1992

MISSING DATA IN REPEATED MEASUREMENT STUDIES

Kejian Niu

Virginia Commonwealth University
School of Basic Health Sciences

This is to certify that the thesis prepared by Kejian Niu entitled "Missing Data In Repeated Measurement Studies" has been approved by his committee as satisfactory completion of the thesis requirement for the degree of Master of Science.

[redacted]

Pippa M. Simpson, Director of Thesis

[redacted]

Sung C. Choi, School of Basic Health Sciences

[redacted]

Linda A. Corey, School of Basic Health Sciences

[redacted]

Teresa P. Germanson, University of Virginia

[redacted]

Walter H. Carter, Department Chair

[redacted]

S. Gaylen Bradley, Dean, School of Basic Health Sciences
Chairman, MCV Graduate Committee

10 December 1992

Date

MISSING DATA IN REPEATED MEASUREMENT STUDIES

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at the Medical College of Virginia, Virginia Commonwealth University

BY

Kejian Niu
B.S., The University of Science and Technology of China, 1986
M.S., Wuhan Institute of Physics, Science Academy of China, 1989

Director: Pippa M. Simpson, Ph. D.
Assistant Professor, Department of Biostatistics

Virginia Commonwealth University
Richmond, Virginia

December, 1992

Acknowledgement

I would like to express my appreciation to the members of my graduate committee, Dr. Pippa M. Simpson, Dr. Sung C. Choi, Dr. Teresa P. Germanson, and Dr. Linda A. Corey for their time and efforts, especially to my advisor, Dr. Pippa M. Simpson for all the help and encouragement she gave me during the course of my studies at MCV.

I would also thank Laura L. Truskowski for providing me with the data analyzed in this thesis, Jennifer L. Hodge for her help with proofreading.

A very special thanks goes to my wife, Bei, without her love and support the completion of this thesis would have been difficult, and to my parents, Wenlan Xu and Geng Niu, although I have not seen them since I came to the United States I can always feel their love, understanding and expectation.

TABLE OF CONTENTS

Chapter 1

Missing Data in Repeated Measurement

1.1 The Repeated Measurement Study

Repeated measurement data or longitudinal data occur often in statistical applications. For example, in a clinical trial comparing the efficacy of a new treatment with that of a standard treatment, rather than measuring the main response variable only once on each patient, or subject, we can take several measurements over time on each subject.

A Repeated measurement study differs from a longitudinal study. The latter generally refers to any study in which one or more response variables are repeatedly measured over time. The former usually imposes some restrictions on the data. One common restriction is that each response variable must be measured at the same time points.

In this thesis, the discussion will be restricted to a repeated measurement study, which is defined as follows: a repeated measurement study is a study in which a univariate response variable is repeatedly measured at the same time points on each subject. It should be pointed out, however, that many of the methods discussed here can also be applied to more general longitudinal studies.

The analysis of repeated measurement data involves two major difficulties. The first problem is the dependence among successive observations made on the same subject. Multivariate methods modeling the joint distribution of the repeated measures over time have been developed to solve this difficulty. The other, probably the more severe problem is missing data. In repeated measurement studies, the data are collected over a period of time, which in some studies could be many years. Therefore, complete control over the circumstances under which measurements are obtained is not possible. The occurrence of missing data is more likely in repeated than in non-repeated measure studies, and is sometimes unavoidable.

In recent years, many methods for coping with the missing data problem in repeated measurement studies have emerged from various applications. The purpose of this thesis is to review and summarize these methods, apply some of them to a practical problem, and identify the needs of further research.

1.2 The Data Structure for Repeated Measurements

For a repeated measurement study, suppose there are T pre-specified time points. Each subject is repeatedly measured at the same T time points. If the design is balanced and there is no missing data, each subject will have the same number of observations.

Let the responses on the ith subject be a T x 1 vector y_i, where $i = 1, \ldots, n$. Let $\mathbf{X}_i$ be a T x p design matrix for the ith subject, and it could be defined by both the subject covariates and the within-subject-over-time relationships. Let β be a vector of unknown parameters, and let $f(y_i|\mathbf{X}_i, \beta)$ be the multivariate density of y_i conditional on $\mathbf{X}_i$ and β. Usually inferences about the unknown parameter vector β, or part of its components are of most interest statistically.

In the following sections in this chapter, the index i for the ith subject is dropped for convenience of notation.

1.3 Models for Non-response

A model for missing data, or non-response, is introduced here in order to discuss the missing data problem. For repeated measurements with the data structure introduced in the last section, let R be a T x 1 vector of indicator variables for the ith subject, where:

$R_{it} = 1$ if y_{it} is observed

$R_{it} = 0$ if y_{it} is missing, or having non-response.

Let **Z** be a T x q matrix of covariates determining the non-response process, and ϕ the vector of parameters of the non-response model. The multivariate density of R given y, **X**, **Z** and ϕ can be expressed as $f(R|y,\boldsymbol{X},\boldsymbol{Z},\phi)$.

When the non-response indicator R is given, the response vector y can be partitioned into $y^T=(y^T{}_o,y^T{}_m)$, where y_o is an M_i vector of observed components of y, and $M_i < T$, and y_m is a T - M_i vector of missing observations. M_i varies with each subject according to R.

Only y_o and R can actually be observed. The density of these observed data can be written as:

$$f(y_o,R|\boldsymbol{X},\boldsymbol{Z},\beta,\phi) = \int f(y|\boldsymbol{X},\beta)\, f(R|y,\boldsymbol{X},\boldsymbol{Z},\phi)\, dy_m \qquad (1.1)$$

where the integration is over the sample space of y_m.

1.4 Missing Data Hierarchy

When analyzing data with missing observations it is very important to differentiate among different types of missing data mechanisms. Using the data structure and the model for non-response introduced in the previous two sections, the missing data hierarchy, originated by Rubin (1977), and further discussed by Little and Rubin (1987) and Laird (1988), can be introduced.

Data is said to be missing completely at random (MCAR) if the non-response mechanism depends on neither the responses y, nor the design matrix **X**, i.e.

$$f(R|y,\mathbf{Z},\mathbf{X},\phi) = f(R|\mathbf{Z},\phi) \tag{1.2}$$

When the probability of non-response depends on the design matrix X, but not on the responses y, the data is defined to be missing at random (MAR):

$$f(R|y,\mathbf{Z},\mathbf{X},\phi) = f(R|\mathbf{Z},\mathbf{X},\phi) \tag{1.3}$$

The non-response mechanism is said to be *ignorable* if the probability of R depends on the observed responses y_o, but not on the missing responses y_m, i.e.,

$$f(R|y,\mathbf{Z},\mathbf{X},\phi) = f(R|y_o,\mathbf{Z},\mathbf{X},\phi) \tag{1.4}$$

This type of non-response is called *ignorable* only if the likelihood based approaches is employed to draw inferences. When methods other than likelihood based approaches are used, this type of non-response mechanism can not be ignored. This difference will be explained in the following section.

Finally, if the probability of R depends on the missing values of y, y_m, the non-response mechanism is called non-ignorable, or missing not at random.

1.5 Approaches Dealing with Repeated Measures with Missing Data

Several approaches are available to deal with different types of missing data mechanisms.

When the missing data mechanism is MCAR, most standard analyses will be valid if the so-called 'Complete Case' analysis is used. Using this analysis any subject with missing data is deleted, and inferences can be drawn from the cases with complete responses left. No bias to inferences will be introduced because the non-response mechanism is not related to the responses. However, the resulting data set may be too small for useful analysis.

In the case of MAR, since the distribution of R takes the form of (1.3), then y_o and R can be shown to be independent. This relation can be illustrated by rewriting equation (1.1) as:

$$f(y, R|\boldsymbol{X}, \boldsymbol{Z}, \beta, \phi) = f(R|\boldsymbol{X}, \boldsymbol{Z}, \phi) \int f(y|\boldsymbol{X}, \beta)\, dy_m \tag{1.5}$$

Since the distribution of R does not depend on y, it can be factored out of the integration. Consequently, the marginal density of y_o is simply the usual marginal density

$$f(y_o|\mathbf{X},\beta) = \int f(y|\mathbf{X},\beta)\,dy_m \tag{1.6}$$

So any method drawing inferences based upon the distributional properties of the marginal distribution of the observed data is valid in dealing with this type of non-response mechanism. In recent years, many semi-parametric and non-parametric methods which can allow for the missing data with missing data mechanism being MAR were developed in repeated measurement analyses (Liang and Zeger, 1986; Stram, Wei and Ware, 1987). This type of method will be discussed further in Chapter 3.

As mentioned before, the non-response mechanism is *ignorable* only in the sense that the inferences are drawn from likelihood based methods which are valid without the need of specifying a non-response model. To illustrate we can write $f(y|\mathbf{X},\beta)$ in the following form:

$$f(y|\mathbf{X},\beta) = f(y_m|y_o,\mathbf{X},\beta)\,f(y_o|\mathbf{X},\beta) \tag{1.7}$$

where $f(y_o|\mathbf{X},\beta)$ is given by (1.6). The distribution of R is shown in (1.4) since the non-response mechanism is *ignorable*. By substituting (1.4) and (1.5) in (1.7), it can be shown that the density of the observed data is:

$$f(y_o,R|\mathbf{X},\mathbf{Z},\beta,\phi) = f(y_o|\mathbf{X},\beta)\,f(R|y_o,\mathbf{X},\mathbf{Z},\phi) \tag{1.8}$$

If the only interest is in inferences about β, then the contribution to the likelihood function of β can be taken as $f(y_o|\mathbf{X},\beta)$. $f(R|y_o,\mathbf{X},\mathbf{Z},\phi)$, in whatever form it takes, can be ignored. The likelihood based approaches, which can be used to make valid inferences about β for *ignorable* missing data, have also been developed since the early 80's (Laird and Ware, 1983; Ware, 1985; Jennrich and Schluchter, 1986; Rochen and Helms, 1988). In Chapter 3, the likelihood based approaches will be discussed in detail, and Chapter 5 will present an example of the application of one of the likelihood based method introduced in Chapter 3.

Ignorable missing data is a broader mechanism than MAR or MCAR, so the methods allowing for *ignorable* non-response mechanism are also valid for data with MAR or MCAR. Since MAR contains MCAR, the methods introduced in Chapter 2 are also valid for data with MCAR. The reverse is not necessarily true. Consider the *ignorable* non-response mechanism as an example. From (1.8) it can be seen that y_o and R are not independent, so the marginal density of y_o is no longer $f(y_o|\boldsymbol{X}, \beta)$, but rather depends upon the non-response model. The methods for MAR non-response mechanism in Chapter 3 are, therefore, not valid for ignorable missing data. If the missing mechanism is non-ignorable, however, the analysis of the data becomes much more complicated. One way to deal with this problem is to model the missing data mechanism directly in the likelihood function to obtain maximum likelihood estimators of the parameters. This idea was used to analyze data with one of the most common non-randomly missing data mechanisms, the informative right censoring. Another way to tackle this problem is to develop some type of protective statistics which are not sensitive to a certain class of non-randomly missing data mechanism. Details of these methods for dealing with non-randomly missing data mechanism will be provided in Chapter 4.

Chapter 2

Likelihood Based Methods

2.1 Introduction

One of the systematic approaches for the missing data problem in repeated measurements is the so-called likelihood based approach.

This method specifies a model to describe the data and estimates the parameters of the model using a maximum likelihood or restricted maximum likelihood approach. The model here allows us to use arbitrary linear models to describe the mean structure and model various types of covariance structure. The likelihood based methods can deal with incomplete data resulting from an ignorable missing mechanism.

Laird and Ware (1982) studied the maximum likelihood estimation procedure for general random-effect models with incomplete data. Ware (1985) discussed maximum likelihood estimation for a similar model with three types of covariance structure: random effects, multivariate, and autoregressive time series. Szatrowski (1983) considered models for incomplete data with a linear covariance structure. Laird, Lange and Stram dealt with the random effect model and an arbitrary (fully parameterized) structure of variance. Rochon and Helms (1988) discussed maximum likelihood estimation for an incomplete repeated measurement model under an ARMA covariance structure. Jennrich and

Schluchter (1986) and Schluchter (1988) gave a general approach of maximum likelihood estimation allowing the variance to take any form of the structures mentioned above except ARMA, plus factor analytic structure and stationary time series structures.

This chapter will discuss the general model introduced by Jennrich and Schluchter (1986). The basic model is introduced in Section 2. The EM algorithm for maximum likelihood estimation in repeated measurements is discussed in Section 3. The various variance structures which can be used in the incomplete data model are discussed in Section 4. Section 5 follows with a brief discussion of the testing of hypotheses of general interests.

2.2 The Model

Let y_i be a T x 1 vector containing the response for the ith subject, where i = 1, ..., n. The general model assumed for y_i can be written as:

$$y_i = \boldsymbol{X_i}\beta + e_i \qquad (2.1)$$

where X_i is a T x p design matrix, β is a p x 1 vector of unknown regression parameters, and the e_i are assumed to be independently distributed as $N(0, \Sigma_i)$. It is further assumed that the elements of each covariance matrix, Σ_i, are known functions of a set of q unknown parameters contained in a vector θ, that is $\Sigma_i=\Sigma_i(\theta)$ for i = 1, ..., n. The regression parameters β are assumed to be independent of the covariance parameters θ.

Since the parameters β and θ are independent, model (2.1) can be viewed as two separate models: the model for the expected values which includes the regression parameters β and the model for the covariance structure which involves the covariance parameters θ.

The expected values of responses can be modelled by the regression format in (2.1). The most used models could be either the ANOVA models or the Linear Growth Curve models. The ANOVA models take the usual form of ANOVA:

$$y_{it} = \mu+\alpha_{g_i}+\tau_t+(\alpha\tau)_{g_it}+e_{it} \tag{2.2}$$

where y_{it} is the response variable for the ith subject at time point t, i=1,2,...,n, t=1,2,...,T, and g_i indexes the group to which subject i belongs, with the constraints $\sum_g \alpha_g=\sum_t \tau_t=\sum_g (\alpha\tau)_{gt}=\sum_t (\alpha\tau)_{gt}=0$. The Linear Growth Curve models are in the following form:

$$y_{it} = a_i+b_ix_t+e_{it} \tag{2.3}$$

where a_i and b_i are unknown parameters with fixed underlying values, and x_t indexes the independent variable for the ith subject at time t.

The covariance model can take various forms. Details about different types of structures for $\Sigma(\theta)$ that can be used in incomplete data models are discussed in section 4.

2.3 EM Algorithm

To obtain the values of maximum likelihood estimates of the parameters, the most often used algorithms are the Newton-Raphson algorithm and the Fisher scoring algorithm. However, to get maximum likelihood estimates from incomplete data, the EM algorithm, which was first introduced by Dempster, Laird and Rubin (1977), has to be employed.

The EM algorithm is for the two steps in each iteration, the E step and the M step. The E step, or the expectation step, finds the expectation of some functions of the missing data given the observed data and current estimated parameters, and then substitutes these expectations for some functions of missing data, which appear in the complete data log-likelihood . The M step, or maximization step, simply perform maximum likelihood estimation of the parameters just as if there were no missing data, i.e. as if the functions of missing data, required in the complete maximum likelihood, had been filled. For details of the properties of the EM algorithm, see Dempster, Laird and Rubin (1977) and Little and Rubin (1987).

The Hybrid EM Scoring algorithm of maximum likelihood introduced by Jennrich and Schluchter (1986) will be briefly discussed below. See also Jennrich and Schluchter (1986) for the discussion of a Generalized EM algorithm for the restricted maximum likelihood.

2.3.1 Hybrid EM Scoring Algorithm

Using the model described in the previous section, the log-likelihood λ for y_1, ..., y_n is

$$\lambda = Const. - \frac{1}{2}\sum_{1}^{n} \log|\Sigma_i| - \frac{1}{2}\sum_{1}^{n} (y_i - \boldsymbol{X}_i\beta)^T \Sigma_i^{-1} (y_i - \boldsymbol{X}_i\beta) \qquad (2.4)$$

The maximum likelihood estimations of β and θ are obtained by maximizing (2.4).

For each iteration of the Hybrid EM Scoring algorithm, the first step, which is equivalent to the M step in an ordinary EM algorithm, uses general least squares to update β. The second part includes three steps which are equivalent to one E step in the generalized EM (GEM) algorithm to update θ.

Let $e_i^* \sim N(0, \Sigma_i)$ be a vector of complete data, and $e_i = y_i - X_i\beta$ be a sub-vector of observed data for the ith subject. The steps of the algorithm are as follows:

(1) Compute updated estimates $\tilde{\beta}$ of β:

$$\tilde{\beta} = (\sum_{i=1}^{n} X_i'\Sigma_i^{-1}X_i)^{-1}(\sum_{i=1}^{n} X_i'\Sigma_i^{-1}y_i) \tag{2.5}$$

(2) Compute $\hat{e}_i^*$ and A_i, while taking updated estimates $\tilde{\beta}$ as β and current estimates of the parameters of the θ:

$$\hat{e}_i^* = E(e_i^*|e_i) \tag{2.6}$$

and

$$A_i = cov(e_i^*|e_i) \tag{2.7}$$

(3) Using $\hat{e}_i^*$ and A_i, compute the matrix

$$S = \frac{1}{n}\sum_{i=1}^{n} (\hat{e}_i^*\hat{e}_i^{*\prime} + A_i) \tag{2.8}$$

(4) Compute the updated θ. When Σ is an unstructured dispersion matrix this is simply:

$$\hat{\Sigma} = S \tag{3.9}$$

When $\Sigma = \Sigma(\theta)$ is structured, a 'Scoring step' is used to obtain $\Delta\theta$:

$$\Delta\theta = I^{-1}s \tag{2.10}$$

where I and s are matrices with general element of the following form:

$$[s]_l = \frac{1}{2} tr(\Sigma^{-1}(S-\Sigma)\Sigma^{-1}\dot{\Sigma}_l) \tag{2.11}$$

$$[I]_{lm} = \frac{1}{2} tr(\Sigma^{-1}\dot{\Sigma}_l\Sigma^{-1}\dot{\Sigma}_m) \tag{2.12}$$

and $\dot{\Sigma}_l = \partial\Sigma/\partial\theta_l$, and then update θ as:

$$\hat{\theta} = \theta+\Delta\theta \tag{2.13}$$

2.3.2 Standard Error of the Parameter Estimates

For the parameter estimates $\hat{\beta}$ and $\hat{\theta}$, the standard error estimates can be obtained from the maximum likelihood algorithm. They can be computed either from the inverse of the Fisher information matrix or from the inverse of the empirical information matrix (Jennrich and Schluchter, 1986). However, if the EM algorithm is used, the information matrix of θ is not computed. Hence the standard error estimates for θ are not available. Since the usual point of interest is in inference about β, the lack of standard error estimates of θ often does not matter, and under this situation, the standard error estimates of β, whether computed from the Fisher information matrix or empirical information matrix, has the same form:

$$(\sum_{i=1}^{n} \boldsymbol{X}_i' \Sigma_i^{-1} \boldsymbol{X}_i)^{-1} \qquad (2.14)$$

2.4 Models for the Variance Structure

As stated earlier, the likelihood based methods have the flexibility to model a variety of covariance structures. This flexibility allows numerous choices for appropriate covariance structure based on our understanding of the true underlying physical or biological process. Even when the main interest is in the parameter β, a proper choice of covariance structure can greatly improve the efficiency of the algorithm. In this section, several common structures of the covariance for the incomplete data will be discussed.

2.4.1 The Incomplete Data Model

The model introduced in Section 2.2 is the general model for complete repeated measurement data. For incomplete data, a similar model can be introduced as long as the observed measurements for each subject can be considered as a subset of a larger number of T measurements.

Let y_i be a M_i x 1 vector of observed data for the ith subject, and let it be thought of as a subvector of a T x 1 complete data vector y_i^*. The observed covariance matrix Σ_i then can be considered as the appropriate submatrix of a larger T x T matrix $\Sigma=\Sigma(\theta)$. This is the incomplete data model, and the EM algorithm discussed in section 2.3 can be used to obtain the parameter estimates for the model, as long as the mechanism of the missing data is *ignorable*, or MAR, or MCAR as discussed in Chapter 1.

Among the various covariance structures, several important types can be used in the incomplete data models and will be discussed here. Some other structures not described by the incomplete data model were discussed by Schluchter (1988).

2.4.2 Unstructured Model

The unstructured model, or fully parameterized model, has T(T+1)/2 parameters in θ. It does not impose any special feature further than the basic symmetry required for a covariance matrix. For example, if let T=4, the structure of the covariance is

$$\begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} & \sigma_{34} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_{44} \end{bmatrix} \tag{2.15}$$

and the vector $\theta=(\sigma_{11},\sigma_{12},\sigma_{13},\sigma_{14},\sigma_{22},\sigma_{23},\sigma_{24},\sigma_{33},\sigma_{34},\sigma_{44})$.

Since this model does not specify any particular structure, it can be chosen when little or no information about the structure of the covariance is available. Furthermore, the unstructured model is often used as a reference to evaluate the goodness of fit of other models of the covariance structures.

When T is very large and the data are highly incomplete, the unstructured model of Σ may cause inefficiency in estimating the parameters. In this case other models for the covariance structures may need to be considered.

2.4.3 Time Series Models

Time series models treat the terms in a subject's error vector as a short time series following a stationary autoregressive or moving-average process. Generally, this model requires that the T measurements in the complete data vector y_i^* be equally spaced across time.

The simplest of the time series models may be the first order autoregressive model, which is called AR(1). An example of AR(1) model for covariance when T=4 is

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix} \qquad (2.16)$$

Another simple time series model is the first order moving-average model, which is called MA(1). An example of MA(1) for covariance when T=4 is

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho & 0 & 0 \\ \rho & 1 & \rho & 0 \\ 0 & \rho & 1 & \rho \\ 0 & 0 & \rho & 1 \end{bmatrix} \qquad (2.17)$$

All the stationary time series models, including AR(1) and MA(1), are special cases of the general autoregressive structure, or banded structure, which has a separate parameter for each of the lag-correlations. An example of the general autoregressive structure when T=4 is

$$\Sigma = \begin{bmatrix} \theta_1 & \theta_2 & \theta_3 & \theta_4 \\ \theta_2 & \theta_1 & \theta_2 & \theta_3 \\ \theta_3 & \theta_2 & \theta_1 & \theta_2 \\ \theta_4 & \theta_3 & \theta_2 & \theta_1 \end{bmatrix} \qquad (2.18)$$

where the parameter vector $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$.

Intuitively, the times series models make sense when the correlation between repeated measurements is only functions of length of time between two measurements.

2.4.4 Random Effects Model

The random effect model is another general class of models that can be considered for the covariance structure. When the data are complete, the random effects model can be written as:

$$y_i^* = \boldsymbol{X}_i^* \beta + \boldsymbol{Z} b_i + u_i^* \qquad (2.19)$$

where y_i^* is the complete data vector, **Z** is a T x k known matrix, and b_i and u_i^* are independent random vectors with $b_i \sim N(0, \Phi)$, and $u_i^* \sim N(0, \sigma^2 \boldsymbol{I})$.

For incomplete data, the model for the observed data y_i is

$$y_i = \boldsymbol{X}_i \beta + \boldsymbol{Z}_i b_i + u_i \qquad (2.20)$$

where $\mathbf{Z}_i$, $\mathbf{X}_i$ and u_i contain, respectively, corresponding rows of $\mathbf{Z}$, $\mathbf{X}_i^*$ and u_i^* which are observed.

The above random effects models is indeed a special case of the incomplete data model where $\Sigma = \boldsymbol{Z} \Phi \boldsymbol{Z}' + \sigma^2 \boldsymbol{I}$.

A linear random effects growth curve model is introduced here as a simple example of the random effects model. Let T=4, x_t be the independent variable at time point t, and y_{it} and ξ_{it} be the response and error, respectively, of ith subject at time point t. The model is

$$y_{it} = a_i + b_i(x_t - \overline{x}) + \xi_{it} \qquad (2.21)$$

Where $\overline{x} = \sum_t x_t/4$, and ξ_{it} are independently distributed as $N(0, \sigma_e^2)$, $t=1,2,3,4$, a_i is the true underlying mean of y when $x_t = \overline{x}$ and b_i is the slope for the ith subject, which are assumed to follow a multivariate normal distribution with different means for the G group and common covariance matrix $\Phi = \{\phi_{ij}\}$:

$$\begin{bmatrix} a_{ig} \\ b_{ig} \end{bmatrix} \sim \left[\begin{bmatrix} \alpha_g \\ \beta_g \end{bmatrix}, \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix}\right] \qquad (2.22)$$

Obviously this random effects linear growth curve model is a special case of the general model (2.1) where

$$\Sigma = \boldsymbol{Z}\boldsymbol{\Phi}\boldsymbol{Z}' + \sigma_e^2 \boldsymbol{I} \qquad (2.23)$$

and the form of **Z** is

$$\boldsymbol{Z} = \begin{bmatrix} 1 & (x_1 - \overline{x}) \\ 1 & (x_2 - \overline{x}) \\ 1 & (x_3 - \overline{x}) \\ 1 & (x_4 - \overline{x}) \end{bmatrix} \qquad (2.24)$$

The simplest random effects model is the compound symmetry model, in which **Z** is a vector of 1's and $\Phi = \{\phi\}$. This model implies that measurements have a constant variance and a common correlation. An example of compound symmetry structure when T=4 is

2.5 Inferences about the Model Parameters

$$\Sigma = \begin{bmatrix} \sigma_e+\phi & \phi & \phi & \phi \\ \phi & \sigma_e+\phi & \phi & \phi \\ \phi & \phi & \sigma_e+\phi & \phi \\ \phi & \phi & \phi & \sigma_e+\phi \end{bmatrix} \qquad (2.25)$$

When data are complete, the Newton-Raphson algorithm or Fisher scoring algorithm can usually be used to get maximum likelihood estimates of β and θ, as well as the estimates of their standard errors, so some tests on these parameters are easily constructed.

For the incomplete data, however, the EM algorithm is used to obtain the maximum likelihood estimators of β and θ, but only the estimate of the standard error of β is available. Fortunately, though in many applications the only interest is on the regression parameters β, effective inferences such as hypothesis test about β are still possible to make.

2.5.1 General Form of Hypothesis Test

When the EM algorithm is used, let $\hat{\theta}$ be the maximum likelihood estimate of θ, hence $\hat{\Sigma}_i=\Sigma_i(\hat{\theta})$, and the estimate of β takes the following form:

$$\hat{\beta} = (\sum_{i=1}^{n} X_i'\hat{\Sigma}_i^{-1}X_i)^{-1}(\sum_{i=1}^{n} X_i'\hat{\Sigma}_i^{-1}y_i) \qquad (2.26)$$

and the estimated covariance matrix :

$$C = (\sum_{i=1}^{N} X_i'\hat{\Sigma}_i^{-1}X_i)^{-1} \qquad (2.27)$$

To test a hypothesis of the form:

$$H_0: \boldsymbol{L}\beta = q \tag{2.28}$$

where L is a k x p matrix of rank k, and q is a k x 1 vector. The Wald test statistics can be written as:

$$W = (\boldsymbol{L}\beta - q)'(\boldsymbol{LCL}')^{-1}(\boldsymbol{L}\beta - q) \tag{2.29}$$

Under H_0 (2.28), W has an asymptotic chi-square distribution with k degree of freedom.

The form of the estimated covariance matrix of β (2.27) is obtained from the expected information matrix. The validity of using it depends on the mechanism causing the missing data. If the missing data are caused by MAR or MCAR as defined in Chapter 1, then (2.27) is a consistent estimator of the standard error of β and can be used. If the missing data mechanism is *ignorable* but not MAR, i.e. the probability of missing depends in some wa upon the observed data, the covariance matrix obtained from (2.27) is not in general consistent. This is because the expectation step in obtaining the Fisher information is done under the assumption that the probability of missing depends neither on observed nor missing responses. See Jennrich and Schluchter (1986) for an alternative method for estimating **C.**

2.5.2 Hypothesis Tests of Usual Interests

The most often used model for the expected values of the parameters β, are probably those mentioned in Section 2.2: the ANOVA model (2.2), and the Linear Growing Curve model (2.3).

For the ANOVA model (2.2), the following three tests are usually of interest:

(1) H_0: No group effect, $\alpha_1 = . . . = \alpha_G = 0$.

(2) H_0: No time effect, $\tau_1 = \ldots = \tau_T = 0$.

(3) H_0: No Group X Time interaction, $(\alpha\tau)_{gt} = 0$ for all g, t.

Wald test statistics, labeled W_G, W_T and W_{GT}, respectively, can be constructed for the above three hypotheses. Asymptotically, they have (G-1), (T-1) and (G-1)(T-1) degree of freedom respectively, under the H_0.

For the Linear Growth Curve model (2.3), if the a_i and b_i in the model are considered as unknown fixed parameters, then the usual hypotheses of interest are

(1) H_0: The a_i are the same in all G groups, $a_1 = a_2 = \ldots = a_G$

(2) H_0: The average slope is 0, $(1/G)\sum b_g = 0$.

(3) H_0: The slope are the same in all G groups, $b_1 = b_2 = \ldots = b_3$.

And if the model is the Random Effects Linear Growth Curve model described in 2.4.4, and a_i and b_i are considered as random variables in (2.22), the above three hypotheses become

(1) H_0: $\alpha_1 = \alpha_2 = \ldots = \alpha_G$.

(2) H_0: $(1/G)\sum \beta_g = 0$.

(3) H_0: $\beta_1 = \beta_2 = \ldots = \beta_G$.

Asymptotic Wald tests can be constructed similarly for these hypotheses.

Notice that all the above Wald type tests are asymptotic tests which are only valid when the sample size n is large. The small sample adjustments to these tests are discussed by Schluchter and Elashoff (1990). Schluchter and Elashoff gave the corrections to each of the above tests and concluded that the most appropriate type of small-sample correction depends upon the form of the assumed covariance structure, which ML procedures are used, and whether the test is a 'between groups' or 'within subject' test.

Chapter 3

Semi-Parametric Methods

3.1 Introduction

The classic likelihood based methods discussed in the previous chapter use parametric models to analyze incomplete continuous repeated measurements data which can be assumed to have a multivariate normal distribution. However, in many applications such as biomedical applications, categorical data particularly binary or ordered categorical data or continuous but extremely non-normal data often occur in repeated measurements and are liable to have missing data. Likelihood based methods can not be applied to these types of data.

The earliest attempt to analyze categorical repeated measurements data was made by Koch et al (1977), using weighted least square methodology. Woolson and Clark (1984) extended this method to analyze incomplete categorical repeated measurement data. This method, however, requires that the covariates be categorical, and in addition, the sample size for the marginal responses at each time point with each category must be sufficiently large that the responses can be considered approximately multivariate normal. This qualification, of course, limits the applicability of this method.

Recently, a class of methods based on extension of the generalized linear model has been developed to analyze categorical repeated measurements data. This class of methods, named semi-parametric methods by Davis (1991), can be applied to repeated measurements with both continuous and categorical responses, as long as the quasi-likelihood formulation, such as those of Normal, Binomial, and Poisson response variables, is appropriate for the marginal distribution of the responses. Because this class of methods can allow either categorical or continuous, and even time-dependent covariates, it eliminates the limitations of the weighted least square method. More important, all these methods can incorporate missing data automatically, but they require the missing data mechanism to be MAR or MCAR. This is a stronger requirement than that of likelihood based methods, which allow for ignorable missing data mechanism.

The backbone of semi-parametric methods is the generalized estimating equations method (GEE) by Liang and Zeger (Liang and Zeger, 1986; Zeger and Liang, 1986). Other methods quite similar to GEE but extending it in some sense include Wei and Stram (1988), Moulton and Zeger (1989), Prentice (1988), Zhao and Prentice (1990), Liang, Zeger and Qaqish (1990) and Lipsitz (1991).

In this chapter, the GEE method will be introduced. In Section 3.2, the partial distribution model for GEE is introduced. In Section 3.3, the construction of GEE, the algorithms used to solve GEE, and the properties of its solutions will be discussed. Further discussions follows in Section 3.4.

3.2 The Model

The primary obstacle in the analysis of non-normal repeated measurements data is the lack of a class of models, such as multivariate normal model, for the joint distribution of the responses. The generalized estimating equation (GEE) introduced by Liang and Zeger (1986) is actually an extension of the generalized linear model estimating equation for multivariate responses. Instead of modeling the joint distribution of the responses, it models the marginal distribution of the multivariate responses. This class of methods is called semi-parametric methods because the estimating equation can be derived without fully specifying the joint distribution of the responses.

Let $y_i=(y_{i1},y_{i2},\ldots,y_{iT})'$ be a T x 1 vector of the responses for the ith subject, where T is assumed the same for every subject without losing generality, and let $\mathbf{X}_i=(X_{i1},\ldots,X_{iT})'$ be the T x p matrix of covariates for the ith subject, $i=1,2,\ldots,n$. It is assumed that the marginal density of y_{it} has the following form:

$$f(y_{it}) = \exp[\{y_{it}\theta_{it}-a(\theta_{it})+b(y_{it})\}\phi] \tag{3.1}$$

where ϕ is nuisance scale parameter. The first two moments of y_{it} are

$$E(y_{it}) = a'(\theta_{it}) = \mu_{it} \tag{3.2}$$

$$var(y_{it}) = a''(\theta_{it})/\phi \tag{3.3}$$

where $a'(\theta_{it})$ and $a''(\theta_{it})$ are the first and second derivatives of $a(\theta_{it})$, respectively.

First, the mean of marginal response can be related to a linear combination of the covariates by a link function g:

$$\eta_{it} = g(\mu_{it}) = \boldsymbol{x}_i'\beta \tag{3.4}$$

where $\beta=(\beta_1,\ldots,\beta_p)'$ is a p x 1 vector of parameters.

The link function g(.) is assumed to be monotone and differentiable, and should be chosen so that it maps the expectation space onto the real line. The most often used link functions are

- logit function $g(x) = \log(x/(1-x))$ for binary responses.
- log function $g(x) = \log(x)$ for counts responses.
- identity function $g(x) = x$ for continuous responses.

Second, the variance of y_{it} can be described as a function of the mean:

$$var(y_{it}) = k(\mu_{it})\phi \tag{3.5}$$

where the function k is a known variance function.

- for binary responses, $k(\mu_{it})=\mu_{it}(1-\mu_{it})$, $\phi=1$.
- for Poisson responses, $k(\mu_{it})=\mu_{it}$, $\phi=1$.
- for continuous responses, $k\equiv1$, $var(y_{it})=\phi$.

3.3 Generalized Estimating Equation

3.3.1 'Working' Correlation Matrix

Using the model specified above, if the observations from a subject are assumed independent of each other, the estimating equations can be readily obtained (Liang and Zeger, 1986). Since the repeated measures from a subject are almost always correlated, the dependencies among the successive responses from the same subject have to be taken into account.

Since no particular multivariate model is assumed, and only the marginal distributions of the responses are specified, the covariance usually depends on the mean.

Liang and Zeger (1986) introduced a 'working' correlation matrix $\boldsymbol{R}(\boldsymbol{\alpha})$. To meet the requirement of being a correlation matrix, let $\boldsymbol{R}(\boldsymbol{\alpha})$ be a T x T symmetric matrix, and let $\boldsymbol{\alpha}$ be an s x 1 vector of parameters which fully characterizes $\boldsymbol{R}(\boldsymbol{\alpha})$. The (t, t') element of $\boldsymbol{R}(\boldsymbol{\alpha})$ is the hypothesized correlation between y_{it} and $y_{it'}$.

The structure of the 'working' correlation matrix can be chosen among a variety of different forms. The simplest choice is the independent working model in which $\boldsymbol{R}(\boldsymbol{\alpha})$ is equal to the identity matrix. Thus, solving GEE is the same as fitting the usual regression model for independent data. Another choice is to completely specify the correlation matrix, i.e., let $\boldsymbol{R}(\boldsymbol{\alpha})=\boldsymbol{R}_0$. When the specified correlation matrix $\mathbf{R}_0$ is very close to the true correlation, this model will yield great efficiency. However, it is usually not clear what kind of structure the true correlation matrix has. When the correlation structure is totally unspecified, the full parameterized model can be used. There are T(T-1)/2 parameters to be estimated in this model. When T is very large or when there are too many missing data, however, it is difficult to use this model because the estimates of $\boldsymbol{\alpha}$ may not be positive.

See Liang and Zeger (1986), Zeger and Liang (1986) for discussions of other possible correlation models.

3.3.2 GEE

Let $\boldsymbol{A}_i = diag\{a''(\theta_{it})\}$ be a T x T matrix, $\boldsymbol{\Delta}_i = diag(d\theta_{it}/d\eta_{it})$ a T x T matrix, and $\boldsymbol{D}_i = d[a_i'(\theta)]/d\beta = \boldsymbol{A}_i\boldsymbol{\Delta}_i\boldsymbol{X}_i$, also let $S_i = Y_i - a_i'(\theta)$.

Define

$$V_i = A_i^{1/2} R(\alpha) A_i^{1/2} / \phi \tag{3.6}$$

$\mathbf{V}_i$ will be cov(y_i) if $R(\alpha)$ is the true correlation between y's.

The general estimating equations is defined to be:

$$\sum_{i=1}^{n} D_i' V_i^{-1} S_i = 0 \tag{3.7}$$

If $R(\alpha)$ is specified as identity matrix, (3.7) reduced to the usual independent equations.

As stated in the beginning of this paper, what is interested in is the effect between groups, and α can be treated as nuisance parameters. Hence only the inferences about β need to be drawn. If the α in (3.7) is replaced by $\hat{\alpha}(y, \beta, \phi)$, a $n^{1/2}$-consistent estimator of α when $\beta,\ and\ \phi$ are known, and in addition, if ϕ in (3.7), which is generally unknown, is replaced by $\hat{\phi}(y, \beta)$, a $n^{1/2}$-consistent estimator of ϕ when β is known, then equation (3.7) becomes

$$\sum_{i=1}^{n} U_i[\beta, \hat{\alpha}(\beta, \hat{\phi}(\beta))] = 0 \tag{3.8}$$

where $U_i(\beta, \alpha) = D_i' V_i^{-1} S_i$ and $\hat{\beta}_G$ is defined to be the solution of equation (3.8).

Under mild regularity conditions and given that

(1) $n^{1/2}(\hat{\alpha} - \alpha) = O_p(1)$, i.e. $\hat{\alpha}$ is $n^{1/2}$-consistent.

(2) $n^{1/2}(\hat{\phi} - \phi) = O_p(1)$, i.e. $\hat{\phi}$ is $n^{1/2}$-consistent.

(3) $|\partial\hat{\alpha}(\beta, \phi) / \partial\phi| \le H(y, \beta)$ which is $O_p(1)$.

then $n^{1/2}(\beta_G - \beta)$ is asymptotically multivariate normal with zero mean and covariance matrix $\mathbf{V}_G$ where:

$$\boldsymbol{V_G} = \lim_{n\to\infty} n\left(\sum_{i=1}^{n} \boldsymbol{D}_i'\boldsymbol{V}_i^{-1}\boldsymbol{D}_i\right)^{-1}\left\{\sum_{i=1}^{n} \boldsymbol{D}_i'\boldsymbol{V}_i^{-1}\mathrm{cov}(y_i)\,\boldsymbol{V}_i^{-1}\boldsymbol{D}_i\right\}\left(\sum_{i=1}^{n} \boldsymbol{D}_i'\boldsymbol{V}_i^{-1}\boldsymbol{D}_i\right)^{-1} \qquad (3.9)$$

The estimate $\hat{\boldsymbol{V}}_G$ of $\boldsymbol{V}_G$ can be obtained by replacing cov(y_i) by S_iS_i', and β, ϕ, α by their respective estimates.

It can be shown that the consistency of $\hat{\beta}_G$ and $\hat{\boldsymbol{V}}_G$ depends only on the correct choice of the mean, not on the correct choice of $\boldsymbol{R}(\alpha)$. That is, the above estimates are robust in the sense that they are consistent even if the correlation matrix $\boldsymbol{R}(\alpha)$ is misspecified, given that the mean structure is correctly specified.

For binary repeated measures data, Liang, Zeger and Qaqish (1990) and Lipsitz (1991) recommended that it is better to parameterize in terms of odds ratio. The estimating equations constructed this way are called GEE2 because it is the second order extension of the GEE in that β and α are estimated simultaneously. Under certain conditions, the estimates obtained here are consistent and likely more efficient than those obtained by the original GEE. See Chinchilli (1991) for more discussions on GEE2.

3.3.3 Solving the GEE

To compute $\hat{\beta}_G$, Liang and Zeger suggested a modified Fisher scoring algorithm. Given the current estimates of the nuisance parameters $\hat{\alpha}$ and $\hat{\phi}$, the following iteration procedure is used to obtain the estimates of β:

$$\hat{\beta}_{j+1} = \hat{\beta}_j\left\{\sum_{i=1}^{n} \boldsymbol{D}_i'(\hat{\beta}_j)\,\tilde{\boldsymbol{V}}_i^{-1}(\hat{\beta}_j)\,\boldsymbol{D}_i(\hat{\beta}_j)\right\}^{-1}\left\{\sum_{i=1}^{n} \boldsymbol{D}_i'(\hat{\beta}_j)\,\tilde{\boldsymbol{V}}_i^{-1}(\hat{\beta}_j)\,S_i(\hat{\beta}_j)\right\} \qquad (3.10)$$

where $\tilde{V}_i(\beta)=V_i[\beta,\hat{\alpha}(\beta,\hat{\phi}(\beta))]$.

Define $D=(D_1',\ldots,D_n')'$, and $S=(S_1',\ldots,S_n')$ and let $\tilde{V}$ be a nT x nT block diagonal matrix with $\tilde{V}_i$ on the diagonal. Define the modified dependent variable $Z=D\beta-S$, and the iterative procedure (3.10) is equivalent to performing an iteratively reweighted least square (IRLS) algorithm.

Liang and Zeger (1986) also suggested that the estimates of α and ϕ at a given iteration can be obtained by using the current standardized Pearson residuals defined by

$$\hat{\gamma}_{it} = [y_{it}-a'(\hat{\theta}_{it})]/[a''(\hat{\theta}_{it})]^{1/2} \tag{3.11}$$

where $\hat{\theta}_{it}$ depends on the current value of β.

ϕ can be estimated by:

$$\hat{\phi}^{-1} = \sum_{i=1}^{n}\sum_{t=1}^{T}\hat{\gamma}_{it}^{2}/(nT-p) \tag{3.12}$$

It can be shown that this estimate is $n^{1/2}$-consistent if the fourth moment of y_{it} is finite.

The estimation of α, which also involves the Pearson residuals, depends upon the choice of $R(\alpha)$. The general approach is to estimate α by:

$$\hat{R}_{uv} = \sum_{i=1}^{n}\hat{\gamma}_{iu}\hat{\gamma}_{iv}/(nT-p) \tag{3.13}$$

Liang and Zeger (1986) gave several examples of estimating α under different choices of $R(\alpha)$.

Prentice (1988) and Zhao and Prentice (1990) suggested making the GEE estimation of β and α simultaneously rather than updating the estimator of α within each iteration of GEE. See Chinchilli (1991) for a brief discussion.

3.4 Discussion

Compared with the likelihood based methods discussed in Chapter 2, the semi-parametric methods are a class of methods which can analyze a broader class of data. While the likelihood based methods can allow only for continuous responses with the assumption that the responses have multivariate normal distribution, the GEE method can allow for both continuous response even when the assumption of multivariate normal responses are not correct, and categorical responses. When the responses are multivariate normal, GEE method reduces to the likelihood based method, and equation (3.7) reduces to the Fisher scoring equations in maximum likelihood.

However, the GEE method, as well as the other semi-parametric methods extended from GEE, gains this advantage of allowing for a broader class of data at the expense of restricting its ability to handle the missing data. As mentioned in the introduction of this chapter, GEE can incorporate the missing data, but requires that the missing data mechanism be MAR or MCAR. The reason is discussed in Section 1.5, where it is shown, in (1.5) and (1.6), that when the missing mechanism is MAR or MCAR, the marginal density of the observed responses is independent of the distribution of nonresponse. Hence, any inference based on the marginal distributions of the observed responses, such as GEE, is independent of the missing data mechanism. Section 1.5 also explains why methods dependent upon the marginal distributions of observed responses cannot deal with the ignorable missing data mechanism, which can be handled by likelihood based methods.

The specification of the structure of the 'working' correlation matrix in GEE is not as important as the specification of the structure of covariance matrix in the likelihood based methods. This is because in GEE method, the assumption about the joint distribution of the responses is avoided by assuming only a functional form of the marginal distribution at each time, and treating the covariance structure as a nuisance. Relying on independence across subjects, consistent estimates of β and its covariance can be obtained even when the 'working' correlation matrix is misspecified. However, a proper specification of the structure of the correlation, which is close to the true unknown correlations, will greatly increase the efficiency of calculation.

In some applications, especially in sociology or economics, where the growth curve across time for each subject, or the correlation within subject, is of primary interests, modelling the marginal distribution and treating the correlation as a nuisance, as in GEE, may be inappropriate. However, as stated in Chapter 1, the primary interest of many biomedical applications is to compare the between group effects, rather than the within subject correlations. Under this situation, the way the models are built in GEE is appropriate.

Although it is assumed throughout this chapter that the data are strictly repeated measurements as defined in Chapter 1, GEE method can also be applied to general longitudinal data in which T may vary from subject to subject. Wei and Stram (1988) gave a model that fits a separate univariate regression to the data at each time point, so the regression parameters may be different at different time points.

Although GEE method can handle a much broader class of data, it is limited when the responses are continuous but extremely non-normal, in which case condition (3.1) is no longer satisfied. In such situations, non-parametric methods, which also appear in recent literature, may be used to tackle the problem. No details about non-parametric methods will be given here. See Davis (1991) for a review.

Chapter 4

Missing Not At Random

4.1 Introduction

When the missing data in repeated measurements are missing not at random, it may not be able to get consistent estimators using the methods discussed in previous chapters. A general approach that can allow the for missing not at random mechanism in repeated measurements has not yet been developed.

The existing approaches usually restrict the problem to a specific type of non-randomly missing data mechanism. Two different ways to deal with non-randomly missing data mechanism in repeated measurement are found in literature. The first way is to model directly the missing data mechanism and include this model in the likelihood function to obtain maximum likelihood estimators of the parameters. One of the most common non-random missing data mechanisms is the 'informative' right censoring, i.e. the censoring, caused by death or withdrawal, which depends on the parameters to be estimated. However, in many cases, even if the missing data are suspected to be missing not at random, we ordinarily cannot know exactly what the missing mechanism is. And some missing data mechanisms are difficult to classify and hence difficult to model properly.

Another method uses the concept of protective statistics. Like the robust statistics which are not sensitive to deviation from the normal distribution assumption, protective statistics are called protective in the sense that they are not sensitive to deviation from the missing at random mechanism. Although in general these statistics are not protective against any kind of non-randomly missing data mechanism, they usually provide satisfactory estimators in a certain type of non-randomly missing data mechanisms.

This chapter will introduce the above two approaches. Sections 2, 3 and 4 will discuss the methods introduced by Margaret C. Wu et al (Wu and Carroll, 1988; Wu and Bailey, 1988; Wu and Bailey, 1989), which analyze the change of rate in the presence of informative right censoring by directly modeling the censoring process. The protective statistic suggested by C. Hendricks Brown (Brown, 1990) to protect against a wide class of non-randomly missing data mechanisms will be introduced in Sections 5 and 6. Section 7 concludes the chapter with some discussions.

4.2 Right Censoring Problem

Measurements in repeated measurement studies are often right censored by a subject's death or withdrawal. Right censoring means that all the measurements after a certain time point are missing. If the probability of censoring does not depend on the values of the missing responses, and hence on the underlying parameters, right censoring is treated as a special case of ignorable missing data mechanism and the methods in Chapter 2 can be employed. However, if the probability of censoring depends on the underlying parameters, the measurement is said to be informative right censored.

Wu and Carroll (1988) proposed an approach to account for censoring using a general right censoring probability distribution that depend on the underlying parameters for the responses. They developed a likelihood ratio test to test the 'informativeness' of the right censoring. The also derived pseudo-maximum likelihood estimates (PMLE) for the response parameters under a probit right censoring model.

Another method proposed by Wu and Bailey (1989) is the conditional linear model. They derived two simple non-iterative estimators of the parameters: linear minimum variance unbiased estimator (LMVUB) and linear minimum mean squared errors estimator (LMMSE).

Wu and Bailey (1988) compared the above two methods using a variety of types of treatment effect and different censoring probability models.

Assume that there are two treatment groups of sample size n_k, for $k=1,2$. The total sample size $n=n_1+n_2$. Let there be J identical measure time points, t_j, for each subject, with one baseline measure at $t_1=0$, and J-1 follow-up time points t_j, $j=2, 3, \ldots J$. t_J is the length of the study. Let $y_i=(y_{i1}, y_{i2}, \ldots, Y_{ij_i})'$ be the vector containing the measurements actually made for the ith subject, where $j_i \leq J$. If $j_i=J$ there is no data missing. On the other hand, if $j_i<J$ then there are some data missing. The missing data may be caused by right censoring as well as other mechanisms, it is assumed that any missing data caused by mechanisms other than right censoring are ignorable.

A random effect linear model is used to model the responses, in which it is assumed that the serial measurements of the response variable follow a linear function of time. Let $\beta_i'=(\beta_{i1}, \beta_{i2})$ be the unobservable vector representing the true initial value and slope of the response variable for the ith subject in the combined sample. For $i \in k$, i.e. the ith subject in the combined sample belongs to the kth treatment group, where k=1,2, the model of y_i is

$$y_i = X_i \beta_i + \varepsilon_i \tag{4.1}$$

where

$$\beta_i \sim N(B_k, \Sigma_\beta) \tag{4.2}$$

$$\varepsilon_i \sim N(0, \sigma_\varepsilon^2 I) \tag{4.3}$$

and

$$X_i' = \begin{bmatrix} 1 & \dots & 1 \\ t_{i1} & \dots & t_{ij_i} \end{bmatrix} \tag{4.4}$$

$$\Sigma_\beta = \begin{bmatrix} \sigma_{\beta_1}^2 & \sigma_{\beta_1\beta_2}^2 \\ \sigma_{\beta_1\beta_2}^2 & \sigma_{\beta_2}^2 \end{bmatrix} \tag{4.5}$$

$$B_k' = (B_{k1}, B_{k2}) \tag{4.6}$$

The objective of the study is to estimate and compare the mean slopes of different treatment groups, i.e. B_{12} and B_{22}. Two approaches mentioned in section 4.1 for informative right censoring in repeated measurement will be discussed in the next two sections.

4.3 Modeling the Censoring Distribution

Wu and Carroll (1988) introduced the direct modeling of the informative right censoring process. They proposed a general censoring probability function $M(\alpha_{0j}, \alpha'\beta)$

that depends on the individual initial value and slope. Here $\alpha=(\alpha_1,\alpha_2)'$ denotes a vector of censoring coefficients for initial value and slope, and α_{0j} for j=2, ..., J are censoring-time parameters. The logical candidates for the distribution function are the Cox's proportional hazards model, logistic model, and probit model. Wu and Carroll used the probit censoring model:

$$Pr(j_i \le j|\beta) = M(\alpha_{0j},\alpha'\beta) = \Phi(\alpha_{0j}+\alpha'\beta) \tag{4.7}$$

where Φ is the cumulative probability of a standard normal variable. Usually σ_ϵ^2 and Σ_β are unknown, but their unbiased estimates can be substituted for them in calculating the likelihood function. Hence the maximum likelihood procedure used here is actually a pseudo-maximum likelihood procedure.

Using the pseudo-maximum likelihood procedure, Wu and Carroll derived the estimations of B_k and α under this probit model, and referred to them as probit pseudo-maximum likelihood estimates (PPMLE).

Based on the pseudo-likelihood procedure, a hypothesis testing of the informativeness of the censoring procedure can be constructed. First the following hypotheses about α can be constructed:

H_0: $\alpha_1 = \alpha_2 = 0.$

H_1: $\alpha_2 = 0$ *and* $\alpha_1 \neq 0.$

Likelihood ratio tests can be performed for the hypothesis H_0 versus H_1. When H_0 is true, the right censoring will be non-informative with respect to B_{k2} for k=1,2. It can be shown, however, that when H_1 is true the right censoring process will usually be informative with respect to B_{k2}, k=1,2.

Wu and Carroll also showed that, when the censoring is non-informative, the maximum likelihood estimate of B_k is the weighted least squares estimate (WLE):

$$WLE(B_k) = [\sum_{i\in k} W_i^{-1}]^{-1} \sum_{i\in k} (W_i^{-1}\beta_i) \tag{4.8}$$

where i(k) denotes that the ith subject of the combined sample belongs to the kth treatment group, and W_i is the covariance of β. When all subjects have complete observations, it can be shown that the maximum likelihood estimate is the unweighted least squares estimate of B_k:

$$UWLE(B_k) = \sum_{i\in k} \beta_i / n \tag{4.9}$$

4.4 Conditional Linear Model

Wu and Bailey (1989) showed that the conditional expectation of the response variable slope, given the censoring time, is a monotone increasing (decreasing) function of the censoring time when the probit censoring coefficient for the response variable slope is negative (positive). So they proposed using a class of increasing functions to model the conditional expectations of the individual slopes with respect to censoring time. The simplest conditional model is a conditional linear model for the individual least square estimated slopes, in which a linear function is assumed.

The form of the conditional linear model can be written as:

$$(\beta_{i2}|j_i=j) = \gamma_{0k}+\gamma_1 t_j+\varepsilon_{kj} \tag{4.10}$$

for subject i who is from the kth treatment group, where γ_{0k} and γ_1 are unknown parameters, ε_{kj} are random variables with $E(\varepsilon_{kj})=0$ and $Var(\varepsilon_{kj})=\sigma^2_{kj}$. The objective of the study is usually to estimate and compare the group slope means:

$$B_{k2} = E_{i\epsilon k}(\beta_{i2}|j_i) = \gamma_{0k}+\gamma_1 E_{i\epsilon k}(t_{j_i}) \tag{4.11}$$

where $E_{i\epsilon k}(.)$ is the expectation taken within kth group. Notice that (4.11) will include the information on censoring time in the between-group comparisons.

Two methods were proposed by Wu and Bailey to estimate expected slopes B_{12} and B_{22}. The first one is named the linear minimum variance unbiased (LMVUB) estimator. This method simply estimates γ_1 and γ_{0k} (k=1,2) by weighted least squares and substitutes the estimates into (4.11):

$$LMVUB(B_{k2}) = \hat{\gamma}_{0k}+\hat{\gamma}_1\bar{t}_k \tag{4.12}$$

where $\bar{t}_k=\sum_{i\epsilon k} t_{j_i}/n_k$ is an estimate of $E_{i\epsilon k}(t_{j_i})$. The conditional variance of the LMVUB estimate is

$$\sigma^2[LMVUB(\hat{B}_{k2}] = [\textstyle\sum_j n_{kj}\sigma^2_{2kj}]^{-1}+[\textstyle\sum_j n_{kj}\sigma^{-2}_{2kj}(t_j-\hat{t}_{kw})^2]^{-1}(\hat{t}_{kw}-\hat{t}_k)^2 \tag{4.13}$$

where $\hat{t}_{kw}=\sum_j n_{kj}\sigma^{-2}_{2kj}t_j/(\sum_j n_{kj}\sigma^{-2}_{2kj})^{-1}$ is the weighted mean of censoring time.

The other method to estimate the expected slopes, linear minimum mean squared error (LMMSE) estimates, is a linear combination of the individual least squares slope estimates:

$$LMMSE(B_{k2}) = \sum_j W_{kj}(\breve{\beta}_{k2|j}) \qquad (4.14)$$

where

$$\breve{\beta}_{k2}|j_i = \sum_{i\epsilon k} (\hat{\beta}_{i2}|j)/n_{kj} \qquad (4.15)$$

with n_{kj} denoting the number of participants in the kth group who were censored after they had j measurements of the response variable. The weights W_{kj} are chosen to minimize the mean squared error under the linear model (4.10). The variance of the LMMSE estimate can be written as

$$\sigma^2[LMMSE(\hat{B}_{k2}] = \sum_j n_{kj}^{-1} W_{kj}^2 \sigma_{2kj}^2 \qquad (4.16)$$

Wu and Bailey (1988, 1989) compared of the PPMLE, LMVUB, and LMMSE, together with the traditional likelihood based methods introduced in Chapter 2 using simulated data with a missing not at random mechanism. Their results show that though the traditional likelihood methods generate biased estimates, all the three conditional linear estimates produce better results. The performance of these conditional linear estimates depends upon the 'true' underlying models of treatment effects.

4.5 Protecting Against Nonrandomly Missing Data

As mentioned in 4.1, the exact form of the missing data mechanism in most problems is difficult or even impossible to specify. But more often than not, some qualitative information about the missing data mechanism is available. This qualitative information is often helpful in constructing protective estimators.

Two simple protective estimators are proposed by Brown (1990). These estimators are consistent if the missing data mechanism is ignorable, and each is consistent under a broad class of non-random mechanisms as well. Only one of the two protective estimators will be discussed below.

In the rest of this section we will introduce a class of nonrandomly missing data mechanisms called generalized censoring mechanism (GCM), against which the protective estimates can protect. In the next section the specific forms of the protective estimator will be discussed.

Let y be a T-dimensional vector representing the T repeated measures from a subject. For simplicity the subject index is omitted. The objective of the study is to make inference about the first and second moments of y, based on the incomplete data from n subjects. In order to indicate which components are missing, as before an indicator vector of k-dimension R is introduced. For the ith subject, $R_j=1$ if y_j is observed, and $R_j=0$ if y_j is missing. It is assumed that (y,R) for each subject are drawn independently with joint density:

$$f(y,z) = n(y;\mu,\Sigma)\,\omega(R=z|y) \tag{4.17}$$

where n(.;.) is the normal density and z is any k-dimensional vector of zeros and ones, and $\omega(.|.)$ is the missing data mechanism.

Notice that in model (4.17) the marginal distribution of y is specified as the multivariate normal, but the missing data mechanism is left unspecified. This formulation is based on the fact that we generally know little about the exact form of the missing data mechanism in most repeated measurements studies.

As before, let $y^{(o)}$ denote the components of y that are observed, and $y^{(m)}$ denote the components of y that are missing. A very useful mechanism defined earlier in

Chapter 2 is the ignorable missing data mechanism:

$$\omega(R=z|y) = \omega(R=z|y^{(o)}) \tag{4.18}$$

This class of mechanism is very restrictive, however. A broader class of missing data mechanism introduced by Brown is the generalized censoring mechanism (GCM). The GCM, which allows missingness related to y, is defined:

$$\omega(R=(z_1,\ldots,z_k)'|y=(y_1,\ldots,y_k)') = h(z|y) = \prod_{j=1}^{k} h_j(z_j|y_j) \tag{4.19}$$

where $h_j(.|.)$, j=1,..,k, are bounded between 0 and 1.

Under the GCM, missingness on each variable depends on that particular variable alone. GCM is only one class of nonrandomly missing mechanism.

Under the GCM, (4.17) becomes:

$$f(y,z) = n(y;\mu,\Sigma)h(z|y) \tag{4.20}$$

Under this mechanism, it is usually require that the first variable, i.e. the baseline measure, be always observed.

4.6 Protective Estimators

Under the GCM, the form of h's is usually unknown. If maximum likelihood estimates with no or few restrictions on the h's are used, the results are equivalent to the estimates obtained by ignoring the mechanism. This leads to Brown's proposal of methods other than maximum likelihood.

Brown (1990) proposed the use of method of moments to obtain estimates of the first and second moments of y. Since the distributions of the consistent estimators obtained from method of moments do not depend on the missing data mechanism, intuitively they might be useful when the mechanism is unknown.

For simplicity, only the case of T=3 is considered: three repeated measurements are assumed for each subject. Since it is required y_1 must be observed, $\overline{y}_1$ and s_{11} will be used to estimate the first and second moments of y_1. Unlike $\overline{y}_1$, however, the mean of all observed y_2 does depend upon the missing data mechanism since by (4.20)

$$f(z_2|z_2 observed) \propto n(z_2;\mu,\Sigma)h_2(1|z_2) \qquad (4.21)$$

However, the conditional distribution of y_1 given y_2, y_2 observed, is

$$f(y_1|y_2,z_2=1) = \frac{n(y_1|y_2;\mu,\Sigma)h_2(1|y_2)}{\int n(y_1|y_2;\mu,\Sigma)h_2(1|y_2)\,dy_1} = n(y_1|y_2;\mu,\Sigma)$$

$$(4.22)$$

and it does not depend on the missing mechanism. Similarly, it can be derived that the distributions of both $y_1|(y_3,z_3=1)$ and $y_1|(y_2,y_3,z_2=z_3=1)$ do not depend on the missing mechanism. Brown (1990) showed that the marginal distribution of y_1 and the above three conditional distributions, all of which do not depend on the missing mechanism, can lead to identifiability of μ and Σ.

Practically, by using method of moments, eight out of nine estimates can be obtained in the explicit forms:

$$\hat{\mu}_1 = \overline{y}_1^{(1..)}$$

$$\hat{\mu}_2 = \frac{\overline{y}_1^{(1..)} - \overline{y}_1^{(11.)}}{b_{12}^{(11.)}} + \overline{y}_2^{(11.)}$$

$$\hat{\mu}_3 = \frac{\overline{y}_1^{(1..)} - \overline{y}_1^{(1.1)}}{b_{13}^{(1.1)}} + \overline{y}_3^{(1.1)}$$

$$\hat{\sigma}_{11} = s_{11}^{(1..)}$$

$$\hat{\sigma}_{22} = \frac{\hat{\sigma}_{11} - s_{11.2}^{(11.)}}{(b_{12}^{(11.)})^2} \tag{4.23}$$

$$\hat{\sigma}_{33} = \frac{\hat{\sigma}_{11} - s_{11.3}^{(1.1)}}{(b_{13}^{(1.1)})^2}$$

$$\hat{\sigma}_{12} = b_{12}^{(11.)}\hat{\sigma}_{22}$$

$$\hat{\sigma}_{13} = b_{13}^{(1.1)}\hat{\sigma}_{33}$$

where the superscripts refer to various subsets of the sample within which the related calculations were carried out. For example, $\overline{y}_1^{(1.1)}$ is the mean of y_1, computed in the subset of the sample in which both y_1 and y_3 are observed, while y_2 is either observed or not observed, and $b_{12}^{(11.)}$ is the least squares regression coefficient of y_1 on y_2, computed in the subset of the sample in which both y_1 and y_2 are observed, while y_3 is either observed or not observed.

Estimate of another parameter, σ_{23} , is obtained by minimizing a residual variance expression for y_1 given y_2 and y_3:

$$\sum_{j:z_{1j}=z_{2j}=z_{3j}=1} (y_{1j} - M_{1.23}(y_{2j}, y_{3j}; \sigma_{23}^*))^2 \tag{4.24}$$

where

$$M_{1.23}(y_2, y_3; \sigma_{23}^{*}) = \hat{\mu}_1 + \frac{\hat{\sigma}_{12}\hat{\sigma}_{33} - \hat{\sigma}_{12}\sigma_{23}^{*}}{\hat{\sigma}_{22}\hat{\sigma}_{33} - \sigma_{23}^{*2}}(z_2 - \hat{\mu}_2) + \frac{\hat{\sigma}_{13}\hat{\sigma}_{22} - \hat{\sigma}_{12}\sigma_{23}^{*}}{\hat{\sigma}_{22}\hat{\sigma}_{33} - \sigma_{23}^{*2}}(y_3 - \hat{\mu}_3) \tag{4.25}$$

The estimate $\hat{\sigma}_{23} = \sigma_{23}^{*}$ can be estimated by minimizing (4.25).

The above procedure can be extended to problems with T > 3, i.e. problems with more than three repeated measures for each subject. The mean and variance of the baseline and all covariances involving the baseline are identifiable. The mean and variance of y_j, $j > 1$, are identifiable if $\rho_{1j} \neq 0$. And σ_{jk}, for $j, k > 1$, are identifiable if $\rho_{1j} \neq 0$ and $\rho_{1k} \neq 0$.

Brown (1990) also proposed another protective estimators which deals with a class of mixed mechanism.

4.7 Discussion

Most of the statistical methods for missing data problems can allow only for the randomly missing mechanism or ignorable missing mechanism. The methods that can deal with nonrandomly missing mechanism are still not fully developed. There is no general approach for repeated measurement with non-randomly missing data. Although the two approaches introduced in this chapter are similar in that they restrict the problem to a certain type of non-randomly missing data, they represent two different possible directions in attempting to solve the non-randomly missing data problem.

The approaches proposed by Wu et al for the informative right censoring data directly model the right censoring mechanism and incorporate the mechanism into the likelihood function. Using simulated informative right censoring data, Wu and colleagues have shown that their estimates are better than the traditional likelihood based methods which treat missing mechanism as ignorable. Since right censoring is one of the missing data mechanisms that most likely to occur in repeated measurement studies, and more often than not, right censoring is expected to be related to responses, this approach provides a good alternative to the traditional likelihood based method in dealing with the right censoring problem.

However, as stated above, except for very few case like the right censoring, there is no complete definition and classification of the nonrandomly missing data mechanism, so it is difficult to model the missing mechanisms directly. In fact, the most difficult aspect of nonrandomly missing data problems is that in most cases the exact missing mechanism for a specific data is unknown. This difficulty may explain the need for the protective estimators proposed by Brown. Brown's approach does not specify the missing data mechanism explicitly but imposes some mild conditions on the form of the mechanism so that a certain class of nonrandomly missing data mechanism can be covered. Brown used method of moments to obtain estimates of the first two moments that are protective against the deviation from randomly missing mechanism. This approach, however, also has its own restrictions. Although it can handle a relatively broad class of the nonrandomly missing mechanisms, it cannot protect against any kind of nonrandomly missing mechanism. Because it uses method of moments, it can only deal with a single sample at a time and, hence, loses some flexibility of incorporating several different treatment groups into one model.

Generally speaking, the increasing ability to handle more complex missing data mechanisms is obtained at the expense of imposing other assumptions on the data. The semi-parametric methods can only handle the missing at random mechanism, but they do not need the assumption of the multivariate normality of the data, and only require that the marginal distributions of the observed responses have a distribution of exponential family. The likelihood based methods, which can handle more complex ignorable mechanism, requires that the data be multivariate normal, although the estimates obtained by likelihood based methods are usually not very sensitive to the deviation from normality. The two approaches discussed in this chapter, which can handle some nonrandomly missing mechanisms, not only require the data be multivariate normally distributed, but also are very sensitive to deviation from normality assumption.

Chapter 5

The Transcranial Doppler Data

5.1 Introduction

In this chapter, some of the methods discussed in previous chapters will be applied to analyze practical data from a clinical trial. Different methods will be compared and contrasted, and possible future research topics will be discussed.

Section 5.2 will present the data set, the transcranial ultrasonic Doppler data from a clinical trail of a potential new treatment of cerebral vasospasm following aneurysmal subarachnoid hemorrhage (SAH). The severe missing data problem in this data will be discussed.

In Section 5.3, two of the approaches discussed in previous chapters, the likelihood based method discussed in Chapter 2 and protective estimators of Brown discussed in Chapter 4, will be used to analyze the data presented in Section 5.2.

Section 5.4 contains a discussion of the results from Section 5.3. The difficulties with the existing methods for repeated measurement data with missing data problem will also be addressed, and some possible future research topics arising from the need in application will be examined.

5.2 The Transcranial Doppler Data

The data used here is from a clinical trial of Nicardipine, a medicine newly developed to treat delayed cerebral vasospasm following aneurysmal subarachnoid hemorrhage (SAH). The clinical trial was a prospective, randomized, double-blind and placebo-controlled multi-center trial. A total of 906 patients with recent aneurysmal SAH (0 to 7 days after SAH) hospitalized in over 40 North American neurosurgical centers were involved in the study. The patients were randomized to two treatment groups (Nicardipine and placebo), with 449 patients in one group and 457 patients in the other. The treatment was administered to each patient daily up to 14 days following SAH. The primary endpoints for the study were the percentage of patients achieving Good Recovery on the Glasgow Outcome Scale at 3 months following the SAH and incidence of ischemic deficits due to vasospasm (symptomatic vasospasm) during the treatment period (from the day the patient received the first dose to 14 days post SAH).

A secondary endpoint in the study was the transcranial ultrasonic Doppler measurements, which can be used as confirmation of vasospasm. A high reading of transcranial Doppler indicates symptomatic vasospasm. Ideally, the transcranial Doppler measurements should have been recorded once a day for each patient from the day he or she entered the study to the 14th day after SAH or the day the patient was dropped from the study. If a patient entered the study on the day of SAH, and was discharged on the 14th day after SAH, 15 transcranial Doppler measurements should have been taken, one baseline measurement recorded on the day of SAH and one measurement daily from the 1st day to the 14th day post SAH. However, at the time the study was carried out (in the late 80's), transcranial Doppler equipment was not standard in many participating centers, so none of the patients treated in those centers without transcranial Doppler equipment had Doppler measurements. Furthermore, in those centers equipped with transcranial Doppler, its availability to patients was limited. As a result, the data set contains a large portion of missing data. Table 5.1 shows the frequency of each missing category. We can see that 63.2% of the patients do not have even one measurement and 11.2% of the patients only have one measurement. Only one patient has complete fifteen measurements.

Table 5.1 The Frequency of Missing Data

	No measures taken	Only one measurement	Two or more measures
Number of Patients	573	101	232
Percent	63.2%	11.2%	25.6%

This data set includes three types of missing data. First, the data are left censored because patients were allowed to be entered anytime from 0 to 7 days after SAH. If a patient entered the study at the fourth day following SAH, his transcranial Doppler from day 0 to day 3 after SAH could not have been obtained. Second, there is right censoring caused by death or withdrawing from the study due to recovery. Third, there are other types of missing data besides the left and right censoring, i.e., even if a patient entered the study at the SAH day and remained in the study until the 14th day after SAH, he or she might still have had missing data due to the other reasons stated above.

The missing data mechanism in this case is suspected to be missing not at random. Patients treated in centers without transcranial Doppler equipment can be eliminated from the data set and it can be assumed that Doppler equipment among the centers is missing completely at random (MCAR). However, we suspected that in those centers with transcranial equipment a patient may have more complete Doppler measurements simply because he/she was diagnosed, by means of other procedures, to be more likely to have vasospasm. Patients in stable condition often have a single measurement (often at the baseline) and have missing data for the rest of the days in the treatment period. Since the patients in better condition are more likely to have a larger portion of missing data, while the patients with more complete observations usually were in worse condition, the missing data appears to depend on the value of the responses, and, thus, the missing mechanism is missing not at random.

One way to cope with this non-randomly missing mechanism is to divide the patients left in the data set (patients with at least one measurement) into two subgroups according to the number of measurements missing. If a patient has only one measurement we assume that this patient was in relatively good condition and classify him or her in the better group; a patient with more than one measurement will be classified in the worse group. Then the patients with only one measurement can be deleted from the data set and the analysis results can be claimed valid only for the subgroup of patients in worse condition. In this subgroup there are only 232 patients left, with 103 patients in one treatment group and 129 patients in the other.

Even in this subgroup, more than 50% of the data is missing. Furthermore, the missing mechanism within this subgroup is still not ignorable. Because patients with left censoring entered the study late, they received proper treatments later than other patients, so their conditions can be assumed worse than that of other patients. Patients with right censoring were dropped from the study either because of death (extremely bad case) or because of recovery (extremely good case). In both left censoring and right censoring, the missing of a measurement seems to depend on the value of the responses, so the censoring is very likely to be informative. In order to reduce the percentage of missing data, and to reduce the effect of the informative left and right censoring, the maximum value among the measurements in day 0 through day 6 was used as a single baseline variable, y_B, and the maximum among measurements in day 12 through day 14 was used as a single follow up measurement, y_F. Since it is known that day 7 to day 11 is the critical period for the occurrence of symptomatic vasospasm, the measurements taken on these days were kept as they were in the original data set. In the next two sections, this modified data set is used as a 'working' data set, which will be analyzed with two different methods discussed in previous chapters.

5.3 The Application of Likelihood Based Methods

It is suspected that even in the 'working' data set the missing data mechanism may still not be missing at random, but since no general algorithm is available to deal with any type of non-randomly missing data, one way to tackle this problem is to assume the missing data mechanism in the 'working' data is missing at random and use the likelihood based methods introduced in Chapter 2 to analyze it.

As stated in Chapter 2, the likelihood based method models the means and the covariance matrix separately. Many situations can be modeled by combining different choices of mean and covariance structures. Six different situations are considered here, and the likelihood ratio tests were employed to test the goodness of fit, and to determine which model will be used.

Table 5.2 lists the models we have examined, and the results of the goodness of fit tests. For each model fitted, the table lists a description of the model, number of parameters in the model, -2λ, chi-square statistics, the degree of freedom for the goodness of fit test, and Arkaike Information Criteria (AIC). AIC is also a criteria for the fitness of the model, the greater the value of AIC, the better the model fits the data.

The first model in Table 5.2 is the unstructured ANOVA model. A different mean is modeled for each treatment group at each time point, therefore, the means model has fourteen parameters. No specific structure is assumed for the covariance matrix, thus the total number of parameters in the covariance model is 28. This is the most complex model. Every other model in the table tries to simplify the model while not increasing the square sum of residuals too much. Detailed discussions of the other models can be found in Chapter 2.

Actually, for each model from model 2 to model 6 the likelihood ratio test is used to test the null hypothesis that this model holds for the data versus the alternative hypothesis that model 1 holds. The results listed in Table 5.2 show that the null hypothesis is rejected in every situation. It is also shown that the AIC of model 1 is the maximum among the six models. These results lead to the conclusion that the only model which fits the data is model 1, the unstructured covariance ANOVA model.

Table 5.2 Summary of models fitted.

no.	Model Description	No.of Para.	-2λ	χ^2	df	AIC
1	ANOVA mean model, unstructured covariance matrix	42	9502.8	---	--	-4779.4
2	ANOVA mean model, compound symmetry covariance matrix	16	9625.6	123.2	26	-4814.8
3	Linear mean model, unstructured covariance matrix	32	9524.6	18.2	10	-4790.3
4	Linear mean model, compound symmetry covariance matrix	6	9655.2	152.4	36	-4829.6
5	Linear mean model, 1st order AR covariance structure	6	9665.2	162.4	36	-4834.6
6	Linear mean, random effect model	8	9613.8	110.0	34	-4810.9

Using model 1, the model of the means has the form:

$$y_i = a + b*TRT + \sum_{j=1}^{6} c_j*DAY_j + \sum_{j=1}^{6} d_j*TRTDAY_j \qquad (5.1)$$

where TRT is a binary variable which equals 1 when the patient is in one treatment group and -1 if in the other, and DAY_1 through DAY_6 are six dummy variables, their values assigned as:

Measure Time	DAY_1	DAY_2	DAY_3	DAY_4	DAY_5	DAY_6
Baseline(YB)	1	0	0	0	0	0
7th Day	0	1	0	0	0	0
8th Day	0	0	1	0	0	0
9th Day	0	0	0	1	0	0
10th Day	0	0	0	0	1	0
11th Day	0	0	0	0	0	1
Follow up(YF)	-1	-1	-1	-1	-1	-1

$TRTDAY_1$ through $TRTDAY_6$ are six dummy variables indicating the interactions between treatments and days, the value of $TRTDAY_j$ is the product of TRT and DAY_j for $j=1$ to 6. Y_i is the Doppler measure for the patient, and a, b, c_1 through c_6, and d_1 through d_6 are fourteen parameters. The covariance part of the model does not assume any structure for the covariance matrix and estimates each element in the matrix as an independent parameter.

The estimates of the covariance matrix are

$$\begin{bmatrix} 3017.5 & & & & & & \\ 1440.9 & 3560.4 & & & & & \\ 1599.4 & 3092.2 & 3763.8 & & & & \\ 1869.1 & 3010.5 & 2827.2 & 4664.6 & & & \\ 1894.9 & 2727.5 & 2685.3 & 3512.3 & 4857.0 & & \\ 2937.5 & 2299.2 & 2486.8 & 3054.8 & 3880.2 & 4783.4 & \\ 1685.0 & 2468.2 & 2295.8 & 3426.0 & 3513.3 & 3805.5 & 4534.7 \end{bmatrix} \tag{5.2}$$

and the estimates of the parameters in the mean model (5.1) are listed below.

Parameter	Estimate	Asymptotic SE
a	152.84	3.71
b	-13.38	3.71
c_1	-21.79	3.26
c_2	0.53	3.04
c_3	5.93	3.31
c_4	3.20	3.52
c_5	-0.47	3.38
c_6	3.31	3.71
d_1	10.38	3.26
d_2	-6.14	3.04
d_3	-7.16	3.31
d_4	1.52	3.52
d_5	3.52	3.38
d_6	2.83	3.71

Since the study's primary purpose is to discover if significant differences in Doppler measures exist between the two treatment groups, two hypotheses are of interests. The first one tests the null hypothesis that no interactions exist between the treatment effects and time effects; the second one tests the null hypothesis that there is no difference in Doppler measures between the two treatment groups. The results of the above two tests are:

Test	DF	Chi-Square	P-Value
TRT	1	13.01	0.000
TRT*DAY	6	20.53	0.002

The null hypotheses were rejected in both tests. For the test of treatment effect, the difference of Doppler measurements between the two treatments are highly significant, this is generally in agreement with the conclusion reached after analyzing other endpoints in the study (C. Haley), although the significance of the interaction between treatment effect and time effect makes interpretation of the difference between the two treatment groups complicated.

5.4 The Applications Of Brown's Method

Brown's protective estimators discussed in Section 4.6 can provide consistent estimators of first and second moments when the missing mechanism is the generalized censoring mechanism (GCM). Although we do not know whether or not the missing mechanism in our data is GCM, we can apply Brown's Method to the data, and compare the results with the results from Jennrich's likelihood based method. If the missing mechanism here is non-random, and is GCM, the results from the above two approaches should be different because Jennrich's method will usually generate inconsistent estimates in this situation.

As discussed in Section 4.6, Brown's Method requires that every patient have a baseline measurement. In order to meet this requirement, we only include those patients with baseline not missing in our 'working' data set. To further simplify the calculation we let the baseline measure y_B be y_1, the maximum value among the measurements taken on 7th day through 11th day after SAH be y_2, and the follow up measure y_F be y_3, so that we only have three time points for each patient, in which case equations (4.23), (4.24) and (4.25) can be used directly to calculate estimates for each treatment group. For comparison, we also analyze the same data set with Jennrich's likelihood based method.

The results from Brown's method are

Group 1

Variables	Means	Covariance		
y_1	127.53	2661.9		
y_2	154.24	3513.7	4638.8	
y_3	141.49	2847.2	3241.0	3758.9

Group 2

Variables	Means	Covariance		
y_1	134.70	3381.6		
y_2	195.72	5131.0	7786.1	
y_3	180.92	4723.6	6959.0	7167.9

While the results from Jennrich's method are

Variables	Group1	Group2	Covariance		
y_1	127.53	134.70	3025.5		
y_2	151.84	197.74	1687.7	4333.5	
y_3	139.69	179.76	1630.0	3510.5	4529.5

Using Brown's method, we obtain separate estimates of the covariance for each treatment group, while with Jennrich's method we have only one set of estimates of covariance matrix for the combined population. We can see that the estimates of means by the two methods are very close to each other, but the estimates of covariance matrix by Brown's method tend to be larger than those by Jennrich's method.

Both methods reject the null hypothesis that the Doppler measures from the two treatment groups are the same, but Jennrich's method also reveals that the interactions between treatment effect and time effect are significant, while Brown's method cannot test those interactions.

Because the results of the two approaches are very similar, we can deduce that the missing mechanism here is either just ignorable, or non-random but not GCM. For the reasons stated in Section 5.2, we suspect that the missing mechanism for our 'working' data is non-random rather than ignorable.

Our conclusion that the missing mechanism is non-random but not GCM makes sense intuitively. For example, if a patient's baseline Doppler measure is low, he or she might remain in good condition throughout the study period, and have a missing measure for y_2 or y_3. The missing of y_2 (or y_3) for this patient depends not only on the value of y_2 (or y_3) but also on the value of y_1; hence the missing mechanism is not GCM.

5.5 Discussion And Summary

The missing data problem is very common, and sometimes unavoidable in repeated measurement studies. A variety of statistical approaches has been developed recently to deal with this problem.

When the mechanism of missing data in a repeated measurement study is missing at random (MAR) or missing completely at random (MCAR), the semi-parametric methods discussed in Chapter 3, which model the marginal distribution of the multivariate responses instead of the joint distribution of the responses, can be used to solve the missing data problem. There are few restrictions on the semi-parametric methods, which can be used for continuous or categorical data, as long as the marginal density of the responses is from the exponential family.

The ignorable missing data mechanism can be handled by the likelihood based methods discussed in Chapter 2. While this approach can deal with ignorable missing data mechanism, as well as MAR and MCAR, the responses must be continuous and multivariate normally distributed.

However, as has been seen in the transcranial Doppler data, the missing data mechanism in a repeated measurement study is most likely missing not at random, a much more complicated class of missing data mechanism. A general approach, such as the likelihood based methods for ignorable mechanism and the semi-parametric methods for MAR and MCAR, has not been developed for this mechanism. Although there have been some attempts to solve this problem, as discussed in Chapter 4, numerous problems remains to be solved.

First, Brown's protective estimators claim to be able to obtain consistent estimators of first and second moments under the GCM mechanism, which is a broader class than ignorable mechanism, but it is found that in many biomedical applications, such as in the transcranial Doppler data, the missing of a response does not depend only on the value of that response itself, i.e., in many applications the missing data mechanism is more complicated than GCM. Even for GCM, Brown's method requires that every subject have a baseline observed, which in many applications, such as in the evaluation of transcranial Doppler data, is not satisfied.

Second, while much effort has been devoted to the informative right censoring problem, the left censoring problem, which is also common in biomedical applications has not been studied closely. In the transcranial Doppler trials, many patients were sent to small, local hospitals first, and only transferred to the participating neurosurgical centers when their conditions worsened. Thus these patients might have entered the study several days after SAH. The lest censoring caused by the delay is very likely informative. Since this scenario is common in clinical trials, methods dealing with informative left censoring also need to be developed.

Third, the current approaches are mostly 'deal-one-type-at-a-time'. Wu's methods for right censoring, for example, assume that all the other missing values in the data except the right censoring are ignorable. But the real world hardly conforms the 'one-type-at-a-time' approach. In the transcranial Doppler data, a typical real life application in biomedical researches, there is informative right censoring, informative left censoring, and missing other than left or right censoring also suspected missing not at random. So approaches able to handle complex missing data problems with several, mixed missing mechanisms are highly desirable.

Finally, all the approaches, except semi-parametric methods, assume that the data are continuous and normally distributed. Categorical data or continuous but not normally distributed data in repeated measurement studies with missing data problem can be properly analyzed only if the mechanism of the missing data are MAR or MCAR. The missing data problem for categorical data or continuous but not normally distributed data in repeated measurement studies remains a challenge to statisticians.

REFERENCE

REFERENCE

Brown, C. H. 'Protecting Against Nonrandomly Missing Data in Longitudinal Studies', *Biometrics*, **46**, 143-155. (1990)

Davis, C.S. 'Semi-parametric and Non-parametric Methods for the Analysis od Repeated Measurements with Applications on Clinical Trials'. *Statistics in Medicine*, **10**, 1959-1980. (1991)

Diggle, P.J. 'An Approach to the Analysis of Repeated Measurements', *Biometrics*, **44**, 959-971. (1988)

Diggle, P.J. 'Testing for Random Dropouts in Repeated Measurement Data', *Biometrics*, **45**, 1255-1258. (1989)

Haley, E. C. Jr., Kassell, Neal F., Torner, James C., etc. 'A Randomized Controlled Trial of High-dose Intravenous Nicardipine in Aneurysmal Subarachnoid Hemorrhage: A Report of the Cooprative Aneurysm Study', *Jounnal of Neurosurgery*, to be published.

Jennrich, R.I., Schluchter, M.D. 'Unbalanced Repeated-Measures Models with Structured Covariance Matrices', *Biometrics*, **42**, 805-820. (1986)

Laird, N.M. and Ware, J.H. 'Random-effects models for longitudinal data', *Biometrics*, **38**, 963-947. (1982)

Laird, N.M., Lange, N. and Stram, D. 'Maximum Likelihood Computations with Repeated Measures: Application of the EM Algorithm'. *Journal of the American Statistical Association*, **82**, 97-105. (1987)

Laird, N.M. 'Missing Data in Longitudinal Studies', *Statistics in Medicine*, **7**, 305-315. (1988)

Liang, K.Y. and Zeger, S.L. 'Longitudinal Data Analysis Using Generalized Linear Models'. *Biometrika*, **73**, 13-22. (1986)

Little, R.J.A and Rubin, D.B. Statistical Analysis with Missing Data. John Wiley & Sons

(1987)

Moulton, L.H. and Zeger, S.L. 'Analzsing Repeated Measures on Generalized Linear Model via Bootstrap'. *Biometrics*, **45**, 381-394. (1989)

Schluchter, M.D. 'Analysis of Incomplete Multivariate Data Using Linear Models with Structured Covariance Matrices', *Statistics in Medicine*, **7**, 317-327. (1988)

Schluchter, M.D. and Elashoff, J.D. 'Small-sample Adjustments to tests with unbalanecd Repeated Measures Assuming Several Covariance Structures', *Jounal of Statistical Computation and Simulation.* **37**, 60-87. (1990)

Ware, James H. 'Linear Models for the Analysis of Longitudinal Studies', *The American Statistician*, **39**, No.2, 95-101. (1985)

Wei, L.J. and Stram, D.O. 'Analysing Repeated Measurements with Possibly Missing by Modeling Marginal Distributions'. *Statistics in Medicine*, **7**, 139-148. (1988)

Wu, M.C. and Bailey K. 'Analysis Changes in the Presence of Informative Right Censoring Caused by Death and Withdrawal', *Statistics in Medicine*, **7**, 337-346. (1988)

Wu, M.C. and Bailey, K. 'Estimation and Comparison of Changes in the Presence of Informative Censoring: Conditional Linear Model', *Biometrics*, **45**, 939-955. (1989)

Wu, M.C. and Carroll, R.J. 'Estimation and Comparison of Changes in the Presence of Informative Right Censoring by Modeling the Censoring Process', *Biometrics*, **44**, 175-188. (1988)

Zeger, S.L. and Liang, K.Y. 'Longitudinal Data Analysis for discrete and continuous outcomes'. *Biometrics*, **42**, 121-130. (1986)

Vita

[redacted]