McMaster University
DigitalCommons@McMaster

Open Access Dissertations and Theses | Open Dissertations and Theses

10-17-2012

Residuals in the growth curve model with applications to the analysis of longitudinal data

WEILIANG HUANG
McMaster University, huangwl@math.mcmaster.ca

Follow this and additional works at: http://digitalcommons.mcmaster.ca/opendissertations

Part of the Statistical Methodology Commons

Recommended Citation
HUANG, WEILIANG, "Residuals in the growth curve model with applications to the analysis of longitudinal data" (2012). *Open Access Dissertations and Theses.* Paper 7545.

This Thesis is brought to you for free and open access by the Open Dissertations and Theses at DigitalCommons@McMaster. It has been accepted for inclusion in Open Access Dissertations and Theses by an authorized administrator of DigitalCommons@McMaster. For more information, please contact scom@mcmaster.ca.

Residuals in the growth curve model with applications to the analysis of longitudinal data

By

WeiLiang Huang

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfilment of the Requirements

for the Degree

Master of Science

McMaster University

© Copyright by WeiLiang Huang, September 2012

MASTER OF SCIENCE (2012) McMaster University

(Statistics) Hamilton, Ontario

TITLE: Residuals in the Analysis of Longitudinal Data

AUTHOR: WeiLiang Huang

(McMaster University, Canada)

SUPERVISOR: Dr. J. Hamid

NUMBER OF PAGES: x, 66

Acknowledgements

I would like to take this opportunity to express my deepest gratitude towards my supervisor Dr.Jemila S Hamid for all her support and encouragement through out this project. My thesis would not have been possible without her helpful comment, infinite patience as well as inspiration.

I would like to express my great gratitude towards Dr. N. Balakrishnan and Dr. Joseph Beyene for agreeing to be my examination committee members, thanks for their time in reading my thesis and all the helpful remarks they made.

I would like to thank the Department of Mathematics and Statistics at McMaster University for providing me with such a great learning environment during my graduate study.

I would like to thank all my friends and colleagues, especially Xiaofeng Liu, for their motivation and making my graduate study a valuable experience.

Last but not least, I would like to thank my family, especially my parents, for their endless love and support that gave me the faith in continuing my study.

Abstract

Statistical models often rely on several assumptions including distributional assumptions on outcome variables and relational assumptions where we model the relationship between outcomes and independent variables. Further assumptions are also made depending on the complexity of the data and the model being used. Model diagnostics is, therefore, a crucial component of any model fitting problem. Residuals play important roles in model diagnostics. Residuals are not only used to check adequacy of model fit, but they also are excellent tools to validate model assumptions as well as identify outliers and influential observations. Residuals in univariate models are studied extensively and are routinely used for model diagnostics. In multivariate models residuals are not commonly used to assess model fit, although a few approaches have been proposed to check multivariate normality. However, in the analysis of longitudinal data, the resulting residuals are correlated and are not normally distributed. It is, therefore, not clear as to how ordinary residuals can be used for model diagnostics. Under sufficiently large sample size, a transformation of ordinary residuals are proposed to check the normality assumption. The transformation is based solely on removing correlation among the residuals. However, we show that these transformed residuals fail in the presence of model mis-specification. In this thesis, we investigate residuals in the analysis of longitudinal data. We consider ordinary residuals, Fitzmaurice's transformed (uncorrelated) residuals as well as von Rosen's decomposed residuals. Using simulation studies, we show how the residuals behave under multivariate normality and when this assumption is violated. We also investigate their properties under correct fitting as well as wrongly fitted models. Finally, we propose new residuals by transforming von

Rosen's decomposed residuals. We show that these residuals perform better than Fitzmourice's transformed residuals in the presence of model mis-specification. We illustrate our approach using two real data sets.

Keywords: Decomposition of linear spaces, growth curve model, Fitzmaurice's transformation, model diagnostics, Small's graphical method, residuals, von Rosen's decomposed residuals.

Contents

Acknowledgements **iii**

Abstract **iv**

1 Introduction **1**

2 Methodology **4**

2.1 Residual Decompositions 4

2.2 Transformation of residuals 9

3 Simulation **12**

3.1 Residuals for checking multivariate normality: normal data 13

3.2 Residuals for checking multivariate normality: non normal data 16

3.3 Residuals in longitudinal data: normal error 19

3.4 Decomposed Residuals in the GCM: normal error 21

3.5 Decomposed Residuals in GCM: non normal error 25

4 Real Data Application **28**

4.1 Dental data 28

4.2 Glucose data 31

5 Discussion **35**

Appendices **37**

A Table **37**

A.1 Residuals analysis setting of non-structured GCM 37

A.2 Simulation settings description of Residuals in bi-linear structured GCM 38

B R Code **39**

B.1 Small's graphical method (Small, 1978) 39

B.2 Examination of random multivariate data 40

B.3 Examination of behavior of residuals in GCM under normality assumption . . . 44

B.4 Examination of behavior of residuals in GCM under non-normality assumption . 50

B.5 Examination of behavior of residuals in the Potthoff-Roy Dental Data 56

B.6 Examination of behavior of residuals in the Glucose Data 60

Bibliography **64**

List of Figures

2.1 *Residuals for univariate and multivariate analysis of variance models* 6

2.2 *Residuals in the growth curve model as defined by von Rosen (1995)* 8

3.1 *Normal Probability Plot of independent (left) and correlated (right) multivariate normal data (n=50, p=7)* 14

3.2 *Normal Probability Plot of independent (left) and correlated (right) multivariate normal data after Fitzmaurice's transformation to remove correlation (n=50, p=7)* 15

3.3 *Beta Probability Plot of independent (left) and correlated (right) multivariate normal data after Small's transformation to remove correlation (n=50, p=7)* . . 16

3.4 *Normal Probability Plot of independent (left) and correlated (right) multivariate log-normal data (n=50, p=7)* 17

3.5 *Normal Probability Plot of independent (left) and correlated (right) multivariate log-normal data after Fitzmaurice's transformation to remove correlation (n=50, p=7)* 17

3.6 *Normal Probability Plot of independent (left) and correlated (right) multivariate log-normal data after Small's transformation to remove correlation (n=50, p=7)* 18

3.7 *Normal Quantile Plot of residuals* $\mathbf{R}$ *in perfectly fitted (left) and mis-specified (right) GCM (n=50, p=7)* 19

3.8 *Normal Quantile Plot of Fitzmaurice's residuals* $\mathbf{R}$ *in perfectly fitted (left) and mis-specified (right) GCM (n=50, p=7)* 20

3.9 *Beta Quantile Plot of Small's residuals* $\mathbf{R}$ *in perfectly fitted (left) and mis-specified (right) GCM (n=50, p=7)* 21

3.10 *Scatter plot of decomposed residuals* $\mathbf{R}_3$ *in perfectly fitted (left) and mis-specified (right) GCM (n=50, p=7)* 22

3.11 *Normal Quantile Plot of residuals* $\mathbf{R}_1 + \mathbf{R}_2$ *in perfectly fitted (left) and mis-specified (right) GCM (n=50, p=7)* 23

3.12 *Quantile Plot of Fitzmaurice's (top left) and Small's (bottom left)* $\mathbf{R}_1 + \mathbf{R}_2$ *in perfectly fitted GCM, and Quantile Plot of Fitzmaurice's (top right) and Small's (bottom right)* $\mathbf{R}_1 + \mathbf{R}_2$ *in mis-specified GCM. (n=50, p=7)* 24

3.13 *Scatter plot of decomposed residuals* $\mathbf{R}_3$ *in perfectly fitted (left) and mis-specified (right) log-normal GCM (n=50, p=7)* 25

3.14 *Normal Quantile Plot of Fitzmaurice's residuals* $\mathbf{R}_1 + \mathbf{R}_2$ *in perfectly fitted (left) and mis-specified (right) log-normal GCM (n=50, p=7)* 26

3.15 *Beta Quantile Plot of Small's residuals* $\mathbf{R}_1 + \mathbf{R}_2$ *in perfectly fitted (left) and mis-specified (right) log-normal GCM (n=50, p=7)* 27

4.1 *Profile plot of Dental data (left) and group average (right)* 29

4.2 *Normal quantile plot of ordinary residual (left) and scatter plot of decomposed* $\mathbf{R}_3$ *(right) of Dental data* 30

4.3 *Quantile plot of Fitzmaurice's (left) and Small's (right) decomposed residuals* $\mathbf{R}_1 + \mathbf{R}_2$ *of Dental data* 30

4.4 *Quantile plot of Fitzmaurice's (left) and Small's (right) decomposed residuals* $\mathbf{R}_1 + \mathbf{R}_2$ *of Dental data without outliers* 31

4.5 *Profile plot of Glucose data (left) and group average (right)* 32

4.6 *Normal quantile plot of ordinary residual (left) and scatter plot of decomposed* $\mathbf{R}_3$ *(right) of Glucose data* 33

4.7 *Scatter plot of decomposed* $\mathbf{R}_3$ *for quadratic (left) and third degree polynomials (right) fitting* 33

4.8 *Quantile plot of Fitzmaurice's (left) and Small's (right) decomposed residuals* $\mathbf{R}_1 + \mathbf{R}_2$ *of Glucose data* 34

Chapter 1

Introduction

Statistical modeling is commonly used in a wide range of applications, from financial banking and weather forecasting to clinical medicine and public health (Taylor, 2007; Gneiting and Raftery, 2005; Altman, 1990; Bland, 2002). In health and biological research in particular, modeling has been demonstrated to be an essential tool for understanding a variety of common as well as rare diseases affecting the public. It plays important roles in disease diagnosis, prognosis, management as well as prevention and health promotion (Altman, 1990; Bland, 2002; Cook, 2008). Statistical models are also commonly used in identifying risk factors associated with diseases and hence allowing effective diagnosis, treatment as well as prevention mechanisms (Demchuk et al., 2007). The terms evidence-based medicine, evidence-based diagnosis and evidence-based decision making highlight the importance of statistical methods in areas of medicine and public health (Straus et al., 2005; Newman and Kohn, 2009; Gray, 2001; Etzioni and Kadane, 1995).

In many cases, statistical models rely on several assumptions, where distributional assumptions are made for the outcome variables and relational assumptions are made to model the relationship between the outcomes and independent variables. Additional assumptions are further made depending on the nature of the data and the complexity of the model. Model diagnostics, to assess model fit and check the validity of assumptions, is therefore a crucial step in any

model fitting problem. It can be argued that model fitting is not complete without validating the appropriateness of the model and assumptions being made.

One can not talk about model diagnostics without mentioning residuals. Residuals are defined as the difference between what is observed and what is estimated using the model. i.e. $r_i = y_i - \hat{y}_i$, where y_i is a vector of observed values and $\hat{y}_i$ is a vector of estimated values. Residuals play important roles in model fitting as they represents what is left unexplained after fitting the model to data (Draper and Smith, 1998; Sen and Srivastava,1990; Cook, 1977;1982). Residuals are not only used to check adequacy of the model fitting, they are also an excellent tool to validate model assumptions and identify outliers and influential observations (Draper and Smith, 1998; Sen and Srivastava, 1990; Cook, 1977). In univariate modeling, the resulting residuals are also univariate and it is relatively straight forward to study them. In fact, many studies have been done regarding residuals in such models and several approaches (both graphical and formal tests), based on the residuals, have been developed to assess model fit and check the validity of assumptions. Moreover, many different types of residuals have been proposed and studied, among them are studentized residuals and standardized residuals (Sen and Srivastava, 1990; Cook, 1977; 1982).

However, in the analysis of longitudinal data using multivariate modeling, the resulting residuals are also multivariate and it is not obvious as to how these residuals may be used to investigate model fitting and check the validity of assumptions (Srivastava and Carter, 1983; von Rosen, 1995a; Hamid and von Rosen, 2006). There are also many challenges when analyzing longitudinal data using univariate approaches such as linear mixed models (Diggle et al., 2002; Fitzmaurice, 2004). This is particularly the case when there is a clear pattern in time, which is often the case in practice.

In this thesis, we examine different types of residuals in the analysis of longitudinal data with

continuous outcome, with the aim of identifying residuals that are useful for assessing model fitting and validating normality assumption. We consider:

- Ordinary residuals as defined as the difference between the observed and estimated values,
- Transformed (uncorrelated) residuals as proposed by Fitzmaurice et al.(2004),
- Decomposed residuals in the growth curve model as proposed by von Rosen (1995),
- We also propose new residuals by transforming von Rosen's decomposed residuals where the resulting residuals are uncorrelated

We propose two types of transformations for von Rosen's residuals and investigate which residual is useful for assessing model fitting and which one is more appropriate to validate the normality assumptions. The first transformation is done using the vector function where we consider the vectored form of the residuals. However, this approach results in correlated residuals. To overcome this challenge, we propose a transformation similar to Fitzmaurices's where the resulting residuals are uncorrelated. The second transformation is done using Small's graphical approach where a distance matrix is defined based on the multivariate residuals.

The thesis consists of 5 chapters and is organized as follows. In Chapter 2, we will provide important methodological background. We will describe the simulation and present the results of the simulation in Chapter 3. We will provide numerical illustrations using two real data sets in Chapter 4. Discussions are provided in Chapter 5. Finally, the lists of the detail simulation setting, as well as the R code for the simulation and the real data analysis are provided in the Appendix.

Chapter 2

Methodology

2.1 Residual Decompositions

Suppose that we have m different groups where repeated measurements are taken from a given individual at p different time points. Suppose also that the mean for the ith group follows a polynomial curve of degree q over time, which can be described as

$$b_{0,i} + b_{1,i}t + +b_{2,i}t^2 \cdots + b_{q,i}t^q.$$

The growth curve model (GCM), can be formulated as

$$\mathbf{X} = \mathbf{ABC} + \mathbf{E}, \tag{2.1}$$

where,

$$\mathbf{A}' = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ t_1 & t_2 & t_3 & \cdots & t_p \\ t_1^2 & t_2^2 & t_3^2 & \cdots & t_p^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ t_1^q & t_2^q & t_3^q & \cdots & t_p^q \end{pmatrix} \qquad \mathbf{C} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0}_{n_2} & \mathbf{0}_{n_3} & \cdots & \mathbf{0}_{n_m} \\ \mathbf{0}_{n_1} & \mathbf{1}_{n_2} & \mathbf{0}_{n_3} & \cdots & \mathbf{0}_{n_m} \\ \mathbf{0}_{n_1} & \mathbf{0}_{n_2} & \mathbf{1}_{n_3} & \cdots & \mathbf{0}_{n_m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{n_1} & \mathbf{0}_{n_2} & \mathbf{0}_{n_3} & \cdots & \mathbf{1}_{n_m} \end{pmatrix},$$

$$\mathbf{B} = \begin{pmatrix} b_{01} & b_{02} & b_{03} & \cdots & b_{0m} \\ b_{11} & b_{12} & b_{13} & \cdots & b_{1m} \\ b_{21} & b_{22} & b_{23} & \cdots & b_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b_{q1} & b_{q2} & b_{q3} & \cdots & b_{qm} \end{pmatrix}.$$

$(\bullet)'$ represent the transpose of a matrix, $\mathbf{X} : p \times n$ is the known observation matrix, $\mathbf{B} : (q+1) \times m$ is the unknown parameter matrix, $\mathbf{A} : p \times (q+1)$ and $\mathbf{C} : m \times n$ are the within and between individual design matrices respectively, $q \leq p, Rank(C) + p \leq n(n = n_1 + n_2 + \ldots + n_m)$, and the columns of $\mathbf{E} : p \times n$ are assumed to be independently p-variate normally distributed with mean zero and an unknown positive definite covariance matrix Σ.

Note that, when $\mathbf{A} = \mathbf{I}_{m\times m}$, the GCM reduces to the classical multivariate analysis of variance (MANOVA) model, i.e. $\mathbf{X} = \mathbf{BC} + \mathbf{E}$. In the univariate or classical multivariate analysis of variance (MANOVA) models (this is also true for regression models), residuals are given by

$$\mathbf{R}_M = \mathbf{X}(\mathbf{I} - \mathbf{C}'(\mathbf{CC}')^{-1}\mathbf{C}),$$

indicating that residuals are obtained by projecting the observation matrix $\mathbf{X}$ onto the space orthogonal to the column space of $\mathbf{C}'$, where $\mathbf{C}$ in this case is the within individual design matrix. A description of the projection space for the estimated model and the residuals is provided in Figure 2.1.

However, in the GCM the space has a bilinear structure and we have two design matrices, $\mathbf{A}$ and $\mathbf{C}$. Therefore the ordinary residuals in this case, are obtained by projecting the observation matrix $\mathbf{X}$ onto the orthogonal complement to the tensor product $C(\mathbf{C}') \otimes C_{\mathbf{S}}(\mathbf{A})$, where $C(\bullet)$ represent the column space of a matrix, and the $\mathbf{S}(2.2)$ in $C_{\mathbf{S}}$ indicates that the inner product is defined with $\mathbf{S}^{-1}$, where $\mathbf{S}$ (given in (2.2) below) is the sample covariance matrix, i.e. $< x, y >= x'\mathbf{S}^{-1}y$. Therefore,

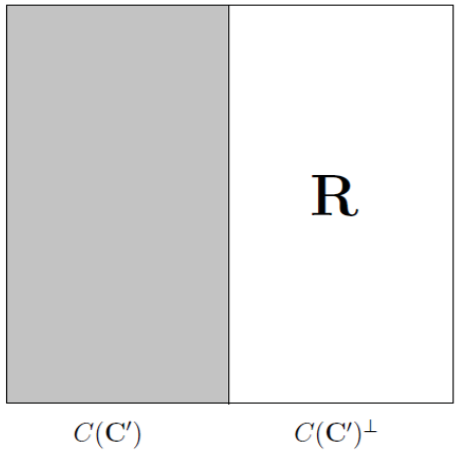

Figure 2.1: *Residuals for univariate and multivariate analysis of variance models*

the residuals in the GCM are not as obvious to understand and interpret as in the univariate or the classical MANOVA case, and new residuals are needed. von Rosen decomposed the linear space where the residuals are defined on and provided new residuals (von Rosen, 1995b). A general framework, using the vector operator, for residuals in GMANOVA models (including the GCM and extended GCM) is provided by Hamid (2006).

Suppose the design matrices $\mathbf{A}$ and $\mathbf{C}$ are of full rank. The maximum likelihood estimator for the parameter matrix B in the GCM is given by Khatri (1966) as

$$\hat{\mathbf{B}} = (\mathbf{A}'\mathbf{S}^{-1}\mathbf{A}')^{-1}\mathbf{A}\mathbf{S}^{-1}\mathbf{X}\mathbf{C}'(\mathbf{C}\mathbf{C}')^{-1},$$

where

$$\mathbf{S} = \mathbf{X}(\mathbf{I} - \mathbf{C}'(\mathbf{C}\mathbf{C}')^{-1}\mathbf{C})\mathbf{X}', \tag{2.2}$$

Therefore, the estimated model can be expressed as

$$\mathbf{A}\hat{\mathbf{B}}\mathbf{C} = \mathbf{A}(\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1}\mathbf{A}'\mathbf{S}^{-1}\mathbf{X}\mathbf{C}'(\mathbf{C}\mathbf{C}')^{-1}\mathbf{C},$$

The ordinary residuals, therefore, are given by

$$\mathbf{R} = \mathbf{X} - \mathbf{A}(\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1}\mathbf{A}'\mathbf{S}^{-1}\mathbf{X}\mathbf{C}'(\mathbf{C}\mathbf{C}')^{-1}\mathbf{C},$$

which can not be explicitly written in a simple formula similar to univariate and multivariate models. That is one of the reasons why one should study ordinary residuals in repeated measurement and longitudinal analysis more carefully. Due to the bilinear structure in the model, the estimated mean depends on two design matrices which correspond to the between and within individual structure respectively. Different components of the residuals may have a counter effect on others, and as a result would lead to inappropriate conclusion about the model fit. Therefore it is important to establish a different look at the ordinary residuals to better understand their properties, and eventually to be able to use them appropriately for checking model fit and validating model assumptions. von Rosen (1995) proposed new residuals by decomposing the space the residuals are defined on, which is the orthogonal complement of the space generated by the two design matrices. The general idea is give as following

$$\begin{aligned}
\mathbf{R} &= \mathbf{X} - \mathbf{A}(\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1}\mathbf{A}'\mathbf{S}^{-1}\mathbf{X}\mathbf{C}'(\mathbf{C}\mathbf{C}')^{-1}\mathbf{C} \\
&= \mathbf{X} - \mathbf{P}_A\mathbf{X}\mathbf{P}_C
\end{aligned}$$

Then, taking the vec operator for both side give

$$\begin{aligned}
vec\mathbf{R} &= vec(\mathbf{X} - \mathbf{P}_A\mathbf{X}\mathbf{P}_C) \\
&= vec\mathbf{X} - (\mathbf{P}_C \otimes \mathbf{P}_A)vec\mathbf{X} \\
&= vec\mathbf{X} - \mathbf{P}vec\mathbf{X} \\
&= (\mathbf{I} - \mathbf{P})vec\mathbf{X}
\end{aligned}$$

Therefore, von Rosen's decomposed residuals are given by

$$\mathbf{R}_1 = (\mathbf{I} - \mathbf{A}(\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1}\mathbf{A}'\mathbf{S}^{-1})\mathbf{X}(\mathbf{I} - \mathbf{C}'(\mathbf{C}\mathbf{C}')^{-1}\mathbf{C}), \tag{2.3}$$

$$\mathbf{R}_2 = (\mathbf{A}(\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1}\mathbf{A}'\mathbf{S}^{-1})\mathbf{X}(\mathbf{I} - \mathbf{C}'(\mathbf{C}\mathbf{C}')^{-1}\mathbf{C}), \tag{2.4}$$

$$\mathbf{R}_3 = (\mathbf{I} - \mathbf{A}(\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1}\mathbf{A}'\mathbf{S}^{-1})\mathbf{X}\mathbf{C}'(\mathbf{C}\mathbf{C}')^{-1}\mathbf{C}, \tag{2.5}$$

where $\mathbf{R}_1$ is obtained from the space $C(\mathbf{C}')^{\perp} \otimes C_{\mathbf{S}}(\mathbf{A})^{\perp}$, $\mathbf{R}_2$ from $C(\mathbf{C}')^{\perp} \otimes C_{\mathbf{S}}(\mathbf{A})$, $\mathbf{R}_3$ from $C(\mathbf{C}') \otimes C_{\mathbf{S}}(\mathbf{A})^{\perp}$, and $\mathbf{R} = \mathbf{R_1} + \mathbf{R_2} + \mathbf{R_3}$. A graphical representation of von Rosen's residuals is provided in Figure 2.2.

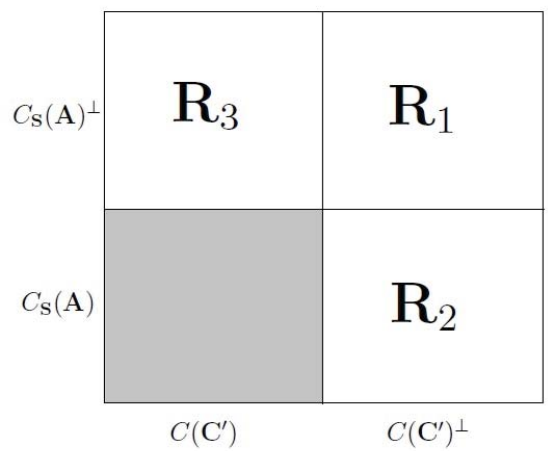

Figure 2.2: *Residuals in the growth curve model as defined by von Rosen (1995)*

To better interpret the resulting decomposed residuals, let us first consider $\mathbf{R}_1 + \mathbf{R}_2$,

$$\begin{aligned}\mathbf{R}_1 + \mathbf{R}_2 = &(\mathbf{I} - \mathbf{A}(\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1}\mathbf{A}'\mathbf{S}^{-1})\mathbf{X}(\mathbf{I} - \mathbf{C}'(\mathbf{C}\mathbf{C}')^{-1}\mathbf{C}) \\ &+(\mathbf{A}(\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1}\mathbf{A}'\mathbf{S}^{-1})\mathbf{X}(\mathbf{I} - \mathbf{C}'(\mathbf{C}\mathbf{C}')^{-1}\mathbf{C}),\end{aligned}$$

which simplifies to

$$\mathbf{R}_1 + \mathbf{R}_2 = \mathbf{X}(\mathbf{I} - \mathbf{C}'(\mathbf{C}\mathbf{C}')^{-1}\mathbf{C}). \tag{2.6}$$

This is identical to the residual in the univariate and classical MANOVA models. The function in (2.6) represents the difference between the observations from their group mean, precisely the between individual assumption. Moreover, it is a linear function of a multivariate random variable $\mathbf{X}$ and a known design matrix $\mathbf{C}$, and hence is distributed as multivariate normal random variable with mean zero and covariance matrix $\Sigma^* = (\mathbf{I} - \mathbf{C}'(\mathbf{C}\mathbf{C}')^{-1}\mathbf{C})\Sigma(\mathbf{I} - \mathbf{C}'(\mathbf{C}\mathbf{C}')^{-1}\mathbf{C})$. We can, therefore, use $\mathbf{R_1} + \mathbf{R_2}$ to check the normality assumption.

On the other hand, $\mathbf{R}_3$ (2.5), can be re-written as

$$\begin{aligned}
\mathbf{R}_3 &= (\mathbf{I} - \mathbf{A}(\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1}\mathbf{A}'\mathbf{S}^{-1})\mathbf{X}\mathbf{C}'(\mathbf{C}\mathbf{C}')^{-1}\mathbf{C}, \\
&= \mathbf{X}\mathbf{C}'(\mathbf{C}\mathbf{C}')^{-1}\mathbf{C} - \mathbf{A}(\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1}\mathbf{A}'\mathbf{S}^{-1}\mathbf{X}\mathbf{C}'(\mathbf{C}\mathbf{C}')^{-1}\mathbf{C}, \\
&= \mathbf{X}\mathbf{C}'(\mathbf{C}\mathbf{C}')^{-1}\mathbf{C} - \mathbf{A}\hat{\mathbf{B}}\mathbf{C}.
\end{aligned}$$

Hence, $\mathbf{R}_3$ represents the difference between the observed mean and the estimated mean. It can, therefore, be used to check the within individual (across time) assumption, i.e. to check the adequacy of the model fitted to describe the time dependency.

2.2 Transformation of residuals

Cholesky factorization

As described in the previous Chapter, residuals obtained in analyzing longitudinal and repeated measure data are often correlated and do not always have constant variance, and hence it is not reasonable to use these residuals to either check the normality assumption or the homogeneity of variances. One suggestion, made by Fitzmaurice et al. (2004), is to tackle the problem by applying a transformation to "de-correlate" the residuals. They state that there are many ways to transform the residuals, and one particular transformation recommended is called the Cholesky factorization. The general idea of the procedure is presented as follows.

Let $\mathbf{R}$ be the ordinary residuals obtained from the analysis of longitudinal data using MIXED models

or the vectorised form of the residuals obtained from the GCM, i.e.,

$$\mathbf{R} = \mathbf{Y}_i - \hat{\mathbf{Y}}_i.$$

Consider the covariance matrix of the residuals, which can be approximated as

$$Cov(\mathbf{R}) \approx Cov(\mathbf{E}) = \hat{\Sigma}.$$

The Cholesky factorization is used to generate a lower triangular matrix $\mathbf{L}$, such that

$$\hat{\Sigma} = \mathbf{L}_i \mathbf{L}_i{}'.$$

The correlated residuals with heterogeneous variances can then be transformed to a set of residuals that are uncorrelated with unit variance by using a transformation using $\mathbf{L}^{-1}$,

$$\mathbf{R}^* = \mathbf{L}^{-1}\mathbf{R} = \mathbf{L}^{-1}(\mathbf{Y}_i - \hat{\mathbf{Y}}_i).$$

Fitzmaurice et al.(2004) recommend these transformed residuals to check the normality assumption. However, we will show in the next section that the Cholesky factorization or Small's transformation(Small, 1978) of $\mathbf{R}_1 + \mathbf{R}_2$ are better when checking the normality assumption, in particular when there is model mis-specification.

Small's graphical method

Another way of checking the assumption of normality is through the Small's graphical method, see Small(1978) and Srivastava (2002). Such graphical approaches also help us to see if there are outliers in the data. The idea behind Small's graphical approach is to reduce the multivariate data to a univariate one. Since we are considering analysis of longitudinal data where there is a bilinear structure, Small's transformation of $\mathbf{R}_1 + \mathbf{R}_2$ along with the ordinary residuals $\mathbf{R}$ will be considered in our investigation. The Small's transformation is briefly presented below.

Suppose $x_1, x_2, \ldots, x_n$ are independently distributed as $N_p(\mu, \Sigma)$. Then the statistics

$$c_i = n(n-1)^{-2}(x_i - \bar{x})'\mathbf{S}^{-1}(x_i - \bar{x}), i = 1, 2, \ldots, n,$$

where

$$\bar{x} = n^{-1}\sum_{j=1}^{n} x_j,$$

$$\mathbf{S} = (n-1)^{-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})'$$

has a beta distribution with parameters $\alpha = \frac{1}{2}p$ and $\beta = \frac{1}{2}(n-p-1)$ (see Small, 1978; Srivastava, 2002), where n is the number of individuals while p is the number of repeated measurements. Asymptotically, c_i's are independently distributed.

Chapter 3

Simulation

We start the simulation by considering simple scenarios (Sections 3.1 and 3.2) and expand to include more complex longitudinal data with structured means (Sections 3.3 - 3.5). The simple scenarios included in our simulation provide important empirical evidence for checking multivariate normality assumptions using vectored form of ordinary residuals and Fitzmaurice's modified residuals, in situations where there is no systematic or modeling error. However, in order to better examine properties of ordinary residuals, modified residuals as well as decomposed residuals, and compare performance of the different residuals, we considered more extensive simulations under various settings. We use the growth curve model (2.1) to generate longitudinal data with different time dependency and different covariance structures. Without loss of generality, we assume that data comes from two groups, although the results and conclusions made are valid for single samples as well as when there are more than two groups. Similarly, we assume p repeated measurements are taken from n individuals, where $n_1 = n_2 = n/2$. We considered data with linear and quadratic means over time to investigate the effect of model mis-specification. The design matrices for the simulations are as follows:

$$\mathbf{A}' = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ t_1 & t_2 & t_3 & \cdots & t_p \\ t_1^2 & t_2^2 & t_3^2 & \cdots & t_p^2 \end{pmatrix} \qquad \mathbf{C} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0}_{n_2} \\ \mathbf{0}_{n_1} & \mathbf{1}_{n_2} \end{pmatrix}, \tag{3.1}$$

and the parameter matrix is:

$$\mathbf{B} = \begin{pmatrix} b_{01} & b_{02} \\ b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}. \tag{3.2}$$

Recall that $\mathbf{A}$ and $\mathbf{C}$ are the within and between individual design matrices, respectively, and B is the parameter matrix. Our data is, therefore, generated using the GCM, $\mathbf{X} = \mathbf{ABC} + \mathbf{E}$ where the values of $\mathbf{B}$ are generated using random number from 1 to 10, and $\mathbf{E}$ is generated from a multivariate normal distribution with mean zero and covariance matrix Σ.

We considered sample sizes $n = 20, 30, 40, 50$, and 100. For each of the sample sizes, we considered the number of time points $p = 4, 7, 8$. Without lost of generality, we will provide the result of the case when $n = 50$ and $p = 7$. We also considered different covariance structures: from identity (independent data) and small correlations(ranging from 0.1 to 0.29) to moderate(ranging from 0.3 to 0.69) and strong correlations(ranging from 0.7 to 1). To show how the residuals behave under non-normality, we considered scenarios where we transformed the error term $\mathbf{E}$ using the exponential function.

3.1 Residuals for checking multivariate normality: normal data

Here we provide the results of the simple scenarios of the simulation we mentioned in the beginning of the chapter. We generated data according to the GCM where $\mathbf{B} = 0$, that is, $\mathbf{X} = \mathbf{E}$. For the covariance matrix, we considered both independent (i.e. $\mathbf{E} \sim MN(0, \mathbf{I})$) as well as correlated data (i.e. $\mathbf{E} \sim MN(0, \Sigma^*)$, where we used the sample covariance matrix of the "mouse" data from Pan and Fang (2002) as our Σ). The data has a sample size of n50 and time point p=7, and is strongly correlated. Figure 3.1 shows the normal quantile plot (QQ-plot) of independent and correlated data, where we converted the multivariate data into a vector.

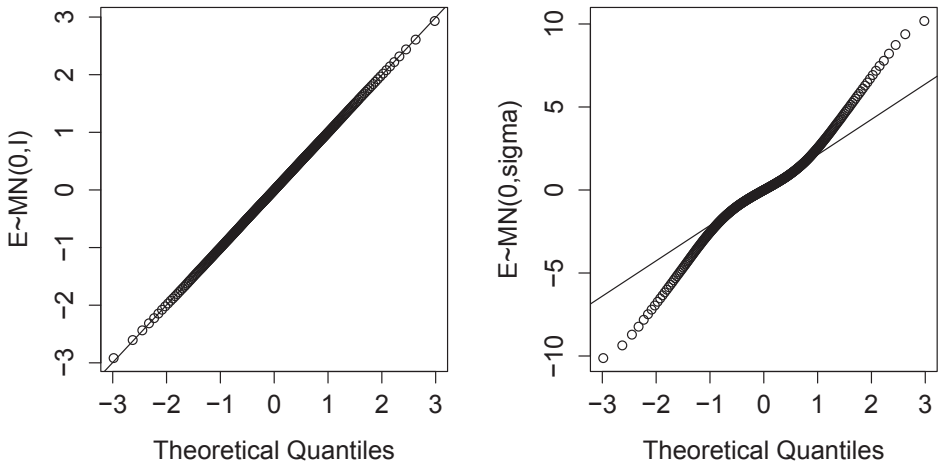

Figure 3.1: *Normal Probability Plot of independent (left) and correlated (right) multivariate normal data (n=50, p=7)*

As we can see from the QQ-plots, when there is no correlation, the probability plot shows a perfect straight line, correctly reflecting the normality assumption of the error terms. On the other hand, in the presence of correlation, the probability plot clearly shows a deviation from straight line leading to an incorrect conclusion that data does not follow multivariate normal distribution, when in fact it does. It is, therefore, important to use uncorrelated sample to properly validate the normality assumption. We transformed the vectored form of both data (independent and correlated data) using Fitzmaurice's approach presented in Chapter 2. The results are presented in Figure 3.2.

As can be seen from Figure 3.2, both data sets resulted in near perfect straight line quantile plots indicating that data are from multivariate normal distribution. We can also observe that the two figures are identical indicating that Fitzmaurice's transformation has successfully removed the correlation in the second data set that previously resulted in a quantile plot that was evidently deviated from the expected straight line.

An alternative approach to Fitzmaurice's transformation is transformation of the multivariate data into univariate. One such approach is Small's transformation presented in Chapter 2. We transformed

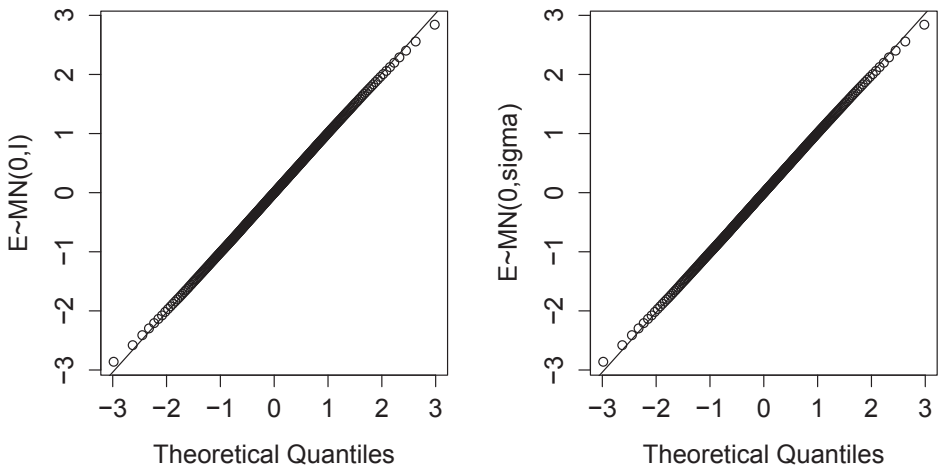

Figure 3.2: *Normal Probability Plot of independent (left) and correlated (right) multivariate normal data after Fitzmaurice's transformation to remove correlation (n=50, p=7)*

two of our simulated data sets (as describe above) using Small's transformation. The results are presented in Figure 3.3 below.

Clearly, the Small's transformation has also done a great job in eliminating the correlation present in the residuals. The two probability plots are clearly observed to be identical and are perfectly linear, regardless of correlation in the second data, indicating that the two data sets come from multivariate normal distributions. Therefore, we conclude that the Small's transformed data can is preferred over the ordinary residuals that are correlated.

It is important to mention that the deviation of the probability plot for correlated data we observed depends on the covariance matrix of Σ, where we observed larger deviations for highly correlated data than data with small correlations. Therefore, not all correlation structure will result in such a clear deviation from normality. For instance, in one of our simulations (data not shown), we used a covariance matrix the same as estimated covariance matrix of the Glucose Data (Pan and Fang, 2002) (which has a small to moderate correlation), we observed that the probability plot of vectored data is approximately straight.

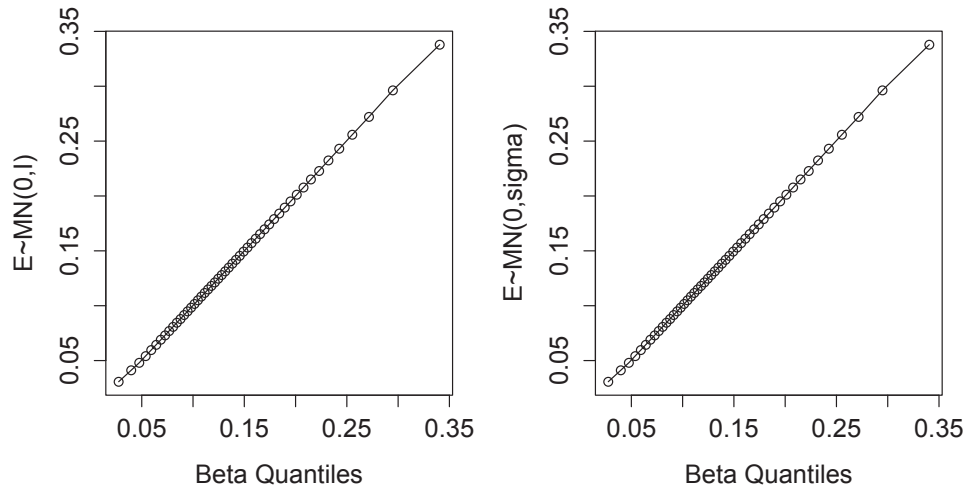

Figure 3.3: *Beta Probability Plot of independent (left) and correlated (right) multivariate normal data after Small's transformation to remove correlation (n=50, p=7)*

3.2 Residuals for checking multivariate normality: non normal data

From the previous section, we have shown that the Fitzmaurice et al's and Small's transformed residuals successfully lead to the right conclusion by not rejecting the correct assumption that data comes from a multivariate normal distribution. Here, we would like to investigate how these residuals perform when data come from a non normal distribution, that is, when the normality assumption is violated. Again, the simple scenarios ($\mathbf{B} = 0$) were considered, however, this time we will take the error term $\mathbf{E}^* = \exp(\mathbf{E})$, so that $\mathbf{E}^*$ has a multivariate log-normal distribution with mean

$$E[\mathbf{E}^*]_i = e^{\mu_i + \frac{1}{2}\Sigma_{ii}},$$

and covariance matrix

$$Var[\mathbf{E}^*]_{ij} = e^{\mu_i + \mu_j + \frac{1}{2}(\Sigma_{ii} + \Sigma_{jj})}(e^{\Sigma_{ij}} - 1),$$

see Kotz et al (2000). Figure 3.4 shows the normal quantile plots for independent and correlated multivariate log-normal data, where we converted the multivariate data into a vector.

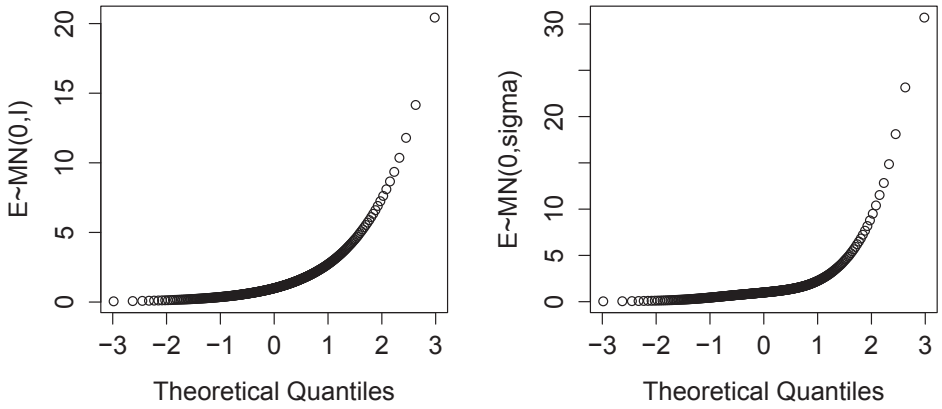

Figure 3.4: *Normal Probability Plot of independent (left) and correlated (right) multivariate log-normal data (n=50, p=7)*

As can be seen, both of the quantile plots were able to reject the normality assumption of the data. However, we again observe that the correlation structure can affect the distribution, where the QQ-plot for correlated data showed more skewness than the uncorrelated data.

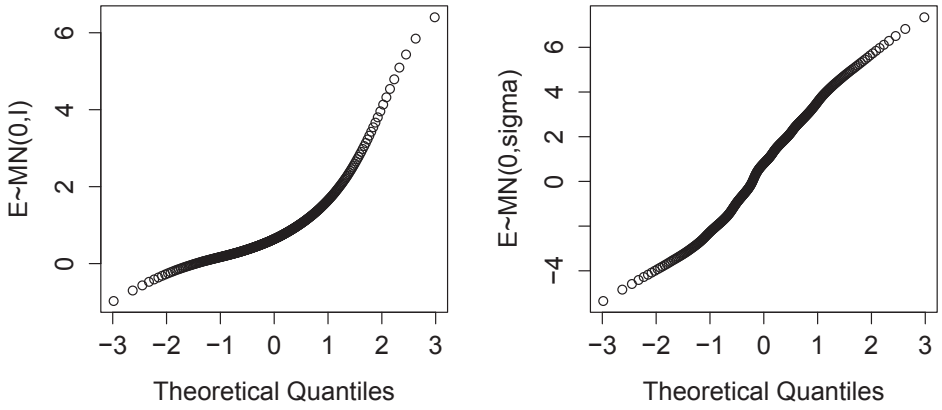

Figure 3.5: *Normal Probability Plot of independent (left) and correlated (right) multivariate log-normal data after Fitzmaurice's transformation to remove correlation (n=50, p=7)*

Figure 3.5 provide the normal quantile plots of independent and correlated multivariate log-normal data using Fitzmaurice's transformation method. As we can see from the Figure, even though both of them were able to provide enough information towards rejecting the normality assumption, the Fitzmaurice's transformation did not do a good job in solely removing the correlation. In fact, the transformation seems to convert the data near normality leading to the wrong conclusion that data are normal when in fact they are not. We, therefore, recommend caution in interpreting Fitzmaurice's uncorrelated residuals when data follow a non-normal distribution.

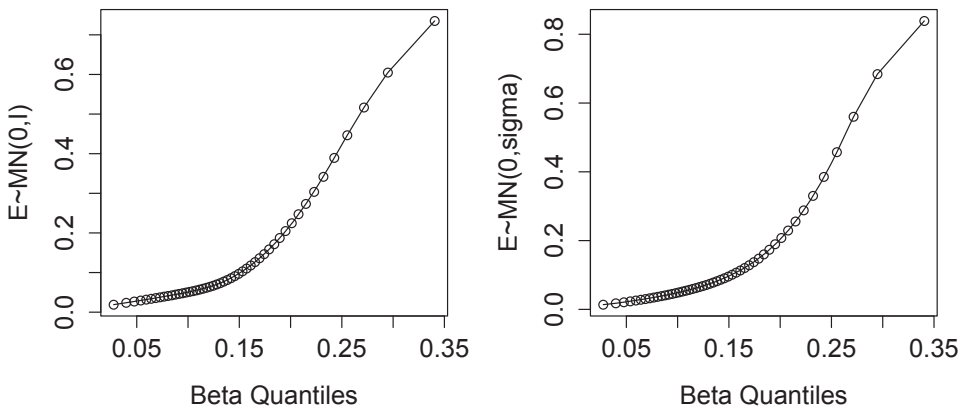

Figure 3.6: *Normal Probability Plot of independent (left) and correlated (right) multivariate log-normal data after Small's transformation to remove correlation (n=50, p=7)*

Figure 3.6 shows the normal quantile plots of independent and correlated multivariate log-normal data, where the residuals are transformed using Small's approach. As we can observe clearly from the quantile plots, the Small's transformation has done an excellent job of removing the correlation structure, while keeping everything else relatively stable. Moreover, both of the quantile plots were able to provide correct conclusion of rejecting the normality assumption. We, therefore, after found out the same pattern across a range of scenarios, would recommend the use of Small's transformation for checking normality, especially when data come from a non-normal distribution.

3.3 Residuals in longitudinal data: normal error

This section consists of simulation results for ordinary residuals in the analysis of longitudinal data where data assumed to be normally distributed with structured mean (time dependent) and covariance matrix $\mathbf{\Sigma}$. Data were generated using the GCM where we assumed linear as well as quadratic curves to represent the mean structure. For quadratic fitting, the design and parameter matrices presented in (3.1) and (3.2) will be used, whereas for the linear fitting, we will consider $\mathbf{A}_l$ and $\mathbf{B}_l$ consisting only of the first two rows of the original $\mathbf{A}^T$ and $\mathbf{B}$:

$$\mathbf{A}_l{}' = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ t_1 & t_2 & t_3 & \cdots & t_p \end{pmatrix} \qquad \mathbf{B}_l = \begin{pmatrix} b_{01} & b_{02} \\ b_{11} & b_{12} \end{pmatrix}. \tag{3.3}$$

The error term was generated using multivariate normal distribution, where $\mathbf{E} \sim MN(0, \Sigma^*)$. Figure 3.7 shows the normal quantile plot of the vector converted ordinary residuals, for both the perfectly fitted and the mis-specified model (i.e. the model was fitted linearly when in fact it has a quadratic mean structure).

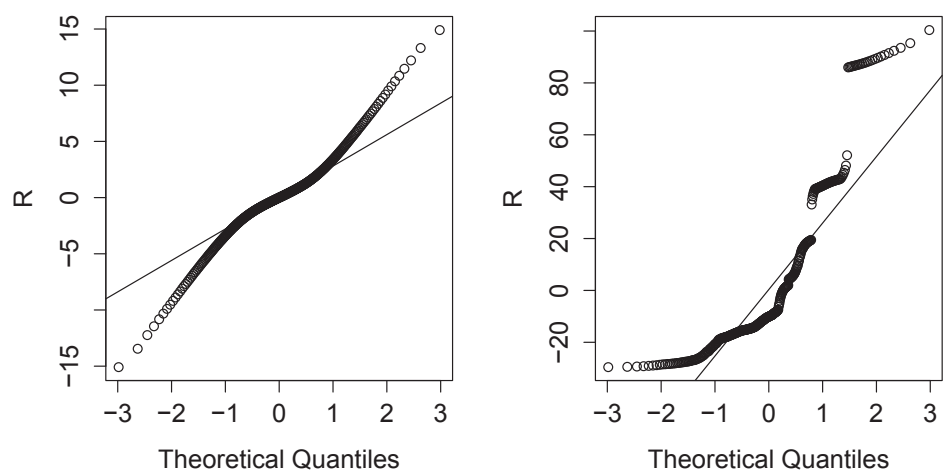

Figure 3.7: *Normal Quantile Plot of residuals* $\mathbf{R}$ *in perfectly fitted (left) and mis-specified (right) GCM (n=50, p=7)*

As we can see from the QQ plot, when the model was perfectly fitted, and there is no systematic error in the model, the quantile plots show exactly the same behaviors as the multivariate normal data

we considered in the previous section, where residual was affected again by the correlation structure. However, when the model was mis-specified using a linear mean structure (instead of the true quadratic mean structure), we observe a discontinuous quantile plot which suggest a strong evidence towards rejecting the true null hypothesis of normality. Therefore, both quantile plots lead to an incorrect conclusion to reject the normality assumption when the data in fact is from the normal distribution.

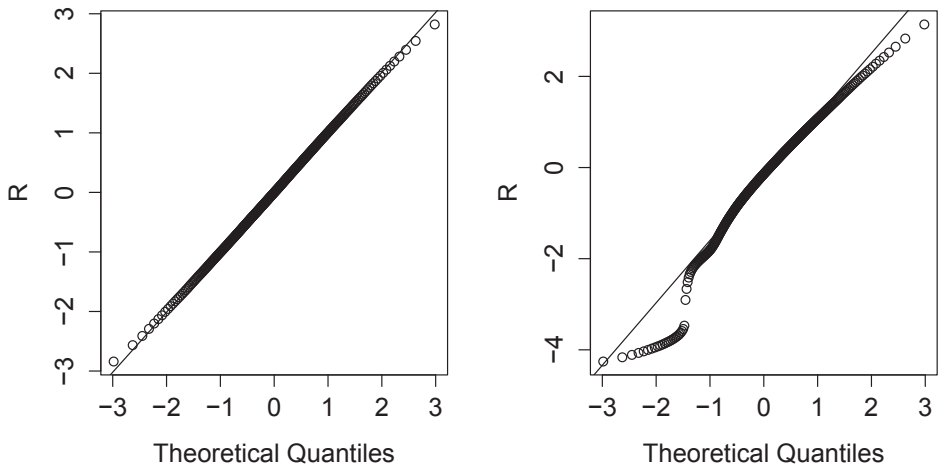

Figure 3.8: *Normal Quantile Plot of Fitzmaurice's residuals* **R** *in perfectly fitted (left) and mis-specified (right) GCM (n=50, p=7)*

As we can see from Figure 3.8, when there is no systematic error in the model fitting, the Fitzmaurice's residuals were again able to successfully remove the correlation and provide a correct decision of accepting the normality assumption. However, in the presence of model mis-specification, we still observe a deviation in the quantile plot, even after the correlation was removed using the suggested transformation. Therefore, the Fitzmaurice's residual failed to provide accurate conclusion in investigating the normality assumption.

Figure 3.9 shows the beta quantile plot of ordinary residuals after Small's transformation for both the perfectly fitted and mis-specified GCM. After the removal of the correlation structure by the Small's method, as we expected, the corresponding normal quantile plot lie right on the straight line, which

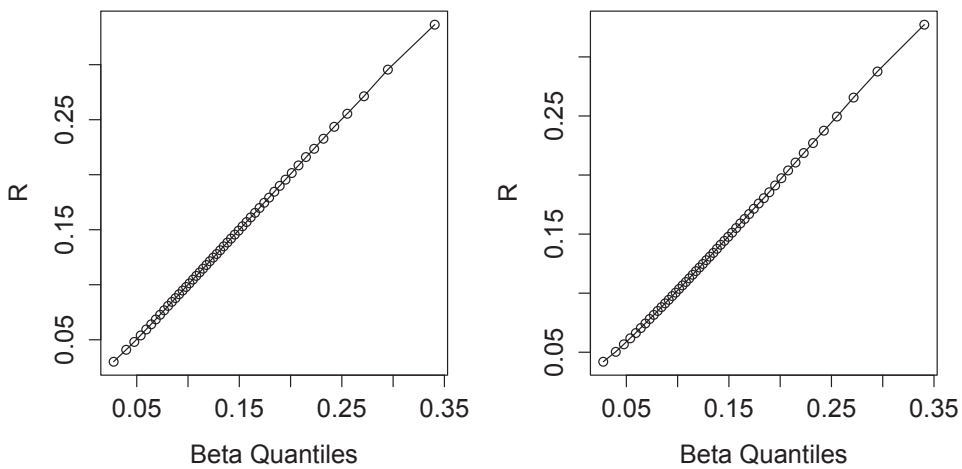

Figure 3.9: *Beta Quantile Plot of Small's residuals* $\mathbf{R}$ *in perfectly fitted (left) and mis-specified (right) GCM (n=50, p=7)*

confirms that data came from the normal distribution. On the other hand, in the case where the GCM was mis-specified by a linear mean structure, the normal quantile plot remain the same linear shape, which also lead to a correct conclusion about the normality assumption. Therefore, we think that the Small's method not only removed the correlation structure of the data, it may have also eliminated systematic error from the residuals. However, after noticing the same pattern over a wide range of simulation scenarios, we suggest that further theoretical investigation must be done in order to check that this is not just because of the nature of these particular data sets.

3.4 Decomposed Residuals in the GCM: normal error

In the previous section, we have observed how systematic errors influence the residuals in the analysis of longitudinal data. However, using QQ plots, it is not clear if the deviation from the straight line is due to non-normality or systematic error. In this section, we provide the results of the simulation for von Rosen's decomposed residuals. The simulated data has a quadratic mean structure which were generated using the design and parameter matrix as described in (3.1) and (3.2), and the error term is generated as $\mathbf{E} \sim MN(0, \Sigma^*)$. We fitted both linear (which is a mis-specification) and quardatic (correct fit) to the data. In the linear fitting of the GCM, the design and parameter matrices are generated according to (3.3).

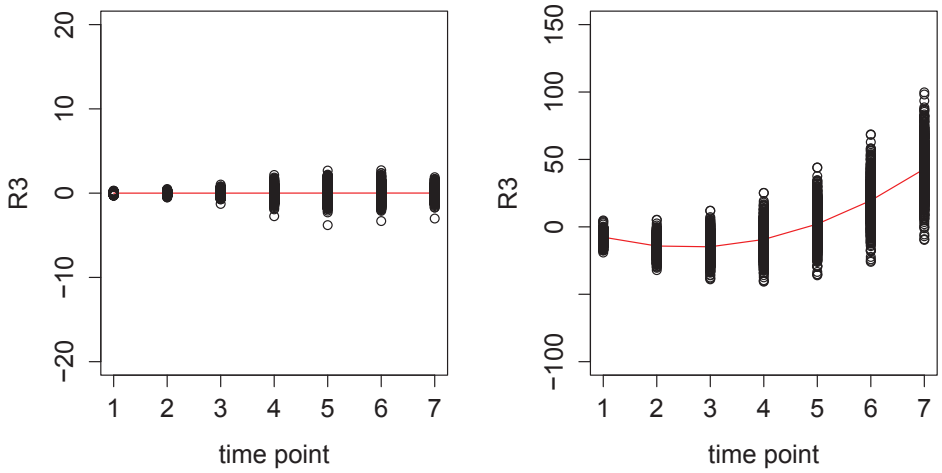

Figure 3.10: *Scatter plot of decomposed residuals* $\mathbf{R}_3$ *in perfectly fitted (left) and mis-specified (right) GCM (n=50, p=7)*

Figure 3.10 provides the scatter plot of the decomposed residual $\mathbf{R}_3$ in both the perfectly fitted and the mis-specified GCM, where all the observation of the 1000 simulations were plotted in its respective scatter plot, and the solid line indicate the average of the simulations. As we can clearly see from the plots, when the model is perfectly fitted, points in the scatter plot lie randomly around zero, reflecting no systematic error; on the other hand, when the GCM was fitted with a linear mean structure, the scatter plot shows a clear quadratic trend indicating the left over information that were being unexplained by the fitted model. Therefore, it confirms von Rosen's (1995) and Hamid and von Rosen's (2006) suggestion that $\mathbf{R}_3$ may be used to identify systematic error in model fitting. We have done several other scenarios with different mean structures, different sample sizes as well as various number of time points, and our empirical evidence is consistent for all the scenarios.

Figure 3.11 shows the normal quantile plot of the vector residuals $\mathbf{R}_1 + \mathbf{R}_2$, for both the perfectly fitted and mis-specified model. We can see that with $\mathbf{R}_3$ removed, the quantile plots of $\mathbf{R}_1 + \mathbf{R}_2$ show identical pattern regardless of the model fitting. Hence, we concluded that $\mathbf{R}_1 + \mathbf{R}_2$ may not be influenced by the systematic component of modeling. However, due to the effect of correlation

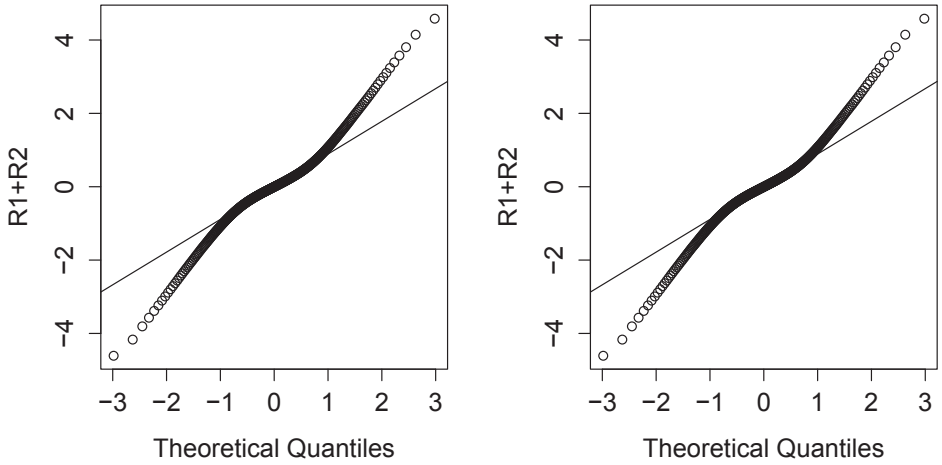

Figure 3.11: *Normal Quantile Plot of residuals* $\mathbf{R}_1+\mathbf{R}_2$ *in perfectly fitted (left) and mis-specified (right) GCM (n=50, p=7)*

structure, the quantile plots are still showing deviation from the normal straight line. Therefore, we again suggest transformation of the $\mathbf{R}_1+\mathbf{R}_2$ to remove correlation, in order to obtain accurate results.

Figure 3.12 provide the quantile plot of the Fitzmaurice's as well as the Small's residuals $\mathbf{R}_1+\mathbf{R}_2$ in both perfectly fitted and mis-specified GCM. As we expected, with the decomposed $\mathbf{R}_3$ and the correlation structure removed, all of quantile plots are correctly indicating the normality of the GCM, as they all lie right on the normal straight line regardless of the existence of the systematic error. Therefore, our results again confirm that the decomposed residual $\mathbf{R}_1+\mathbf{R}_2$ does not depend on the model fitting, and hence is a better choice over the ordinary residual for checking normality, in particular in the presence of model mis-specified.

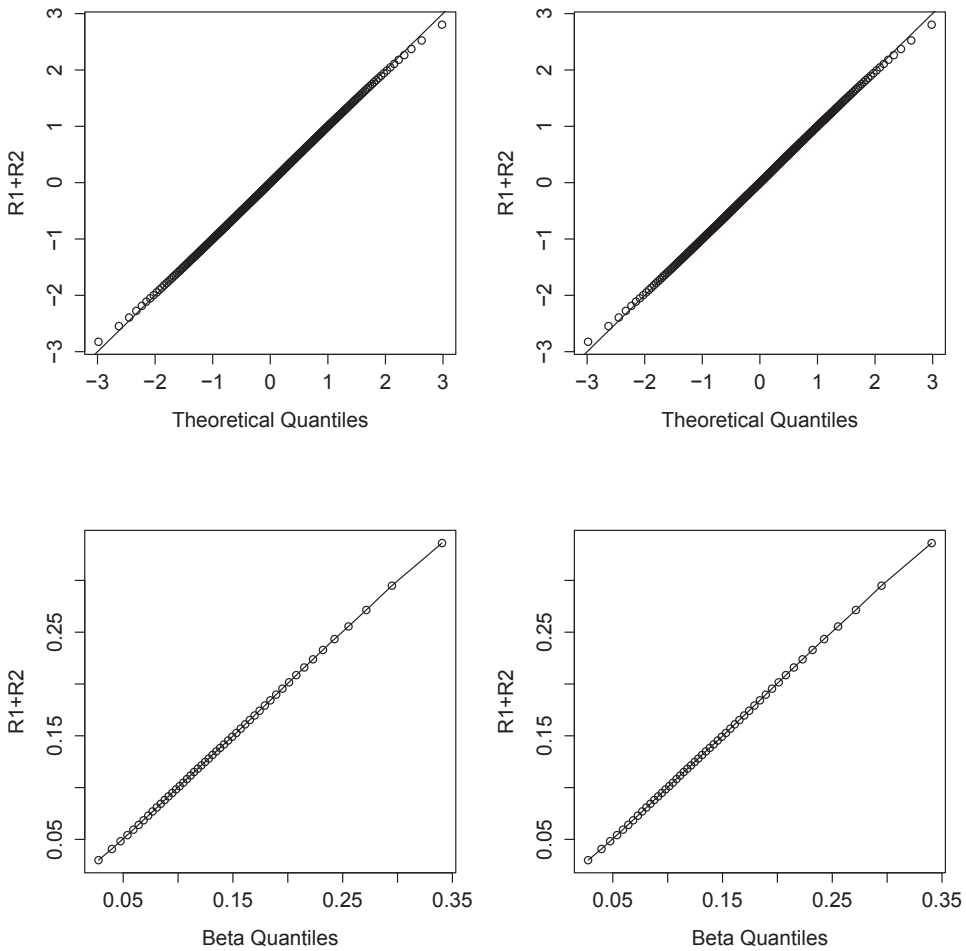

Figure 3.12: *Quantile Plot of Fitzmaurice's (top left) and Small's (bottom left)* $\mathbf{R}_1 + \mathbf{R}_2$ *in perfectly fitted GCM, and Quantile Plot of Fitzmaurice's (top right) and Small's (bottom right)* $\mathbf{R}_1 + \mathbf{R}_2$ *in mis-specified GCM. (n=50, p=7)*

3.5 Decomposed Residuals in GCM: non normal error

In the previous sections, we used data from multivariate normal distribution and showed the advantage of using decomposed residuals over the ordinary residuals, in terms of identifying systematic error and removing it to provide more appropriate residuals for checking the normality assumption. In this section, we extend our investigation to non-normal data, to see if the residuals are still able to detect systematic errors as well as able to reject the null hypothesis that data are distributed according to the normal distribution. We generate $\mathbf{E}^* = \exp(\mathbf{E})$, where $\mathbf{E}^*$ has a multivariate log-normal distribution (see Kotz (2000)). For quadratic fitting, we used design and parameter matrices according to (3.1) and (3.2), and for the linear fitting we used the setting provided in (3.3).

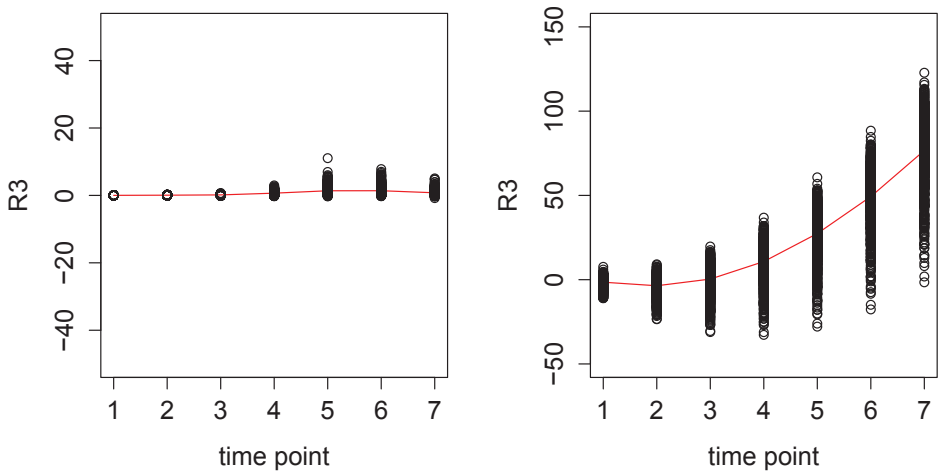

Figure 3.13: *Scatter plot of decomposed residuals* $\mathbf{R}_3$ *in perfectly fitted (left) and mis-specified (right) log-normal GCM (n=50, p=7)*

Figure 3.13 shows the scatter plot of the decomposed residual $\mathbf{R}_3$ in both the perfectly fitted and mis-specified GCM, where the GCM follow a multivariate log-normal distribution. All the observation of

the 1000 simulations were plotted in its respective scatter plot, and the solid line indicate the average of the simulations. we can clearly see that, regardless of the normality assumption of the GCM, the decomposed $\mathbf{R}_3$ successfully identifies the systematic error and displays the left over information with a quadratic trend in the presence of mis-specified. Therefore, the usefulness of $\mathbf{R}_3$ for identifying the systematic error in the GCM is valid not only for normal errors, but also true under non-normality. Again, we have simulated several other scenarios under different model parameters and the findings are consistent. However, one can observe that there is a variation in the scatter plot, that may be due to the fact that, when we took the exponential transformation of the error term $\mathbf{E}$, it also effect the mean of the error term.

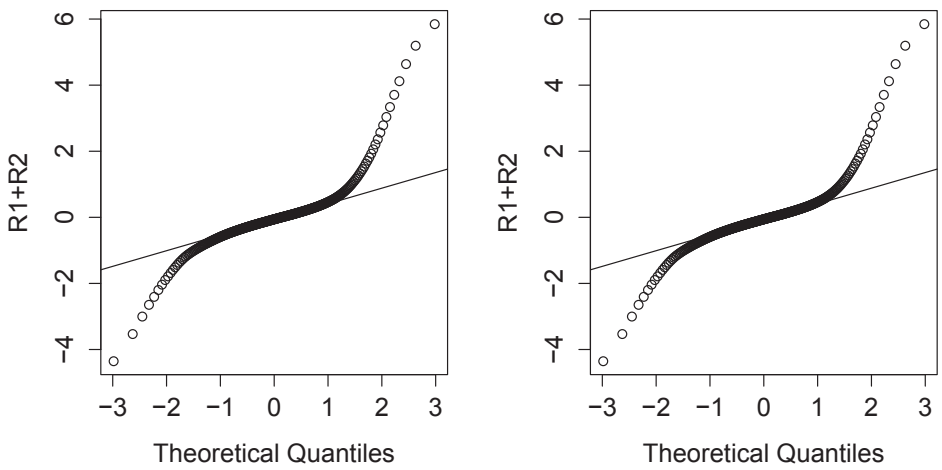

Figure 3.14: *Normal Quantile Plot of Fitzmaurice's residuals* $\mathbf{R}_1 + \mathbf{R}_2$ *in perfectly fitted (left) and mis-specified (right) log-normal GCM (n=50, p=7)*

Figure 3.14 compares the normal quantile plots of the Fitzmaurice's decomposed $\mathbf{R}_1 + \mathbf{R}_2$ for perfectly fitted GCM with the GCM with that of the mis-specified model, where the error terms for the models follow multivariate log-normal distribution. Again, the resulting plots clearly shows that, with the removal of $\mathbf{R}_3$, the decomposed $\mathbf{R}_1 + \mathbf{R}_2$ no longer depend on the systematic error resulting from misfitting, as the respectively probability plots show identical pattern. Moreover, the two plots provide a strong evidence towards rejecting the null hypothesis that data came from the normal distribution, leading to the correct conclusion.[]

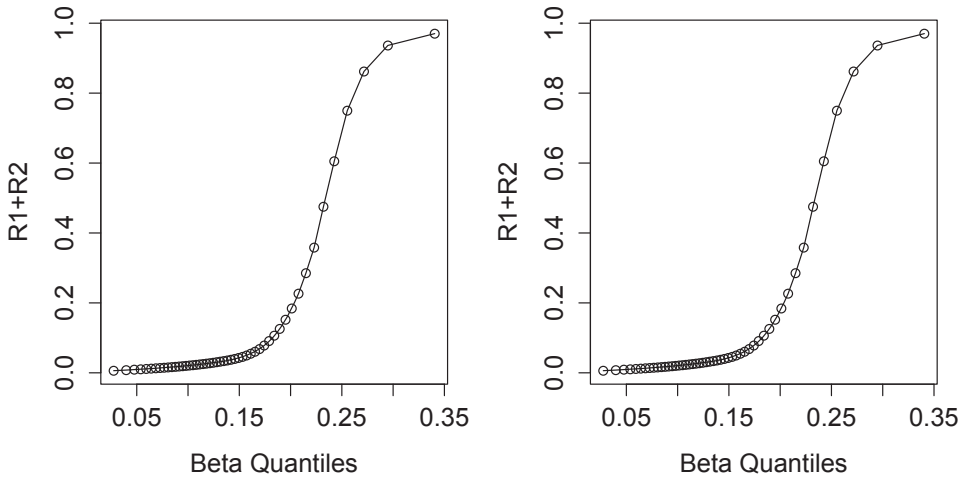

Figure 3.15: *Beta Quantile Plot of Small's residuals* $\mathbf{R}_1 + \mathbf{R}_2$ *in perfectly fitted (left) and mis-specified (right) log-normal GCM (n=50, p=7)*

Figure 3.15 provides the comparison of the beta quantile plots of the Small's decomposed $\mathbf{R}_1 + \mathbf{R}_2$ for perfectly fitted GCM with that of the mis-specified GCM, where the error terms follow multivariate log-normal data. As we can see, the quantile plots show exactly the same pattern, once again confirming that these residuals are independent of the systematic component of the model. The plots also show deviation from normality leading to a correct decision towards rejecting the null hypothesis of multivariate normality.

Chapter 4

Real Data Application

In this chapter, we provide real data examples to illustrate the decomposed residuals including the transformed versions of the decomposed residuals. We consider Potthoff and Roys' (1964) dental data and the glucose data from Zerbe (1979). We start by looking at the profile plot of the data in order to have a basic feeling of their behaviors. Then, scatter plot of the decomposed $\mathbf{R}_3$ is used to determine its ability to detect any systematic error in model fitting. Finally, we examine the quantile plot of both the Fitzmaurice's and Small's transformed $\mathbf{R}_1 + \mathbf{R}_2$, to check the normality assumption.

4.1 Dental data

This section provides the analysis of the decomposed residual in the dental data. This data first appeared in Potthoff and Roy (1964) and subsequently investigated by many, including Lee and Geisser (1975), Rao (1987), Lee (1988a, 1991) and Pan and Fang (2002). The data consists of distance (in mm) from the center of the pituitary to the pterygomaxillary fissure, and is taken from 11 girls and 16 boys at ages 8, 10, 12, and 14 years. If we fit the GCM linearly, the design matrix can be written as:

$$\mathbf{A}' = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 8 & 10 & 12 & 14 \end{pmatrix} \qquad \mathbf{C} = \begin{pmatrix} \mathbf{1}_{11} & \mathbf{0}_{16} \\ \mathbf{0}_{11} & \mathbf{1}_{16} \end{pmatrix},$$

where we have group $m = 2$, sample size $n_1 = 11$, $n_2 = 16$, and time point $p = 4$. Note that this dental data has two outliers in the boy measurement group according to previous study (Pan and Fang (2002)).

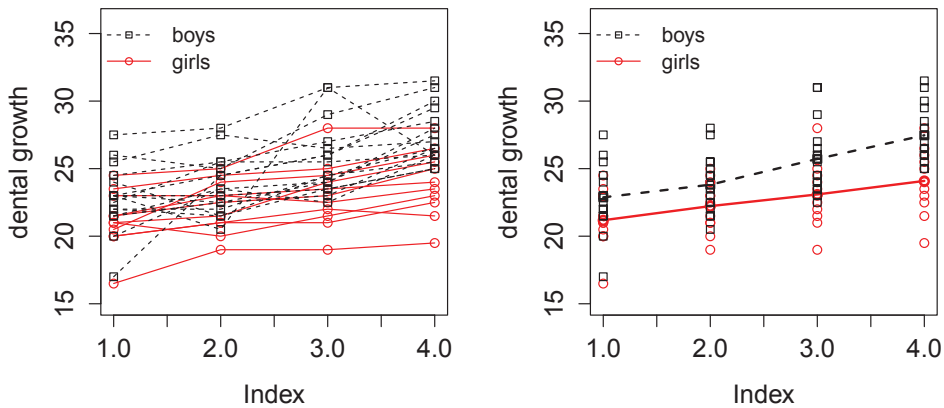

Figure 4.1: *Profile plot of Dental data (left) and group average (right)*

Figure 4.1 shows the profile plots of the dental data as well as their group mean. The red solid line represent the measurements for the girls while the black dashed line represent the boy's. As we can seen from the plot, the data shows a overall linear structure except measurements from one boy which fluctuate dramatically over time.

Both the normal quantile plot of the ordinary residual $\mathbf{R}$ and the scatter plot of the decomposed $\mathbf{R}_3$ are shown in Figure 4.2. As we can see in the scatter plot, the results of the decomposed $\mathbf{R}_3$ lie randomly and closely around zero, which is consistent with our observation that the data has a linear mean structure. On the other hand, with the limited systematic effect, the normal quantile plot of the ordinary residual shows a relatively straight line and therefore suggest a normal behavior of the data. However, it is important to note that the quantile plot might be affected by the correlation structure, and does not provide any information about the outliers in the data.

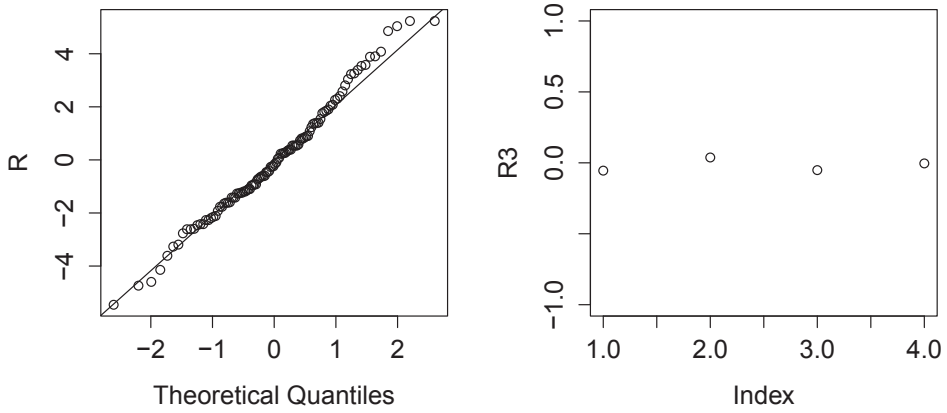

Figure 4.2: *Normal quantile plot of ordinary residual (left) and scatter plot of decomposed* $\mathbf{R}_3$ *(right) of Dental data*

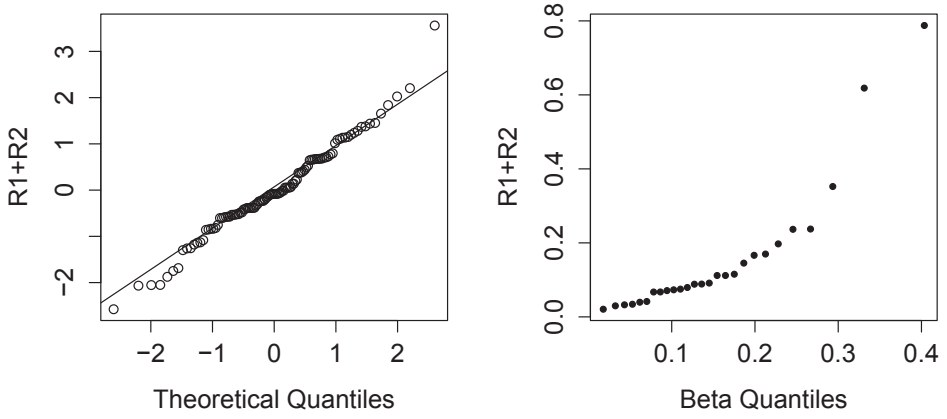

Figure 4.3: *Quantile plot of Fitzmaurice's (left) and Small's (right) decomposed residuals* $\mathbf{R}_1 + \mathbf{R}_2$ *of Dental data*

Figure 4.3 represent the normal quantile plot of the Fitzmaurice's transformed $\mathbf{R}_1 + \mathbf{R}_2$ as well as the beta quantile plot of the $\mathbf{R}_1 + \mathbf{R}_2$ by Small's method. As shown in the plots, both of them confirm the normality assumption, and while they both suggest the existence of possible outliers in the data, only the Small's method was able to provide a clear identification of the two outliers. Figure 4.4 shows the QQ-polts after the outliers have been removed.

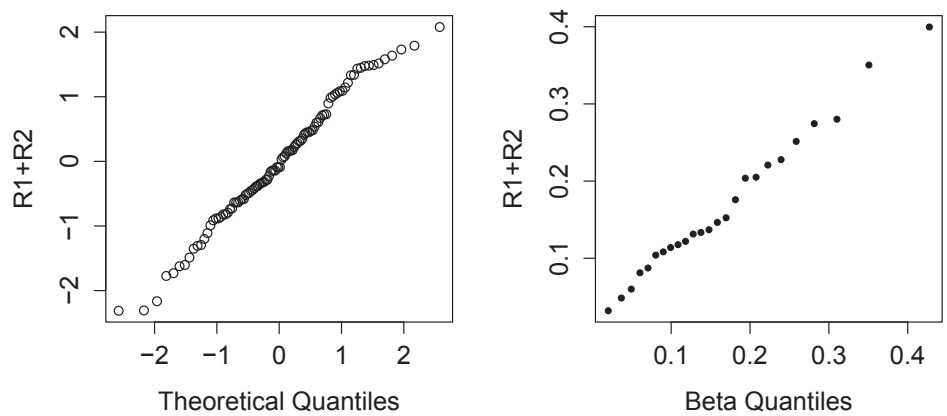

Figure 4.4: *Quantile plot of Fitzmaurice's (left) and Small's (right) decomposed residuals* $\mathbf{R}_1 + \mathbf{R}_2$ *of Dental data without outliers*

4.2 Glucose data

This section provides the analysis of the decomposed residuals for the glucose data. This data was first presented by Zerbe (1979), and has been studied extensively by Chi and Reinsel (1989), and Keramidas and Lee (1995). Data comes from a standard glucose tolerance test which measure the plasma inorganic phosphate, and consists of a repeated measurements on 13 control and 20 obese patients at 0, 0.5, 1, 1.5, 2, 3, 4, and 5 hours after the standard dose-oral glucose is administered. If we fit the GCM linearly, the design matrix can be written as:

$$\mathbf{A}' = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0.5 & 1 & 1.5 & 2 & 3 & 4 & 5 \end{pmatrix} \qquad \mathbf{C} = \begin{pmatrix} \mathbf{1}_{13} & \mathbf{0}_{20} \\ \mathbf{0}_{13} & \mathbf{1}_{20} \end{pmatrix},$$

here we have group $m = 2$, sample size $n_1 = 13$, $n_2 = 20$, and time point $p = 8$. Note that according to previous study (see Pan and Fang 2002), a measurement of at least the second degree should be chosen to fit this GCM. Therefore, by fitting the GCM with a linear mean structure, we would expect to observe a systematic error in our residuals analysis.

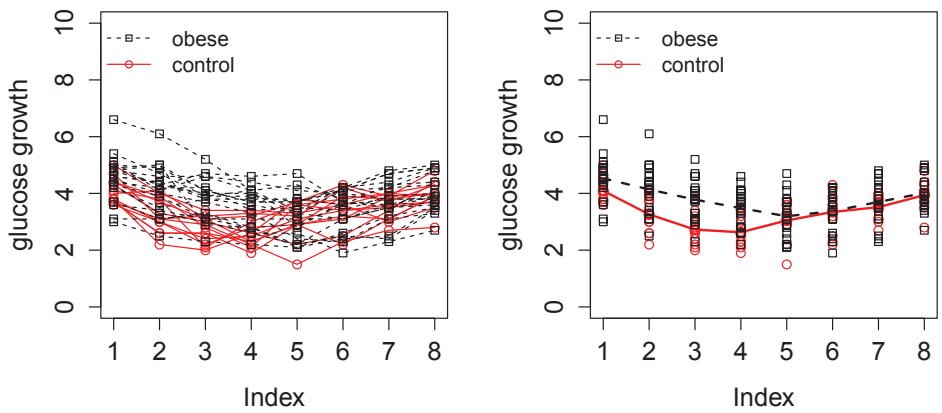

Figure 4.5: *Profile plot of Glucose data (left) and group average (right)*

The profile plots of the glucose data and their group mean are shown in Figure 4.5. Observation of the patients in the control and obese group is represented by the red solid and black dashed line, respectively. As can be seen in the profile plot, the observations shows a concave up curve which reflect that the data has at least a second order polynomial mean structure.

Figure 4.6 provides the normal quantile plot of the ordinary residual $\mathbf{R}$ and the scatter plot of the decomposed $\mathbf{R}_3$ of the Glucose data. Recall that in the previous studies, the data was fitted using second degree polynomial mean structure. However, we intentionally fitted a linear growth curve with the aim of identifying this systematic error. As can be seen in the scatter plot in 4.6, $\mathbf{R}_3$ shows a clear quadratic pattern, therefore successfully identify the systematic error of our model fit. The plot of $\mathbf{R}_3$ after modeling quadratic and third degree polynomial for the glucose data are presented in Figure 4.7.

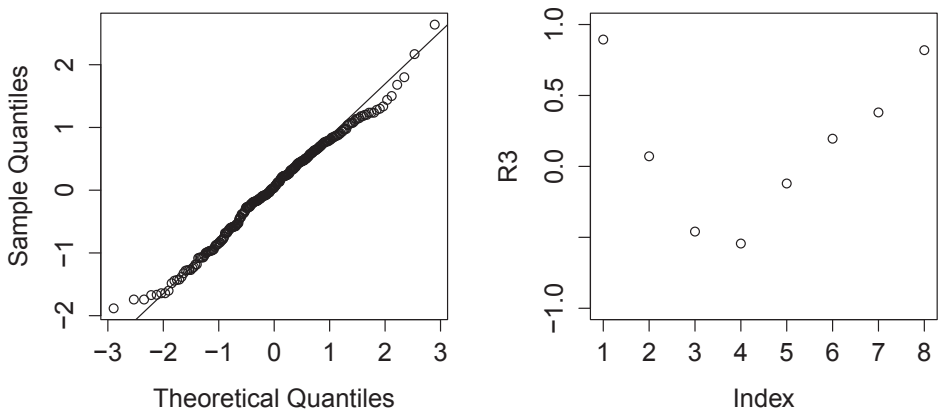

Figure 4.6: *Normal quantile plot of ordinary residual (left) and scatter plot of decomposed* $\mathbf{R}_3$ *(right) of Glucose data*

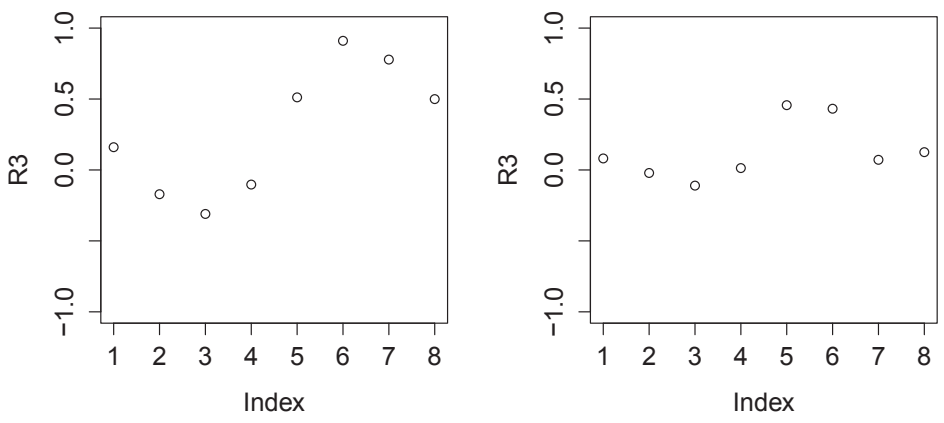

Figure 4.7: *Scatter plot of decomposed* $\mathbf{R}_3$ *for quadratic (left) and third degree polynomials (right) fitting*

As one can see, with the increase of the polynomial fitting, the magnitude of the error decrease. However, both scatter plots are still showing a clear pattern, one could therefore suggest different method of fitting. On the other hand, despite the effect of systematic error and the correlation structure, we observe that the normal quantile plot of the ordinary residuals (Figure 4.6) lie closely on the straight line which give no evidence to reject the normality assumption.

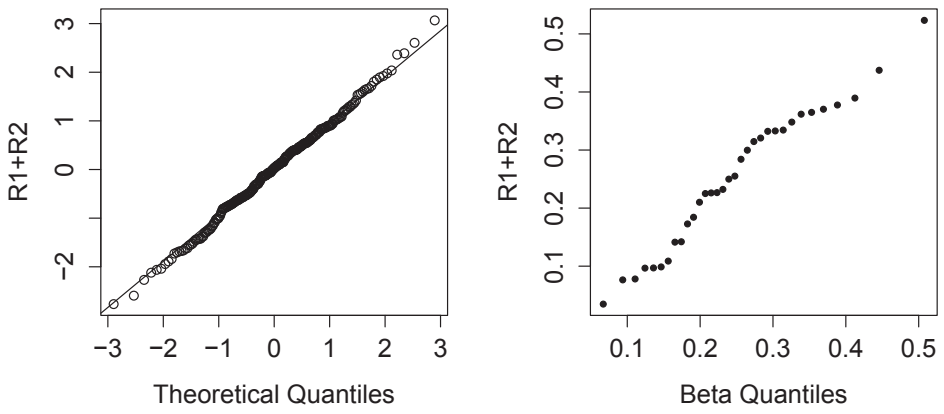

Figure 4.8: *Quantile plot of Fitzmaurice's (left) and Small's (right) decomposed residuals* $\mathbf{R}_1 + \mathbf{R}_2$ *of Glucose data*

Figure 4.8 shows both the normal quantile plot of the Fitzmaurice's transformed $\mathbf{R}_1 + \mathbf{R}_2$ and the beta quantile plot of the $\mathbf{R}_1 + \mathbf{R}_2$ using Small's approach for the Glucose data. As shown in the plots, with the removal of the correlation structure and the systematic error, residuals in both quantile plots lie closely around a straight line, therefore we concluded that the data follows a multivariate normal distribution.

Chapter 5

Discussion

To summarize, we have demonstrated that von Rosen's decomposed residuals are preferable over the ordinary residuals in model diagnostics for longitudinal data. The scatter plot of $\mathbf{R}_3$ is an excellent tool to identify systematic error of the model fitting; by removing the correlation structure using the Fitzmaurice's transformation or the Small's graphical method, the decomposed $\mathbf{R}_1 + \mathbf{R}_2$ can provide reliable analysis for checking the normality assumption. However, as the Small's graphical method is able to provide information about the true distribution of the data, we suggest its use instead of Fitzmaurice's transformation.

In this thesis, we focused on verifying the normality assumption and model mis-specification using graphical approaches by visualizing scatter and quantile plots. However, conclusions that come from these graphical approaches might vary from person to person, and hence arguably subjective. In order to overcome this subjectivity, we suggest that one may use a formal test that will give a clear indicator (e.g. p-value) to determine whether there is evidence to reject the normality. One such a test, for instance is, using Kolmogorov-Smirnovs test of equality of two distributions.

We would like to note that we transformed the error term $\mathbf{E}$ using the exponential function to show the behavior of residuals under non-normality, where the error term $\mathbf{E}$ is distributed as a multivariate normal random variable. The transformed data therefore, follows the multivariate log-normal distribution. However, it is important to note that $E[\mathbf{E}^*]_i = e^{\mu_i + \frac{1}{2}\Sigma_{ii}}$. Indicating that the transformation

we used also effect the mean of the error term, leading to an expected value of the error term $\mathbf{E}^*$ which is no long equal to zero. This might affect the behaviour of the residuals and hence may lead us to conclude the presence of model mis-specification when in fact the systematic component of the data has been appropriately modeled. Therefore, we recommend further investigation using nonnormal error with mean zero. One, for instance, can standardize our transformed data to ensure the mean of the error term equal to zero.

Furthermore, decomposed residuals in the Growth Curve Model are defined by projecting the observation matrix on the space orthogonal to the space generated by the design matrices. In fact, this idea can be extended to Extended Growth Curve Model (EGCM)

$$\mathbf{X} = \sum_{i=1}^{m} \mathbf{A}_i \mathbf{B}_i \mathbf{C}_i + \mathbf{E}. \tag{5.1}$$

This model was first introduced by von Rosen (1989), while the canonical form of the model was considered by Banken (1984). A paper by Hamid and von Rosen (2006) provided decomposed residuals for such bilinear models for the case when $m = 2$. As a result the ordinary residuals were decomposed into 4 parts: $\mathbf{R}_1$, $\mathbf{R}_2$, $\mathbf{R}_3$, and $\mathbf{R}_4$. The authors suggested that $\mathbf{R}_1 + \mathbf{R}_2$ can be used to check the normality assumption, as in the case of GCM; while $\mathbf{R}_3$ and $\mathbf{R}_4$ can be used to check the systematic fitting of model $\mathbf{A}_1\mathbf{B}_1\mathbf{C}_1$; and $\mathbf{A}_2\mathbf{B}_2\mathbf{C}_2$. The authors also provided a general framework for defining residuals in the more general EGCM (5.1). Our empirical investigation, using simulation, can be extended to residuals in the EGCM. The residuals can also be further decomposed to investigate different components of the systematic modeling. Moreover, the transformation we proposed in this thesis can be used to transform the decomposed residuals into uncorrelated residuals. This thesis laid the ground for further investigation toward the decomposed residuals in the general EGCM, which is useful in the analysis of longitudinal data when groups are assumed to have mean growth curves with different shapes (eg. one group linear and the other group quadratic).

Appendix A

Table

A.1 Residuals analysis setting of non-structured GCM

Simulation	Sample size n	♯ of time point p	covriance strcture
Sim204I, Sim204E	20	4	**I**, Pott and Roy
Sim207I, Sim207E	20	7	**I**, Mouse
Sim208I, Sim208E	20	8	**I**, Glucose
Sim304I, Sim304E	30	4	**I**, Pott and Roy
Sim307I, Sim307E	30	7	**I**, Mouse
Sim308I, Sim308E	30	8	**I**, Glucose
Sim404I, Sim404E	40	4	**I**, Pott and Roy
Sim407I, Sim407E	40	7	**I**, Mouse
Sim408I, Sim408E	40	8	**I**, Glucose
Sim504I, Sim504E	50	4	**I**, Pott and Roy
Sim507I, Sim507E	50	7	**I**, Mouse
Sim508I, Sim508E	50	8	**I**, Glucose
Sim1004I, Sim1004E	100	4	**I**, Pott and Roy
Sim1007I, Sim1007E	100	7	**I**, Mouse
Sim1008I, Sim1008E	100	8	**I**, Glucose

Note: **I** represent the identity matrix, while Pott and Roy, Mouse, and Glucose represent the covariance matrix of the corresponding data set in Pan and Fang(2002), respetively.

A.2 Simulation settings description of Residuals in bilinear structured GCM

Simulation setting	Sample size n	♯ of time point p	Covriance strcture	Design Matirx A
Sim204R-l	20	4	Pott and Roy	Linear
Sim204R-q	20	4	Pott and Roy	Quardratic
Sim207R-l	20	7	Mouse	Linear
Sim207R-q	20	7	Mouse	Quardratic
Sim208R-l	20	8	Glucose	Linear
Sim208R-q	20	8	Glucose	Quardratic
Sim304R-l	30	4	Pott and Roy	Linear
Sim304R-q	30	4	Pott and Roy	Quardratic
Sim307R-l	30	7	Mouse	Linear
Sim307R-q	30	7	Mouse	Quardratic
Sim308R-l	30	8	Glucose	Linear
Sim308R-q	30	8	Glucose	Quardratic
Sim404R-l	40	4	Pott and Roy	Linear
Sim404R-q	40	4	Pott and Roy	Quardratic
Sim407R-l	40	7	Mouse	Linear
Sim407R-q	40	7	Mouse	Quardratic
Sim408R-l	40	8	Glucose	Linear
Sim408R-q	40	8	Glucose	Quardratic
Sim504R-l	50	4	Pott and Roy	Linear
Sim504R-q	50	4	Pott and Roy	Quardratic
Sim507R-l	50	7	Mouse	Linear
Sim507R-q	50	7	Mouse	Quardratic
Sim508R-l	50	8	Glucose	Linear
Sim508R-q	50	8	Glucose	Quardratic
Sim1004R-l	100	4	Pott and Roy	Linear
Sim1004R-q	100	4	Pott and Roy	Quardratic
Sim1007R-l	100	7	Mouse	Linear
Sim1007R-q	100	7	Mouse	Quardratic
Sim1008R-l	100	8	Glucose	Linear
Sim1008R-q	100	8	Glucose	Quardratic

Note: Pott and Roy, Mouse, and Glucose represent the covariance matrix of the corresponding data set in Pan and Fang(2002), respetively.

Appendix B

R Code

B.1 Small's graphical method (Small, 1978)

```
fix(Small)
function (xx)
{
    n<-ncol(xx)
    p<-nrow(xx)
### calculating equation (2.8) ###
    xbar<-c(0)
    for(i in 1:p)
    {
        xbar[i]<-sum(xx[i,])/n
    }
### calculating equation (2.9) ###
    M<-vector(mode="list", length=n)
    for(j in 1:n)
    {
        M[[j]]<-((xx[,j] - xbar) %*% t(xx[,j] - xbar))
    }
```

```
    S<-matrix(0,p,p)
    for(j in 1:n)
    {
        S<-(S + M[[j]])
    }
    S<-S/(n-1)
### calculating equation (2.7) ###
    c<-c(0)
    for (k in 1:n)
    {
        c[k]<-(n/(n-1)^2) * t(xx[,k] - xbar) %*% solve(S) %*% (xx[,k] - xbar)
    }
c
}
```

B.2 Examination of random multivariate data

```
### workspace set up ###
library(MASS)
library(Matrix)
E<-vector(mode="list", length=1000)
E_vec<-vector(mode="list", length=1000)
L<-vector(mode="list", length=1000)
E_star<-vector(mode="list", length=1000)
E_small<-vector(mode="list", length=1000)
E_vec_sum<-c(0)
E_star_sum<-c(0)
E_small_sum<-c(0)
```

```
### simulation of 1000 ###
for (i in 1:1000)
{
  ### generating multivariate data ###
  set.seed(99+i)
  E[[i]]<-mvrnorm(n=50, rep(0, 7), diag(7))
  ### transformation using simple vectorizing and sorting ###
  E_vec[[i]]<-sort(as.vector(E[[i]]))
  ### transformation using Cholesky factorization ###
  L[[i]]<-t(chol(cov(E[[i]])))
  E_star[[i]]<-solve(L[[i]]) %*% t(E[[i]])
  E_star[[i]]<-sort(as.vector(E_star[[i]]))
  ### transformation using Small's graphcal method ###
  E_small[[i]]<-sort(Small(t(E[[i]])))
  ### sum of the 1000 simulation ###
  E_vec_sum<-(E_vec_sum + E_vec[[i]])
  E_star_sum<-(E_star_sum + E_star[[i]])
  E_small_sum<-(E_small_sum + E_small[[i]])
}

### average of the 1000 simulation ###
E_vec_avg<-E_vec_sum/1000
E_star_avg<-E_star_sum/1000
E_small_avg<-E_small_sum/1000

### Normal probability plot of average in vector form ###
qqnorm(E_vec_avg,main="Normal Probability Plot - Vector",
       ylab="E~MN(0,I)" )
qqline(E_vec_avg)
```

```
### Normal probability plot of average of Cholesky factorization ###
qqnorm(E_star_avg,main="Normal Probability Plot - Cholesky",
       ylab="E~MN(0,I)" )
qqline(E_star_avg)

### Normal probability plot of average of Small's graphical method ###
q<-c(0)
for(i in 1:n)
{
  q[i]<-qbeta((i-0.5)/n,0.5 * p,0.5 * (50 - 7 - 1))
}
plot(q,E_small_avg,main="Beta Probabilty Plot - Small",xlab="Beta
           Quantiles", ylab="E~MN(0,I)",type="o")
```

For generating multivariate data with covariance matrix Σ^*, replace the command of "generating multivariate data" by

```
### workspace set up ###
library(MASS)
library(Matrix)
E<-vector(mode="list", length=1000)
E_vec<-vector(mode="list", length=1000)
L<-vector(mode="list", length=1000)
E_star<-vector(mode="list", length=1000)
E_small<-vector(mode="list", length=1000)
E_vec_sum<-c(0)
E_star_sum<-c(0)
E_small_sum<-c(0)
e1<-c(0.0005,0.0007,0.0006,0.0010,0.0007,0.0010,0.0010)
e2<-c(0.0007,0.0012,0.0014,0.0024,0.0024,0.0030,0.0025)
```

```
e3<-c(0.0006,0.0014,0.0030,0.0050,0.0060,0.0061,0.0053)
e4<-c(0.0010,0.0024,0.0050,0.0105,0.0127,0.0128,0.0106)
e5<-c(0.0007,0.0024,0.0060,0.0127,0.0179,0.0172,0.0142)
e6<-c(0.0010,0.0030,0.0061,0.0128,0.0172,0.0188,0.0151)
e7<-c(0.0010,0.0025,0.0053,0.0106,0.0142,0.0151,0.0142)
sigma<-cbind(e1,e2,e3,e4,e5,e6,e7)

### simulation of 1000 ###
for (i in 1:1000)
{
  ### generating multivariate data ###
  set.seed(99+i)
  E[[i]]<-mvrnorm(n=50, rep(0, 7), 100*sigma)
  ### transformation using simple vectorizing and sorting ###
  E_vec[[i]]<-sort(as.vector(E[[i]]))
  ### transformation using Cholesky factorization ###
  L[[i]]<-t(chol(cov(E[[i]])))
  E_star[[i]]<-solve(L[[i]]) %*% t(E[[i]])
  E_star[[i]]<-sort(as.vector(E_star[[i]]))
  ### transformation using Small's graphcal method ###
  E_small[[i]]<-sort(Small(t(E[[i]])))
  ### sum of the 1000 simulation ###
  E_vec_sum<-(E_vec_sum + E_vec[[i]])
  E_star_sum<-(E_star_sum + E_star[[i]])
  E_small_sum<-(E_small_sum + E_small[[i]])
}

### average of the 1000 simulation ###
E_vec_avg<-E_vec_sum/1000
```

```
E_star_avg<-E_star_sum/1000
E_small_avg<-E_small_sum/1000

### Normal probability plot of average in vector form ###
qqnorm(E_vec_avg,main="Normal Probability Plot - Vector",
       ylab="E~MN(0,sigma)" )
qqline(E_vec_avg)

### Normal probability plot of average of Cholesky factorization ###
qqnorm(E_star_avg,main="Normal Probability Plot - Cholesky",
       ylab="E~MN(0,sigma)" )
qqline(E_star_avg)

### Normal probability plot of average of Small's graphical method ###
q<-c(0)
for(i in 1:n)
{
  q[i]<-qbeta((i-0.5)/n,0.5 * p,0.5 * (50 - 7 - 1))
}
plot(q,E_small_avg,main="Beta Probabilty Plot - Small",xlab="Beta
           Quantiles", ylab="E~MN(0,sigma)",type="o")
```

B.3 Examination of behavior of residuals in GCM under normality assumption

```
### Loading R package and workspace set up ###
library(MASS)
library(Matrix)
```

```
X<-vector(mode="list", length=1000)
E<-vector(mode="list", length=1000)
X_hat<-vector(mode="list", length=1000)
R1<-vector(mode="list", length=1000)
R1_vec<-vector(mode="list", length=1000)
R1_small<-vector(mode="list", length=1000)
R2<-vector(mode="list", length=1000)
R2_vec<-vector(mode="list", length=1000)
R2_small<-vector(mode="list", length=1000)
R3<-vector(mode="list", length=1000)
L<-vector(mode="list", length=1000)
L12<-vector(mode="list", length=1000)
R_small<-vector(mode="list", length=1000)
R_star<-vector(mode="list", length=1000)
R1R2_small<-vector(mode="list", length=1000)
R1R2_star<-vector(mode="list", length=1000)

### 1000 simulation of X = A*B*C + E (n=50, p=7) ###
n=50
p=7
### design matrix ###
a<-c(rep(1,p),seq(1, p, by = 1),1,4,9,16,25,36,49)
A<-matrix(a,p)
c<-c(rep(1,n/2),rep(0,n/2),rep(0,n/2),rep(1,n/2))
C<-matrix(c,2,byrow=T)
### parameter matrix ###
B<-matrix(sample(1:10,6,replace=T),3)

### generation of E ###
```

```
e1<-c(0.0005,0.0007,0.0006,0.0010,0.0007,0.0010,0.0010)
e2<-c(0.0007,0.0012,0.0014,0.0024,0.0024,0.0030,0.0025)
e3<-c(0.0006,0.0014,0.0030,0.0050,0.0060,0.0061,0.0053)
e4<-c(0.0010,0.0024,0.0050,0.0105,0.0127,0.0128,0.0106)
e5<-c(0.0007,0.0024,0.0060,0.0127,0.0179,0.0172,0.0142)
e6<-c(0.0010,0.0030,0.0061,0.0128,0.0172,0.0188,0.0151)
e7<-c(0.0010,0.0025,0.0053,0.0106,0.0142,0.0151,0.0142)
sigma<-cbind(e1,e2,e3,e4,e5,e6,e7)
for (i in 1:1000)
{
  set.seed(500+i)
  E[[i]]<-mvrnorm(n=n, rep(0, p), 100*sigma)
}

### generation of the GCM ###
for (j in 1:1000)
{
  X[[j]]<-A %*% B %*% C+ t(E[[j]])
}

### estimation of (X_hat = A1*B_hat*C) ###
A1<-matrix(c(rep(1,p),seq(1, p, by = 1)),p)
for (k in 1:1000)
{
  S<-X[[k]] %*% (diag(n)-t(C) %*% solve(C %*% t(C)) %*% C) %*% t(X[[k]])
  B_hat<-solve(t(A1) %*% solve(S) %*% A1) %*% t(A1) %*% solve(S)
         %*% X[[k]] %*% t(C) %*% solve(C %*% t(C))
  X_hat[[k]]<-A1 %*% B_hat %*% C
```

```
  ### decomposing R1 R2 R3 ###
  R1[[k]]<-(diag(p) - A1 %*% solve(t(A1) %*% solve(S) %*% A1) %*% t(A1) %*%
           solve(S)) %*% X[[k]] %*% (diag(n)-t(C) %*% solve(C %*% t(C)) %*% C)
  R1_vec[[k]]<-sort(as.vector(R1[[k]]))
  R2[[k]]<-A1 %*% solve(t(A1) %*% solve(S) %*% A1) %*% t(A1) %*% solve(S) %*%
           X[[k]] %*% (diag(n)-t(C) %*% solve(C %*% t(C)) %*% C)
  R2_vec[[k]]<-sort(as.vector(R2[[k]]))
  R3[[k]]<-(diag(p) - A1 %*% solve(t(A1) %*% solve(S) %*% A1) %*% t(A1) %*%
           solve(S)) %*% X[[k]] %*% t(C) %*% solve(C %*% t(C)) %*% C

  ### Small's transformation of R, R1+R2 ###
  R_small[[k]]<-sort(Small(R1[[k]] + R2[[k]] + R3[[k]]))
  R1R2_small[[k]]<-sort(Small(R1[[k]] + R2[[k]]))

  ### Cholesky factorization of R, R1+R2 ###
  L[[k]]<-t(chol(cov(t(R1[[k]]+R2[[k]]+R3[[k]]))))
  R_star[[k]]<-solve(L[[k]]) %*% (R1[[k]]+R2[[k]]+R3[[k]])
  R_star[[k]]<-sort(as.vector(R_star[[k]]))
  L12[[k]]<-t(chol(cov(t(R1[[k]]+R2[[k]]))))
  R1R2_star[[k]]<-solve(L12[[k]]) %*% (R1[[k]]+R2[[k]])
  R1R2_star[[k]]<-sort(as.vector(R1R2_star[[k]]))
}

### average of the 1000 simulation ###

X_sum<-matrix(0,p,n)
X_hat_sum<-matrix(0,p,n)
R1_sum<-c(0)
R2_sum<-c(0)
```

```
R3_sum<-matrix(0,p,n)
R_small_sum<-c(0)
R1R2_small_sum<-c(0)
R_star_sum<-c(0)
R1R2_star_sum<-c(0)
for(l in 1:1000)
{
  X_sum<-(X_sum + X[[l]])
  X_hat_sum<-(X_hat_sum + X_hat[[l]])
  R1_sum<-(R1_sum + R1_vec[[l]])
  R2_sum<-(R2_sum + R2_vec[[l]])
  R3_sum<-(R3_sum + R3[[l]])
  R_small_sum<-(R_small_sum + R_small[[l]])
  R1R2_small_sum<-(R1R2_small_sum + R1R2_small[[l]])
  R_star_sum<-(R_star_sum + R_star[[l]])
  R1R2_star_sum<-(R1R2_star_sum + R1R2_star[[l]])
}
X_avg<-X_sum/1000
X_hat_avg<-X_hat_sum/1000
R1_avg<-R1_sum/1000
R2_avg<-R2_sum/1000
R3_avg<-R3_sum/1000
R_small_avg<-R_small_sum/1000
R1R2_small_avg<-R1R2_small_sum/1000
R_star_avg<-R_star_sum/1000
R1R2_star_avg<-R1R2_star_sum/1000

### Normal probability plot of average of Cholesky factorization ###
qqnorm(R_star_avg, ylab="R")
```

```
qqline(R_star_avg)
qqnorm(R1R2_star_avg,main=" ",
       ylab="R1+R2")
qqline(R1R2_star_avg)

### Normal probability plot of average of Small's graphical method ###
q<-c(0)
for(i in 1:n)
{
  q[i]<-qbeta((i-0.5)/n,0.5 * p,0.5 * (50 - 7 - 1))
}
plot(q,R_small_avg,main=" ",xlab="Beta Quantiles", ylab="R",type="o")
plot(q,R1R2_small_avg,main=" ",xlab="Beta Quantiles", ylab="R1+R2",type="o")


### Scatter plot of average R3 ###
plot(R3_avg[,1],ylim=c(-50,50),main="Scatter plot of R3",ylab="R3",type="o",col="red")
for (k in 1:1000)
{
  points(R3[[k]][,1],)
}
```

For the GCM that has a linear within individual structure, replace the command of "design matrix" and "parameter matrix" by

```
### design matrix ###
a<-c(rep(1,p),seq(1, p, by = 1))
A<-matrix(a,p)
c<-c(rep(1,n/2),rep(0,n/2),rep(0,n/2),rep(1,n/2))
C<-matrix(c,2,byrow=T)
```

```
### parameter matrix ###
B<-matrix(sample(1:10,4,replace=T),3)
```

B.4 Examination of behavior of residuals in GCM under non-normality assumption

```
### Loading R package and workspace set up ###
library(MASS)
library(Matrix)

X<-vector(mode="list", length=1000)
E<-vector(mode="list", length=1000)
X_hat<-vector(mode="list", length=1000)
R1<-vector(mode="list", length=1000)
R1_vec<-vector(mode="list", length=1000)
R1_small<-vector(mode="list", length=1000)
R2<-vector(mode="list", length=1000)
R2_vec<-vector(mode="list", length=1000)
R2_small<-vector(mode="list", length=1000)
R3<-vector(mode="list", length=1000)
L<-vector(mode="list", length=1000)
L12<-vector(mode="list", length=1000)
R_small<-vector(mode="list", length=1000)
R_star<-vector(mode="list", length=1000)
R1R2_small<-vector(mode="list", length=1000)
R1R2_star<-vector(mode="list", length=1000)

### 1000 simulation of X = A*B*C + E (n=50, p=7) ###
n=50
```

```
p=7
### design matrix ###
a<-c(rep(1,p),seq(1, p, by = 1),1,4,9,16,25,36,49)
A<-matrix(a,p)
c<-c(rep(1,n/2),rep(0,n/2),rep(0,n/2),rep(1,n/2))
C<-matrix(c,2,byrow=T)
### parameter matrix ###
B<-matrix(sample(1:10,6,replace=T),3)

### generation of E ###
e1<-c(0.0005,0.0007,0.0006,0.0010,0.0007,0.0010,0.0010)
e2<-c(0.0007,0.0012,0.0014,0.0024,0.0024,0.0030,0.0025)
e3<-c(0.0006,0.0014,0.0030,0.0050,0.0060,0.0061,0.0053)
e4<-c(0.0010,0.0024,0.0050,0.0105,0.0127,0.0128,0.0106)
e5<-c(0.0007,0.0024,0.0060,0.0127,0.0179,0.0172,0.0142)
e6<-c(0.0010,0.0030,0.0061,0.0128,0.0172,0.0188,0.0151)
e7<-c(0.0010,0.0025,0.0053,0.0106,0.0142,0.0151,0.0142)
sigma<-cbind(e1,e2,e3,e4,e5,e6,e7)
for (i in 1:1000)
{
  set.seed(500+i)
  E[[i]]<-mvrnorm(n=n, rep(0, p), 100*sigma)
  E[[i]]<-exp(E[[i]])
}

### generation of the GCM ###
for (j in 1:1000)
{
  X[[j]]<-A %*% B %*% C+ t(E[[j]])
```

```
}

### estimation of (X_hat = A1*B_hat*C) ###
A1<-matrix(c(rep(1,p),seq(1, p, by = 1)),p)
for (k in 1:1000)
{
  S<-X[[k]] %*% (diag(n)-t(C) %*% solve(C %*% t(C)) %*% C) %*% t(X[[k]])
  B_hat<-solve(t(A1) %*% solve(S) %*% A1) %*% t(A1) %*% solve(S)
         %*% X[[k]] %*% t(C) %*% solve(C %*% t(C))
  X_hat[[k]]<-A1 %*% B_hat %*% C

  ### decomposing R1 R2 R3 ###
  R1[[k]]<-(diag(p) - A1 %*% solve(t(A1) %*% solve(S) %*% A1) %*% t(A1) %*%
           solve(S)) %*% X[[k]] %*% (diag(n)-t(C) %*% solve(C %*% t(C)) %*% C)
  R1_vec[[k]]<-sort(as.vector(R1[[k]]))
  R2[[k]]<-A1 %*% solve(t(A1) %*% solve(S) %*% A1) %*% t(A1) %*% solve(S) %*%
           X[[k]] %*% (diag(n)-t(C) %*% solve(C %*% t(C)) %*% C)
  R2_vec[[k]]<-sort(as.vector(R2[[k]]))
  R3[[k]]<-(diag(p) - A1 %*% solve(t(A1) %*% solve(S) %*% A1) %*% t(A1) %*%
           solve(S)) %*% X[[k]] %*% t(C) %*% solve(C %*% t(C)) %*% C

  ### Small's transformation of R, R1+R2 ###
  R_small[[k]]<-sort(Small(R1[[k]] + R2[[k]] + R3[[k]]))
  R1R2_small[[k]]<-sort(Small(R1[[k]] + R2[[k]]))

  ### Cholesky factorization of R, R1+R2 ###
  L[[k]]<-t(chol(cov(t(R1[[k]]+R2[[k]]+R3[[k]]))))
  R_star[[k]]<-solve(L[[k]]) %*% (R1[[k]]+R2[[k]]+R3[[k]])
  R_star[[k]]<-sort(as.vector(R_star[[k]]))
```

```
  L12[[k]]<-t(chol(cov(t(R1[[k]]+R2[[k]]))))
  R1R2_star[[k]]<-solve(L12[[k]]) %*% (R1[[k]]+R2[[k]])
  R1R2_star[[k]]<-sort(as.vector(R1R2_star[[k]]))
}

### average of the 1000 simulation ###

X_sum<-matrix(0,p,n)
X_hat_sum<-matrix(0,p,n)
R1_sum<-c(0)
R2_sum<-c(0)
R3_sum<-matrix(0,p,n)
R_small_sum<-c(0)
R1R2_small_sum<-c(0)
R_star_sum<-c(0)
R1R2_star_sum<-c(0)
for(l in 1:1000)
{
  X_sum<-(X_sum + X[[l]])
  X_hat_sum<-(X_hat_sum + X_hat[[l]])
  R1_sum<-(R1_sum + R1_vec[[l]])
  R2_sum<-(R2_sum + R2_vec[[l]])
  R3_sum<-(R3_sum + R3[[l]])
  R_small_sum<-(R_small_sum + R_small[[l]])
  R1R2_small_sum<-(R1R2_small_sum + R1R2_small[[l]])
  R_star_sum<-(R_star_sum + R_star[[l]])
  R1R2_star_sum<-(R1R2_star_sum + R1R2_star[[l]])
}
X_avg<-X_sum/1000
```

```
X_hat_avg<-X_hat_sum/1000
R1_avg<-R1_sum/1000
R2_avg<-R2_sum/1000
R3_avg<-R3_sum/1000
R_small_avg<-R_small_sum/1000
R1R2_small_avg<-R1R2_small_sum/1000
R_star_avg<-R_star_sum/1000
R1R2_star_avg<-R1R2_star_sum/1000

### Normal probability plot of average of Cholesky factorization ###
qqnorm(R_star_avg, ylab="R")
qqline(R_star_avg)
qqnorm(R1R2_star_avg,main=" ",
       ylab="R1+R2")
qqline(R1R2_star_avg)

### Normal probability plot of average of Small's graphical method ###
q<-c(0)
for(i in 1:n)
{
  q[i]<-qbeta((i-0.5)/n,0.5 * p,0.5 * (50 - 7 - 1))
}
plot(q,R_small_avg,main=" ",xlab="Beta Quantiles", ylab="R",type="o")
plot(q,R1R2_small_avg,main=" ",xlab="Beta Quantiles", ylab="R1+R2",type="o")


### Scatter plot of average R3 ###
plot(R3_avg[,1],ylim=c(-50,50),main="Scatter plot of R3",ylab="R3",type="o",col="red")
for (k in 1:1000)
```

```
{
  points(R3[[k]][,1],)
}
```

For the GCM that has a linear within individual structure, replace the command of "design matrix" and "parameter matrix" by

```
### design matrix ###
a<-c(rep(1,p),seq(1, p, by = 1))
A<-matrix(a,p)
c<-c(rep(1,n/2),rep(0,n/2),rep(0,n/2),rep(1,n/2))
C<-matrix(c,2,byrow=T)

### parameter matrix ###
B<-matrix(sample(1:10,4,replace=T),3)
```

Normal Probability plot of a univariate log-normal distribution can be computed by

```
### workspace setup ###
ln<-vector(mode="list", length=1000)
ln_sum<-c(0)

### generation of 1000 random univariate log-normal distributed data ###
for (i in 1:1000)
{
  set.seed(500+i)
  ln[[i]]<-sort(rlnorm(100,0,1))
}

### obtain the average of the 1000 simulation ###
for (i in 1:1000)
{
```

```
  ln_sum<-ln_sum+ln[[i]]
}
ln_avg<-ln_sum/1000

### Normal Probability plot of the average sequence ###
qqnorm(ln_avg)
```

B.5 Examination of behavior of residuals in the Potthoff-Roy Dental Data

```
### Importing data and workspace setup ###
roy <- read.table("C:/Users/Will/Desktop/thesis/roy.txt", quote="\"")
roy<-as.data.frame(roy)
roy<-cbind(roy[,3:6])
roy<-as.matrix(roy)
roy<-t(roy)

### Average of profile plot ###
girl_sum<-c(0)
for (i in 1:11)
{
  girl_sum<-girl_sum+roy[,i]
}
girl_avg<-girl_sum/11
boy_sum<-c(0)
for (i in 12:27)
{
  boy_sum<-boy_sum+roy[,i]
}
```

```
boy_avg<-boy_sum/16

### Linear fitting using maximum likelihood method ###
A_roy<-matrix(c(rep(1,4),seq(from=8, to=14, by = 2)),4)
c<-c(rep(1,11),rep(0,16),rep(0,11),rep(1,16))
C_roy<-matrix(c,2,byrow=T)
S_roy<-roy %*% (diag(27)-t(C_roy) %*% solve(C_roy %*% t(C_roy)) %*% C_roy) %*% t(roy)
B_roy<-solve(t(A_roy) %*% solve(S_roy) %*% A_roy) %*% t(A_roy) %*% solve(S_roy) %*%
       roy %*% t(C_roy) %*% solve(C_roy %*% t(C_roy))
roy_hat<-A_roy %*% B_roy %*% C_roy

### Decomposed residuals calculation ###
R1_roy<-(diag(4) - A_roy %*% solve(t(A_roy) %*% solve(S_roy) %*% A_roy) %*% t(A_roy)
        %*%solve(S_roy)) %*% roy %*% (diag(27)-t(C_roy) %*% solve(C_roy %*% t(C_roy))
        %*% C_roy)
R2_roy<-A_roy %*% solve(t(A_roy) %*% solve(S_roy) %*% A_roy) %*% t(A_roy) %*%
        solve(S_roy)%*% roy %*% (diag(27)-t(C_roy) %*% solve(C_roy %*% t(C_roy))
        %*% C_roy)
R3_roy<-(diag(4) - A_roy %*% solve(t(A_roy) %*% solve(S_roy) %*% A_roy)%*%t(A_roy)
        %*%solve(S_roy)) %*% roy %*% t(C_roy) %*% solve(C_roy %*% t(C_roy))%*%C_roy

### Profile plots of all the individuals in the Dental data ###
plot(roy[,1],main="profile plot of Roy data",ylim=c(15,32), ylab="roy",type="o")
for (k in 2:27)
{
  lines(roy[,k])
}

### Scatter plot of the decomposed R3 ###
```

```
plot(R3_roy[,1],main="scatterplot of R3",ylim=c(-1,1),ylab="R3")

### Beta Probability Plot of the Small's R1 + R2 ###
q<-c(0)
for(i in 1:27)
{
  q[i]<-qbeta((i-0.5)/27,0.5 * nrow(R1_roy),0.5 * (ncol(R1_roy) -nrow(R1_roy) - 1))
}
plot(q,sort(Small(R1_roy + R2_roy)),main="Quantile Plot of R1+R2",ylab="R1+R2",
     xlab="Beta Quantiles")

### Normal Probability Plot of the Cholesky's R1 + R2 ###
L_roy<-t(chol(cov(t(R1_roy + R2_roy))))
R12_star_roy<-solve(L_roy) %*% (R1_roy + R2_roy)
qqnorm(R12_star_roy,main="QQ plot of (R1+R2)*", ylab="R1+R2")
```

For the analysis with the outliers removed, replace the setting of design matrices A and C by

```
A_roy<-matrix(c(rep(1,4),seq(from=8, to=14, by = 2)),4)
c<-c(rep(1,11),rep(0,14),rep(0,11),rep(1,14))
C_roy<-matrix(c,2,byrow=T)

Therefore
S_roy<-roy %*% (diag(25)-t(C_roy) %*% solve(C_roy %*% t(C_roy)) %*% C_roy) %*% t(roy)

And the plots can be generated by
plot(roy[,1],ylim=c(15,36), ylab="measurement(mm)",type="o",col="red")
for (k in 2:11)
{
  lines(roy[,k],type="o",col="red")
}
```

```
for (k in 12:25)
{
  lines(roy[,k],type="o",pch=22,lty=2)
}
legend("topleft", c("boys","girls"), cex=0.8,
       col=c("black","red"), pch=c(22,21), lty=c(2,1),bty="n")

plot(girl_avg,ylim=c(15,36), ylab="dental growth",type="o",col="red",lwd=2)
lines(boy_avg,type="o",pch=22,lty=2,lwd=2)
for (k in 1:11)
{
  points(roy[,k],type="p",col="red")
}
for (k in 12:25)
{
  points(roy[,k],type="p",pch=22,lty=2)
}
legend("topleft", c("boys","girls"), cex=0.8,
       col=c("black","red"), pch=c(22,21), lty=c(2,1),bty="n")

plot(R3_roy[,1],main="scatterplot of R3",ylim=c(-1,1),ylab="R3")

qqnorm(R1_roy + R2_roy + R3_roy, ylab="R")

L_roy<-t(chol(cov(t(R1_roy + R2_roy))))
R12_star_roy<-solve(L_roy) %*% (R1_roy + R2_roy)
qqnorm(R12_star_roy,main="Fitzmaurice's R1+R2",ylab="R1+R2")

q<-c(0)
```

```
for(i in 1:25)
{
  q[i]<-qbeta((i-0.5)/25,0.5 * nrow(R1_roy),0.5 * (ncol(R1_roy) -nrow(R1_roy) - 1))
}
plot(q,sort(Small(R1_roy + R2_roy)),main="Small's R1+R2",ylab="R1+R2",xlab="Beta Quantiles"
     ,pch=20)
```

And average of profile plots can be computed by

```
girl_sum<-c(0)
for (i in 1:11)
{
  girl_sum<-girl_sum+roy[,i]
}
girl_avg<-girl_sum/11
boy_sum<-c(0)
for (i in 12:25)
{
  boy_sum<-boy_sum+roy[,i]
}
boy_avg<-boy_sum/14
```

B.6 Examination of behavior of residuals in the Glucose Data

```
### Importing data and workspace setup ###
glucose<-read.table("C:/Users/Will/Desktop/thesis/glucose.txt", quote="\"")
glucose<-as.data.frame(glucose)
glucose<-cbind(glucose[,2:9])
glucose<-as.matrix(glucose)
```

```
glucose<-t(glucose)

### Average of profile plots ###
control_sum<-c(0)
for (i in 1:13)
{
  control_sum<-control_sum+glucose[,i]
}
control_avg<-control_sum/13
obese_sum<-c(0)
for (i in 14:33)
{
  obese_sum<-obese_sum+glucose[,i]
}
obese_avg<-obese_sum/20

### Linear fitting using maximum likelihood method ###
A_glucose<-matrix(c(rep(1,8),0,0.5,1,1.5,2,3,4,5),8)
c<-c(rep(1,13),rep(0,20),rep(0,13),rep(1,20))
C_glucose<-matrix(c,2,byrow=T)
S_glucose<-glucose %*% (diag(33)-t(C_glucose) %*% solve(C_glucose %*% t(C_glucose))
           %*% C_glucose) %*% t(glucose)
B_glucose<-solve(t(A_glucose) %*% solve(S_glucose) %*% A_glucose) %*% t(A_glucose)
           %*% solve(S_glucose) %*% glucose %*% t(C_glucose) %*% solve(C_glucose %*%
           t(C_glucose))
glucose_hat<-A_glucose %*% B_glucose %*% C_glucose

### Decomposed residuals calculation ###
R1_glucose<-(diag(8) - A_glucose %*% solve(t(A_glucose) %*% solve(S_glucose) %*%
```

```
            A_glucose) %*% t(A_glucose) %*% solve(S_glucose)) %*% glucose %*%
            (diag(33)-t(C_glucose)%*%solve(C_glucose%*%t(C_glucose))%*%C_glucose)
R2_glucose<-A_glucose %*% solve(t(A_glucose) %*% solve(S_glucose) %*% A_glucose)
            %*% t(A_glucose) %*% solve(S_glucose) %*% glucose %*% (diag(33)-
            t(C_glucose)%*% solve(C_glucose %*% t(C_glucose)) %*% C_glucose)
R3_glucose<-(diag(8) - A_glucose %*% solve(t(A_glucose) %*% solve(S_glucose)
            %*% A_glucose)%*% t(A_glucose) %*% solve(S_glucose)) %*% glucose
            %*% t(C_glucose) %*%solve(C_glucose %*% t(C_glucose)) %*% C_glucose

### Profile plots of all the individuals in the Glucose data ###
plot(glucose[,1],main="profile plot of glucose data",ylim=c(1.3,6.8), ylab=
     "glucose",type="o")
for (k in 2:33)
{
  lines(glucose[,k])
}

### Scatter plot of the decomposed R3 ###
plot(R3_glucose[,1],main="scatterplot of R3",ylim=c(-1,1),ylab="R3")

### Beta Probability Plot of the Small's R1 + R2 ###
q<-c(0)
for(i in 1:33)
{
  q[i]<-qbeta((i-0.5)/33,0.5 * nrow(R1_glucose),0.5 *
        (ncol(R1_glucose) -nrow(R1_glucose) - 1))
}
plot(q,sort(Small(R1_glucose + R2_glucose)),main="Quantile Plot of
     R1+R2",ylab="R1+R2",xlab="Beta Quantiles")
```

```
### Normal Probability Plot of the Cholesky's R1 + R2 ###
L_glucose<-t(chol(cov(t(R1_glucose + R2_glucose))))
R12_star_glucose<-solve(L_glucose) %*% (R1_glucose + R2_glucose)
qqnorm(R12_star_glucose,main=" QQ plot of (R1+R2)*",ylab="R1+R2")
```

The quadratic and third degrees polynomial fitting can be computed using design matrix A

```
A_glucose<-matrix(c(rep(1,8),0,0.5,1,1.5,2,3,4,5,0,0.25,1,2.25,
           4,9,16,25),8)
A_glucose<-matrix(c(rep(1,8),0,0.5,1,1.5,2,3,4,5,0,0.25,1,2.25,
           4,9,16,25,0,0.125,1,3.375,8,27,64,125),8)
```

Bibliography

[1] Altman, D. G., (1990). *Practical Statistics for Medical Research.* Chapman and Hall, UK.

[2] Banken, L., (1984). *Eine Verallgemeinerung des GMANOVA Modells.* Dissertation, University of Trier, Trier, Germany.

[3] Bland, J. M., (2002). *An Introduction to Medical Statistics, Third edition.* Oxford Medical Publications, UK.

[4] Chi, E. M. and Reinsel, G. C., (1989). Models for longitudinal data with random effects and AR(1) errors. *Journal of the American Statistical Association* **84**, 452-459.

[5] Cook, R. D., (1977). Detection of influential observation in linear regression. *Technometrics* **19**, 15-18.

[6] Cook, R. D., (1982). *Residuals and Influence in Regression.* Chapman and Hall, New York.

[7] Cook, R. N., (2008). Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. *Clinical Chemistry* **54**, 17-23.

[8] Demchuk, E., Yucesoy, B., Johnson, V. J., Andrew, M., Weston, A., Germolec, D. R., De Rosa, C. T. and Michael, I. L., (2007). A statistical model for assessing genetic susceptibility as a risk factor in multifactorial diseases: lessons from occupational asthma. *Environmental Health Perspectives* **115**, 231-234.

[9] Diggle, P. J., Heagerty, P., Liang, K. Y., and Zeger, S. L., (2002). *Analysis of Longitudinal Data.* Oxford University Press, Oxford.

[10] Draper, N. R. and Smith, H., (1998). *Applied Regression Analysis.* Wiley, New York.

[11] Etzioni, R. D. and Kadane, J. B., (1995). Bayesian statistical methods in public health and medicine. *Annual Review of Public Health* **16**, 23-41.

[12] Fizmaurice, M. G., Laird, M. N. and Ware, H. J., (2004). *Applied Longitudinal Analysis.* Wiley, New York.

[13] Gneiting, T. and Raftery A. E., (2005). Weather forecasting with ensemble methods. *The Science Journal* **310**, 248-249.

[14] Gray, S. M., (2001). *Evidence Based Health Care and Public Health, Third edition.* Elsevier, UK.

[15] Hamid, S. J. and Von Rosen, D., (2006). Residuals in the extended growth curve model. *Scandinavian Journal of Statistics* **33**, 121-138.

[16] Keramidas, E. M. and Lee, J. C., (1995). Selection of a covariance structure for growth-curves. *Biometrical Journal* **37**, 783-797.

[17] Khatri, C. G., (1966). A note on a MANOVA model applied to the problems in growth curve. *Annals of the Institute of Statistical Mathematics* **18**, 75-86.

[18] Kotz, S., Balakrishnan, N. and John, N. L., (2000). *Continuous Multivariate Distributions, Second edtion.* John Wiley and Sons, USA.

[19] Lee, J. C. and Geisser, S., (1975). Applications of growth curve prediction. *Sankhya Series* A **37**, 239-256.

[20] Lee, J. C., (1988). Prediction and estimation of growth curve with special covariance structure. *Journal of the American Statistical Association* **83**, 432-440.

[21] Lee, J. C. and Geisser, S., (1991). Tests and model selection for the general growth curve model. *Biometrics* **47**, 147-159.

[22] Newman, T. B. and Kohn, M. A., (2009). *Evidence-Based Diagnosis.* Cambridge University Press, New York.

[23] Pan, J. X. and Fang, K. T., (2002). *Growth Curve Model and Statistical Diagnostics.* Springer, New York.

[24] Potthoff, R. F. and Roy, S. N., (1964). A generalized multivariate analysis of variance model useful especially for growth curves. *Biometrika* **51**, 313-326.

[25] Rao, C. R., (1987). Prediction of future observation in polynomial growth curve models. *Statistical Science* **2**, 434-471.

[26] Sen, A. and Srivastava, M. S., (1990). *Regression Analysis, Theory and Applications.* Springer, New York.

[27] Small, N. J. H., (1978). Ploting squared radii. *Biometrics* **65**, 675-678.

[28] Srivastava, M. S. and Carter, E. M., (1983). *An Introduction to Applied Multivariate Statistics.* North-Holland, New York.

[29] Srivastava, M. S., (2002). *Methods of Multivariate Statistics.* Wiley, New York.

[30] Straus, S. E., Richardson, W. S., Glasziou, P., Haynes, R. B. and Strauss, S. E., (2005). *Evidence Based Medicine, Third edition.* Churchill Livingstone, USA.

[31] Taylor, S. J., (2007). *Modelling Financial Time Series, Second edition.* Lancaster University, UK.

[32] von Rosen, D., (1989). Maximum likelihood estimators in multivariate linear normal models. *Journal of Multivariate Analysis* **31**, 187-200.

[33] von Rosen, D., (1995a). Influential observations in multivariate linear models. *Scandinavian Journal of Statistics* **22**, 207-222.

[34] von Rosen, D., (1995b). Residuals in the growth curve model. *Annals of the Institute of Statistical Mathematics* **47**(1), 129-136.

[35] Zerbe, G. O., (1979). Randomization analysis of the completely randomized design extended to growth and response curves. *Journal of the American Statistical Association* **74**, 215-221.