Statistical Analysis of Longitudinal Data with a Case Study

STATISTICAL ANALYSIS OF LONGITUDINAL DATA WITH A CASE STUDY

BY

KAI LIU, B.Sc.

A THESIS

SUBMITTED TO THE DEPARTMENT OF MATHEMATICS & STATISTICS

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

Master of Science (2014) McMaster University

(Mathematics & Statistics) Hamilton, Ontario, Canada

TITLE:	Statistical Analysis of Longitudinal Data with a Case Study
AUTHOR:	Kai Liu B.Sc., (Mathematics & Statistics) McMaster University, Hamilton, Canada
SUPERVISOR:	Prof. Narayanaswamy Balakrishnan
NUMBER OF PAGES:	ix, 73

Abstract

Preterm birth is the leading cause of neonatal mortality and long-term morbidity. Neonatologists can adjust nutrition to preterm neonates to control their weight gain so that the possibility of long-term morbidity can be minimized. This optimization of growth trajectories of preterm infants can be achieved by studying a cohort of selected healthy preterm infants with weights observed during day 1 to day 21. However, missing values in such a data poses a big challenge in this case. In fact, missing data is a common problem faced by most applied researchers. Most statistical softwares deal with missing data by simply deleting subjects with missing items. Analyses carried out on such incomplete data result in biased estimates of the parameters of interest and consequently lead to misleading or invalid inference. Even though many statistical methods may provide robust analysis, it will be better to handle missing data by imputing them with plausible values and then carry out a suitable analysis on the full data. In this thesis, several imputation methods are first introduced and discussed. Once the data get completed by the use of any of these methods, the growth trajectories for this cohort of preterm infants can be presented in the form of percentile growth curves. These growth trajectories can now serve as references for the population of preterm babies. To find out the explicit growth rate, we are interested in establishing predictive models for weights at days 7, 14 and 21. I have used

both univariate and multivariate linear models on the completed data. The resulting predictive models can then be used to calculate the target weight at days 7, 14 and 21 for any other infant given the information at birth. Then, neonatologists can adjust the amount of nutrition given in order to preterm infants to control their growth so that they will not grow too fast or too slow, thus avoiding later-life complications.

Acknowledgements

I would like to express my appreciation to my supervisor Prof. Narayanaswamy Balakrishnan. It was a wonderful experience to work under his supervision. I would also like to sincerely thank Dr. Christoph Fusch for providing me an opportunity to apply my statistical knowledge to problems in health science and Dr. Niels Rochow for his ongoing medical guidance and for taking the time out of his busy schedule to discuss ideas pertinent to this project. I want to thank the supervisory committee members, Drs. Joseph Beyene and Shui, Feng for their support and recommendations, my parents for their constant encouragement, and my friends Tian Feng, Hon Yiu So, Xiaojun Zhu, Yang Ye, Sandip Barui and all the other research students for their kind assistance and moral support.

Contents

List of Figures

Chapter 1

Introduction

1.1 Background

Infants born at gestational age less than 37 weeks are defined as preterm infants. Roughly, 15 million babies are born preterm every year and this number is increasing (WHO, 2014). Preterm birth is the major cause for newborn deaths during the first four weeks of life with a mortality rate over 10% (WHO, 2014). Luckily, great improvements have been made to increase survival rate of preterm infants recently Iams *et al.* (2008). Survived preterm babies, however, still have subsequent complications in later life (Saigal and Doyle, 2008; Larroque *et al.*, 2008; Barker *et al.*, 1993). Therefore, current research works focus on prevention and reduction of long-term morbidity of the survived preterm infants.

Morbidity has an inverse relationship with gestational age. That is, the shorter the gestational age is, the higher the morbidity rate is. A follow-up study on 5-year-old children born preterm show that about 50% of survival children born at 24-28 weeks

of gestation have a neurodevelopmental disability, while only about 33% of survival children born at 29-32 weeks of gestation have a neurodevelopmental disability (Larroque *et al.*, 2008). Birth weight is another factor associated with mobidity. Low birth weight is related to an increased risk of developing the risk factors of cardiovascular disease in later life, including blood pressure and insulin resistance (Barker *et al.*, 1993). Even though interventions before and during pregnancy can lower the possibility of occurrence of birth weight extremes and small gestational age (Iams *et al.*, 2008), the incidence of low birth weight preterm births can not be avoided.

Weight gain during early postnatal life, reflecting growth of preterm infants, is associated with adult-onset disease. Slow weight gain during early postnatal life may lead to chronic lung disease, infection and poor neurodevelopment (Ho *et al.*, 2003; Latal Hajnal *et al.*, 2003), while rapid growth during early postnatal life of preterm infants may increase the risk of later-life adiposity, insulin resistance, cardiovascular disease and metabolic syndrome (Jain and Singhal, 2012; Steward, 2012). Despite the neonatologists control the weight gain of infants by adjusting the amount of nutrition given to them, the optimal growth trajectory a preterm infant should adjust to after postnatal adaptation is unknown (Fisch *et al.*, 2014).

1.2 Aims of Growth Trajectories of Healthy Preterm Infants (GTHPI) Study

Our project here is designed to figure out the optimal growth trajectories of preterm infants based on a longitudinal study. This is achieved by characterizing the growth

trajectories of a cohort of healthy preterm infants in terms of quantile plots and then establishing predictive models for weight at days 7, 14 and 21. Hence, target weights at days 7, 14 and 21 for any weak infant can be drawn, given the information at birth, by the predictive models so that neonatologists can adjust nutrition accordingly to control the infant's weight gain.

1.3 Scope of the Thesis

In chapter 2, I will briefly introduce longitudinal data and then describe the GTHPI data that will be analyzed in this thesis. The standard statistical analysis models that can be used to study the longitudinal GTHPI data are explained in Chapter 3. The challenge of the study based on this data set is that there is a large proportion of missing values in the data and for this reason, various reasons for missingness and approaches for dealing with missing data are then detailed in Chapter 4. In Chapter 5, different imputation methods are applied to the GTHPI data and compared and the consequent results of the statistical analysis of the completed data are described in detail in Chapter 6. Further discussion and some possible directions of work are finally described in Chapter 7. The R-codes developed for the thesis are all presented in the Appendix.

Chapter 2

Longitudinal Data

In health science, longitudinal studies are frequently designed to investigate changes over time. In contrast to cross-sectional data wherein measurements are required at only a single time point, longitudinal data have repeated measurements of outcome through a period of time (Fitzmaurice *et al.*, 2012). The measurements are usually made at a set of common time points for all subjects. There are often some single-valued variables associated with each subjects, termed as covariates. Longitudinal studies are not only interested in overall within-individual changes in response, but also how the relationship between response changes with the covariates. In this Chapter, I will describe briefly some features of longitudinal data and highlight some main aspects of their analyses.

2.1 Structures of Longitudinal Data

Longitudinal data are usually recorded in two forms. Most longitudinal data are structured in a format with a single row for each individual. Each row contains

multiple covariates and a series of repeated response values. The covariate values are fixed in these data. Another format is one in which repeated measurements of response are recorded in a long format. That is, each individual has multiple rows of record with one measurement of response in one row along with corresponding covariates. In this form, covariate values are not necessarily same for each individual at distinct time points, and they may vary with respect to response values. This form of longitudinal data is like a cross-sectional data wherein repeated response values are all recorded in one column and one more covariate is added to indicate the time of measurement. The wide format of longitudinal data can be transformed into long format easily by function "melt" in the R package "reshape", in which the covariates are repeated multiple rows for each subject. The structures are important when we carry out statistical analysis. Most standard statistical analysis methods deal with longitudinal data in the long format.

2.2 Descriptive Analysis of GTHPI Data

Our data were collected from neonatal intensive care units of five academic hospitals in Canada and Germany (Fisch *et al.*, 2014), namely, McMaster University Medical Centre (Hamilton, Canada), St. Joseph's Healthcare (Hamilton, Canada), St. Michael's Hospital (Toronto, Canada), Greifswald University Hospital (Greifswald, Germany), and Heidelberg University Hospital (Heidelberg, Germany).

Table 2.1 gives a listing and description of all the variables collected in the data. Daily measurements of body weight from day of life 1 to 21 of 1202 infants with

Table 2.1: Description of Variables in GTHPI Data

Variables	Description
id	anonymous identifier
centre	the academic hospital where each preterm infant is admitted
yoa	year of admission to hospital
gender	gender
gaw	completed weeks of gestational age
gawexact	exact gestational age in weeks
gad	exact gestational age in days
bw	birth weight
defstart	day of life when infants are start to be fed
dtpnstop	day of life when protein stops
md	mode of delivery
eth	ethnic
preg	number of pregnancy, i.e., singleton, twins or triplet
w1	weight at day of life 1, equivalent to birth weight
w2	weight at day of life 2
w3	weight at day of life 3
.	.
.	.
.	.
w20	weight at day of life 20
w21	weight at day of life 21

Table 2.2: Summary of GTHPI Data

completed week of gestation	25	26	27	28	29	30	31	32	33	34
male	3	20	19	43	37	21	40	86	117	165
female	7	17	21	25	38	16	37	59	78	132
Total	10	37	40	68	75	37	77	145	195	297

unimpaired postnatal adaptation admitted to one of the participating hospitals between 2008 to 2012 were the subjects in our study. Briefly, unimpaired adaptation was defined as preterm babies with no maternal diabetes/substance use, nosocomial sepsis, respiratory distress, feeding intolerance or major congenital malformations (Fisch *et al.*, 2014). Infants' demographic characteristics such as gestational age, gender and ethnicity were collected. Our data are recorded in a wide format.

Outliers with extreme large or small birth weight for infants grouped in gestational age were checked by boxplot and excluded from the study. The final data set consisted of 981 preterm infants and were recorded in a wide format. Table 2.2 provides a summary of the frequency of infants with respect to completed weeks of gestation.

Chapter 3

Linear Regression Models

Linear regression is an approach quite commonly used to investigate relationship between certain continuous dependent variable and one or more independent variables (called covariates). A linear model then specifies that the dependent variable can be expressed as a linear combination of the covariates and some unknown parameters. The linear regression is not only useful in modelling relationships between response variable and covariates, but also useful in predicting missing response values given the values of corresponding covariates.

3.1 Multiple Linear Regression

Suppose a data set of m subjects with observed response values and covariates is given, and that the response is denoted by y_i and the p covariates are denoted by $x_1, x_2, \ldots, x_p$ for the i^{th} subject. Then, a set of equations can be written as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, \ldots, m. \tag{3.1}$$

3.1.1 Assumptions

There are three assumptions made in the analysis of linear regression models. First of all, it is assumed that the response variable can be expressed as a linear combination of covariates and some unknown parameters as written in (3.1). Implicit in this assumption is that all the covariates, which can potentially affect the response variable, have been included in the model. But, this may generally be not true. Yet, the model can still hold because the unobserved covariates can be considered as part of the error term inserted into the model.

The next assumption made is that the covariates are linearly independent, meaning that any one of the covariates can not be written as a linear combination of the other covariates. If this assumption does not hold, then there will be some problems encountered when estimating the unknown parameters.

The final assumption is that the error terms are identically and independently normally distributed with mean 0 and variance σ^2. That is, the variance of the error terms is constant at σ^2 regardless of the values of the covariates.

To state these formally, we have the response variable Y_i as follows:

- $E(Y_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$;
- $\text{Var}(Y_i) = \sigma^2$;
- Y_i follow a normal distribution, for all i and are independent.

3.1.2 Least-Squares Estimates

If we know the value of the parameters, the expected value of Y_i is $\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$. The error between the observed value and the expected value is given by

$$\epsilon_i = y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip}, \quad i = 1, \ldots, m. \tag{3.2}$$

Intuitively, we would like to make these error terms as small as possible. Since negative error terms can offset positive error terms if we sum up all the error terms, we would like to use the sum of squared errors (SSE) to be an overall measure of the fit of models, and it is given by

$$S(\beta_0, \beta_1, \cdots, \beta_p) = \sum_{i=1}^{m} \Big[y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \Big]^2. \tag{3.3}$$

The linear regression model can be equivalently written in a matrix form as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{3.4}$$

where $\boldsymbol{Y}$ is the $m \times 1$ response vector, $\boldsymbol{X}$ is the $m \times (p+1)$ design matrix, $\boldsymbol{\beta}$ is the $(p+1) \times 1$ parameter vector, and $\boldsymbol{\epsilon}$ is the $m \times 1$ error vector, written as follows:

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{m1} & x_{m2} & \cdots & x_{mp} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_m \end{pmatrix}.$$

Thus, the expectation and variance of $\boldsymbol{Y}$ are given by

$$E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta},\ \mathrm{Var}(\boldsymbol{Y}) = \sigma^2\boldsymbol{I}, \tag{3.5}$$

and the sum of squared errors can be written as

$$S(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}). \tag{3.6}$$

Taking derivatives of $S(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ and setting them to be 0, we obtain the least-squares estimator of $\boldsymbol{\beta}$ as

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}. \tag{3.7}$$

3.1.3 Fitted Values, Residuals and Estimation of σ^2

The fitted values or expected values of response variable can be expressed as

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}. \tag{3.8}$$

Let $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. Then P is a $n \times n$ matrix called the hat matrix. The reason why $\boldsymbol{P}$ is called the hat matrix is that when $\boldsymbol{Y}$ is multiplied by $\boldsymbol{P}$, we obtain $\hat{\boldsymbol{Y}}$; that is, (3.8) can be expressed in terms of $\boldsymbol{P}$ as

$$\hat{\boldsymbol{Y}} = \boldsymbol{P}\boldsymbol{Y}. \tag{3.9}$$

After getting the fitted values, the residuals, which give the difference between the observed response and the fitted response are then calculated as

$$\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{P}\boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}. \tag{3.10}$$

Then, the variance of the residuals can be obtained as

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{e}) &= \mathrm{Var}((\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}) \\
&= (\boldsymbol{I} - \boldsymbol{P})'\mathrm{Var}(\boldsymbol{Y})(\boldsymbol{I} - \boldsymbol{P}) \\
&= (\boldsymbol{I} - \boldsymbol{P})'\sigma^2\boldsymbol{I}(\boldsymbol{I} - \boldsymbol{P}) \\
&= \sigma^2(\boldsymbol{I} - \boldsymbol{P} - \boldsymbol{P}' + \boldsymbol{P}'\boldsymbol{P}) \\
&= \sigma^2(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' - (\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')' \\
&\quad + (\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')'(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')) \\
&= \sigma^2(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \\
&\quad + \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}') \\
&= \sigma^2(\boldsymbol{I} - \boldsymbol{2X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}') \\
&= \sigma^2(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}') \\
&= \sigma^2(\boldsymbol{I} - \boldsymbol{P}).
\end{aligned} \tag{3.11}$$

The estimated variance of the residuals is then given by

$$\widehat{\mathrm{Var}}(\boldsymbol{e}) = \hat{\sigma}^2(\boldsymbol{I} - \boldsymbol{P}), \tag{3.12}$$

and consequently the residuals can be standardized as

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1-p_{ii}}}, \tag{3.13}$$

where p_{ii} is the i^{th} diagonal element of the matrix $\boldsymbol{P}$ and $\hat{\sigma}$ is the estimator of standard deviation. The unbiased estimator of σ^2 can be found readily from sum of squared residuals as

$$\widehat{\sigma}^2 = \frac{SSE}{m-(p+1)} = \frac{\boldsymbol{e'e}}{m-(p+1)}. \tag{3.14}$$

3.1.4 Sampling Distribution and Confidence Intervals for Regression Parameters

The mean and variance-covariance matrix of the parameter estimators can be easily derived as follows:

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) &= E((\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}|\boldsymbol{X}) \\ &= (\boldsymbol{X'X})^{-1}\boldsymbol{X'}E(\boldsymbol{Y}|\boldsymbol{X}) \\ &= (\boldsymbol{X'X})^{-1}\boldsymbol{X'}(\boldsymbol{X\beta}) \\ &= \boldsymbol{\beta}, \end{aligned} \tag{3.15}$$

$$\begin{aligned} \mathrm{Var}(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) &= \mathrm{Var}((\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}|\boldsymbol{X}) \\ &= (\boldsymbol{X'X})^{-1}\boldsymbol{X'}((\boldsymbol{X'X})^{-1}\boldsymbol{X'})'\sigma^2\boldsymbol{I} \\ &= (\boldsymbol{X'X})^{-1}\boldsymbol{X'X}(\boldsymbol{X'X})^{-1}\sigma^2 \\ &= \sigma^2(\boldsymbol{X'X})^{-1}. \end{aligned} \tag{3.16}$$

Then, the estimated variance-covariance matrix for $\hat{\boldsymbol{\beta}}$ is simply given by

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1}, \tag{3.17}$$

in which the diagonal elements are the estimated variances of the estimates of β_j's, denoted by $\mathrm{Var}(\hat{\beta}_j)$. The standard deviations are then denoted by $\mathrm{se}(\hat{\beta}_j)$, called the standard errors. The sampling distribution of $\hat{\beta}_j$'s are given by

$$\frac{\hat{\beta}_j - \beta_j}{\mathrm{se}(\hat{\beta}_j)} \sim t_{m-(p+1)}, \quad j = 0, 1, \ldots, p, \tag{3.18}$$

where t_v denotes a central t-distribution with v degrees of freedom. So, confidence interval for β_j can be obtained as

$$\hat{\beta}_j \pm t_{(m-p-1;\alpha/2)}\mathrm{se}(\hat{\beta}_j), \tag{3.19}$$

where $t_{(m-p-1;\alpha/2)}$ denotes the upper $\alpha/2$ percentage point of t_{m-p-1}.

3.1.5 Diagnostics for Regression Models

As mentioned earlier, linear models are established with several assumptions. It is important to check if all these assumptions are satisfied before developing inference, otherwise the developed inference will be invalid. Checking of model assumptions can be numerical or graphical. We prefer graphical diagnostics as they provide visual and clear results.

Residuals play a key role in checking these assumptions because patterns in residual

plots can reflect situations such as non-linearity, non-normality or non-independence for linear models as detailed below:

- Residual vs. fitted plot is a plot of residuals against the fitted values. We seek a random scatter along the horizontal line $y = 0$ with a constant variance. Any pattern in this plot indicates that at least the linearity assumption is violated. If linearity assumption is violated, one may consider using some transformation. Transformation can linearize at least approximately a non-linear relationship between the response variable and the covariates.

- Normal Q-Q plot is a plot of quantiles of ordered standardized residuals against the standard normal quantiles. We expect a pattern of $y = x$ if normality assumption holds. If there are curvatures in the tail, it means violation of normality assumption.

The above plots also show outliers, which are points that deviate considerably from the model. Outliers will generally have large absolute values of residuals and standardized residuals. However, outliers in response or in covariates may or may not affect the parameter estimation in linear models. The points that do affect are influential points. Influential points can significantly change regression models and so may lead to misleading inference. Cook (2000) proposed a measure of influence of points. It measures the difference between the fitted value with and without the i^{th} observation, and is given by

$$C_i = \frac{\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_{j(i)})^2}{\hat{\sigma}^2(p+1)}, \quad i = 1, \ldots, n,$$

where $\hat{y_j}$ is the j^{th} fitted value with the full data and $\hat{y}_{j(i)}$ is the j^{th} fitted value without the i^{th} observation.

- Plot of C_i against i shows influential Influent points with large Cook's distance standing out and in this case further action needs to be taken such as transformation or addition of interaction terms. Usually the cut-off value to examine influential points is $C_i > 1$.

3.1.6 Weakness of Classical Linear Regression

As mentioned earlier, when dealing with longitudinal data, one will convert wide format data structure into long format. In this way, the repeated responses will be treated as a single variable along with a time covariate and then standard statistical analytic methods will be applied. The classical linear models described above, however, assume that the observations are independent of one another. This assumption is not reasonable in these case of longitudinal data since the response of an individual at a time point will be dependent on the response of the same individual at a future time, so that there will exist correlation between responses of the same individual at different time points.

Multivariate linear models extend linear models by assuming the multivariate response variable in the longitudinal data to have a multivariate normal distribution, so that appropriate models for the analysis of longitudinal data can be developed in this manner by a combination of methods for linear models and the multivariate normal distribution.

3.2 Multivariate Linear Models

Let $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in})^T$ denote the response vector for the i^{th} individual, $\boldsymbol{X}_i$ denote the design matrix for the same individual given by

$$\boldsymbol{X}_i = \begin{pmatrix} 1 & X_{i11} & \cdots & X_{i1p} \\ 1 & X_{i12} & \cdots & X_{i2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{in1} & \cdots & X_{inp} \end{pmatrix},$$

and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ denote the parameter vector, for $i = 1, 2, \ldots, N$. Then, the multivariate linear model is given by

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, m, \tag{3.20}$$

where the response vector $\boldsymbol{Y}_i$ is assumed to have a multivariate normal distribution with

$$E(\boldsymbol{Y}_i) = \boldsymbol{X}_i\boldsymbol{\beta}, \quad \text{Var}(\boldsymbol{Y}_i) = \boldsymbol{\Sigma}_i, \tag{3.21}$$

3.2.1 Parameter Estimation

Assuming that the distribution of the multivariate response variable $\boldsymbol{Y}_i$ is multivariate normal, we have the probability density function (p.d.f.) as

$$f_{\boldsymbol{Y}_i}(\boldsymbol{y}_i) = (2\pi)^{-\frac{n}{2}}|\boldsymbol{\Sigma}_i|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})\right\}. \tag{3.22}$$

So, the log-likelihood function is given by

$$l = -\frac{mn}{2}\log(2\pi) - \sum_{i=1}^{m}\log(|\boldsymbol{\Sigma}_i|) - \frac{1}{2}\sum_{i=1}^{m}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}). \tag{3.23}$$

To get the maximum likelihood estimate of $\boldsymbol{\beta}$, we take the derivative of log-likelihood function with respect to $\boldsymbol{\beta}$ and set it equal to 0. Solution of the resulting equation gives the estimator of $\boldsymbol{\beta}$ as

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{m}\boldsymbol{X}_i'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{X}_i\right)^{-1}\sum_{i=1}^{m}\boldsymbol{X}_i'\boldsymbol{\Sigma}_{\boldsymbol{i}}^{-\mathbf{1}}\boldsymbol{y}_{\boldsymbol{i}}. \tag{3.24}$$

Usually $\boldsymbol{\Sigma}_i$ are unknown, and so by replacing $\boldsymbol{\Sigma}_i$ by its estimator $\hat{\boldsymbol{\Sigma}}_i$, we obtain the estimator of $\boldsymbol{\beta}$ as

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{m}\boldsymbol{X}_i'\hat{\boldsymbol{\Sigma}}_i^{-1}\boldsymbol{X}_i\right)^{-1}\sum_{i=1}^{m}\boldsymbol{X}_i'\hat{\boldsymbol{\Sigma}}_i^{-1}\boldsymbol{y}_i. \tag{3.25}$$

This is an unbiased estimator of $\boldsymbol{\beta}$ since

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}) &= \left(\sum_{i=1}^{m}\boldsymbol{X}_i'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{X}_i\right)^{-1}\sum_{i=1}^{m}\boldsymbol{X}_i'\boldsymbol{\Sigma}_{\boldsymbol{i}}^{-1}E(\boldsymbol{Y}_i) \\ &= \left(\sum_{i=1}^{m}\boldsymbol{X}_i'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{X}_i\right)^{-1}\sum_{i=1}^{m}\boldsymbol{X}_i'\boldsymbol{\Sigma}_{\boldsymbol{i}}^{-1}\boldsymbol{X}_i\boldsymbol{\beta} \\ &= \left(\sum_{i=1}^{m}\boldsymbol{X}_i'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{X}_i\right)^{-1}\left(\sum_{i=1}^{m}\boldsymbol{X}_i'\boldsymbol{\Sigma}_{\boldsymbol{i}}^{-1}\boldsymbol{X}_i\right)\boldsymbol{\beta} \\ &= \boldsymbol{\beta}. \end{aligned} \tag{3.26}$$

3.2.2 Sampling Distribution of β's

The variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ is given by

$$\begin{aligned}
\mathrm{Cov}(\hat{\boldsymbol{\beta}}) &= \left(\sum_{i=1}^{m} \boldsymbol{X}_i'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{X}_i\right)^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i'\boldsymbol{\Sigma}_{\boldsymbol{i}}^{-1}\mathrm{Var}(\boldsymbol{Y}_i)(\boldsymbol{X}_i'\boldsymbol{\Sigma}_i^{-1})' \left(\left(\sum_{i=1}^{m} \boldsymbol{X}_i'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{X}_i\right)^{-1}\right)' \\
&= \left(\sum_{i=1}^{m} \boldsymbol{X}_i'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{X}_i\right)^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Sigma}_i(\boldsymbol{\Sigma}_i^{-1})'\boldsymbol{X}_i \left(\sum_{i=1}^{m} \boldsymbol{X}_i'(\boldsymbol{\Sigma}_i^{-1})'\boldsymbol{X}_i\right)^{-1} \\
&= \left(\sum_{i=1}^{m} \boldsymbol{X}_i'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{X}_i\right)^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i'(\boldsymbol{\Sigma}_i^{-1})'\boldsymbol{X}_i \left(\sum_{i=1}^{m} \boldsymbol{X}_i'(\boldsymbol{\Sigma}_i^{-1})'\boldsymbol{X}_i\right)^{-1} \\
&= \left(\sum_{i=1}^{m} \boldsymbol{X}_i'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{X}_i\right)^{-1}.
\end{aligned} \tag{3.27}$$

So, the estimated variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ is given by

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{m} \boldsymbol{X}_{\boldsymbol{i}}'\hat{\boldsymbol{\Sigma}}_{\boldsymbol{i}}^{-\mathbf{1}}\boldsymbol{X}_{\boldsymbol{i}}\right)^{-1}, \tag{3.28}$$

where the diagonal elements are the estimated variance of $\hat{\beta}_j$'s. The sampling distribution of $\hat{\beta}_j$ is given by

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_j)}} \sim N(0,1), \quad j = 0, 1, \ldots, p,$$

using which the confidence intervals for β_j's can be readily obtained.

3.3 Generalized Linear Models and Generalized Estimating Equations

Generalized linear models (GLMs) are generalizations of multiple linear models. Multiple linear models assume the response variable to have a normal distribution with its expectation being a linear combination of a set of independent variables. However, many response variables are not necessarily continuous and may not even be normally distributed, and GLMs enable the analysis of such diverse types of univariate responses. Also, as in the case of linear models, a GLM links the expectation of response to a linear combination of covariates through some function, and can be described as follows:

- $\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$, where η_i is called a linear predictor;
- Y_i follows a certain distribution from the exponential family, for $i = 1, \cdots, m$, and are independent;
- $g(E(Y_i)) = \eta_i$, or equivalently, $E(Y_i) = g^{-1}(\eta_i)$, where $g(.)$ is called the link function;
- $\text{Var}(Y_i) = V(E(Y_i))$, where $V(.)$ is the variance function depending on the distribution and is a diagonal matrix.

In general, when repeated response variables in longitudinal studies do not necessarily have a multivariate normal distribution, but the correlation of the responses are considered, one can make use of generalized estimating equations (GEEs), proposed by Liang and Zeger (1986), for the estimation of parameters, which is generalization of multivariate linear models.

Let $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in})^T$ denote the response vector for the i^{th} individual, $\boldsymbol{X}_i$ denote the design matrix for the same individual given by

$$\boldsymbol{X}_i = \begin{pmatrix} 1 & X_{i11} & \cdots & X_{i1p} \\ 1 & X_{i12} & \cdots & X_{i2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{in1} & \cdots & X_{inp} \end{pmatrix},$$

and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ denote the parameter vector, for $i = 1, 2, \ldots, m$. Assume that the expectation of the response $E(Y_{ij}) = \mu_{ij}$ depends on the covariates through a known link function $g(.)$

$$g(\mu_{ij}) = \eta_{ij} = \boldsymbol{X}_{ij}^T \boldsymbol{\beta}. \tag{3.29}$$

The variance of Y_{ij} depends on the mean by

$$\text{Var}(Y_{ij}) = \phi v(\mu_{ij}), \tag{3.30}$$

where $v(\mu_{ij})$ is a known variance function and ϕ is a scale parameter. Then, the covariance matrix can be decomposed into the standard deviations and correlations as

$$\boldsymbol{V}_i = \boldsymbol{A}_i^{\frac{1}{2}} \text{Corr}(\boldsymbol{Y}_i) \boldsymbol{A}_i^{\frac{1}{2}}, \tag{3.31}$$

where A_i is a diagonal matrix with $\text{Var}(Y_{ij})$ along the diagonal and $\text{Corr}(\boldsymbol{Y}_i)$ is the correlation matrix as a function of the set of within-subject association parameters

$\boldsymbol{\alpha}$. The GEE estimator of $\boldsymbol{\beta}$ can be derived by minimizing the function

$$\sum_{i=1}^{m} \left(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\right) T\boldsymbol{V}_i^{-1} \left(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\right), \tag{3.32}$$

It is equivalent to solve the following generalized estimating equations:

$$\sum_{i=1}^{m} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = 0, \tag{3.33}$$

where $\boldsymbol{V}_i$ is the working covariance matrix and $\boldsymbol{D}_i = \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta}$ is the derivative matrix. R provides a package "multgee" for solving such GEEs. For more details, one may refer to Liang and Zeger (1986) on GEEs.

Chapter 4

Missing Values and Imputation

It is desirable to have complete data for longitudinal analysis, but missing values is a common problem in longitudinal data. The existence of missing data causes some problems in statistical inference. First, the data set will be unbalanced when data are missing and so some statistical analytic methods that require balanced data can not be used, like those described in the last chapter. Secondly, when missing values occur in a data set, there will be some loss of information that may result in a reduction of precision and will make the consequent analysis less effective. Finally, missing values can also result in an increase in bias for inference (Fitzmaurice *et al.*, 2012).

Many methods of handling missing data have been discussed in the literature. The easiest way is the so-called listwise deletion, which simply deletes the subjects with missing variables and is applied in most of the statistical softwares when missing data occur. This method is effective when there are only a few missing values in a large data and the deleted cases do not have any special characteristics other than what the remaining individuals have. But, the deletion method will lead usually to biased

estimates of parameters of interest. It can also be problematic when the proportion of missing values is large as it reduces the sample size, resulting in a reduction in the statistical power of the analysis. Furthermore, the listwise deletion wastes available information of individuals who have observed values for some of the variables.

Another similar approach for dealing with missing values is the so-called pairwise deletion. It excludes the subjects where missing values occur in variables included in the analysis. Even though this approach has the same drawbacks as listwise deletion that it results in bias and reduction of statistical power of analysis, it is more efficient than listwise deletion because it makes as much use of available information as possible. However, when pairwise deletion is applied, the sample sizes for analyzing different aspects of a study would be different and so might result in inconsistency. For example, if we are interested in the predictive models of body weight of preterm babies at day 7 and day 21, the models probably can not be representative for the same populations because of the distinct sample sizes.

In order to avoid wasting information caused by deletion, one may not be willing to delete the missing subjects, but rather be willing to substitute reasonable values for missing values. This procedure of dealing with missing values is called imputation. Imputation results in complete and balanced data so that standard statistical analytic methods can then be used. Further, it maintains the sample size so that the the statistical power of analysis can be preserved and also in improving the bias possibly.

4.1 Missingness in GTHPI data

In longitudinal data, there exists two missingness patterns. Response variables of some subjects may not be measured at some time point, but measured at later time points, creating intermittent missingness pattern. On the other hand, some subjects may drop out of a study and never return to participate in the study, leading to monotone missingness pattern. The GTHPI data set assesses both intermittent missingness and monotone missingness patterns in the response variable. The intermittent missingness may occur when preterm infants are unstable and not suitable for measurements when they are kept under care in neonatal intensive care units, while monotone missingness will happen when preterm infants are discharged from the hospital and never come back for further observation. Figure 4.1 shows the missing pattern of measurements of body weight of healthy preterm infants with observed values in black and missing values in white. In this case, about 9% of the measurements are missing intermittently while 6% are missing monotonically. Note that the explanatory variables are all observed for each infant in this data set.

4.2 Missingness Mechanisms

Before we deal with missing values, we need to figure out why the data are missing. It is unrealistic to record accurately all potential causes for missingness in a data set, but missingness may be intrinsically related to the data (Schafer and Graham, 2002). One can use a binary indicator variable I to describe the missingness of a response variable Y, that is, I takes the value 1 when Y is observed and 0 when Y is missing. Missingness mechanisms can then be considered as a joint distribution

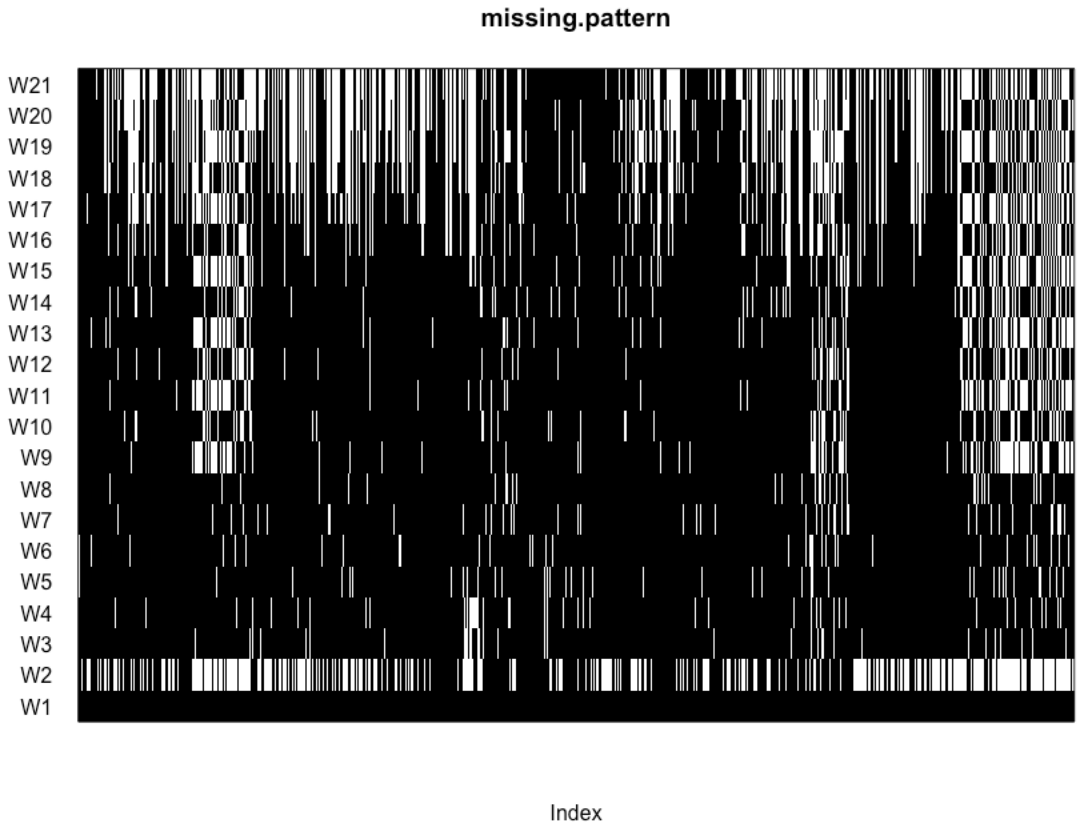

Figure 4.1: Missing pattern of postnatal growth for healthy preterm infants: white are missing while black are observed

of the set of response indicator variables, and it is best to describe the relationship between the missingness and the missing values (Schafer and Graham, 2002). The reason that a value is missing can be summarized into three types - missing completely at random (MCAR), missing at random (MAR) (Rubin, 1976), and not missing at random (NMAR) (Little and Rubin, 2002). Missingness mechanisms is important because it determines whether we need to complete the incomplete data by imputation or we can ignore the missing values, and which statistical method is appropriate for the analysis. Inappropriate method of analysis would increase bias and result in invalid inference.

4.2.1 Notation

Let $\boldsymbol{I}$ be a $m \times n$ indicator matrix corresponding to $\boldsymbol{Y}$ given by

$$\boldsymbol{I} = \begin{pmatrix} I_{11} & \cdots & I_{1j} & \cdots & I_{1n} \\ \vdots & & \vdots & & \vdots \\ I_{i1} & \cdots & I_{ij} & \cdots & I_{in} \\ \vdots & & \vdots & & \vdots \\ I_{m1} & \cdots & I_{mj} & \cdots & I_{mn} \end{pmatrix} = \begin{pmatrix} \boldsymbol{I}_1 \\ \vdots \\ \boldsymbol{I}_i \\ \vdots \\ \boldsymbol{I}_m \end{pmatrix},$$

where I_{ij} takes the value 1 when Y_{ij} is observed and 0 otherwise. The individual response variables $\boldsymbol{Y}_i$ can be partitioned into observed components and missing components, denoted by $\boldsymbol{Y}_i^o$ and $\boldsymbol{Y}_i^m$, respectively.

4.2.2 Missing Completely At Random (MCAR)

Data are MCAR when the probabilities that a response will be missing are all equal, independent of both the observed variables and the missing value itself. That is, the data is MCAR when $\boldsymbol{I}_i$ is independent of $\boldsymbol{Y}_i^o$, $\boldsymbol{Y}_i^m$ and $\boldsymbol{X}_i$, i.e.,

$$P(\boldsymbol{I}_i|\boldsymbol{Y}_i, \boldsymbol{X}_i) = P(\boldsymbol{I}_i).$$

MCAR is as if we determine whether to measure a response or not from the toss of a fair die. For example, if a 6 turns up, we do not measure the response value. The complete observed data can then be thought of as a random sample from the whole data. As a result, the distribution of the complete observed data $\boldsymbol{Y}_i$ given $\boldsymbol{X}_i$ is same as the distribution of the complete data. If the complete data are representative of the target population, the reduced complete observed data are sufficient enough to represent the target population. Analysis restricted on the completers will not increase the bias and any method of analysis yielding valid inference based on complete data can therefore yield valid inference on the complete observed data if the data are indeed MCAR.

On the other hand, the observed components for individuals with missing values have similar results under the assumption that data are MCAR. That is, the distribution of the observed components of incompleters $\boldsymbol{Y}_i^o$ given $\boldsymbol{X}_i$ have the same distribution as the corresponding components of completers. Moreover, the distribution of the missing components of an individual also have the same distribution as the corresponding components of completers. That is, $\boldsymbol{Y}_i^m$ given $\boldsymbol{X}_i$ is identically distributed as the same components of subjects with no missing responses. Hence,

the distribution of the observed components $\boldsymbol{Y}_i^o$ coincides with the distribution of the same components in the target population. Method of analysis using all the available data can also give valid inference in this case and do not increase the bias under this assumption.

4.2.3 Missing At Random (MAR)

Data are MAR when the probability that the missingness of response variables occur depends on observed values, but unrelated to the missing values themselves. That is, data are MAR when $\boldsymbol{I}_i$ is conditionally dependent of $\boldsymbol{Y}_i^o$ and covariates $\boldsymbol{X}_i$, i.e.,

$$P(\boldsymbol{I}_i|\boldsymbol{Y}_i, \boldsymbol{X}_i) = P(\boldsymbol{I}_i|\boldsymbol{Y}_i^o, \boldsymbol{X}_i).$$

As in MAR, all the factors that affect the missingness of response variables should be included as covariates; otherwise, the assumption of MAR would not hold.

Since the missingness of response variables depends on the observed values, the distribution of $\boldsymbol{Y}_i$ given $\boldsymbol{X}_i$ of the completers is not the same as the distribution of $\boldsymbol{Y}$ given $\boldsymbol{X}$ in the target population. Analysis on the data of completers is not appropriate and may lead to biased estimates of parameters in this case.

The distribution of the observed components $\boldsymbol{Y}_i^o$ of $\boldsymbol{Y}_i$ is not the same as the distribution of the same components of completers. However, the distribution of the missing components $\boldsymbol{Y}_i^m$ conditioned on the observed components $\boldsymbol{Y}_i^o$ is the same as the distribution of the corresponding components conditioned on the same values as $\boldsymbol{Y}_i^o$ of the completers. Consequently, missing values can be predicted by using the

observed data and a model derived from the complete cases with the same observed components.

Under MAR, the joint distribution of $\boldsymbol{I}_i|(\boldsymbol{Y}_i, \boldsymbol{X}_i)$ is not needed to develop likelihood-based analysis, only the joint distribution of $\boldsymbol{Y}_i|\boldsymbol{X}_i$ is needed. MCAR is a special case of MAR and so it also has this property. These missingness mechanisms are said to be ignorable since they do not depend on $P(\boldsymbol{I}_i|\boldsymbol{Y}_i, \boldsymbol{X}_i)$.

4.2.4 Not Missing At Random (NMAR)

If the probabilities that response variables are missing depend on the missing values themselves, the missing values are said to be MNAR. That is, the distribution of $\boldsymbol{I}_i$ is related to Y_i^m and depends on at least one of the elements of $\boldsymbol{Y}_i^m$, i.e.,

$$P(\boldsymbol{I}_i|\boldsymbol{Y}_i, \boldsymbol{X}_i) = P(\boldsymbol{I}_i|\boldsymbol{Y}_i^m, \boldsymbol{Y}_i^o, \boldsymbol{X}_i).$$

This missing mechanism is said to be non-ignorable since the distribution of $\boldsymbol{I}_i|(\boldsymbol{Y}_i, \boldsymbol{X}_i)$ gives information about the distribution of the missing observations. The distribution of $\boldsymbol{Y}_i^m$ depends on $\boldsymbol{Y}_i^o$ and $P(\boldsymbol{I}_i|\boldsymbol{Y}_i, \boldsymbol{X}_i)$.

4.2.5 General Rule

It is hard to identify the missingness mechanism for a data unless analysts know the data collection procedure, especially in the case of NMAR data. In most analysis, data are assumed to be MAR since some analytic methods under this assumption are also suitable for MNAR data. Due to the special nature of a longitudinal data in

that they possess correlation between response variables at different time points, the missing data are assumed to be MCAR.

4.3 Single Imputation

4.3.1 Last Observation Carried Forward (LOCF)

Last observation carried forward is specific for longitudinal data, imputing missing values with the last observed values for each individual. It is unrealistic as this method assumes that the measurements of outcome variable remain unchanged for the period when measurements are missing, especially when the missingness is caused by dropouts in clinical trials.

4.3.2 Regression Imputation

The procedure of regression imputation in longitudinal data is that a series of regression models are established by taking a response variable at a time to be dependent variable and all the previous responses and covariates to be independent variables. The regression model is fitted with individuals containing fully observed values for those variables required in the models. The fitted model is then used to predict expected values where the response is missing and the predicted values are substituted for the missing values. This approach takes the relationship between variables into account and so the estimation of parameters of interest are less biased than under LOCF.

Specifically, suppose $\boldsymbol{Y} = (\boldsymbol{Y}^o, \boldsymbol{Y}^m)$ is a $m \times 1$ vector, partitioned into two sub-vectors containing observed components and missing components, respectively. Let $\boldsymbol{X}$ be a $m \times (p+1)$ design matrix whose entries are all observed. $\boldsymbol{X}$ can be partitioned into $\boldsymbol{X}^o$ and $\boldsymbol{X}^m$ corresponding to $\boldsymbol{Y}^o, \boldsymbol{Y}^m$, respectively. The procedure can then be summarized in three steps:

- **Step 1:** Models are established as

$$\boldsymbol{Y}^o = \boldsymbol{X}^o\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon}|\boldsymbol{X}^o \sim N(0, \sigma^2\boldsymbol{I});$$

- **Step 2:** Use of least-squares method gives the estimates of the coefficients as $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{oT}\boldsymbol{X}^o)^{-1}\boldsymbol{X}^{oT}\boldsymbol{Y}^o$;

- **Step 3:** Then, the missing values can be imputed as

$$\hat{\boldsymbol{Y}}^{\boldsymbol{m}} = \boldsymbol{X}^m\hat{\boldsymbol{\beta}}.$$

4.3.3 Stochastic Regression Imputation

Regression imputation, which simply replaces missing values with expected values by linear regression models, underestimates the variability of imputation models as imputed values are exactly along the regression line without deviation. To correct the lack of error term, stochastic regression adds error terms to the predicted values by linear models. The choice of error terms is that they are randomly selected from a normal distribution with mean 0 and standard deviation $\hat{\sigma}$, where $\hat{\sigma}$ is the estimated deviation of the observed residuals.

The procedure is similar to the above regression imputation except that in the third step the missing values are imputed as

$$\hat{\boldsymbol{Y}}^{\boldsymbol{m}} = \boldsymbol{X}^m\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\epsilon}}, \quad \hat{\boldsymbol{\epsilon}} \sim N(0, \hat{\sigma}^2\boldsymbol{I}).$$

4.3.4 Advantages and Disadvantages of Single Imputation

Single imputation retains the statistical power without a reduction in sample size and is also easy to implement. However, it generates only a complete data set and treat the imputed values as the real observed values and do not account for uncertainty caused by missing values.

4.4 Multiple Imputation

Rubin (2004) proposed multiple imputation to preserve the uncertainty in missing values. The procedure of statistical analysis based on multiple imputation can be summarized in three steps as follows:

- First, each missing value is imputed with k plausible values to create k complete data sets. k may be taken to be any value between 2 to 5;
- Second, each complete data is analyzed by using standard statistical methods to obtain k set of parameters or statistics of interest;
- Third, the multiple results are appropriately combined to draw a single final result, together with standard errors that reflect the inherent uncertainty in missing data. For example, suppose we are interested in the coefficient $\boldsymbol{\beta}$ of

the model, with the k sets of the estimated regression parameter obtained from the k completed data sets, denoted by $\hat{\boldsymbol{\beta}}^{(i)}$ for $i = 1, 2, \ldots, k$, the multiple imputation estimate of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = \frac{1}{k} \sum_{i=1}^{k} \hat{\boldsymbol{\beta}}^{(i)},$$

and its estimated variance is

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = \boldsymbol{W} + \left(1 + \frac{1}{k}\right) \boldsymbol{B},$$

where $\boldsymbol{W} = \frac{1}{k} \sum_{i=1}^{k} \widehat{\mathrm{Var}} \left(\hat{\boldsymbol{\beta}}^{(i)}\right)$ and $\boldsymbol{B} = \frac{1}{k-1} \sum_{i=1}^{k} \left(\hat{\boldsymbol{\beta}}^{(i)} - \hat{\boldsymbol{\beta}}\right) \left(\hat{\boldsymbol{\beta}}^{(i)} - \hat{\boldsymbol{\beta}}\right)^{T}$. Multiple imputation is a flexible technique for handling missing data and many software packages are available in statistical softwares such as in SAS, R and SPSS. Each multiple imputation package may, however, have different methods inside it. One may have to understand the imputation procedure used in different software packages. Here, I only focus on the "mi" package in R to impute missing data (Su *et al.*, 2011). "mi" uses an algorithm known as a chained equation approach (Van Buuren and Oudshoorn, 2000; Raghunathan *et al.*, 2001) which requires a series of regression models to be specified for each response variable with missing data. The models for imputation are not necessarily same as models used for the statistical analysis. Missing data in each variable are then imputed sequentially by using the specified model. We now describe the procedure of multiple imputation in "mi" package in R in detail.

4.4.1 Methodology

Let $\boldsymbol{X} = (X_1, \ldots, X_p)$ be a random covariate vector, $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ be a random response vector, and $\boldsymbol{Y}_{-j} = (Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_n)$ be the vector containing all the response variables except Y_j. It is assumed that missing data occur only in response variables and that covariates are fully observed. A series of regression models for Y_j are established, given by

$$Y_j = \beta_0 + \sum_{k=1}^{j-1} \beta_k Y_k + \sum_{k=j+1}^{n} \beta_k Y_k + \sum_{r=1}^{p} \alpha_r X_r + \epsilon_j, \ \epsilon_j \sim N(0, \sigma_j^2).$$

Then, the mean and variance of the conditional distribution of Y_j, given $\boldsymbol{X}, \boldsymbol{Y}_{-j}$, are given by

$$E(Y_j|\boldsymbol{X}, \boldsymbol{Y}_{-j}) = \beta_0 + \beta_1 Y_1 + \cdots + \beta_{j-1} Y_{j-1} + \beta_{j+1} Y_{j+1} + \cdots + \beta_n Y_n, \ \mathrm{Var}(Y_j|\boldsymbol{X}, \boldsymbol{Y}_{-j}) = \sigma_j^2.$$

That is, the conditional distribution of Y_j, given $\boldsymbol{X}, \boldsymbol{Y}_{-j}$, is then assumed to be normal with mean $\beta_0 + \beta_1 Y_1 + \cdots + \beta_{j-1} Y_{j-1} + \beta_{j+1} Y_{j+1} + \cdots + \beta_n Y_n$ and variance σ_j^2, denoted by $Y_j|(\boldsymbol{X}, \boldsymbol{Y}_{-j}) \sim N(\beta_0 + \beta_1 Y_1 + \cdots + \beta_{j-1} Y_{j-1} + \beta_{j+1} Y_{j+1} + \cdots + \beta_n Y_n, \sigma_j^2)$. Note that the sets of regression parameters for each regression model are different.

Imputation by chained equations is actually a Gibbs sampler process. It starts with a completed data in which missing values are imputed with randomly selected values from the observed data in the same response variable. The regression model for the conditional mean of the response at j^{th} occasion are then fitted by using the completed data of subjects who do not have missing values at that occasion iteratively

and sequentially and missing values in predictors are replaced by their imputed values from the previous imputation. Each missing value is imputed with a random selection from its conditional distribution. Specifically, for $t = 2, \ldots, T$, where T is a predetermined value,

$$
\begin{aligned}
&\text{Select } Y_1^{(t)} \text{ from } N(\hat{\beta}_0 + \hat{\beta}_2 Y_2^{(t-1)} + \cdots + \hat{\beta}_n Y_n^{(t-1)} + \hat{\alpha}_1 X_1 + \cdots + \hat{\alpha}_p X_p, \hat{\sigma}_1^2),\\
&\text{Select } Y_2^{(t)} \text{ from } N(\hat{\beta}_0 + \hat{\beta}_1 Y_1^{(t)} + \hat{\beta}_3 Y_3^{(t-1)} + \cdots + \hat{\beta}_n Y_n^{(t-1)} + \hat{\alpha}_1 X_1 + \cdots + \hat{\alpha}_p X_p, \hat{\sigma}_2^2),\\
&\vdots\\
&\text{Select } Y_j^{(t)} \text{ from } N(\hat{\beta}_0 + \hat{\beta}_1 Y_1^{(t)} + \cdots + \hat{\beta}_{j-1} Y_{j-1}^{(t)} + \hat{\beta}_{j+1} Y_{j+1}^{(t-1)} + \cdots + \hat{\beta}_n Y_n^{(t-1)}\\
&\quad + \hat{\alpha}_1 X_1 + \cdots + \hat{\alpha}_p X_p, \hat{\sigma}_j^2),\\
&\vdots\\
&\text{Select } Y_{n-1}^{(t)} \text{ from } N(\hat{\beta}_0 + \hat{\beta}_1 Y_1^{(t)} + \cdots \hat{\beta}_{n-2} Y_{n-2}^{(t)} + \hat{\beta}_n Y_n^{(t-1)} + \hat{\alpha}_1 X_1 + \cdots + \hat{\alpha}_p X_p, \hat{\sigma}_{n-1}^2),\\
&\text{Select } Y_n^{(t)} \text{ from } N(\hat{\beta}_0 + \hat{\beta}_1 Y_1^{(t)} + \cdots + \hat{\beta}_{n-1} Y_{n-1}^{(t)} + \hat{\alpha}_1 X_1 + \cdots + \hat{\alpha}_p X_p, \hat{\sigma}_n^2)
\end{aligned}
$$

for the missing data. Each time, the imputed missing data are overwritten by subsequent imputing values and the set of imputation in the last iteration is used to form a completed data set. The same procedure is then repeated multiple times to obtain the k completed data sets.

4.4.2 Diagnostics

Diagnostics is an important part of any statistical procedure. Imputation procedure using regression models is in fact the same as statistical analysis procedure and so the fit of models used for imputation should be checked. "mi" provides three diagnostic plots to check the fit of imputation models (Su *et al.*, 2011). The first plot is a

histogram of the observed, the imputed and the completed values of a variable to be imputed. If the imputed values are all within reasonable range of the observed values, then the imputed values are acceptable. The second plot is the binned residual plot (Gelman *et al.*, 2000). It is a modification of the usual residual plot by partitioning the scatter plot into several bins with respect to expected values. There are approximately equal numbers of points in each bin. The average of residuals and the average of expected values for each bin are calculated and plotted. The binned residual plot plays a similar role as the usual residual plot. The average residual points are preferred to fall within the 95% error bounds. If there are a lot of points falling outside the error bounds, improvement for imputation is needed. The improvement can be achieved by transformation to response variable. The third plot is the scatterplot of the observed against the predicted values of the observed and imputed values. The scatterplot demonstrates the similarity of the imputed data to the observed data.

4.4.3 Advantages and Disadvantages of Multiple Imputation

Multiple imputation shares the advantages of single imputation. It completes the incomplete data set so that one can use standard methods of statistical analysis that require balanced data set. It retains the sample size so that the statistical power is not reduced.

Multiple imputation is advantageous than single imputation to some extent. With multiple imputation, each missing value is replaced with multiple plausible values to generate multiple completed data sets, ensuring that the uncertainty related to the

imputed values can be taken into account. Besides, multiple imputation is advantageous when covariates which are predictive of either the probability of missingness or the responses are excluded from the model for analysis. These covariates can be introduced in the imputation process to improve the imputation of missing values. Indeed, inclusion of any variates that are highly related to the response would increase precision of imputation and so would increase the accuracy of statistical analysis. It implies that the imputation model and the model for analysis are necessarily the same, and usually the model for analysis is simpler than the imputation model. Moreover, multiple imputation can reflect the distribution of each missing value if a large number of completed data sets are generated.

However, there are obvious disadvantages of multiple imputation. First, it requires more computation than single imputation as it needs to create multiple completed data sets. Second, more work is needed to analyze the multiple data sets individually to draw results for each data set. This disadvantage should not be of great concern when k is modest. Generally, it is adequate to choose k between 2 to 5. But, when the proportion of missing data is large, a large k is required, which would result in a burdensome computational task.

4.5 Evaluation Methods

After imputation, we still do not know how accurate the imputed values are since we do not know the real values of the missing ones. Cross-validation is a technique that can be used to test the accuracy of imputation. It treats the completed data to be the real data and assumes that some of the observed values are missing, then the

imputation methods are applied to these "artificial" data and the imputed values are compared with the observed values.

4.5.1 Cross-validation

The detailed procedure of cross-validation technique can be summarized as follows:

- First, we randomly select 10% of measurements that are observed to be "missing". This chosen data are called the validation data set. The remaining data, including imputed values, are called the training data;
- Second, we reapply the imputation method to the training data set to obtain imputed values for the "missing" data;
- Third, compare them with the observed values to see how well the models performed. If the predicted values of the validation data are very close to the observed values, it means that the imputation method resulted in accurate estimates of the missing values.

4.5.2 Measure of Evaluation

The sum of squared errors of the observed and imputed values of the validation data can serve as an overall measure to compare different imputation models. The smaller the sum of squared errors is, the better the imputation model is.

Chapter 5

Results of Imputation

Several single imputation methods and multiple imputation methods have been described earlier. Imputation is widely used in practice since standard analytic methods can be applied to the completed data set. Any imputation method, replacing missing data with plausible values to complete the incomplete data set, can retain statistical power by retaining the sample size. Both single imputation and multiple imputation have their own advantages and disadvantages, as pointed out earlier. Single imputations which generate only one completed data set need less work and can therefore save time. However, they do not take the uncertainty of imputed values into account. Multiple imputation can reveal this inherent uncertainty by generating multiple different completed data sets. But, it would take more time and effort to implement multiple imputation and the corresponding analysis. In practice, when imputations are applied to data, some problem may arise according to the particular data set that is being analyzed. To better understand each of the imputation methods described earlier, imputations are applied to GTHPI data.

LOCF imputes the missing responses individual-by-individual to preserve the characteristics of body weight for each infant. The missing responses are imputed sequentially with a series of regression models for each response variable by the regression-based imputation. LOCF method only generates fixed values for missing values and so does the regression imputation once the regression models are determined. Meanwhile, stochastic imputation and multiple imputation can generate different results each time we implement them due to the randomness involved in these two methods. This randomness can not be avoided, but one can set seed while doing the computation to keep different imputation methods consistent.

Before we impute the missing values, we need to check the validation of the models for regression imputation and multiple imputation. Let us consider the imputation for day 21, for example. Figure 5.2 provides the diagnostics check of the model for regression imputation/stochastic imputation of missing weights on day 21. We can see from the plots that all the assumptions are satisfied well since the residuals are randomly scattered along the horizontal line $y = 0$ and the QQ plot is approximately linear. However, it is noticed that there are some outliers. Several influential points are seen from the Cook's distance and so the inclusion and exclusion of those points can lead to deviation of the regression line, but they should remain in the sample because those weights are not strange values and that we would rather keep the sample size as large as possible. Similarly, Figure 5.3 shows the diagnostics for the multiple imputation for day 21. The blue ones are for the observed values and red are for imputed values. These show that the imputation models are quite reasonable. The diagnostic results for imputation of the other days are all similar.

5.1 Illustration for One Particular Infant

Completed data set should follow the nature of the exact values observed. That is to say, the imputed values must be positive and the growth of any infant must follow a trend that an initial weight loss is permitted and weight must increase as time goes by. Table 5.3 gives the imputed values of weight from day of life 1 to 21 for the infant with most missing values by each of the imputation method. The LOCF have constant body weight when successive missingness occur, which is unrealistic and will certainly underestimate the weight gain. Regression imputation and stochastic imputation gives reasonable imputed values, consistent with the expected growth of decreasing weight during the first week and increasing weight afterward. But, the multiple imputation is observed not to yield imputed values as we would have expected.

Figure 5.4 visualizes the growth of this infant. It can be seen from this plot that growth trajectory of this infant greatly fluctuates with multiple imputation. Regression imputation and stochastic imputation have a slight fluctuant growth during the first week of life. Even the growth trajectory of data completed by LOCF depicts a gradient increase in weight after an initial weight loss during the first week, but the constant weight from day 10 to day 17 is unrealistic.

5.2 Illustration of Imputation for the Complete Data

Since we do not know the exact values of the missing data, it is not easy to assess how well different imputation methods are performed by studying the imputed data. Cross-validation is an efficient approach to assess how precise the imputed data obtained by different imputation methods are. For this purpose, we can choose a certain proportion of observed data to be "missing" and apply imputation to the thus-created incomplete data set. The difference between the imputed and observed values of the "missing" data can then be used to assess the precision of imputation.

I chose 98 observed values (10% of the subjects) for each response variable to be "missing" and then re-completed the data set with each of the imputation methods. The sum of squared errors obtained from each of the imputation methods are summarized in Table 5.4. As we would expect, LOCF is worse than the regression-based imputations. It is consistent with the cross-validation result since they have very large sum of squared errors. Regression imputation, stochastic imputation and multiple imputation, which are regression-based methods, all relate the missing data to other covariates. These methods have better prediction of missing values. The stochastic imputation is not necessarily better off as it adds randomly selected error terms to the predicted values of missing data which results in pushing the imputed values farther from the actual values. Even though multiple imputation is advantageous in that it account for the uncertainty, it has the same problem as stochastic imputation since missing data are imputed with randomly selected values from its

conditional distribution.

Figure 5.2: Diagnostics for model $W_{21} = \beta_0 + \beta_1 W_1 + \cdots + \beta_{20} W_{20} + \alpha_1 GA + \alpha_2 BW + \epsilon$

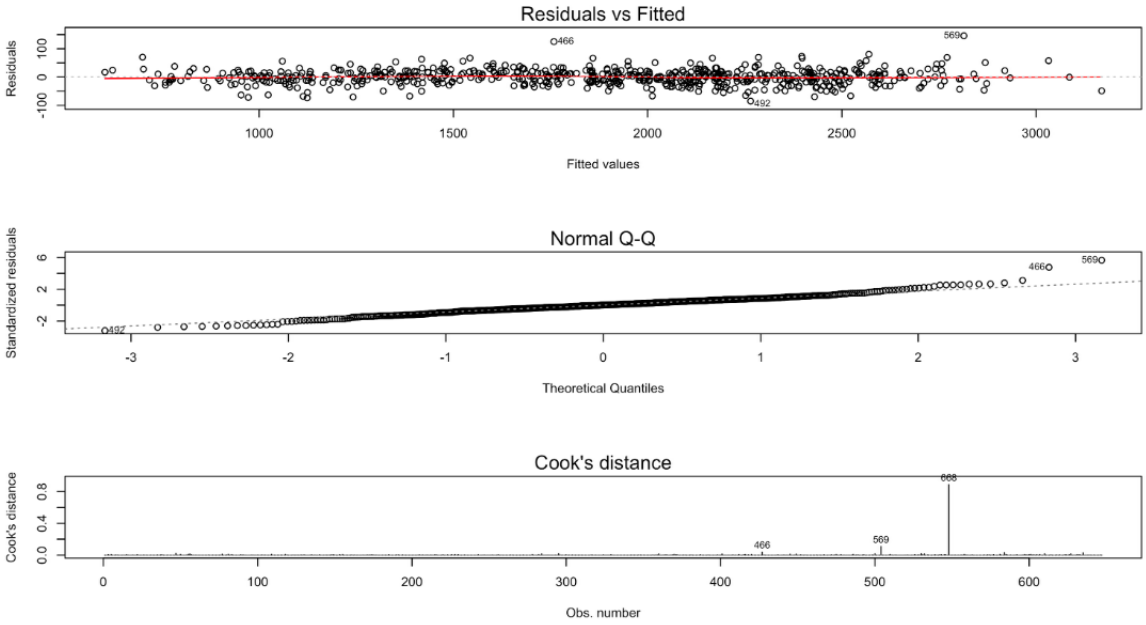

Figure 5.3: Multiple Imputation Diagnostics: Red are imputed and blue are observed

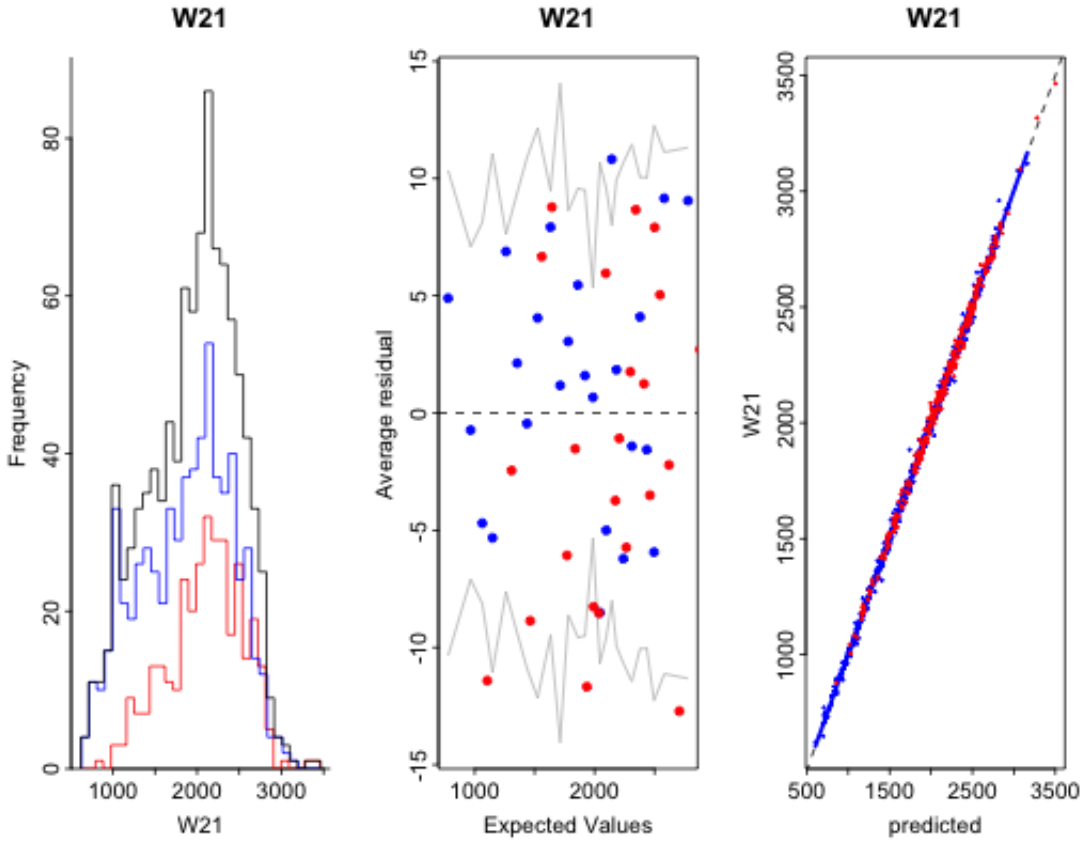

Table 5.3: Imputation results for the infant with most missing values

day	Original	LOCF	Regression Imputation	Stochastic Imputation	Multiple Imputation		
					data1	data2	data3
1	1210	1210	1210	1210	1210	1210	1210
2	NA	1210	1190.8	1192.4	1088.9	1179.7	1101.6
3	870	870	870	870	870	870	870
4	NA	870	917.3	899.3	872.3	889.0	844.7
5	850	850	850	850	850	850	850
6	NA	850	878.8	858.0	837.2	874.7	838.5
7	860	860	860	860	860	860	860
8	852	852	852	852	852	852	852
9	NA	852	886.4	884.2	952.5	1015.0	1057.4
10	907	907	907	907	907	907	907
11	NA	907	929.1	934.8	1151.3	1230.2	1104.4
12	NA	907	946.4	940.5	942.3	943.5	870.5
13	NA	907	960.6	949.6	1439.1	1368.6	1202.2
14	NA	907	982.0	960.4	923.6	1020.7	893.1
15	NA	907	1000.8	988.1	1489.8	1439.7	1199.4
16	NA	907	1021.4	985.3	1017.5	1019.8	900.5
17	NA	907	1041.6	1001.1	1637.6	1433.8	1281.9
18	1001	1001	1001	1001	1001	1001	1001
19	NA	1001	1039.2	1025.0	1348.7	1220.8	1233.1
20	1057	1057	1057	1057	1057	1057	1057
21	NA	1057	1083.8	1037.7	1082.5	1036.7	1077.3

Table 5.4: SSE by Cross-Validation for the Complete Data

LOCF	Regression Imputation	Stochastic Imputation	Multiple Imputation
243,852,939	2,819,639	5,226,838	9,028,473

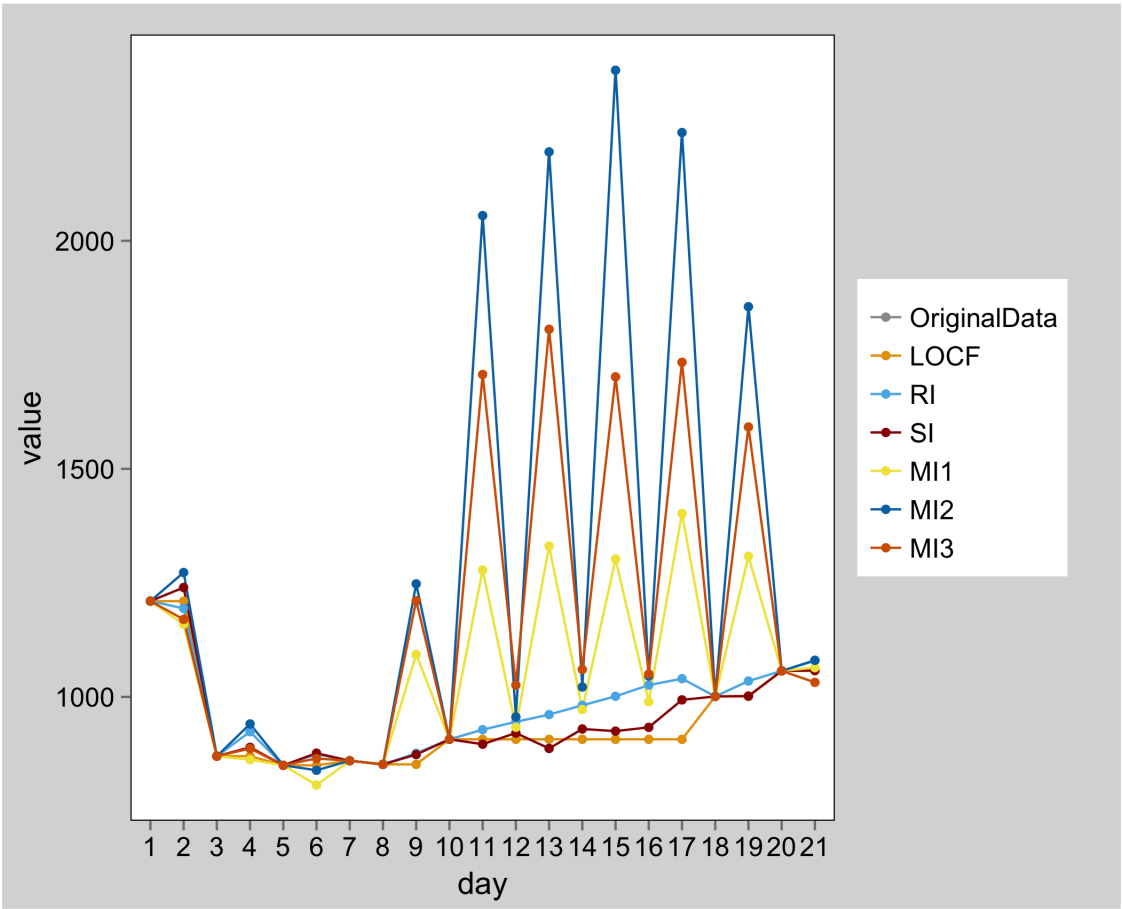

Figure 5.4: Individual Growth Trajectory for the Infant with Most Missing Values

Chapter 6

Subsequent Analysis

So far, we have introduced and examined several imputation methods and observed that the regression-based imputation results in good estimates of the missing values in the GTHPI data. The aim of imputation is not just to find the missing values, but rather to complete the missing data so that standard analysis approaches can then be used. Even though the regression imputation does not take the uncertainty into account as compared to the multiple imputation, it is robust enough for our GTHPI study. As mentioned before, the aim of this study is to characterize the feature of the preterm infants specific for gestational age to find which percentile a preterm infant will adjust to after 21 days of self-adjustment and also to develop predictive models for estimating the weight at days 7, 14 and 21.

6.1 Characterization of GTHPI data

Statistically speaking, the data are characterized in terms of summary of quantiles of weights specific for completed gestational age. Figure 6.5 summarizes percentiles

of weights in groups of completed gestational age visually in which the solid coloured curves are the medians, the shaded areas are corresponding to the 10th and 90th percentiles, and the vertical bars are the 3rd and 97th percentiles. The starting time point for each group is the median of exact gestational ages. The black curves are the Fenton reference curves (Fenton and Kim, 2013), in which the solid curve is the Fenton median curve, the long dash curves are the 10th and 90th percentiles, while the dashed curves are the 3rd and 97th percentiles of weight. It can be seen from Figure 6.5 that the preterm birth results in a gap between the growth trajectories of preterm infants and the reference growth trajectories. The quantile growth trajectories of preterm infants are lower than their corresponding reference quantile curves. Let us consider the growth curves for preterm infants with 34 weeks of completed gestation, as an example. Approximately, the third quantile curve adjusts to the third quantile reference curve, the 10th quantile curve adjusts to the 8th quantile reference curve, the median growth curve adjusts to 20th quantile reference curve, the 90th quantile curve adjusts to 38th quantile reference curve while the 97th quantile curve adjusts to 48th quantile reference curve. The other groups follow similar adjustment. This characterization gives us an overall growth pattern of preterm infants with specific completed gestational age.

6.2 Predictive Models

6.2.1 Univariate Response

We will first treat the response at day 7 or 14 or 21 to be a univariate response,and not care about the correlation between repeated measurements. The responses are

independent among individuals so that linear regression models can be considered. The data is in wide format when it is analyzed with this method.

We have many variables other than weights recorded in the data, and not all the variables need to be covariates in the first place. We interested in predictive models which can be used to predict the weight of preterm infants at certain day of life given the information at birth. That is, the covariates are centre, yoa, gender, gas, gad/gawexact, bw, md, eth and preg. Intuitively, yoa, md, eth and preg may not be significant since later-on weights would not differ much among different levels of these categorical variables. Table 6.5 shows the parameter estimates for the model for weight at day 7. It is evident that md, eth, preg and yoa are not significant covariates as our intuition is suggested. We can also use backward selection method to check that these variables are not significant. The backward selection proceeds as follows,

- **Step 1:** fit the model with all possible covariates;
- **Step 2:** Look at the p-values of the coefficients, and find the variable whose coefficient has largest p-values; if it is larger than 0.05, then it is not significant and can be removed from the model. If it is not, then the variable should remain in the model and the model is determined and the selection is stopped;
- **Step 3:** If there is a removal of a variable in Step 2, then the model is refitted with the remaining covariates and Steps 2 and 3 are repeated until the final model is fitted.

Table 6.6 shows the estimates of the backward selected model for weight at day 7, with the remaining covariates being all significant now. The results are similar for

predictive models for weights at days 14 and 21. There are some weak points in these models since significance of medical centres constrains the prediction of weight of preterm infants in the five participating medical centres. Since our target population of prediction is all preterm infants born during 25 and 34 completed weeks of gestation, this means that these models are not desirable for this reason.

To determine whether centre and gender can be excluded from the model compared to the four-covariates model, we can look at their relative AIC (Akaike Information Criterion) which is the ratio of AIC between the reduced models and the full model. AIC is a measure of relative quality of statistical models when models have different amount of covariates, and is given by

$$AIC = 2k - 2\ln(L),$$

where k is the number of parameters in the model, and L is the maximized value of the likelihood function for the model. Hence, AIC not only rewards goodness of fit, but also penalizes the decreased residual deviation resulted by increased number of covariates among models. Reduced models with gestational age, birth weight and with or without gender are then compared with the full model with all four variables, gestational age, birth weight, medical centre and gender in terms of relative efficiency. It can be seen from Table 6.7 that the relative AIC between reduced models and full models are near 1 showing that the reduced models have close relative AIC values, and so the simpler model are as good as the full model including all four variables.

Table 6.5: Parameter Estimates for the Full Model to Predict Weight at Day 7

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
Intercept	$8.902e+03$	$4.301e+03$	2.070	$0.038816*$
bw	$8.541e-01$	$8.795e-03$	97.114	$< 2e-16 ***$
gaw	$-1.559e+01$	$8.570e+00$	-1.819	0.069234.
genderM	$1.958e+01$	$5.145e+00$	3.807	$0.000152 ***$
centre	$3.595e+00$	$2.182e+00$	1.647	0.099857.
yoa	$-4.669e+00$	$2.141e+00$	-2.181	$0.029457*$
gad	$5.045e+00$	$1.231e+00$	4.098	$4.6e-05 ***$
mdC	$1.009e+02$	$5.607e+01$	1.799	0.072374.
mdCS	$-1.885e+01$	$2.332e+01$	-0.808	0.419150
mdV	$-1.287e+01$	$2.336e+01$	-0.551	0.581755
mdVS	$-2.705e+01$	$5.578e+01$	-0.485	0.627899
ethA	$-1.616e+01$	$3.406e+01$	-0.474	0.635341
ethAf	$5.232e+01$	$4.047e+01$	1.293	0.196480
ethC	$5.218e+00$	$3.176e+01$	0.164	0.869551
ethM	$-3.495e+01$	$3.605e+01$	-0.970	0.332486
ethMe	$-1.411e+01$	$4.196e+01$	-0.336	0.736650
ethN	$-6.327e+00$	$3.125e+01$	-0.202	0.839604
ethNa	$1.495e+01$	$7.827e+01$	0.191	0.848615
ethSA	$8.389e+00$	$5.970e+01$	0.141	0.888280
pregMG	$7.828e+00$	$2.229e+01$	0.351	0.725576
pregS	$2.063e+01$	$2.210e+01$	0.933	0.350853

The final models for the prediction of weights at days 7, 14, 21 are given by

$$W_7 = -533.4 + 0.8624\text{bw} + 21.75\text{gaw},$$

$$W_{14} = -747.899 + 0.87\text{bw} + 33.566\text{gaw},$$

$$W_{21} = -973.5 + 0.896\text{bw} + 46.28\text{gaw}.$$

Table 6.6: Parameter Estimates for the Reduced Model to Predict Weight at Day 7

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
Intercept	$-5.303e+02$	$4.162e+01$	-12.742	$<2e-16***$
bw	$8.617e-01$	$8.008e-03$	107.609	$<2e-16***$
gaw	$2.079e+01$	$1.678e+00$	12.391	$<2e-16***$
genderM	$1.476e+01$	$4.632e+00$	3.187	$0.001483**$
centre	$6.660e+00$	$1.726e+00$	3.859	$0.000121***$

Table 6.7: Relative AIC of Models

Covariates	gaw, bw, centre, gender	gaw, bw, gender	gaw, bw
AIC at day 7	8291.94	8382.32	8389.48
Relative AIC		1.001557	1.002412
AIC at day 14	8891.062	8907.092	8910.557
Relative AIC		1.001803	1.002193
AIC at day 21	9680.453	9704.204	9708.467
Relative AIC		1.002453	1.002894

6.2.2 Multivariate Response

Now let us consider the multivariate model by taking into consideration the correlation between the repeated responses. First, we need to transform the data into a long format and add a time variable. The covariates considered here are gas, bw, gender, centre as well as day of life. The well-known AIC cannot be directly applied since AIC is based on maximum likelihood estimation, while GEE is nonlikelihood based. Pan (2001) proposed Quasi-likelihood Information criterion (QIC) based on AIC by replacing the likelihood with quasi-likelihood given by

$$QIC = -2Q + 2\text{trace}(\hat{\boldsymbol{\Omega}}\hat{\boldsymbol{V}}),$$

where Q is the quasi-likelihood, $\hat{\boldsymbol{\Omega}} = -\frac{\partial^2 Q}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}$ and $\hat{\boldsymbol{V}}$ is the estimator of the variance-covariance matrix of $\boldsymbol{\beta}$. Table 6.8 shows the QIC and the relative QIC of

Table 6.8: QIC of Multivariate Linear Models

Covariates	gaw, bw, centre, gender, day	gaw, bw, gender, day	gaw, bw, day
QIC	191,822	192044	192,109
Relative QIC		1.0012	1.0015

the reduced models compared to the full model. As in the case of linear models above, these results reveal that the QIC ratio of reduced models to the full model are once again very close to 1. Thus, the model with gaw, bw, day as covariates is sufficient. The predictive model fitted by MLM is given by

$$W = -823.703 + 0.888\text{bw} + 26.962\text{gaw} + 19.645\text{day}$$

Therefore, the simpler predictive models for weight at days 7, 14 and 21 are as follows:

$$W_7 = -686 + 0.888\text{bw} + 26.962\text{gaw}$$

$$W_{14} = -549 + 0.888\text{bw} + 26.962\text{gaw}$$

$$W_{21} = -411 + 0.888\text{bw} + 26.962\text{gaw}$$

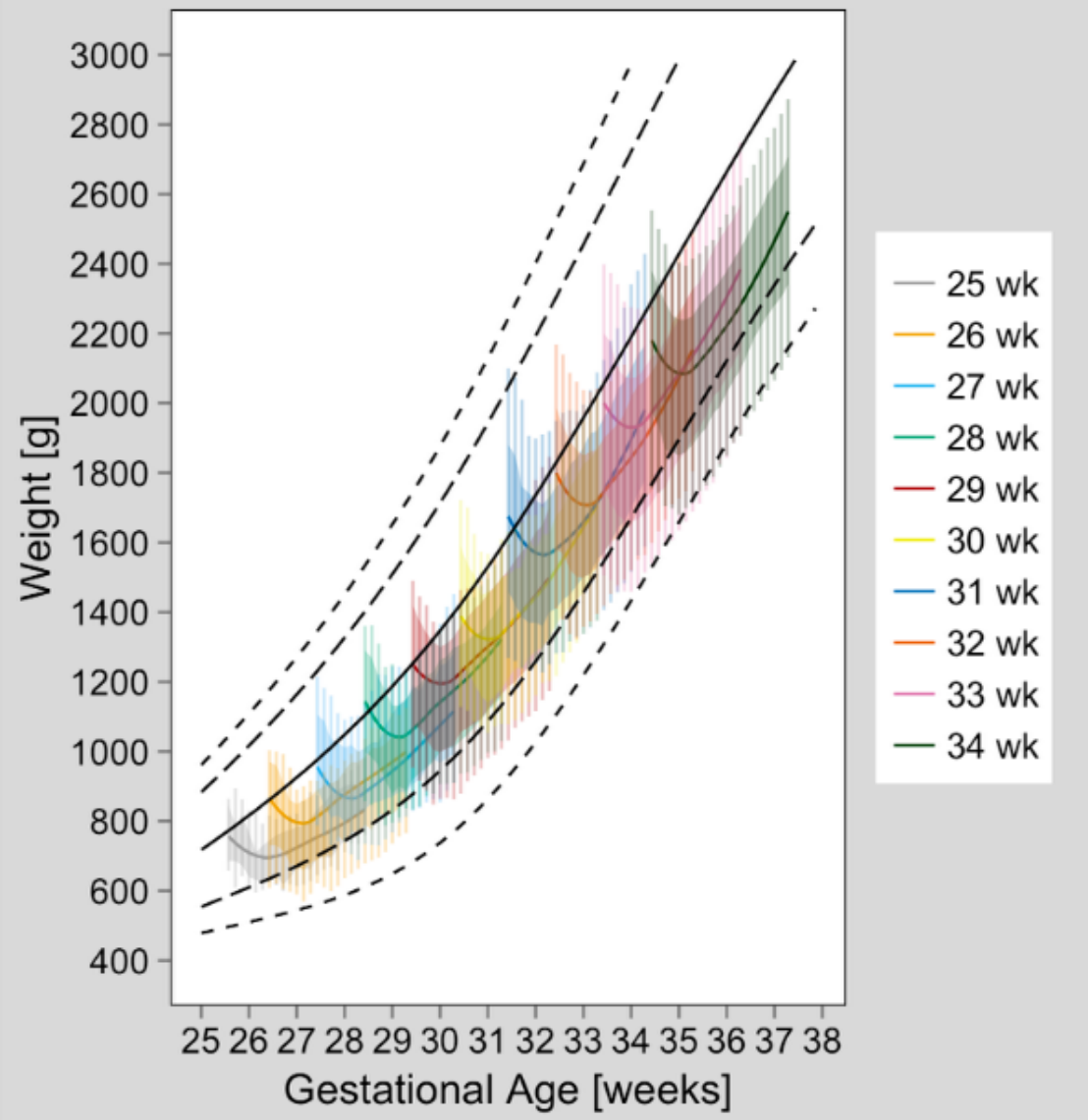

Figure 6.5: Median growth curves

Chapter 7

Discussion and Further Work

In many studies in health science, missing data are inevitable and should be treated carefully when conducting data analysis since improper treatment of missing data would yield misleading results. In fact, the challenge of statistical analysis in applied fields is not the data analysis but the treatment of missing data. Much effort is spent in dealing with missing data.

I have described three missingness mechanisms, namely, MCAR, MAR and NMAR, in terms of the probability distribution of a binary indicator variable used to indicate the missingness of data. The missingness mechanisms are important when we analyze an incomplete data set. Most statistical softwares deal with incomplete data by deleting all the cases having missing values or by deleting cases having missing values in variables included in the analysis. Missing data are ignorable when they are MCAR so that general statistical analysis can be applied to the incomplete data set without any concerns. Deletion methods will decrease the statistical power of the procedure by decreasing the sample size. In practice, most incomplete data are either

MAR or NMAR. These data can not be analyzed directly since the incompleteness can produce biased estimates of parameters of interest and can lead to invalid or misleading inference. Imputation can avoid these problems to some extent.

Several widely used imputation procedures for imputing missing data are described and illustrated. Each method has its own merits and demerits. In general, LOCF replacing missing data with the last observed value for any subject is commonly used in longitudinal data. It yields conservative estimates of parameters of interest. It preserves independence of subjects by dealing with missing data on an individual-by-individual basis. In contrast, regression-based imputations are more reliable when imputation models are properly specified. They are flexible in handling each response variable with missing values by specifying one regression model for the variable. Here, only imputation for continuous variables are discussed. Other type of variables can also be easily imputed by regression-based imputations using generalized linear models. For example, missing data in a binary variable can be imputed by using logistic models. The flexibility of regression-based imputations is also displayed by covariates selection. For example, all the main effect terms and effective interactions among main factors can be included in the imputation models. However, more restrictions are required to preserve the nature of the growth of weights in our study. In other words, the imputed values of missing data should be bounded between previous and latter observed values to preserve the monotonicity in growth.

Single imputations considering imputed data as real data do not account for uncertainty in imputed data. Multiple imputation overcomes this shortcoming of single

imputation by providing multiple plausible values for each missing value. Each completed data is then analyzed separately and results from each data set are then averaged to draw final conclusions. Multiple imputation is computationally a lot more intensive than single imputation as it requires repeat statistical analysis. However, completed data sets resulting from multiple imputation are not necessarily better than single imputation. The reason for this is that multiple imputation is a much more complicated procedure. First of all, the initial values of the missing data, randomly selected from the observed data, can make the imputed values be considerably away from the real values. Constraints for the initial selection of imputed values may be considered when applying multiple imputation. That is, in longitudinal data, the initial values of missing response variables for each infant can be randomly selected from its observed measurements. Second, constraint should also be placed during iteration to follow the monotone growth among observed and imputed weight for each infant.

Once the data are suitably completed, the statistical analysis can be readily carried out. Linear models that assume the repeated responses for each infant to be independent or dependent are used to analyze the GTHPI data. It is observed that the gestational age, birth weight, centre and gender are significantly related to the later-on weight of preterm infants. The AIC or QIC suggests that the simpler models with gestational age and birth weight as covariates are quite adequate for the prediction of weights. However, the coefficients from the two models are quite different.

It can be seen from our analysis that the imputation models and models for analysis

need not necessarily be the same since the primary aim of imputation is only to fill in the missing data with plausible values as close as possible to the real data and not for the ensuing statistical analysis.

Chapter 8

Appendix

```
# calculate quantiles of preterm infants for each day
#  in different gestational age

cal<-function(dat2)
{
  require("plyr")
  t<-ddply(dat2,.(gaw,day),summarize,
           mean=mean(weight),
           sd=sd(weight),
           median=quantile(weight,prob=0.5),
           Q1=quantile(weight,prob=0.25),
           Q3=quantile(weight,prob=0.75),
           one=quantile(weight,prob=0.01),
           ninenine=quantile(weight,prob=0.99),
           tenth=quantile(weight,prob=0.1),
```

```
          nintyth=quantile(weight,prob=0.9))
  t$start<-NA
  start<-ddply(dat2,.(gaw),summarize,gawexact.median=quantile(gawexact,prob=0.5))
  for (i in 1:nrow(t))
    for (j in 1:nrow(start))
    {
      if (t$gaw[i]==start$gaw[j])
      {t$start[i]=start$gawexact.median[j]}
    }
  t$x<-t$start+(as.numeric(t$day)-1)/7
  return(t)
}

# Indentify last observed value
ob_last<-function(dat,i,j)
{
  m<-j-1
  while (m>=1)
  {
    if (!is.na(dat[i,m])) {return(m)}
    else {m<-m-1}
  }
}
# LOCF Imputation
```

```
LOCF<-function(dat)
{source("ob_last.R")
  for (i in 1:nrow(dat))
  {m<-NULL; n<-NULL;
   for (j in 15:ncol(dat))
   {
     if (is.na(dat[i,j]))
     {m<-ob_last(dat,i,j);dat[i,j]<-dat[i,m]}
   }
  }
  return(dat)
}

# Regression Imputation
# W_t = W_1 + ... + W_{t-1} + GA + BW
RegImp_2<-function(dat2)
{
  for (j in 16:35)
  {
    dat2.lm<-cbind(dat2[,c("gaw","bw")],dat2[,15:(j-1)],y=dat2[,j])
    m2<-lm(y~.,!is.na(y),data=dat2.lm)
    summary(m2)
    #png(file=paste("RI2_day", j-14, ".png", sep = ""),
    #    width=10, height=6, units='in',res=300)
```

```
    #par(mfrow = c(3, 1), oma = c(0, 0, 2, 0))
    #plot(m2,which=c(1,2,4))
    #dev.off()
    dat2[is.na(dat2[,j]),j]<-predict(m2,newdata=dat2.lm[is.na(dat2[,j]),])
  }
  return(dat2)
}

# Stochastic Imputation
# # W_t = W_1 + ... + W_{t-1} + GA + BW + epsilon
StoImp_2<-function(dat)
{
  set.seed(2000)
  for (j in 16:35)
  {
    dat.lm<-cbind(dat[,c("gaw","bw")],dat[,15:(j-1)],y=dat[,j])
    m<-lm(y~.,data=dat.lm)
    sigma<-summary(m)$sigma
    summary(m)
    dat[is.na(dat[,j]),j]<-predict(m,newdata=dat.lm[is.na(dat[,j]),])+
    rnorm(sum(is.na(dat[,j])),0,sigma)
  }
  return(dat)
}
```

```
# Multiple Imputation
require(ggplot2)
data<-read.csv("data.csv",head=T)
#data<-data[,c(5:8,15:35)]
#data$gaw<-as.factor(data$gaw)
head(data)
# Multiple imputation
install.packages("mi")
library("mi")
info<-mi.info(data)
info
info<-update(info,"include",
list("id"=F,"yoa"=F,"defstart"=F,"dfef"=F,
"dtpnstop"=F,"md"=F, "eth"=F,"preg"=F,"centre"=F))
#mp.plot(data,gray.scale=T)
#mp.plot(data,y.order=T,gray.scale=T)

imp<-mi(data,info,n.imp=3,n.iter=10,max.minutes=10,seed=1250)

png(file="multipleImputation"),width=10, height=6, units='in',res=300)
plot(imp)
dev.off()
```

```
mi.scatterplot()
d1<-mi.data.frame(imp,m=1)
d2<-mi.data.frame(imp,m=2)
d3<-mi.data.frame(imp,m=3)
write.csv(d3,"MI3.csv",row.names=F)

# Cross-Validation

dat<-read.csv("data.csv",head=T)
n<-round(0.1*nrow(dat),0)
indix<-data.frame(matrix(nrow=n,ncol=20))
set.seed(1000)
for (j in 16:35)
{indix[,j-15]<-sample(which(!is.na(dat[,j])),n,replace=F)}
artificial<-function(d,indix)
{
  for (j in 16:35)
  {d[indix[,j-15],j]<-NA}
  return(d)
}
sse<-function(old,new,indix)
{
  sse<-0
  for (j in 16:35)
```

```
  {temp<-sum(old[indix[,j-15],j]-new[indix[,j-15],j])^2
   sse<-sse+temp}
  return(sse)
}

source("RegImp_2.R")
dat2<-RegImp_2(dat)
source("artificial.R")
temp2<-artificial(dat2,indix)
DAT2<-RegImp_2(temp2)
source("sse.R")
REG2<-sse(dat2,DAT2,indix)


source("StoImp_2.R")
sto2<-StoImp_2(dat)
TEMP2<-artificial(sto2,indix)
STO2<-StoImp_2(TEMP2)
STOsse2<-sse(sto2,STO2,indix)


source("LOCF.R")
locf<-LOCF(dat)
TEMPlocf<-artificial(locf,indix)
```

```
LOCF<-LOCF(TEMPlocf)
SSElocf<-sse(locf,LOCF,indix)

SSEmi<-function(data,complete,indix)
{
library("mi")
info<-mi.info(data)
imp<-mi(data,info,n.imp=3,n.iter=10,max.minutes=10,seed=1250)
d1<-mi.data.frame(imp,m=1)
d2<-mi.data.frame(imp,m=2)
d3<-mi.data.frame(imp,m=3)
sse1<-sse(complete,d1,indix)
sse2<-sse(complete,d2,indix)
sse3<-sse(complete,d3,indix)
sum<-sse1+sse2+sse3
return(sum)
}

D1<-artificial(d1,indix)
source("SSEmi.R")
SSEmi1<-SSEmi(D1,d1,indix)
SSEmi1
# 31751160
D2<-artificial(d2,indix)
```

```
SSEmi2<-SSEmi(D2,d2,indix);SSEmi2 #25117922
D3<-artificial(d3,indix)
SSEmi3<-SSEmi(D3,d3,indix);SSEmi3 # 24387176

(SSEmi1+SSEmi2+SSEmi3)/9

write.csv(d3,"MI3,csv",row.names=F)



# Linear Models
dat<-read.csv("data_complete_final.csv",head=T) # weights

M.full<-lm(X7~bw+gaw+gender+centre,data=dat)
summary(M.full)
M.rd<-lm(X7~bw+gaw+gender,data=dat)
summary(M.rd)
M<-lm(X7~bw+gaw,data=dat)
summary(M)

# Multivariate Linear Models / GEE

dat<-read.csv("data_complete_final.csv",head=T)
require(reshape)
names(dat)[15:35]<-1:21
```

```
d<-melt(dat,id=names(dat)[1:14])
names(d)[15:16]<-c("day","weight")
install.packages("geepack")
require(geepack)
install.packages("MuMIn")
require(MuMIn)
m1<-geeglm(weight~centre+gender+gaw+bw+day,data=d,family=gaussian("identity"),
id=id,corstr="ar1")
summary(m1)
m2<-geeglm(weight~gender+gaw+bw+day,
data=d,family=gaussian("identity"),id=id,corstr="ar1")
summary(m2)
m3<-geeglm(weight~gaw+bw+day,data=d,
family=gaussian("log"),id=id,corstr="ar1")
summary(m3)
c(QIC(m1),QIC(m2),QIC(m3))
```

Bibliography

Barker, D. J., Godfrey, K. M., Gluckman, P. D., Harding, J. E., Owens, J. A., and Robinson, J. S. (1993). Fetal nutrition and cardiovascular disease in adult life. *The Lancet*, **341**(8850), 938–941.

Cook, R. D. (2000). Detection of influential observation in linear regression. *Technometrics*, **42**(1), 65–68.

Fenton, T. R. and Kim, J. H. (2013). A systematic review and meta-analysis to revise the fenton growth chart for preterm infants. *BMC pediatrics*, **13**(1), 59.

Fisch, C., Poeschl, J., Heckmann, M., Campbell, D., Seigel, S., Jahn, A., Goettler, S. A., Rochow, N., Raja, P., Balakrishnan, N., and Liu, K. (2014). Growth trajectories of selected healthy preterm infants. Manuascript.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2012). *Applied Longitudinal Analysis*. John Wiley & Sons, New York.

Gelman, A., Goegebeur, Y., Tuerlinckx, F., and Van Mechelen, I. (2000). Diagnostic checks for discrete data regression models using posterior predictive simulations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **49**(2), 247–268.

Ho, M. Y., Yen, Y. H., Hsieh, M. C., Chen, H. Y., Chien, S. C., and Hus Lee, S. M. (2003). Early versus late nutrition support in premature neonates with respiratory distress syndrome. *Nutrition*, **19**(3), 257–260.

Iams, J. D., Romero, R., Culhane, J. F., and Goldenberg, R. L. (2008). Primary, secondary, and tertiary interventions to reduce the morbidity and mortality of preterm birth. *The Lancet*, **371**(9607), 164–175.

Jain, V. and Singhal, A. (2012). Catch up growth in low birth weight infants: striking a healthy balance. *Reviews in Endocrine and Metabolic Disorders*, **13**(2), 141–147.

Larroque, B., Ancel, P. Y., Marret, S., Marchand, L., André, M., Arnaud, C., Pierrat, V., Rozé, J. C., Messer, J., Thiriez, G., *et al.* (2008). Neurodevelopmental disabilities and special care of 5-year-old children born before 33 weeks of gestation (the epipage study): a longitudinal cohort study. *The Lancet*, **371**(9615), 813–820.

Latal Hajnal, B., von Siebenthal, K., Kovari, H., Bucher, H. U., and Largo, R. H. (2003). Postnatal growth in vlbw infants: significant association with neurodevelopmental outcome. *The Journal of Pediatrics*, **143**(2), 163–170.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**(1), 13–22.

Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons, New York.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics*, **57**(1), 120–125.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, **27**(1), 85–96.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, **63**(3), 581–592.

Rubin, D. B. (2004). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, Hoboken, New Jersey.

Saigal, S. and Doyle, L. W. (2008). An overview of mortality and sequelae of preterm birth from infancy to adulthood. *The Lancet*, **371**(9608), 261–269.

Schafer, J. L. and Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, **7**(2), 147.

Steward, D. K. (2012). Growth outcomes of preterm infants in the neonatal intensive care unit: Long-term considerations. *Newborn and Infant Nursing Reviews*, **12**(4), 214–220.

Su, Y. S., Yajima, M., Gelman, A. E., and Hill, J. (2011). Multiple imputation with diagnostics (mi) in r: Opening windows into the black box. *Journal of Statistical Software*, **45**(2), 1–31.

Van Buuren, S. and Oudshoorn, C. (2000). *Multivariate Imputation by Chained Equations: MICE V1. 0 Users's Manual*. TNO Prevention and Health, Public Health.

WHO (2014). Preterm birth. `http://www.who.int/mediacentre/factsheets/fs363/en/`. Accessed November, 2014.