixed effect models are used to analyze medical data, we are often interested in the effect of a specific treatment or risk factor under a given condition. Hence, the estimates of fixed- and random-effects may in many cases not answer the questions we are interested in.

We use the diet intervention study as an example. This is a full factorial design with two factors, diet and time. We have four combinations of diet and time covariates, $A0$, $A6$, $B0$ and $B6$. These effects are not directly of interest to biologists. To biologists, the contrasts of interest are:

- The effect of diet $A$, $A6 - A0$.
- The effect of diet $B$, $B6 - B0$.
- The difference in effect between diet $A$ and diet $B$, $(B6 - B0) - (A6 - A0)$.

There are several ways of finding the linear functions for these three contrasts. One way is to use the linear functions of the full factorial design, and simply subtract the linear function corresponding to $A0$ from the linear function corresponding to $A6$, and so on.

Another way is to interpret the fixed coefficients in the linear mixed effect model directly as contrasts. For example, in the case where the fixed effect vector is given as

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{intercept} \\ \beta_{time} \\ \beta_{diet} \\ \beta_{time:diet} \end{bmatrix}, \tag{4.1}$$

each of the coefficients can be interpreted directly as contrasts. This is called the *treatment − contrast* parametrization, and can be seen in Table 4.1.

| Coefficient | Contrast | Interpretation |
|---|---|---|
| Intercept | $A0$ | The baseline level |
| Time | $A6 - A0$ | The effect of diet $A$ |
| Diet | $B0 - A0$ | The difference in starting values |
| Time:Diet | $(B6 - B0) - (A6 - A0)$ | The difference in effect between the diets |

Table 4.1: The treatment-contrast parametrization of the diet intervention study.

We notice that the contrast of interest, $B6-B0$, is not directly represented in Table 4.1. $B6 - B0$ can however be extracted as the sum of the time coefficient and the interaction coefficient of the original model. That is, $((B6 - B0) - (A6 - A0)) + (A6 - A0) = B6 - B0$.

In order to get a thorough understanding of how contrasts are estimated and even more interesting, how they are tested, we will now take a closer look at some variations of the resistin measurements' fitted linear mixed effect model, given in Equation (3.7). The different variations of the model will include different fixed effect vectors with different parametrization. Since the focus here is on the fixed part of the LME, we will keep the random effects and the residuals on the simplest forms. We remember from Equation (2.31), that the denominator degrees of freedom is not influenced by the structure of the covariance matrices, $\mathbf{D}$ and $\mathbf{R}_i$.

In $R$, there are several ways of estimating such contrasts for LME-objects. We have chosen to use the function *estimable*, in the package *gmodels* by Warnes (2011). The *estimable* function uses the conditional t-test, given in Equation (2.29), with degrees of freedom equal to the smallest degree of freedom among the parameters used to construct the linear function or contrast being tested.

## 4.1 Within-subject factors

The first variation of the resistin measurements' fitted linear mixed effect model, given in Equation (3.7), only contains two dichotomous within-subject factors and their interaction term in the fixed effect vector, given by Equation (4.1).

Hence, the four possible outcomes of the models fixed effects are given by the following linear functions,

$$\begin{bmatrix} A0 \\ A6 \\ B0 \\ B6 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}.$$

That is, $A0 = \beta_{intercept}$, $A6 = \beta_{intercept} + \beta_{time}$, $B0 = \beta_{intercept} + \beta_{diet}$ and $B6 = \beta_{intercept} + \beta_{time} + \beta_{diet} + \beta_{time:diet}$.

By subtraction, we find that the contrasts of interest are given by

$$\begin{bmatrix} A6 - A0 \\ B6 - B0 \\ (B6 - B0) - (A6 - A0) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

That is, $A6 - A0 = \beta_{time}$, $B6 - B0 = \beta_{time} + \beta_{time:diet}$ and $(B6 - B0) - (A6 - A0) = \beta_{time:diet}$.

The results of all the estimated linear functions and contrasts can be seen in Table 4.2. Here each linear function and contrast is tested with the conditional t-test, given in Equation (2.29). Since all the factors included in the model are within-subject factors, that is, level 2 factors, the degrees of freedom will also be given on level 2 according to Equation (2.31). Hence, since the resistin measurements have no missing values, the degrees of freedom for factors estimated at level 2 is given by

$$DF_{level2} = m_2 - (m_1 + p_2) = 128 - (32 + 3) = 93. \tag{4.2}$$

| Contrast | Estimate | Std. Error | t value | DF | Pr(>\|t\|) |
|---|---|---|---|---|---|
| A0 | 607.50 | 55.36 | 10.97 | 93 | 0.00e+00 |
| A6 | 810.44 | 55.36 | 14.64 | 93 | 0.00e+00 |
| B0 | 566.66 | 55.36 | 10.24 | 93 | 0.00e+00 |
| B6 | 706.44 | 55.36 | 12.76 | 93 | 0.00e+00 |
| A6-A0 | 202.94 | 46.44 | 4.37 | 93 | 3.24e-05 |
| B6-B0 | 139.78 | 46.44 | 3.01 | 93 | 3.37e-03 |
| (B6-B0)-(A6-A0) | -63.16 | 65.68 | -0.96 | 93 | 3.39e-01 |

Table 4.2: Contrasts for the within-subject model, given in Equation (4.1).

In Table 4.2 we observe that all the linear combinations and both the contrast which represents the effect of diet A and the contrast which represents the effect of diet B, are significant on a $\alpha = 0.05$ significance level. However, the contrast which represents the difference in effect between the two diets are not significant.

## 4.2 Between-subject factors

The second variation of the resistin measurements' fitted linear mixed effect model, given in Equation (3.7), contains the all the fixed effects from the previous model and the between-subject factor sex. Hence, the fixed effect vector is given by

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{intercept} \\ \beta_{sex} \\ \beta_{time} \\ \beta_{diet} \\ \beta_{time:diet} \end{bmatrix}. \tag{4.3}$$

This means that we no longer only have factors estimated at level 2, but also a factor estimated at level 1. Since we still have the same number of factors estimated at level 2 as in the model only including within-subject factors, the degrees of freedom for factors estimated at level 2 is still given by Equation (4.2). The degrees of freedom for factors estimated at level 1 however, is given by

$$DF_{level1} = m_1 - (m_0 + p_1) = 32 - (1 + 1) = 30. \tag{4.4}$$

Hence, the linear functions and contrasts have different degrees of freedom depending on whether sex is among the parameters used to construct the linear function or contrast, or not.

Due to the inclusion of the sex factor, we are now able to calculate the linear functions, $A0$, $A6$, $B0$ and $B6$, for male and female separately or together as a mean, or a weighted mean, of all included subjects in the study. There are several ways of doing this. We will now estimate and test the linear functions and contrasts using two different parameterizations of the sex factor, the treatment contrast coding and the sum to zero contrast coding.

### 4.2.1 Treatment contrast coding

We start by coding the sex factor as "treatment", by using $male = 1$ and $female = 0$ in the design matrix for the fixed effects, $\mathbf{X}_i$. This is the default setting in $R$ if a factor has two levels.

#### Female participants

The four possible outcomes of model coefficients for the female participants, when the sex factor is coded as "treatment", are

$$\begin{bmatrix} A0 \\ A6 \\ B0 \\ B6 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \end{bmatrix}.$$

That is, $A0 = \beta_{intercept}$, $A6 = \beta_{intercept} + \beta_{time}$, $B0 = \beta_{intercept} + \beta_{diet}$ and $B6 = \beta_{intercept} + \beta_{time} + \beta_{diet} + \beta_{time:diet}$.

By subtraction, we find that the contrasts of interest are given by

$$\begin{bmatrix} A6 - A0 \\ B6 - B0 \\ (B6 - B0) - (A6 - A0) \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \tag{4.5}$$

That is, $A6 - A0 = \beta_{time}$, $B6 - B0 = \beta_{time} + \beta_{time:diet}$ and $(B6 - B0) - (A6 - A0) = \beta_{time:diet}$.

Here we observe that the sex factor is not included in any of the linear functions, $A0$, $A6$, $B0$ and $B6$, nor contrasts. Hence, all the included factors are at level 2 and all the linear function's and contrast's degrees of freedom are given by Equation (4.2), $DF = DF_{level2} = 93$.

| Contrast | Estimate | Std. Error | t value | DF | Pr(>\|t\|) |
|---|---|---|---|---|---|
| A0 | 673.43 | 79.39 | 8.48 | 93 | 3.30e-13 |
| A6 | 876.37 | 79.39 | 11.04 | 93 | 0.00e+00 |
| B0 | 632.59 | 79.39 | 7.97 | 93 | 3.95e-12 |
| B6 | 772.37 | 79.39 | 9.73 | 93 | 8.88e-16 |
| A6-A0 | 202.94 | 46.44 | 4.37 | 93 | 3.24e-05 |
| B6-B0 | 139.78 | 46.44 | 3.01 | 93 | 3.37e-03 |
| (B6-B0)-(A6-A0) | -63.16 | 65.68 | -0.96 | 93 | 3.39e-01 |

Table 4.3: Contrasts for the female participants for Model (4.3), using the treatment contrast coding for sex.

The results of the estimated linear functions and contrasts can be seen in Table 4.3. Here we observe that all the linear combinations are significant on a $\alpha = 0.05$ significance level. Compared to the model which only included within-subject factor, given in Equation (4.1), in Table 4.2, we observe that the estimates of the linear functions are larger for the female participants.

The contrasts however, have the exact same estimates as they had in Table 4.2. Hence, both the contrasts which represents the effect of diet A and the contrast which represents the effect of diet B, are significant on a $\alpha = 0.05$ significance level. And the contrast which represents the difference in effect between the two diets are not significant for the female participants.

### Male participants

The four possible outcomes of model coefficients for the male participants, when the sex factor is coded as "treatment", are

$$\begin{bmatrix} A0 \\ A6 \\ B0 \\ B6 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

That is, $A0 = \beta_{intercept} + \beta_{sex}$, $A6 = \beta_{intercept} + \beta_{sex} + \beta_{time}$, $B0 = \beta_{intercept} + \beta_{sex} + \beta_{diet}$ and $B6 = \beta_{intercept} + \beta_{sex} + \beta_{time} + \beta_{diet} + \beta_{time:diet}$.

By subtraction, we find that the contrasts of interest are the exact same as for the female participants, given by Equation (4.5). That is, $A6 - A0 = \beta_{time}$, $B6 - B0 = \beta_{time} + \beta_{time:diet}$ and $(B6 - B0) - (A6 - A0) = \beta_{time:diet}$.

In this case, the sex factor is included in all the linear functions, $A0$, $A6$, $B0$ and $B6$, and since the linear functions degrees of freedom is equal to the smallest degree of freedom among the set of parameters included in the linear function, the corresponding degrees of freedom is given by Equation (4.4), $DF = DF_{level1} = 30$.

For the three contrasts however, sex is not included and so the contrast's degrees of freedom are given by Equation (4.2) as for the female participants, $DF = DF_{level2} = 93$.

| Contrast | Estimate | Std. Error | t value | DF | Pr(>\|t\|) |
|---|---|---|---|---|---|
| A0 | 562.39 | 67.59 | 8.32 | 30 | 2.75e-09 |
| A6 | 765.32 | 67.59 | 11.32 | 30 | 2.34e-12 |
| B0 | 521.54 | 67.59 | 7.72 | 30 | 1.31e-08 |
| B6 | 661.32 | 67.59 | 9.78 | 30 | 7.57e-11 |
| A6-A0 | 202.94 | 46.44 | 4.37 | 93 | 3.24e-05 |
| B6-B0 | 139.78 | 46.44 | 3.01 | 93 | 3.37e-03 |
| (B6-B0)-(A6-A0) | -63.16 | 65.68 | -0.96 | 93 | 3.39e-01 |

Table 4.4: Contrasts for the male participants for Model (4.3), using the treatment contrast coding for sex.

The results of the estimated linear functions and contrasts can be seen in Table 4.4. Here we observe that all the linear combinations are significant on a $\alpha = 0.05$ significance level. Compared to the model which only included within-subject factor, Equation (4.1), in Table 4.2, we observe that the estimates of the linear functions are smaller for the male participants.

The contrasts have the exact same estimates as they had for both the within-subject model in Table 4.2 and the female participants' model in Table 4.3.

### All participants

When the sex factor is coded as "treatment", we can find the mean linear functions for all included subjects in the study by using $sex = 0.5$. Hence, the linear functions for the mean participant, when the sex factor is coded as "treatment", are

$$\begin{bmatrix} A0 \\ A6 \\ B0 \\ B6 \end{bmatrix} = \begin{bmatrix} 1 & 0.5 & 0 & 0 & 0 \\ 1 & 0.5 & 1 & 0 & 0 \\ 1 & 0.5 & 0 & 1 & 0 \\ 1 & 0.5 & 1 & 1 & 1 \end{bmatrix}.$$

That is, $A0 = \beta_{intercept} + 0.5\beta_{sex}$, $A6 = \beta_{intercept} + 0.5\beta_{sex} + \beta_{time}$, $B0 = \beta_{intercept} + 0.5\beta_{sex} + \beta_{diet}$ and $B6 = \beta_{intercept} + 0.5\beta_{sex} + \beta_{time} + \beta_{diet} + \beta_{time:diet}$.

By subtraction, we find that the mean contrasts of interest are the same as for the male and female participants, given by Equation (4.5). That is, $A6 - A0 = \beta_{time}$, $B6 - B0 = \beta_{time} + \beta_{time:diet}$ and $(B6 - B0) - (A6 - A0) = \beta_{time:diet}$.

In this case, the sex factor is included in all the mean linear functions, $A0$, $A6$, $B0$ and $B6$, and since the linear functions degrees of freedom is equal to the smallest degree of freedom among the set of parameters included in the linear function, the corresponding degrees of freedom should have been given by Equation (4.4), $DF = DF_{level1} = 30$. However, the results in Table 4.5 shows that the *estimable* function in $R$ weights the degrees of freedom. Hence, the degrees of freedom used for the linear functions, $A0$, $A6$, $B0$ and $B6$, in Table 4.5 is $30 \times 0.5 = 15$, due to the fact that $sex = 0.5$.

For the three mean contrasts however, sex is not included and so the contrast's degrees of freedom are given by Equation (4.2) as for the male and female participants, $DF = DF_{level2} = 93$.

| Contrast | Estimate | Std. Error | t value | DF | Pr(>\|t\|) |
|---|---|---|---|---|---|
| A0 | 617.91 | 55.88 | 11.06 | 15 | 1.31e-08 |
| A6 | 820.85 | 55.88 | 14.69 | 15 | 2.60e-10 |
| B0 | 577.07 | 55.88 | 10.33 | 15 | 3.26e-08 |
| B6 | 716.85 | 55.88 | 12.83 | 15 | 1.73e-09 |
| A6-A0 | 202.94 | 46.44 | 4.37 | 93 | 3.24e-05 |
| B6-B0 | 139.78 | 46.44 | 3.01 | 93 | 3.37e-03 |
| (B6-B0)-(A6-A0) | -63.16 | 65.68 | -0.96 | 93 | 3.39e-01 |

Table 4.5: Contrasts for the mean participant for Model (4.3), using the treatment contrast coding for sex.

The results of the estimated mean linear functions and contrasts can be seen in Table 4.5. Here we observe that the linear combinations are significant on a $\alpha = 0.05$ significance level. We also observe that the estimates of the linear functions are larger than the estimates made from the model only including within-subject factors in Table 4.2. This is due to the fact that we have not taken into account that there are more male than female participants in the diet intervention study.

The contrasts have the exact same estimates as they had for the within-subject model in Table 4.2, the female participants' model in Table 4.3 and the male participants' model in Table 4.4.

### All participants with a weighted mean

By taking into account that there are more male than female participants in the diet intervention study, we can find the weighted mean linear functions and contrasts by using $sex = weight$. Where $weight$ is calculated by,

$$weight = \frac{\text{number of male participants}}{\text{number of male and female participants}} = \frac{19}{32} = 0.59375.$$

Hence, the weighted mean linear functions are

$$\begin{bmatrix} A0 \\ A6 \\ B0 \\ B6 \end{bmatrix} = \begin{bmatrix} 1 & 0.59375 & 0 & 0 & 0 \\ 1 & 0.59375 & 1 & 0 & 0 \\ 1 & 0.59375 & 0 & 1 & 0 \\ 1 & 0.59375 & 1 & 1 & 1 \end{bmatrix}.$$

That is, $A0 = \beta_{intercept} + 0.59375\beta_{sex}$, $A6 = \beta_{intercept} + 0.59375\beta_{sex} + \beta_{time}$, $B0 = \beta_{intercept} + 0.59375\beta_{sex} + \beta_{diet}$ and $B6 = \beta_{intercept} + 0.59375\beta_{sex} + \beta_{time} + \beta_{diet} + \beta_{time:diet}$.

By subtraction, we find that the weighted mean contrasts of interest are the same as for the male, female and the mean participant, given by Equation (4.5). That is, $A6 - A0 = \beta_{time}$, $B6 - B0 = \beta_{time} + \beta_{time:diet}$ and $(B6 - B0) - (A6 - A0) = \beta_{time:diet}$.

In this case, the sex factor is included in all the weighted mean linear functions, $A0$, $A6$, $B0$ and $B6$, and since the linear functions degrees of freedom is equal to the smallest degree of freedom among the set of parameters included in the linear function, the corresponding degrees of freedom should have been given by Equation (4.4), $DF = DF_{level1} = 30$. However, the results in Table 4.6 shows that the *estimable* function in $R$ weights the degrees of freedom. Hence, the degrees of freedom used for the linear functions, $A0$, $A6$, $B0$ and $B6$, in Table 4.6 is $30 \times 0.59375 = 17.81$.

For the three weighted mean contrasts however, sex is not included and so the contrast's degrees of freedom are given by Equation (4.2), $DF = DF_{level2} = 93$.

The results of the estimated weighted mean linear functions and contrasts can be seen in Table 4.6. Here we observe that all estimates are identical to the estimates made from the model only including within-subject factors in Table 4.2. The standard error is however a little smaller for the linear functions, $A0$, $A6$, $B0$ and $B6$. And from Equation (2.29), we know that a smaller standard error gives a larger t-value. Due to the fact that the t-values for the linear functions are so large it does not matter that the degrees of freedom are smaller than they should, because the linear combinations are still significant on a $\alpha = 0.05$ significance level.

The contrasts have the exact same estimates and standard error as they had for the within-subject model in Table 4.2, the female participants' model in Table 4.3, the male participants' model in Table 4.4 and the unweighted mean participant' model in Table 4.5.

| Contrast | Estimate | Std. Error | t value | DF | Pr(>\|t\|) |
|---|---|---|---|---|---|
| A0 | 607.50 | 55.14 | 11.02 | 18 | 2.21e-09 |
| A6 | 810.44 | 55.14 | 14.70 | 18 | 2.13e-11 |
| B0 | 566.66 | 55.14 | 10.28 | 18 | 6.51e-09 |
| B6 | 706.44 | 55.14 | 12.81 | 18 | 2.01e-10 |
| A6-A0 | 202.94 | 46.44 | 4.37 | 93 | 3.24e-05 |
| B6-B0 | 139.78 | 46.44 | 3.01 | 93 | 3.37e-03 |
| (B6-B0)-(A6-A0) | -63.16 | 65.68 | -0.96 | 93 | 3.39e-01 |

Table 4.6: Contrasts for the weighted mean participant for Model (4.3), using the treatment contrast coding for sex

### 4.2.2 Sum to zero contrast coding

Now we code the sex factor as "sum to zero", by using $male = -1$ and $female = 1$ in the design matrix for the fixed effects, $\mathbf{X}_i$. This coding is often used when you want to extract information about the mean, which given there are the same amount of male and female participant, is $sex = 0$.

#### Female participants

The linear functions, $A0$, $A6$, $B0$ and $B6$, for female participants, when the sex factor is coded as "sum to zero", are

$$\begin{bmatrix} A0 \\ A6 \\ B0 \\ B6 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

That is, $A0 = \beta_{intercept} + \beta_{sex}$, $A6 = \beta_{intercept} + \beta_{sex} + \beta_{time}$, $B0 = \beta_{intercept} + \beta_{sex} + \beta_{diet}$ and $B6 = \beta_{intercept} + \beta_{sex} + \beta_{time} + \beta_{diet} + \beta_{time:diet}$.

By subtraction we find that the contrasts of interest are the same as before, given by Equation (4.5). That is, $A6 - A0 = \beta_{time}$, $B6 - B0 = \beta_{time} + \beta_{time:diet}$ and $(B6 - B0) - (A6 - A0) = \beta_{time:diet}$.

In this case, the sex factor is included in all the linear functions, $A0$, $A6$, $B0$ and $B6$, and since the linear functions degrees of freedom is equal to the smallest degree of freedom among the set of parameters included in the linear function, the corresponding degrees of freedom is given by Equation (4.4), $DF = DF_{level1} = 30$.

For the three contrasts however, sex is not included and so the contrast's degrees of freedom are given by Equation (4.2), $DF = DF_{level2} = 93$.

| Contrast | Estimate | Std. Error | t value | DF | Pr(>\|t\|) |
|---|---|---|---|---|---|
| A0 | 673.43 | 79.39 | 8.48 | 30 | 1.82e-09 |
| A6 | 876.37 | 79.39 | 11.04 | 30 | 4.36e-12 |
| B0 | 632.59 | 79.39 | 7.97 | 30 | 6.81e-09 |
| B6 | 772.37 | 79.39 | 9.73 | 30 | 8.65e-11 |
| A6-A0 | 202.94 | 46.44 | 4.37 | 93 | 3.24e-05 |
| B6-B0 | 139.78 | 46.44 | 3.01 | 93 | 3.37e-03 |
| (B6-B0)-(A6-A0) | -63.16 | 65.68 | -0.96 | 93 | 3.39e-01 |

Table 4.7: Contrasts for the female participants for model (4.3), using the sum to zero contrast coding for sex.

The results of the estimated linear functions and contrasts can be seen in Table 4.7. Here we observe that both the linear functions and the contrasts are the exact same as in Table 4.3, for the female participants when sex was coded as "treatment".

### Male participants

The linear functions, $A0$, $A6$, $B0$ and $B6$, for male participants, when the sex factor is coded as "sum to zero", are

$$\begin{bmatrix} A0 \\ A6 \\ B0 \\ B6 \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 1 & -1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 1 & 0 \\ 1 & -1 & 1 & 1 & 1 \end{bmatrix}.$$

That is, $A0 = \beta_{intercept} - \beta_{sex}$, $A6 = \beta_{intercept} - \beta_{sex} + \beta_{time}$, $B0 = \beta_{intercept} - \beta_{sex} + \beta_{diet}$ and $B6 = \beta_{intercept} - \beta_{sex} + \beta_{time} + \beta_{diet} + \beta_{time:diet}$.

By subtraction we find that the contrasts of interest are the exact same as for the female participants, given by Equation (4.5). That is, $A6 - A0 = \beta_{time}$, $B6 - B0 = \beta_{time} + \beta_{time:diet}$ and $(B6 - B0) - (A6 - A0) = \beta_{time:diet}$.

In this case, the sex factor is included in all the linear functions, $A0$, $A6$, $B0$ and $B6$, and since the linear functions degrees of freedom is equal to the smallest degree of freedom among the set of parameters included in the linear function, the corresponding degrees of freedom is given by Equation (4.4), $DF = DF_{level1} = 30$.

For the three contrasts however, sex is not included and so the contrast's degrees of freedom are given by Equation (4.2) as for the female participants, $DF = DF_{level2} = 93$.

The results of the estimated linear functions and contrasts can be seen in Table 4.8. Here we observe that both the linear functions and the contrasts are the exact same as in Table 4.4, for the male participants when sex was coded as "treatment".

| Contrast | Estimate | Std. Error | t value | DF | Pr(>\|t\|) |
|---|---|---|---|---|---|
| A0 | 562.39 | 67.59 | 8.32 | 30 | 2.75e-09 |
| A6 | 765.32 | 67.59 | 11.32 | 30 | 2.34e-12 |
| B0 | 521.54 | 67.59 | 7.72 | 30 | 1.31e-08 |
| B6 | 661.32 | 67.59 | 9.78 | 30 | 7.57e-11 |
| A6-A0 | 202.94 | 46.44 | 4.37 | 93 | 3.24e-05 |
| B6-B0 | 139.78 | 46.44 | 3.01 | 93 | 3.37e-03 |
| (B6-B0)-(A6-A0) | -63.16 | 65.68 | -0.96 | 93 | 3.39e-01 |

Table 4.8: Contrasts for the male participants for Model (4.3), using the sum to zero contrast coding for sex.

### All participants

The linear functions, $A0$, $A6$, $B0$ and $B6$, for the mean participant, when the sex factor is coded as "sum to zero", are

$$\begin{bmatrix} A0 \\ A6 \\ B0 \\ B6 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \end{bmatrix}.$$

That is, $A0 = \beta_{intercept}$, $A6 = \beta_{intercept} + \beta_{time}$, $B0 = \beta_{intercept} + \beta_{diet}$ and $B6 = \beta_{intercept} + \beta_{time} + \beta_{diet} + \beta_{time:diet}$.

However this mean is not weighted, meaning that is does not take into account that there are more male than female participants in the diet intervention study.

By subtraction, we again find that the contrasts of interest given by Equation (4.5). That is, $A6 - A0 = \beta_{time}$, $B6 - B0 = \beta_{time} + \beta_{time:diet}$ and $(B6 - B0) - (A6 - A0) = \beta_{time:diet}$.

In this case, the sex factor is not included in any of the linear functions, $A0$, $A6$, $B0$ and $B6$. Hence, the degrees of freedom, for both the linear functions and the contrasts, are given by Equation (4.2), $DF = DF_{level2} = 93$.

The results of the estimated linear functions and contrasts can be seen in Table 4.9. Here we observe that all estimates are identical to the contrast estimates for the unweighted mean participant, using the treatment contrast coding for sex, in Table 4.5.

| Contrast | Estimate | Std. Error | t value | DF | Pr(>\|t\|) |
|---|---|---|---|---|---|
| A0 | 617.91 | 55.88 | 11.06 | 93 | 0.00e+00 |
| A6 | 820.85 | 55.88 | 14.69 | 93 | 0.00e+00 |
| B0 | 577.07 | 55.88 | 10.33 | 93 | 0.00e+00 |
| B6 | 716.85 | 55.88 | 12.83 | 93 | 0.00e+00 |
| A6-A0 | 202.94 | 46.44 | 4.37 | 93 | 3.24e-05 |
| B6-B0 | 139.78 | 46.44 | 3.01 | 93 | 3.37e-03 |
| (B6-B0)-(A6-A0) | -63.16 | 65.68 | -0.96 | 93 | 3.39e-01 |

Table 4.9: Contrasts for the mean participant for Model (4.3), using the sum to zero contrast coding for sex.

## 4.3 Interaction between within-subject and between-subject factors

We have also estimated contrasts for the loaded model, which is given by Equation (3.1) with the fixed effect vector given by

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{intercept} \\ \beta_{sex} \\ \beta_{time} \\ \beta_{diet} \\ \beta_{sex:time} \\ \beta_{sex:diet} \\ \beta_{time:diet} \\ \beta_{sex:time:diet} \end{bmatrix}. \tag{4.6}$$

The between-subject factor, sex, is still the only factor estimated at level 1. Hence, the degrees of freedom for factors estimated at level 1 is still given by Equation (4.2).

The amount of factors estimated at level 2, has however increased. In the fixed effect vector, (4.6), we count 6 within-subject factors. Hence, the degrees of freedom for factors estimated at level 2 is given by

$$DF_{level2} = m_2 - (m_1 + p_2) = 128 - (32 + 6) = 90. \tag{4.7}$$

The results of the estimated linear functions and contrasts, for the loaded model given by Equation (4.6), can be seen in Appendix A. We have chosen not to include all the results due to the fact that the conclusions are the same as for the other variations of the simplest model, given by Equation (4.1) and (4.3).

## 4.4 Contrast discussion

We have not succeeded in finding theoretical articles on conditional tests based on contrasts, which leaves the probability that we might have overlooked work in this field unintentionally. The only literature on degrees of freedom for contrasts we have found are connected to the following *R* packages, *gmodels* by Warnes (2011), *Design* by Harrell (2009) and *contrasts* by Kuhn, Weston, Wing and Forester (2010). Specially on degrees of freedom, there are variations between the different packages in *R*, which estimates and tests contrasts.

It is clearly undesirable that the linear functions, $A0$, $A6$, $B0$ and $B6$, have different degrees of freedom for male and female dependent on the parametrization of the between-subject factor, sex. For example, when comparing Table 4.3 and 4.7 we observe that the estimates and standard error of the linear functions are the same for the two models, but they have different degrees of freedom. We believe that the degrees of freedom for male or female linear functions, should be equal to the degrees of freedom for the sex factor which is estimated at level 1 and therefore given by Equation (4.4), $DF = DF_{level1} = 30$. The contrasts for male or female are however the difference between to linear functions for either female or male participants, and is therefore independent of the sex factor. Hence, we believe that contrasts should be equal to the degrees of freedom for the within-subject factors which are estimated at level 2 and therefore given by Equation (4.2), $DF = DF_{level2} = 93$.

We also believe that the weighting of degrees of freedom is the wrong approach for calculating degrees of freedom for the linear functions, $A0$, $A6$, $B0$ and $B6$ for the mean and the weighted mean in Table 4.5 and 4.6, respectively. We believe that the degrees of freedom here should be the same as for males and females. That is, the degrees of freedom should be equal to the smallest degrees of freedom among the parameters used to construct the linear function, which is the sex factor estimated at level 1 and therefore given by Equation (4.4), $DF = DF_{level1} = 30$.

Another complicating factor in the estimations of contrasts, is the interpretation of the mean and the weighted mean contrasts for all participants. If we compare the results in Table 4.2 and 4.6 we observe that the estimates are the same for both models, but the standard errors are slightly different. Hence, the weighted mean linear functions, with weights relative to the sample proportions of males and females, and contrasts are the overall mean of this data set.

If we compare the results in Table 4.5 and 4.9, we observe that both the estimates and the standard errors are equal for the two models. Hence, the unweighted mean linear functions and contrasts, using the treatment contrast coding, and the mean linear functions and contrasts, using sum to zero contrast coding, are the mean linear functions and contrasts for the entire population, when it is assumed to be a 50 percent chance of being male and a 50 percent chance of being female. This is however also true when fitting linear models, but for LM there are no within-subject correlation and so the degrees of freedom are equal for all factors in the model.

# Chapter 5

# The implied marginal model

In this chapter we will use the implied marginal variance-covariance matrix, $\mathbf{V}_i$, given by Equation (2.14), to get a deeper understanding of how the correlation in the data is structured in the covariance matrix associated with the random effects, $\mathbf{D}$, and covariance matrix associated with the residuals, $\mathbf{R}_i$.

From the implied marginal linear model, given by Equation (2.13), we saw that the linear mixed effect model can be written as a linear model with normally distributed residuals with a mean of zero and a implied marginal variance-covariance matrix, given as

$$\mathbf{V}_i = \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^T + \mathbf{R}_i.$$

For the diet intervention study example from Chapter 3, the loaded linear model is given as

$$\mathbf{y}_i^{lm} = \mathbf{X}_i \boldsymbol{\beta}^{lm} + \boldsymbol{\varepsilon}_i^{lm}, \tag{5.1}$$

where $\mathbf{X}_i$ is the design matrix for subject $i$,

$$\boldsymbol{\beta}^{lm} = \begin{bmatrix} \beta_{intercept} \\ \beta_{sex} \\ \beta_{time} \\ \beta_{diet} \\ \beta_{sex,time} \\ \beta_{sex,diet} \\ \beta_{time,diet} \\ \beta_{time,diet,sex} \end{bmatrix}$$

and $\boldsymbol{\varepsilon}_i^{lm} \sim N_{n_i}(0, \mathbf{V})$ is the residuals.

By fitting a loaded linear model to the data we can estimate the empirical variance-covariance matrix, $\hat{\mathbf{V}}$, of the residuals, by

$$\hat{\mathbf{V}} = \begin{bmatrix} \widehat{\mathrm{Var}}(\boldsymbol{\varepsilon}_{A0}) & \widehat{\mathrm{Cov}}(\boldsymbol{\varepsilon}_{A0}, \boldsymbol{\varepsilon}_{A6}) & \widehat{\mathrm{Cov}}(\boldsymbol{\varepsilon}_{A0}, \boldsymbol{\varepsilon}_{B0}) & \widehat{\mathrm{Cov}}(\boldsymbol{\varepsilon}_{A0}, \boldsymbol{\varepsilon}_{B6}) \\ \widehat{\mathrm{Cov}}(\boldsymbol{\varepsilon}_{A6}, \boldsymbol{\varepsilon}_{A0}) & \widehat{\mathrm{Var}}(\boldsymbol{\varepsilon}_{A6}) & \widehat{\mathrm{Cov}}(\boldsymbol{\varepsilon}_{A6}, \boldsymbol{\varepsilon}_{B0}) & \widehat{\mathrm{Cov}}(\boldsymbol{\varepsilon}_{A6}, \boldsymbol{\varepsilon}_{B6}) \\ \widehat{\mathrm{Cov}}(\boldsymbol{\varepsilon}_{B0}, \boldsymbol{\varepsilon}_{A0}) & \widehat{\mathrm{Cov}}(\boldsymbol{\varepsilon}_{B0}, \boldsymbol{\varepsilon}_{A6}) & \widehat{\mathrm{Var}}(\boldsymbol{\varepsilon}_{B0}) & \widehat{\mathrm{Cov}}(\boldsymbol{\varepsilon}_{B0}, \boldsymbol{\varepsilon}_{B6}) \\ \widehat{\mathrm{Cov}}(\boldsymbol{\varepsilon}_{B6}, \boldsymbol{\varepsilon}_{A0}) & \widehat{\mathrm{Cov}}(\boldsymbol{\varepsilon}_{B6}, \boldsymbol{\varepsilon}_{A6}) & \widehat{\mathrm{Cov}}(\boldsymbol{\varepsilon}_{B6}, \boldsymbol{\varepsilon}_{B0}) & \widehat{\mathrm{Var}}(\boldsymbol{\varepsilon}_{B6}) \end{bmatrix}. \tag{5.2}$$

In our analysis of the diet intervention study, we have looked at three potential structures for the the covariance matrix associated with the random effect, $\mathbf{D}$, and three potential structures for the covariance matrix associated with the residuals, $\mathbf{R}_i$. Hence, there are 9 potential structures for the implied marginal variance-covariance matrix, $\mathbf{V}_i$. Our aim in this chapter is to use the estimated empirical variance-covariance matrix, $\hat{\mathbf{V}}$, to predict which of these structures is the best fit for resistin, uric acid, triglyceride and visfatin.

## 5.1 The simplest form of a LME: Resistin

We start by fitting the resistin data to the loaded linear model, given in Equation (5.1), using the *lm* method by R Development Core Team (2010). Further, we estimate the residuals, $\boldsymbol{\varepsilon}_i^{lm}$, for all subjects $i$ and construct an empirical variance-covariance matrix, $\hat{\mathbf{V}}$, given by Equation (5.2). That is,

$$\hat{\mathbf{V}} = \begin{bmatrix} 70885 & 67273 & 49119 & 53353 \\ 67273 & 145457 & 61140 & 73381 \\ 49119 & 61140 & 83061 & 58777 \\ 53353 & 73381 & 58777 & 80372 \end{bmatrix}. \tag{5.3}$$

Here we observe that $\widehat{\text{Var}}(\boldsymbol{\varepsilon}_{A6})$ in element $(2,2)$ of the matrix, is greater than the other variances. Hence, the structure of the covariance matrix of the random effects, $\mathbf{D}$, nor the covariance matrix of the residuals, $\mathbf{R}_i$, are not likely to be associated with time diet or time.

The implied marginal linear model of the fitted resistin model, (3.7), with the fixed effect parameter given by Equation (3.8), is given as

$$\mathbf{Y}_i = \mathbf{X}_i \begin{bmatrix} \beta_{intercept} \\ \beta_{time} \\ \beta_{diet} \end{bmatrix} + \boldsymbol{\varepsilon}_i^{\star},$$

where

$$\boldsymbol{\varepsilon}_i^{\star} \sim N_{n_i}(\mathbf{0}, \mathbf{V}_i)$$

and the implied marginal variance-covariance matrix, $\mathbf{V}_i$, is given as

$$\begin{aligned} \mathbf{V}_i &= \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^T + \mathbf{R}_i \\ &= \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \sigma_{int}^2 [1111] + \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \sigma^2 \\ &= \begin{bmatrix} \sigma_{int}^2 + \sigma^2 & \sigma_{int}^2 & \sigma_{int}^2 & \sigma_{int}^2 \\ \sigma_{int}^2 & \sigma_{int}^2 + \sigma^2 & \sigma_{int}^2 & \sigma_{int}^2 \\ \sigma_{int}^2 & \sigma_{int}^2 & \sigma_{int}^2 + \sigma^2 & \sigma_{int}^2 \\ \sigma_{int}^2 & \sigma_{int}^2 & \sigma_{int}^2 & \sigma_{int}^2 + \sigma^2 \end{bmatrix}. \end{aligned}$$

By inserting the estimates from the fitted resistin model in Table 3.6, we find that the estimated implied marginal variance-covariance matrix, $\tilde{\mathbf{V}}_i$, is given as

$$\tilde{\mathbf{V}}_i = \begin{bmatrix} 98051 & 63565 & 63565 & 63565 \\ 63565 & 98051 & 63565 & 63565 \\ 63565 & 63565 & 98051 & 63565 \\ 63565 & 63565 & 63565 & 98051 \end{bmatrix}, \tag{5.4}$$

for all participants $i$.

In order to get a more thorough understanding of the variance and covariance in the resistin data, we take a closer look at the difference between the empirical variance-covariance matrix, $\hat{\mathbf{V}}$, and the implied marginal variance-covariance matrix, $\tilde{\mathbf{V}}_i$, given as

$$\hat{\mathbf{V}} - \tilde{\mathbf{V}}_i = \begin{bmatrix} -27166 & 3708 & -14446 & -10212 \\ 3708 & 47406 & -2426 & 9816 \\ -14446 & -2426 & -14990 & -4788 \\ -10212 & 9816 & -4788 & -17680 \end{bmatrix}.$$

Here we observe that the variance and covariance in the resistin data are not modeled particularly well. We notice that the variance for $A6$, in element $(2, 2)$ of the matrix, has the worst fit. Perhaps a better fit for the data would be a heterogeneous variance structure of $\mathbf{R}_i$, given as

$$\mathbf{R}_i = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2_{A0} & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & \sigma^2 \end{bmatrix}, \tag{5.5}$$

where $\sigma^2_{A0} > \sigma^2$, which allows the $A0$ resistin measurements to have a higher variance than the other measurements. This structure for the residual covariance matrix, $\mathbf{R}_i$, is however not investigated in this thesis.

## 5.2 Time associated random effect: Uric acid

We fit the uric acid data to the loaded linear model, given in Equation (5.1), using the *lm* method by R Development Core Team (2010). Due to the fact that there is one missing value in the uric acid data, we choose to remove all measurements from the participant with missing values in order to calculate the empirical variance-covariance matrix, $\hat{\mathbf{V}}$. The empirical variance-covariance matrix, $\hat{\mathbf{V}}$, is given as

$$\hat{\mathbf{V}} = \begin{bmatrix} 0.0248 & 0.0133 & 0.0191 & 0.0133 \\ 0.0133 & 0.0290 & 0.0130 & 0.0207 \\ 0.0191 & 0.0130 & 0.0222 & 0.0144 \\ 0.0133 & 0.0207 & 0.0144 & 0.0302 \end{bmatrix}. \tag{5.6}$$

Here we observe that $\widehat{\mathrm{Var}}(\boldsymbol{\varepsilon}_{A6})$ and $\widehat{\mathrm{Var}}(\boldsymbol{\varepsilon}_{B6})$, in element $(2, 2)$ and $(4, 4)$ respectively, are greater than the two other elements on the diagonal, which makes us believe that time most likely should be associated with either the random effects or the residuals. Since we also observe that both $\widehat{\mathrm{Cov}}(\boldsymbol{\varepsilon}_{A0}, \boldsymbol{\varepsilon}_{B0})$ and $\widehat{\mathrm{Cov}}(\boldsymbol{\varepsilon}_{A6}, \boldsymbol{\varepsilon}_{B6})$ are higher than the other covariances, time should most likely be associated with the random effects in a unstructured **D** matrix, given by Equation (2.5).

The implied marginal linear model of the fitted uric acid model, (3.9), with the fixed effect parameter given by Equation (3.10), is given as

$$\mathbf{Y}_i = \mathbf{X}_i \begin{bmatrix} \beta_{intercept} \\ \beta_{sex} \\ \beta_{time} \\ \beta_{diet} \end{bmatrix} + \boldsymbol{\varepsilon}_i^{\star},$$

where

$$\boldsymbol{\varepsilon}_i^{\star} \sim N_{n_i}(\mathbf{0}, \mathbf{V}_i)$$

and the implied marginal variance-covariance matrix, $\mathbf{V}_i$, is given as

$$\begin{aligned}
\mathbf{V}_i &= \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^T + \mathbf{R}_i \\
&= \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \sigma_{int}^2 & \sigma_{int,time} \\ \sigma_{int,time} & \sigma_{time}^2 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \sigma^2 \\
&= \begin{bmatrix} \sigma_{int}^2 + \sigma^2 & \sigma_{int}^2 + \sigma_{int,time} & \sigma_{int}^2 & \sigma_{int}^2 + \sigma_{int,time} \\ \sigma_{int}^2 + \sigma_{int,time} & \sigma_{int}^2 + 2\sigma_{int,time} + \sigma_{time}^2 + \sigma^2 & \sigma_{int}^2 + \sigma_{int,time} & \sigma_{int}^2 + 2\sigma_{int,time} + \sigma_{time}^2 \\ \sigma_{int}^2 & \sigma_{int}^2 + \sigma_{int,time} & \sigma_{int}^2 + \sigma^2 & \sigma_{int}^2 + \sigma_{int,time} \\ \sigma_{int}^2 + \sigma_{int,time} & \sigma_{int}^2 + 2\sigma_{int,time} + \sigma_{time}^2 & \sigma_{int}^2 + \sigma_{int,time} & \sigma_{int}^2 + 2\sigma_{int,time} + \sigma_{time}^2 + \sigma^2 \end{bmatrix}.
\end{aligned}$$

By inserting the estimates from the fitted uric acid model in Table 3.9, we find that the estimated implied marginal variance-covariance matrix, $\tilde{\mathbf{V}}_i$, is given as

$$\tilde{\mathbf{V}}_i = \begin{bmatrix} 0.0249 & 0.0181 & 0.0182 & 0.0181 \\ 0.0181 & 0.0384 & 0.0181 & 0.0317 \\ 0.0182 & 0.0181 & 0.0249 & 0.0181 \\ 0.0181 & 0.0317 & 0.0181 & 0.0384 \end{bmatrix}, \tag{5.7}$$

for all participants $i$.

Hence,

$$\hat{\mathbf{V}} - \tilde{\mathbf{V}}_i = \begin{bmatrix} -1.09e{-}04 & -4.78e{-}03 & 9.66e{-}04 & -4.81e{-}03 \\ -4.78e{-}03 & -9.47e{-}03 & -5.17e{-}03 & -1.10e{-}02 \\ 9.66e{-}04 & -5.17e{-}03 & -2.69e{-}03 & -3.74e{-}03 \\ -4.81e{-}03 & -1.10e{-}02 & -3.74e{-}03 & -8.22e{-}03 \end{bmatrix}.$$

Here we observe that the variances and covariances in the data are modeled very well.

## 5.3 Time associated residual variances: Triglyceride

Fitting the triglyceride data to the loaded linear model, given in Equation (5.1), we find that the empirical variance-covariance matrix, $\hat{\mathbf{V}}$, is given as

$$\hat{\mathbf{V}} = \begin{bmatrix} 0.0968 & 0.0537 & 0.0542 & 0.0496 \\ 0.0537 & 0.1032 & 0.0791 & 0.0636 \\ 0.0542 & 0.0791 & 0.1507 & 0.0758 \\ 0.0496 & 0.0636 & 0.0758 & 0.0894 \end{bmatrix}. \tag{5.8}$$

Here we observe that $\widehat{\mathrm{Var}}(\boldsymbol{\varepsilon}_{B0})$, in element $(3,3)$, is greater than the other three elements on the diagonal, and that both $\widehat{\mathrm{Cov}}(\boldsymbol{\varepsilon}_{A6}, \boldsymbol{\varepsilon}_{B0})$ and $\widehat{\mathrm{Cov}}(\boldsymbol{\varepsilon}_{B0}, \boldsymbol{\varepsilon}_{B6})$ are higher than the other covariances. This does not look like any of the structures we have tested, perhaps a better fit for the triglyceride data would be a unstructured covariance matrix $\mathbf{D}$, associated with measurement $B0$,

$$\mathbf{D} = \begin{bmatrix} \sigma^2_{int} & \sigma_{int,B0} \\ \sigma_{int,B0} & \sigma^2_{B0} \end{bmatrix}. \tag{5.9}$$

This is however not one of our three potential structures of the covariance matrix $\mathbf{D}$. We do however assume that the covariance should be associated to either time or diet in the covariance matrix associated with the residuals, $\mathbf{R}_i$, because $\widehat{\mathrm{Var}}(\boldsymbol{\varepsilon}_{B0})$ is greater than the three other elements on the diagonal. This might however lead to a overestimation of either $\widehat{\mathrm{Var}}(\boldsymbol{\varepsilon}_{A0})$ or $\widehat{\mathrm{Var}}(\boldsymbol{\varepsilon}_{B6})$, respectively.

The implied marginal linear model of the fitted triglyceride model, (3.11), with the fixed effect parameter given by Equation (3.12), is given as

$$\mathbf{Y}_i = \mathbf{X}_i \begin{bmatrix} \beta_{intercept} \\ \beta_{time} \\ \beta_{diet} \end{bmatrix} + \boldsymbol{\varepsilon}^\star_i,$$

where

$$\boldsymbol{\varepsilon}^\star_i \sim N_{n_i}(\mathbf{0}, \mathbf{V}_i)$$

and the implied marginal variance-covariance matrix, $\mathbf{V}_i$, is given as

$$\begin{aligned} \mathbf{V}_i &= \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^T + \mathbf{R}_i \\ &= \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \sigma^2_{int} [1111] + \begin{bmatrix} \sigma^2_{time=0} & 0 & 0 & 0 \\ 0 & \sigma^2_{time=1} & 0 & 0 \\ 0 & 0 & \sigma^2_{time=0} & 0 \\ 0 & 0 & 0 & \sigma^2_{time=1} \end{bmatrix} \\ &= \begin{bmatrix} \sigma^2_{int} + \sigma^2_{time=0} & \sigma^2_{int} & \sigma^2_{int} & \sigma^2_{int} \\ \sigma^2_{int} & \sigma^2_{int} + \sigma^2_{time=1} & \sigma^2_{int} & \sigma^2_{int} \\ \sigma^2_{int} & \sigma^2_{int} & \sigma^2_{int} + \sigma^2_{time=0} & \sigma^2_{int} \\ \sigma^2_{int} & \sigma^2_{int} & \sigma^2_{int} & \sigma^2_{int} + \sigma^2_{time=1} \end{bmatrix}. \end{aligned}$$

By inserting the estimates from the fitted triglyceride model in Table 3.12, we find that the estimated implied marginal variance-covariance matrix, $\tilde{\mathbf{V}}_i$, is given as

$$\tilde{\mathbf{V}}_i = \begin{bmatrix} 0.1282 & 0.0634 & 0.0634 & 0.0634 \\ 0.0634 & 0.0956 & 0.0634 & 0.0634 \\ 0.0634 & 0.0634 & 0.1282 & 0.0634 \\ 0.0634 & 0.0634 & 0.0634 & 0.0956 \end{bmatrix}, \tag{5.10}$$

for all participants $i$.

Hence,

$$\hat{\mathbf{V}} - \tilde{\mathbf{V}}_i = \begin{bmatrix} -3.14e{-}02 & -9.68e{-}03 & -9.17e{-}03 & -1.37e{-}02 \\ -9.68e{-}03 & 7.55e{-}03 & 1.57e{-}02 & 2.61e{-}04 \\ -9.17e{-}03 & 1.57e{-}02 & 2.25e{-}02 & 1.24e{-}02 \\ -1.37e{-}02 & 2.61e{-}04 & 1.24e{-}02 & -6.20e{-}03 \end{bmatrix}.$$

Here we observe that the variances and covariances in the data are modeled fairly well.

## 5.4 Diet associated random effect and residual variances: Visfatin

Fitting the visfatin data to the loaded linear model, given in Equation (5.1), we have to remove all participants with missing values in order to calculate the empirical variance-covariance matrix, $\hat{\mathbf{V}}$. The empirical variance-covariance matrix, $\hat{\mathbf{V}}$, is given as

$$\hat{\mathbf{V}} = \begin{bmatrix} 1.1744 & 0.4241 & 0.2231 & 0.2244 \\ 0.4241 & 1.2766 & -0.0201 & -0.0901 \\ 0.2231 & -0.0201 & 0.4919 & 0.3904 \\ 0.2244 & -0.0901 & 0.3904 & 0.5485 \end{bmatrix}. \tag{5.11}$$

Here we observe that $\widehat{\text{Var}}(\boldsymbol{\varepsilon}_{A0})$ and $\widehat{\text{Var}}(\boldsymbol{\varepsilon}_{A6})$, in element $(1,1)$ and $(2,2)$ respectively, are greater than the other two elements on the diagonal. Hence, we assume that diet should be associated with either the random effects or the residuals. Since $\widehat{\text{Cov}}(\boldsymbol{\varepsilon}_{A0}, \boldsymbol{\varepsilon}_{A6})$ and $\widehat{\text{Cov}}(\boldsymbol{\varepsilon}_{B0}, \boldsymbol{\varepsilon}_{B6})$ are greater than the other covariances in the matrix, we assume that diet should be associated with the random effects in a unstructured $\mathbf{D}$ matrix, given in Equation (2.5).

The implied marginal linear model of the fitted visfatin model, (3.13), with the fixed effect parameter given by Equation (3.14), is given as

$$\mathbf{Y}_i = \mathbf{X}_i \begin{bmatrix} \beta_{intercept} \\ \beta_{time} \\ \beta_{diet} \end{bmatrix} + \boldsymbol{\varepsilon}_i^{\star},$$

where

$$\boldsymbol{\varepsilon}_i^{\star} \sim N_{n_i}(\mathbf{0}, \mathbf{V}_i)$$

and the implied marginal variance-covariance matrix, $\mathbf{V}_i$, is given as

$$\begin{aligned}
\mathbf{V}_i &= \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T + \mathbf{R}_i \\
&= \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \sigma^2_{int} & \sigma_{int,diet} \\ \sigma_{int,diet} & \sigma^2_{diet} \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} + \begin{bmatrix} \sigma^2_{diet=A} & 0 & 0 & 0 \\ 0 & \sigma^2_{diet=A} & 0 & 0 \\ 0 & 0 & \sigma^2_{diet=B} & 0 \\ 0 & 0 & 0 & \sigma^2_{diet=B} \end{bmatrix} \\
&= \begin{bmatrix} \sigma^2_{int}+\sigma^2_{diet=A} & \sigma^2_{int} & \sigma^2_{int}+\sigma_{int,diet} & \sigma^2_{int}+\sigma_{int,diet} \\ \sigma^2_{int} & \sigma^2_{int}+\sigma^2_{diet=A} & \sigma^2_{int}+\sigma_{int,diet} & \sigma^2_{int}+\sigma_{int,diet} \\ \sigma^2_{int}+\sigma_{int,diet} & \sigma^2_{int}+\sigma_{int,diet} & \sigma^2_{int}+2\sigma_{int,diet}+\sigma^2_{diet}+\sigma^2_{diet=B} & \sigma^2_{int}+2\sigma_{int,diet}+\sigma^2_{diet} \\ \sigma^2_{int}+\sigma_{int,diet} & \sigma^2_{int}+\sigma_{int,diet} & \sigma^2_{int}+2\sigma_{int,diet}+\sigma^2_{diet} & \sigma^2_{int}+2\sigma_{int,diet}+\sigma^2_{diet}+\sigma^2_{diet=B} \end{bmatrix}.
\end{aligned}$$

By inserting the estimates from the fitted visfatin model in Table 3.15, we find that the estimated implied marginal variance-covariance matrix, $\tilde{\mathbf{V}}_i$, is given as

$$\tilde{\mathbf{V}}_i = \begin{bmatrix} 1.1972 & 0.3988 & 0.2640 & 0.2640 \\ 0.3988 & 1.1972 & 0.2640 & 0.2640 \\ 0.2640 & 0.2640 & 0.9524 & 0.7678 \\ 0.2640 & 0.2640 & 0.7678 & 0.9524 \end{bmatrix}, \tag{5.12}$$

for all participants $i$.

Hence,

$$\hat{\mathbf{V}} - \tilde{\mathbf{V}}_i = \begin{bmatrix} -2.28e-02 & 2.53e-02 & -4.09e-02 & -3.97e-02 \\ 2.53e-02 & 7.93e-02 & -2.84e-01 & -3.54e-01 \\ -4.09e-02 & -2.84e-01 & -4.60e-01 & -3.77e-01 \\ -3.97e-02 & -3.54e-01 & -3.77e-01 & -4.04e-01 \end{bmatrix}.$$

Here we observe that that the variances and covariances in the data are modeled fairly well. However we do observe that $\widehat{\text{Var}}(\boldsymbol{\varepsilon}_{B0})$ and $\widehat{\text{Var}}(\boldsymbol{\varepsilon}_{B6})$, in element $(3,3)$ and $(4,4)$ respectively, have the lowest difference and is perhaps slightly overestimated.

## 5.5 Discussion

Our motivation for examining the implied marginal variance-covariance matrix, $\mathbf{V}_i$, is the fact that it is hard to get a understanding of $\mathbf{V}_i$ by just studying the covariance matrix associated with the random effects, $\mathbf{D}$, and covariance matrix associated with the residuals, $\mathbf{R}_i$.

We have observed that the empirical variance-covariance matrix, $\hat{\mathbf{V}}$, and the estimated implied marginal variance-covariance matrix, $\tilde{\mathbf{V}}_i$, helps us get insight into the covariance structure of the data.

After examining the empirical variance-covariance matrix, $\hat{\mathbf{V}}$, for the four biomarkers, we have seen that other structures of the covariance matrix associated with the random effects, $\mathbf{D}$, or covariance matrix associated with the residuals, $\mathbf{R}_i$, might be of interest. We therefore believe that it perhaps would be useful to examine the empirical variance-covariance matrix, $\hat{\mathbf{V}}$, prior to choosing which covariance structures to test in the top-down strategy. Since the empirical variance-covariance matrix might give strong indications on the structure, we can save time by not testing structures which fits badly.

# Chapter 6

# Intraclass correlation

The intraclass correlation, called the ICC, is a term often used in biology and medical analysis. According to Shrout and Fleiss (1979) the ICC is the correlation between one measurement on a target and another measurement obtained on that target. The motivation behind finding the ICC is to assess the amount of error due to specific factors. There are several versions of the ICC, and so it is important to report which version one uses. The ICC can also be seen as a generalized correlation coefficient.

For the simplest form for a LME, exemplified by the resistin measurements from the diet intervention study in Chapter 3, the intraclass correlation is given as

$$\text{ICC} = \frac{\sigma_{int}^2}{\sigma_{int}^2 + \sigma^2}. \tag{6.1}$$

In linear mixed effects models with more complex structures of the covariance matrix associated with the random effects, $\mathbf{D}$, and covariance matrix associated with the residuals, $\mathbf{R}_i$, the ICC formula, given in Equation (6.1), can not be used. Using the estimated implied variance-covariance matrix, $\tilde{\mathbf{V}}_i$, we can very easily calculate one version of the ICC. Hence, for the LME we define the ICC as

$$\text{ICC}_{M1,M2} = \frac{\widehat{\text{Cov}}(\boldsymbol{\varepsilon}_{M1}, \boldsymbol{\varepsilon}_{M2})}{\sqrt{\widehat{\text{Var}}(\boldsymbol{\varepsilon}_{M1})}\sqrt{\widehat{\text{Var}}(\boldsymbol{\varepsilon}_{M2})}}, \tag{6.2}$$

where $M1$ is one measurement on a target and $M2$ is another measurement obtained on that target.

Hence, from the estimated implied marginal variance-covariance matrix, $\tilde{\mathbf{V}}_i$, in Chapter 5, we can calculate the intraclass correlations corresponding to the four biomarkers resistin, uric acid, triglyceride and visfatin, in the diet intervention study.

# ICCs for the resistin data

Since resistin has been modeled by the simplest form of a LME, we can easily with the results from Table 3.6 calculate the intraclass correlation given by Equation (6.1). That is,

$$ICC = \frac{63565.29}{63565.29 + 34486.17} = 0.65$$

Due to the structure of the estimated implied marginal variance-covariance matrix, $\tilde{\mathbf{V}}_i$, of the fitted resistin model, given in Equation (5.4), all ICCs are given as

$$\text{ICC}_{M1,M2} = \frac{63565}{\sqrt{98051} \times \sqrt{98051}} = 0.65,$$

where $M1$ and $M2$ are all combinations of $A0$, $A6$, $B0$ and $B6$. Hence, for the simplest model the two versions of the ICC, given in Equation (6.1) and (6.2), gives the same results.

# ICCs for the uric acid data

From Equation (5.7), the intraclass correlations corresponding to the fitted uric acid model (3.9), with the fixed effect vector given by Equation (3.10), are given as

$$\begin{aligned}\text{ICC}_{A0,A6} = \text{ICC}_{A0,B6} = \text{ICC}_{A6,B0} = \text{ICC}_{B0,B6} &= \frac{0.0181}{\sqrt{0.0249} \times \sqrt{0.0384}} = 0.59,\\ \text{ICC}_{A0,B0} &= \frac{0.0182}{\sqrt{0.0249} \times \sqrt{0.0249}} = 0.73\\ \text{and } \text{ICC}_{A6,B6} &= \frac{0.0317}{\sqrt{0.0384} \times \sqrt{0.0384}} = 0.83.\end{aligned}$$

# ICCs for the triglyceride data

From Equation (5.10), the intraclass correlations corresponding to the fitted triglyceride model (3.11), with the fixed effect vector given by Equation (3.12), are given as

$$\begin{aligned}\text{ICC}_{A0,A6} = \text{ICC}_{A0,B6} = \text{ICC}_{A6,B0} = \text{ICC}_{B0,B6} &= \frac{0.0634}{\sqrt{0.1282} \times \sqrt{0.0956}} = 0.57,\\ \text{ICC}_{A0,B0} &= \frac{0.0634}{\sqrt{0.1282} \times \sqrt{0.1282}} = 0.49\\ \text{and } \text{ICC}_{A6,B6} &= \frac{0.0634}{\sqrt{0.0956} \times \sqrt{0.0956}} = 0.66.\end{aligned}$$

# ICCs for the visfatin data

From Equation (5.12), the intraclass correlations corresponding to the fitted visfatin model (3.13), with the fixed effect vector given by Equation (3.14), are given as

$$\begin{aligned}\mathrm{ICC}_{A0,B0} = \mathrm{ICC}_{A0,B6} = \mathrm{ICC}_{A6,B0} = \mathrm{ICC}_{A6,B6} &= \frac{0.2640}{\sqrt{1.1972} \times \sqrt{0.9524}} = 0.25, \\ \mathrm{ICC}_{A0,A6} &= \frac{0.3988}{\sqrt{1.1972} \times \sqrt{1.1972}} = 0.33 \\ \text{and } \mathrm{ICC}_{B0,B6} &= \frac{0.7678}{\sqrt{0.9524} \times \sqrt{0.9524}} = 0.81.\end{aligned}$$

# Chapter 7

# Integrated nested Laplace approximations (INLA)

Integrated nested Laplace approximations, INLA, is a method for Bayesian inference on latent Gaussian models, which combines Laplace approximations and numerical integration in a very efficient manner, according to Rue, Martino and Chopin (2009).

We will now fit the four models for resistin, uric acid, triglycerides and visfatin using INLA and compare the results to the results from the *lme* fit in Chapter 3. For each model parameter we present median in the posterior distribution, the estimate, and the lower and upper 2.5% percentile of the posterior distribution, the credibility interval.

## 7.1 The simplest form of a LME: Resistin

We fit the resistin model (3.7), with the fixed effect vector given by Equation (3.8), using the *inla* function in R by Rue and Martino (2009). The results can be seen in Table 7.1, next to the results we got estimating the same model with *lme*. When comparing the two results, we can see that the estimates are very different.

| | *inla* estimate<br>(95% Credibility interval) | *lme* estimate<br>(95% Confidence interval) |
|---|---|---|
| **Fixed effects** | | |
| Intercept, $\beta_{int}$ | 660.73 (593.72, 727.97) | 623.29 (518.32, 728.25) |
| Time, $\beta_{time}$ | 41.69 (-13.07, 95.97) | 171.36 (106.18, 236.54) |
| Diet, $\beta_{diet}$ | -17.61 (-71.67, 36.53) | -72.42 (-137.60, -7.24) |
| **Random effects** | | |
| Intercept, $\sigma^2_{int}$ | 5.3e-05 (1.5e-05, 0.00061) | 252.12 (189.91, 334.71) |
| **Residuals** | | |
| Intercept, $\sigma^2$ | 99843.84 (79064.67, 128811.89) | 185.70 (160.97, 214.24) |

Table 7.1: Estimated parameters for the resistin model using *inla*, compared to the results from *lme*.

Since the $QQ$-plot of the simplest LME model of the resistin data were quite similar for both the original and natural logarithmic form, we choose to try fitting the resistin measurements on a natural logarithmic scale instead. The results of all the hypothesis' can be seen in Table7.2. Here we see that the best fit for the resistin measurements is still the simplest LME model, (3.7).

| | Hyp 3.3 | Hyp 3.4 | Hyp 3.5 | Hyp 3.6 |
|---|---|---|---|---|
| P-value | 0.5674 | 0.0658 | 0.1447 | 0.2788 |

Table 7.2: P-values of all hypothesis for the resistin data on a natural logarithmic scale.

Following the top-down strategy, we reduce the loaded model by preforming type I F-tests iteratively, using the ANOVA method by R Development Core Team (2010). Here we find that on a significance level $\alpha = 0.05$, only the intercept and time factor should be included as fixed effects in the model. Hence, the fixed effect vector is given by

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{intercept} \\ \beta_{time} \end{bmatrix}. \tag{7.1}$$

We fit the new resistin model (3.7), with the fixed effect vector given by Equation (7.1), using the *inla* method by Rue and Martino (2009). The results can be seen in Table 7.3, next to the results of fitting the same model with *lme*. When comparing the two results, we see that the estimates are almost identical. We find the largest differences in the random-effect term, $\sigma^2_{int}$. But comparing the *lme* 95% confidence interval with the *inla* 95% credibility interval, we see that these intervals overlap for all parameters.

| | *inla* estimate<br>(95% Credibility interval) | *lme* estimate<br>(95% Confidence interval) |
|---|---|---|
| **Fixed effects** | | |
| Intercept, $\beta_{int}$ | 6.2525 (6.0991, 6.4058) | 6.2525 (6.0960, 6.4090) |
| Time, $\beta_{time}$ | 0.2765 (0.1673, 0.3856) | 0.2765 (0.1663, 0.3867) |
| **Random effects** | | |
| Intercept, $\sigma^2_{int}$ | 0.1338 (0.0795, 0.2476) | 0.14958 (0.0836, 0.2676) |
| **Residuals** | | |
| Intercept, $\sigma^2$ | 0.0969 (0.0744, 0.1299) | 0.0987 (0.0742, 0.1311) |

Table 7.3: Estimated parameters for the resistin model on a logarithmic scale using *inla*, compared to the results from *lme*.

It is not completely clear why the *inla* results on the original scale, given in Table 7.1, are so different from the *lme* results, while the results on the natural logarithmic scale, given in Table 7.3, agree very well. In particular, because the LME model on the original scale for the resistin data seem to fit well according to the diagnostic plots in Figure 3.2, 3.3 and 3.4. A possible reason for the differences in Table 7.1 could be that there are more than one solution. That is if the model is unidentifiable.

## 7.2 Time associated random effect: Uric acid

In Table 7.4 we observe the results from fitting the uric acid model (3.9), with the fixed effect vector given by Equation (3.10), using both *inla* and *lme* in R. Comparing the two results, we see that the estimates of the fixed effects are very similar, but that there are some differences in the random effects as well as in the residuals.

For the fixed effects all *lme* 95% confidence interval and *inla* 95% credibility interval overlap. This is also the case for the estimated residual variance, $\sigma^2$. The estimated correlation parameter, $\rho_{int,time}$ in the covariance matrix corresponding to the random effects, $\mathbf{D}$, is in both *lme* and *inla* negative and the *lme* 95% confidence interval and *inla* 95% credibility interval overlap. But for the estimated variance of intercepts, $\sigma^2_{int}$, and time, $\sigma^2_{time}$, the differences are larger. This might suggest a possible non uniqueness in partitioning of variance for intercepts, $\sigma^2_{int}$, and time, $\sigma^2_{time}$.

| | *inla* estimate<br>(95% Credibility interval) | *lme* estimate<br>(95% Confidence interval) |
|---|---|---|
| **Fixed effects** | | |
| Intercept, $\beta_{int}$ | 5.5473 (5.3926, 5.7014) | 5.5643 (5.4851, 5.6436) |
| Sex, $\beta_{sex}$ | 0.2100 (0.0163, 0.4037) | 0.1814 (0.0802, 0.2825) |
| Time, $\beta_{time}$ | -0.0962 (-0.1776, -0.0150) | -0.0962 (-0.1464, -0.0460) |
| Diet, $\beta_{diet}$ | 0.0463 (0.0182, 0.0743) | 0.0463 (0.0173, 0.0752) |
| **Random effects** | | |
| Intercept, $\sigma^2_{int}$ | 0.0736 (0.0462, 0.1280) | 0.0182 (0.0100, 0.0332) |
| Time, $\sigma^2_{time}$ | 0.0460 (0.0296, 0.0766) | 0.0137 (0.0064, 0.0294) |
| Intercept:Time, $\rho_{int,time}$ | -0.3660 (-0.6336, -0.0634) | -0.2530 (-0.6187, 0.2029) |
| **Residuals** | | |
| Intercept, $\sigma^2$ | 0.0062 (0.0046, 0.0088) | 0.0067 (0.0047, 0.0095) |

Table 7.4: Estimated parameters for the uric acid model using *inla*, compared to the results from *lme*.

## 7.3 Time associated residual variances: Triglyceride

We fit Model 3.11, with the fixed effect vector given by Equation 3.12, using the *inla* function in R. The results can be seen in Table 7.5 next to the results we got estimating the same model with *lme*. When comparing the two results, we can see that all the estimates are practically the same.

| | *inla* estimate<br>(95% Credibility interval) | *lme* estimate<br>(95% Confidence interval) |
|---|---|---|
| **Fixed effects** | | |
| Intercept, $\beta_{int}$ | -0.0625 (-0.1751, 0.0507) | -0.0623 (-0.1769, 0.0523) |
| Time, $\beta_{time}$ | -0.2228 (-0.2992, -0.1464) | -0.2228 (-0.3001, -0.1454) |
| Diet, $\beta_{diet}$ | -0.1272 (-0.2001, -0.0523) | -0.1277 (-0.2005, -0.0548) |
| **Random effects** | | |
| Intercept, $\sigma^2_{int}$ | 0.0568 (0.0336, 0.1050) | 0.0634 (0.0353, 0.1138) |
| **Residuals** | | |
| Time, $\sigma^2_{Day0}$ | 0.0619 (0.0430, 0.0935) | 0.0649 (0.0435, 0.0967) |
| Time, $\sigma^2_{Day6}$ | 0.0306 (0.0203, 0.0491) | 0.0322 (0.0114, 0.0908) |

Table 7.5: Estimated parameters for the triglyceride model using *inla*, compared to the results from *lme*.

## 7.4 Diet associated random effect and residual variances: Visfatin

We fit Model 3.13, with the fixed effect vector given by Equation 3.14, using the *inla* function in R. The results can be seen in Table 7.6 next to the results we got estimating the same model with *lme*. When comparing the two results, we can see that the estimates of the fixed effects are quite similar, but that there are some differences both in the random effects and in the residuals.

The *lme* 95% confidence interval and *inla* 95% credibility interval overlap for all parameters. But in particular, the estimated variance of diet, $\sigma^2_{diet}$, the differences are quite large. This might suggest a possible non uniqueness in partitioning of variance.

| | *inla* estimate<br>(95% Credibility interval) | *lme* estimate<br>(95% Confidence interval) |
|---|---|---|
| **Fixed effects** | | |
| Intercept, $\beta_{int}$ | 7.8447 (7.5014, 8.1871) | 7.8503 (7.5103, 8.1902) |
| Time, $\beta_{time}$ | -0.2687 (-0.4662, -0.0755) | -0.2744 (-0.4693, -0.0794) |
| Diet, $\beta_{diet}$ | -0.6653 (-1.0213, -0.3106) | -0.6680 (-1.0532, -0.2829) |
| **Random effects** | | |
| Intercept, $\sigma^2_{int}$ | 0.3812 (0.0896, 2.3325) | 0.3988 (0.1222, 1.3017) |
| Diet, $\sigma^2_{diet}$ | 0.2410 (0.0741, 0.8822) | 0.6385 (0.2376, 1.7163) |
| Intercept:Diet, $\rho_{int,diet}$ | -0.6682 (-0.9214, -0.0230) | -0.5292 (-0.8537, 0.0914) |
| **Residuals** | | |
| Diet, $\sigma^2_{DietA}$ | 0.9183 (0.5457, 1.7011) | 0.7985 (0.4695, 1.3578) |
| Diet, $\sigma^2_{DietB}$ | 0.1730 (0.1132, 0.2887) | 0.1846 (0.0523, 0.6519) |

Table 7.6: Estimated parameters for the visfatin model using *inla*, compared to the results from *lme*.

## 7.5 Discussion

With one exception, being the resistin model on the original scale, the fixed effect *lme* parameter estimates and the *inla* parameter estimates, the median of the posterior distribution, agree very well. For two of the data sets fitted, resistin on the natural logarithmic scale and triglyceride, the random effects and the residual parameters agree well. But for the other three data sets, there seem to be more than one way to distribute variances across parameter.

# Chapter 8

# Discussion and conclusion

In this master thesis we have presented and discussed the linear mixed effects model for analyzing repeated measures data. We have seen that the covariance matrices for the random effects and the residuals can vary in complexity, which can make it difficult to interpret the results of a estimated LME.

We have studied several statistical aspects of the linear mixed effects model, including estimation of contrasts, investigations of the structure the implied marginal variance-covariance matrix, the intraclass correlation and using integrated nested Laplace approximations (INLA) to fit a LME.

We have chosen to follow the top-down strategy for model selection, according to Chapter 2 of West et al. (2007). Here only three potential structures of the covariance matrix for the random effects, $\mathbf{D}$, and three potential structures of the covariance matrix for the residuals, $\mathbf{R}_i$, are tested. We have observed that there are several other covariance structures for both $\mathbf{D}$ and $\mathbf{R}_i$, which could have been tested and perhaps given a better fitted model to the biomarkers in the diet intervention study. Perhaps it would be useful to examine the empirical variance-covariance matrix, $\hat{\mathbf{V}}$, prior to choosing which covariance structures to test, since the empirical variance-covariance matrix might give strong indications on the structure.

We have seen that writing down the estimated implied marginal variance-covariance matrix, $\tilde{\mathbf{V}}_i$, and the $\text{ICC}_{M1,M2}$ gain insight into the estimated LME.

We have in this thesis work looked at using the LME to analyze repeated measures data where multiple observations are recorded for each subject, thus a two-level analysis. The simplest solution to the challenge of correlated observations within subjects is to include a random intercept in the covariance matrix for the random effects, $\mathbf{D}$. This solution is easy to understand and for this model the ICC has a clear interpretation. We have looked at fitting more complex structures for the covariance matrix for the random effects, $\mathbf{D}$, and the covariance matrix for the residuals, $\mathbf{R}_i$, leading to a more difficult interpretation of the ICC and the variance-covariance structure of the LME. Is the added complexity by incorporating a more elaborate covariance matrix for the random effects, $\mathbf{D}$, and the residuals, $\mathbf{R}_i$, a sound investment? Does it change the conclusions to inferential questions asked? Is the most important mission of the covariance matrix for the random effects, $\mathbf{D}$, just to model the intercept? Is it necessary to introduce correlation between random effects? We have only looked at a two-level repeated measures data set, perhaps more complex structures of the covariance matrices, $\mathbf{D}$ and $\mathbf{R}_i$ is more useful in LMEs with more nested levels?

When between-subject factors are not present in the LME, hypothesis tests of linear contrasts are to our understanding handled satisfyingly with the *estimable* function in the *gmodels* package in *R*. But, when between-subject factors are included we have seen that the strategy of using the smallest degrees of freedom among the parameter estimates included in the contrast, leads to ambiguous results dependent on the coding of the parameters in the design matrix. This is highly unsatisfying.

In retrospect, there are so many interesting topics connected to the linear mixed effects model, and we see that we have tried to cover too many topics to be able to cover all in great depth. Thus the thesis contains less work than desired on the integrated nested Laplace approximations (INLA) in Chapter 7. Nevertheless, the results from Chapter 7 inspires further research. The Bayesian approach is not often used for LME, but for generalized linear mixed effects models the method has become popular, Fong, Rue and Wakefield (2010). Using INLA for inference in LME would produce credibility intervals for contrasts as easily as for parameters. Thus, the degrees of freedom problems with between-subject factors we encountered using the frequentist approach, would not be an issue.

# Bibliography

Arbo, I., Brattbakk, H.-R., Langaas, M., Kuiper, M., Lindberg, F. A., Kulseng, B. and Johansen, B. (2010). A balanced macronutrient diet induces changes in a host of pro-inflammatory biomarkers, rendering a more healthy phenotype; a randomized cross-over trial, *In preparation* .

Berger, A. (2001). Resistin: a new hormone that links obesity with type 2 diabetes, *The British Medical Journal* **322**: 193.

Chi, K. R. (2007). What is visfatin?, `http://www.the-scientist.com/news/display/53297/`.

Dugdale, D. C. (2009). Uric acid - blood, `http://www.nlm.nih.gov/medlineplus/ency/article/003476.htm`.

Dugdale, D. C. (2010). Triglyceride level, `http://www.nlm.nih.gov/medlineplus/ency/article/003493.htm`.

Fong, Y., Rue, H. and Wakefield, J. (2010). Bayesian inference for generalized linear mixed models, *Biostatistics* **11**(3): 397–412.

Harrell, F. E. (2009). *Design: Design Package.* R package version 2.3-0.

Kuhn, M., Weston, S., Wing, J. and Forester, J. (2010). *contrast: A collection of contrast methods.* R package version 0.13.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. and R Development Core Team (2010). *nlme: Linear and Nonlinear Mixed Effects Models.* R package version 3.1-97.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-Effects Models in S an S-PLUS*, Springer.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rue, H. and Martino, S. (2009). *INLA: Functions which allow to perform a full Bayesian analysis of structured additive models using Integrated Nested Laplace Approximaxion.* R package version 0.0.

Rue, H., Martino, S. and Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations, *Journal of the Royal Statistical Society B* **71**(2): 319–392.

Shrout, P. E. and Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability., *Psychological Bulletin* **86**(2): 420–428.

Verbeke, G. and Molenberghs, G. (2000). *Linear mixed models for longitudinal data*, Springer.

Warnes, G. R. (2011). *gmodels: Various R programming tools for model fitting.* R package version 2.15.1.

West, B. T., Welch, K. B. and Galecki, A. T. (2007). *Linear Mixed Models - A Practical Guide Using Statistical Software*, Chapman and Hall.

# Appendix A

# Contrasts for the loaded model

## Treatment contrast coding

### Female participants

| Contrast | Estimate | Std. Error | t value | DF | Pr(>\|t\|) |
|---|---|---|---|---|---|
| A0 | 679.23 | 86.87 | 7.82 | 90 | 9.62e-12 |
| A6 | 881.23 | 86.87 | 10.14 | 90 | 0.00e+00 |
| B0 | 637.62 | 86.87 | 7.34 | 90 | 9.15e-11 |
| B6 | 756.69 | 86.87 | 8.71 | 90 | 1.38e-13 |
| A6-A0 | 202.00 | 73.99 | 2.73 | 90 | 7.62e-03 |
| B6-B0 | 119.08 | 73.99 | 1.61 | 90 | 1.11e-01 |
| (B6-B0)-(A6-A0) | -82.92 | 104.64 | -0.79 | 90 | 4.30e-01 |

Table A.1: Female contrasts for model 4.6, using the treatment contrast coding.

### Male participants

| Contrast | Estimate | Std. Error | t value | DF | Pr(>\|t\|) |
|---|---|---|---|---|---|
| A0 | 558.42 | 71.86 | 7.77 | 30 | 1.14e-08 |
| A6 | 762.00 | 71.86 | 10.60 | 30 | 1.15e-11 |
| B0 | 518.11 | 71.86 | 7.21 | 30 | 5.03e-08 |
| B6 | 672.05 | 71.86 | 9.35 | 30 | 2.12e-10 |
| A6-A0 | 203.58 | 61.20 | 3.33 | 90 | 1.28e-03 |
| B6-B0 | 153.95 | 61.20 | 2.52 | 90 | 1.37e-02 |
| (B6-B0)-(A6-A0) | -49.63 | 86.55 | -0.57 | 90 | 5.68e-01 |

Table A.2: Male contrasts for model 4.6, using the treatment contrast coding.

### All participants

| Contrast | Estimate | Std. Error | t value | DF | Pr(>\|t\|) |
|---|---|---|---|---|---|
| A0 | 618.83 | 56.37 | 10.98 | 15 | 1.44e-08 |
| A6 | 821.76 | 56.37 | 14.58 | 15 | 2.90e-10 |
| B0 | 577.98 | 56.37 | 10.25 | 15 | 3.59e-08 |
| B6 | 717.76 | 56.37 | 12.73 | 15 | 1.91e-09 |
| A6-A0 | 202.94 | 47.16 | 4.30 | 53 | 7.21e-05 |
| B6-B0 | 139.78 | 47.16 | 2.96 | 53 | 4.53e-03 |
| (B6-B0)-(A6-A0) | -63.16 | 66.69 | -0.95 | 53 | 3.48e-01 |

Table A.3: Mean contrasts for model 4.6, using the treatment contrast coding.

### The weighted mean participant

| Contrast | Estimate | Std. Error | t value | DF | Pr(>\|t\|) |
|---|---|---|---|---|---|
| A0 | 607.50 | 55.37 | 10.97 | 18 | 2.36e-09 |
| A6 | 810.44 | 55.37 | 14.64 | 18 | 2.27e-11 |
| B0 | 566.66 | 55.37 | 10.23 | 18 | 6.94e-09 |
| B6 | 706.44 | 55.37 | 12.76 | 18 | 2.14e-10 |
| A6-A0 | 202.94 | 47.16 | 4.30 | 53 | 7.21e-05 |
| B6-B0 | 139.78 | 47.16 | 2.96 | 53 | 4.53e-03 |
| (B6-B0)-(A6-A0) | -63.16 | 66.69 | -0.95 | 53 | 3.48e-01 |

Table A.4: Weighted contrasts for model 4.6, using the treatment contrast coding.

## Sum to zero contrast coding

### Female participants

| Contrast | Estimate | Std. Error | t value | DF | Pr(>\|t\|) |
|---|---|---|---|---|---|
| A0 | 679.23 | 86.87 | 7.82 | 30 | 1.00e-08 |
| A6 | 881.23 | 86.87 | 10.14 | 30 | 3.28e-11 |
| B0 | 637.62 | 86.87 | 7.34 | 30 | 3.56e-08 |
| B6 | 756.69 | 86.87 | 8.71 | 30 | 1.03e-09 |
| A6-A0 | 202.00 | 73.99 | 2.73 | 90 | 7.62e-03 |
| B6-B0 | 119.08 | 73.99 | 1.61 | 90 | 1.11e-01 |
| (B6-B0)-(A6-A0) | -82.92 | 104.64 | -0.79 | 90 | 4.30e-01 |

Table A.5: Female contrasts for model 4.6, using the sum to zero contrast coding.

### Male participants

| Contrast | Estimate | Std. Error | t value | DF | Pr(>\|t\|) |
|---|---|---|---|---|---|
| A0 | 558.42 | 71.86 | 7.77 | 30 | 1.14e-08 |
| A6 | 762.00 | 71.86 | 10.60 | 30 | 1.15e-11 |
| B0 | 518.11 | 71.86 | 7.21 | 30 | 5.03e-08 |
| B6 | 672.05 | 71.86 | 9.35 | 30 | 2.12e-10 |
| A6-A0 | 203.58 | 61.20 | 3.33 | 90 | 1.28e-03 |
| B6-B0 | 153.95 | 61.20 | 2.52 | 90 | 1.37e-02 |
| (B6-B0)-(A6-A0) | -49.63 | 86.55 | -0.57 | 90 | 5.68e-01 |

Table A.6: Male contrasts for model 4.6, using the sum to zero contrast coding.

### All participants

| Contrast | Estimate | Std. Error | t value | DF | Pr(>\|t\|) |
|---|---|---|---|---|---|
| A0 | 618.83 | 56.37 | 10.98 | 90 | 0.00e+00 |
| A6 | 821.62 | 56.37 | 14.58 | 90 | 0.00e+00 |
| B0 | 577.86 | 56.37 | 10.25 | 90 | 0.00e+00 |
| B6 | 714.37 | 56.37 | 12.67 | 90 | 0.00e+00 |
| A6-A0 | 202.79 | 48.01 | 4.22 | 90 | 5.74e-05 |
| B6-B0 | 136.51 | 48.01 | 2.84 | 90 | 5.52e-03 |
| (B6-B0)-(A6-A0) | -66.28 | 67.90 | -0.98 | 90 | 3.32e-01 |

Table A.7: Mean contrasts for model 4.6, using the sum to zero contrast coding.

# Appendix B

# R code

## B.1 Fitting a LME model, with *lme*

R code for fitting a linear mixed effects model to the resistin, uric acid, triglyceride and visfatin data, using the *lme* method by Pinheiro et al. (2010). Used for all biomarkers are *resp*, which is the biomarker measurements, *id*, which is the identification of all the participants, *sex*, which is a vector of all the participants gender, *diet* and *time*, which is the diet and time at which the biomarker measurement were taken, respectively.

### Resistin

Fitting the resistin model (3.7), with the fixed effect vector given by Equation (3.8).

```
res.data <- data.frame(resp, id, diet, time, sex)
lme.res <- lme(resp~factor(time) + diet, random = ~1 | id,
data = res.data, na.action = na.omit)
```

### Uric acid

Fitting the uric acid model (3.9), with the fixed effect vector given by Equation (3.10)

```
uric.data <- data.frame(resp=log(resp), id, diet, time, sex)
lme.uric <- lme(resp ~ sex + factor(time) + diet,
random = ~ factor(time) | id, data = uric.data, na.action = na.omit)
```

### Triglyceride

Fitting the triglyceride model (3.11), with the fixed effect vector given by Equation (3.12).

```
tri.data <- data.frame(resp=log(resp), time, diet, id)
lme.tri <- lme(resp ~ factor(time) + diet, random = ~ 1 | id,
weights = varIdent(form = ~1 | factor(time)), data = tri.data,
na.action = na.omit)
```

### Visfatin

Fitting the visfatin model (3.13), with the fixed effect vector given by Equation (3.14).

```
vis.data <- data.frame(resp=log(resp), time, diet, id)
lme.vis <- lme(resp ~ factor(time) + diet, random = ~ diet | id,
weights = varIdent(form = ~1 | diet), data = vis.data,
na.action = na.omit)
```

## B.2 Diagnostics

R code for obtaining diagnostic plots. We have chosen to only include the R code for the resistin data.

Normal plot of residuals by diet:

```
qqnorm(lme.res, ~ resid(.) | diet)
```

Plot of observed versus fitted values:

```
plot(lme.res, resp ~ fitted(.) | diet, abline = c(0,1),
ylab="Observed values")
```

Normal plot of random effects:

```
qqnorm(lme.res, ~ ranef(.))
```

## B.3 Contrasts

R code for estimating contrasts for the loaded model, (3.1), with the fixed effect vector given by Equation (4.6). We have chosen to only include the R code for the mean participant, when the sex factor is coded as "sum to zero".

```
new.data <- res.data
sex.sum <- contr.sum(2)
new.data[,"sex"] <- C(new.data[,"sex"], sex.sum)
fit <- lme(resp~sex*factor(time)*diet, random=~1 | id, data=new.data)
A0 <- estimable(fit, c(1,0,0,0,0,0,0,0))
A6 <- estimable(fit, c(1,0,1,0,0,0,0,0))
B0 <- estimable(fit, c(1,0,0,1,0,0,0,0))
B6 <- estimable(fit, c(1,0,1,1,0,0,1,0))
AA <- estimable(fit, c(0,0,1,0,0,0,0,0))
BB <- estimable(fit, c(0,0,1,0,0,0,1,0))
AB <- estimable(fit, c(0,0,0,0,0,0,1,0))
```

## B.4 The variance-covariance matrices

R code for extracting the implied marginal variance-covariance matrix, $\mathbf{V}_i$, given by Equation (2.14). We have chosen to only include the R code for the visfatin data, since this model has the most complicated covariance structure.

```
var1 <- intervals(lme.vis)$reStruct$id[1,2][1]^2
var2 <- intervals(lme.vis)$reStruct$id[2,2]^2
cor <- intervals(lme.vis)$reStruct$id[3,2]
cov <- cor*var1*var2
resA <- intervals(lme.vis)$sigma[2]^2
resB <- (coef(lme.vis$modelStruct$varStruct, uncons = FALSE)*
intervals(lme.vis)$sigma[2])^2
row1 <- c(var1+resA, var1, var1+cov, var1+cov)
row2 <- c(var1, var1+resA, var1+cov, var1+cov)
row3 <- c(var1+cov, var1+cov, var1+2*cov+var2+resB, var1+2*cov+var2)
row4 <- c(var1+cov, var1+cov, var1+2*cov+var2, var1+2*cov+var2+resB)
V.mat <- rbind(row1, row2, row3, row4)
```

For the resistin model we are able to extract the implied marginal variance-covariance matrix, $\mathbf{V}_i$, using the *getVarCor* function by Pinheiro et al. (2010). This function does however not work for models with correlated random effects or a heterogeneous structure for the covariance matrix associated with the residuals, $\mathbf{R}_i$.

R code for constructing the empirical variance-covariance matrix, $\hat{\mathbf{V}}$, of the loaded linear model for the visfatin data.

```
na <- which(is.na(resp))
na.id <- id[na]
una <- unique(na.id)
list <- NULL
for(i in una){
    rem <- which(id==i)
    list <- c(list, rem)
}
sex <- sex[-una]
diet <- diet[-list]
time <- time[-list]
resp <- resp[-list]
new.data <- data.frame(resp, diet, time, sex)
fit.lm <- lm(resp ~ sex*factor(time)*diet, data=new.data)
res <- residuals(fit.lm)
V.emp <- cov(cbind(res[1:26], res[27:52], res[53:78], res[79:104]))
```

## B.5 Fitting a LME model, with *inla*

R code for fitting a linear mixed effects model to the resistin, uric acid, triglyceride and visfatin data, using the *inla* method by Rue and Martino (2009). In addition to the response and the explanatory variables introduced in the R code for fitting a

LME using the *lme* method, we now also use *random* which is another identification vector similar to *id* but with different numbers, used to allow correlated random effects in *inla*.

## Resistin

Fitting the resistin model, (3.7), with the fixed effect vector given by Equation (3.8), using the *inla* method by Rue and Martino (2009).

```
res.data <- data.frame(resp=log(resp), id, diet, time)
inla.res <- inla(resp ~ factor(time) + diet + f(id,model="iid"),
family="gaussian", data=res.data)
```

## Uric acid

Fitting the uric acid model, (3.9), with the fixed effect vector given by Equation (3.10), using the *inla* method by Rue and Martino (2009).

```
uric.data <- data.frame(resp=log(resp), id, diet, time, sex, random)
inla.uric <- inla(resp ~ sex + factor(time) + diet
+ f(id, model="iid2d", n=64) + f(random, factor(time), copy="id",
fixed=T), family = "gaussian", data = uric.data)
```

## Triglyceride

Fitting the triglyceride model, (3.11), with the fixed effect vector given by Equation (3.12), using the *inla* method by Rue and Martino (2009).

```
resp <- log(resp)
temp1 <- rep(NA, 4*n)
temp1[time==0] <- resp[time==0]
temp2 <- rep(NA, 4*n)
temp2[time==1] <- resp[time==1]
resp.mat <- cbind(temp1, temp2)
new.tri.data <- data.frame(resp.mat, time, diet, id)
inla.tri <- inla(resp.mat ~ factor(time) + diet + f(id, model="iid"),
family = c("gaussian", "gaussian"), data = new.tri.data)
```

## Visfatin

Fitting the visfatin model, (3.13), with the fixed effect vector given by Equation (3.14), using the *inla* method by Rue and Martino (2009).

```
resp <- log(resp)
temp1 <- rep(NA, 4*n)
temp1[diet=="A"] <- resp[diet=="A"]
temp2 <- rep(NA, 4*n)
temp2[diet=="B"] <- resp[diet=="B"]
resp.mat <- cbind(temp1, temp2)
vis.data <- data.frame(resp.mat, time, diet, id, random)
```

```
inla.vis <- inla(resp.mat ~ factor(time) + diet
+ f(id, model="iid2d", n=64) + f(random, diet, copy="id", fixed=T),
family = c("gaussian", "gaussian"), data = vis.data)
```