Chapter 1

Introduction to Longitudinal Data Analysis

Longitudinal studies are designed to measure intra-individual change over time. Repeated observations are made on individual subjects, usually at a set of common time points specified by the study protocol. A main objective of longitudinal studies is to relate change over time in individuals to their characteristics (exposure, sex, etc.), or to an experimental condition (drug treatment arm, time since baseline, etc.). In some studies, exposures or experimental conditions may change during the course of the study (as in crossover designs or repeated measures experiments). For example, in the typical crossover design each subject receives every treatment in sequence, with suitable washout periods prior to each treatment. The sequence of treatments is determined a priori by randomization. In repeated measures studies, each subject is measured under a set of pre-specified conditions; differences in the response due to conditions are of primary interest. Outcomes may be measurements, counts, or dichotomous indicators, and we may have multivariate outcomes measured at each of several occasions as well.

In the ideal setting, we will have all subjects measured at the same set of occasions; this greatly facilitates the analysis and interpretation. Some studies may be unbalanced by design, as, for example, when measurements are costly and/or invasive, so only a subset of subjects are measured at all occasions. In other instances it may be very difficult to obtain measurements on all subjects on the same set of occasions. This is especially true when studying human subjects over a long period of follow-up, and when studying clinic populations where illness is a big factor in patient availability. Observations may be mistimed and/or missing,

and subjects may drop out or become unavailable for observation.

When the study has a simple, classical design, meaning all subjects are measured on the same set of occasions, and the only covariates which vary over time do so by design, then standard multivariate methods for the analysis of polynomial growth curves, crossover designs, or repeated measures may be used, when outcomes are multivariate normal (see, e.g., Morrison, 1990, and Johnson and Wichern, 1992). When there is substantial missing data, traditional analytic methods can only be applied if one restricts the analysis to the units with complete observations and their validity relies upon the strong and often unrealistic assumptions about the missing data mechanism. Even when these assumptions hold, the complete-case analysis is unsatisfactory, as it discards the information available in the units with incomplete observations.

To some extent, standard univariate regression models and methods can be used to analyze longitudinal data, provided one uses the proper design matrices and takes into account the fact that the observations on individuals are correlated. This approach is the basis of the General Estimating Equations (GEE) approach (Diggle *et al.*, 1994) which we will take up in detail in Chapters 4 and 6. Selection of the design matrix is a key element in all of the methods we discuss, and involves subtleties of model formulation which are crucial in settings with non-standard designs.

Section 1.1 of this chapter will be concerned with types of design matrices for specifying the expected response when the mean is linear in design variables. We will introduce a Linear Model for Correlated Data (LMCD) which can be used to analyze data from any longitudinal study where the mean response can be characterized as linear in design variables. In this case we think of subjects as the basic sampling unit, and response is measured repeatedly on subjects. The model can also be used for the analysis of clustered data; here the sampling unit is the cluster and the repeated observations are individuals in the cluster. This includes traditional clustered survey data, family studies, and nested experimental designs such as those used in animal or teratology studies. Before turning to model development, we consider a few of the data examples which will be presented in later chapters.

Six Cities Study of Air Pollution and Health

The Six Cities Study of Air Pollution and Health (Dockery, 1985) was designed to characterize pulmonary function growth between the ages of six and eighteen and the factors that affect growth. A cohort of 13,379 children born in or after 1967 was enrolled in six communities in

the US, designed to have a gradient of air pollution from high to low. Most children were enrolled in the first or second grade and participants were seen annually until high school graduation or loss to follow-up. At each examination, spirometry was performed and a respiratory health questionnaire was completed by a parent or guardian.

Pulmonary function measurements obtained from simple spirometry are widely used as a measure of respiratory health in clinical and epidemiological research. The basic maneuver in simple spirometry is maximal inspiration followed by forced exhalation as rapidly as possible into a closed chamber. Many different measures could be computed from the spirometric curve of volume exhaled versus time; two widely used measures are forced vital capacity (FVC), the total volume of air exhaled, and FEV_1, the volume of air exhaled in the first second of the maneuver.

Because the survey was school-based, children moving out of or into the community during the period were either lost to follow-up or late entrants. Children absent from school during periods of measurement had incomplete records as well. This type of imbalance is commonly found in school- or community-based surveys.

A Clinical Trial in Patients Undergoing an Acute Schizophrenic Episode

Schizophrenia is an incurable disorder characterized by periods of acute psychosis of variable length and intensity. Antipsychotic medication is effective in reducing psychotic behavior in many individuals, but can pose significant adverse side effects. An example of a longitudinal clinical trial is the equivalency trial of a new antipsychotic drug for schizophrenia described in Lapierre *et al.* (1990). This clinical trial was a double-blinded study with randomization between four treatments: three doses (low, medium, and high) of an experimental drug and a control drug with known antipsychotic effects as well as known side effects. Initial studies prior to this double-blinded study suggested that the experimental drug had equivalent antipsychotic activity, with less side effects. The primary objective of this study was the determination of a dose-response relationship for efficacy, tolerability, and safety, and the comparison to the control drug. The study was conducted at 13 clinical centers, and a total of 245 patients were enrolled. The primary efficacy parameter was the Brief Psychiatric Rating Scale (BPRS). This scale measures the extent of a total of 18 observed behaviors, reported behaviors, moods, and feelings, and rates each one on a seven point scale, with a higher number reflecting a worse evaluation. The total BPRS score is the sum of the scores on the 18 items. The stated endpoint was change

in BPRS from baseline to six weeks.

Patients were admitted to the hospital for the first four weeks of treatment, and discharged as the clinical condition permitted for the final two weeks. Patients were evaluated at baseline and after one, two, three, four, and six weeks of treatment. One hundred and thirty-four (55%) of the patients completed the study, and eleven additional patients had a six-week evaluation, even though they were technically considered non-completers. The primary reason for discontinuation was a perceived lack of effectiveness of the treatment by the physician; there were also several withdrawals due to side effects. Because patients were hospitalized for the main period of follow-up, there are very few missing observations, except those due to patient removal from protocol. In the event of patient removal, a final measurement was made; then the patient was terminated and no further measurements were made.

AIDS Clinical Trial Data

The AIDS Clinical Trials Group (ACTG) conducts numerous trials designed to evaluate therapies for people infected with Human Immunodeficiency Virus (HIV), which causes AIDS. ACTG Protocol 128 (ACTG, 1993; Brady *et al.*, 1996), hereafter ACTG128, was a multicenter, randomized, double-blinded trial that compared high versus low dose Zidovudine (ZDV) therapy in children born with HIV. Zidovudine, also known as AZT (azidothymidine), inhibits replication of HIV and has been shown to decrease mortality and frequency of opportunistic infection among adults with symptomatic HIV (see ACTG, 1993 and references therein).

One primary endpoint in ACTG128 is neuropsychological development as measured by IQ ratio: children are measured at baseline and every six months for two years (five times total), and some continue to be measured even after being removed from assigned treatment. This practice differs from many trials, where removal from treatment is synonymous with removal from the trial. Reasons for treatment termination prior to two years include undesirable side effects, toxicity, lack of efficacy, and parental decision. Children are not put back on ZDV once removed, but can switch to another treatment regimen. Neuropsychological decline represents advancing disease state, while successful treatment prevents decline.

Muscatine Coronary Risk Factor Study

The Muscatine Coronary Risk Factor Study was a longitudinal study of coronary risk factors in school children (Woolson and Clarke, 1984).

The data set used in our examples contains records on 1,014 children from Muscatine, Iowa, who were 7–9 years old in 1977. Height and weight were measured on each child in three survey years: 1977, 1979, and 1981. Although each child was eligible to participate in all three surveys, data on many children are incomplete. For each survey year, the median weight was calculated for each gender and 1 inch of height. Children with relative weight greater than 110% of the median weight in their respective stratum were classified as obese. This criterion resulted in approximately 20% of the children being described as obese (Woolson and Clarke, 1984).

The repeated binary response of interest is whether the child is described as being obese or not (1 = yes, 0 = no) at each occasion. One of the objectives of this study was to determine trends in obesity in children and the effects of gender and age on risk of obesity in children.

A Comparative trial of a contraceptive drug

Machin *et al.* (1988) discuss problems in the analysis of contraceptive trial data which arise as a result of subject discontinuation or dropout. With modern contraceptive methods, discontinuation due to pregnancy is rare, but it occurs often due to unacceptable side effects, such as irregular bleeding patterns. While time to discontinuation is usually the primary endpoint in these trials, it is also of interest to quantify comparative trends in irregular bleeding patterns, and those patterns which lead to discontinuation of the contraceptive method under study.

A comparative trial of two dosages of depot-medroxyprogesterone acetate (DMPA, 100mg and 50 mg) was conducted by the World Health Organization (1986). Women were administered the drug at three month intervals. The discontinuation rate at the end of the first year of the trial was 40%; more than half of those gave bleeding disturbance as the primary reason for discontinuation.

As part of the trial data collection, women maintained daily diaries recording the presence of any irregular bleeding pattern. The analysis presented by Machin *et al.* (1988) uses a binary indicator of the presence or absence of a specific bleeding disturbance, amenorrhea, in each of the three month periods between injections to study comparative trends in amenorrhea in the two dosage groups, and to study the relationship between trends in amenorrhea and discontinuation.

Familial Aggregation of Lung Cancer in Nonsmokers

Studies of familial aggregation are often designed using the case-control design. Either clinical- or population-based samples of cases and

controls (called probands) are selected, then disease and covariate information is obtained on the relatives of the cases and controls. The objective is to study the risk of disease in relatives given the case-control (disease) status of the proband.

A detailed family history study was designed by Schwartz *et al.* (1996) to evaluate family history of lung cancer in a nonsmoker as a risk factor for lung cancer in first-degree relatives. Briefly, cases and controls were selected from participants in a previous study, which included population-based nonsmoking lung cancer cases, ages 40–84 years, diagnosed November 1, 1984 through June 30, 1987. These cases were originally ascertained through the Metropolitan Detroit Cancer Surveillance system of the Karmanos Cancer Institute, a participant in NCI's SEER seer Program. Nonsmokers were identified as individuals reporting never smoking more than 100 cigarettes in their lifetime. Non-smoking controls were matched to the cases on $\pm$ 5 years, race, sex and county of residence.

Risk factor data collected for the cases and controls included environmental tobacco smoke exposure, occupational history, history of other lung diseases, and family history. The detailed family history collected on relatives included the above-mentioned risk factor data, smoking history, demographics, and occurrence of cancer and other lung diseases among spouses and first-degree relatives of the probands. Questionnaire data for 2,252 family members of nonsmoking cases and 2,408 family members of 247 nonsmoking controls were obtained. Laird *et al.* (1998) used these data to illustrate the use of multivariate methods in the analysis of family risk data.

1.1 A Linear Model for Correlated Data

The data

Consider a sample of N randomly selected units with n_i measurements of response on each unit, $i = 1, \ldots, N$,

$$\underset{n_i \times 1}{Y_i} = \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{in_i} \end{pmatrix}$$

where the Y_i are independent vectors and n_i may or may not be the same for all units i. Associated with the jth measurement on the ith unit is a

$p \times 1$ vector of covariates

$$\underset{p\times 1}{X_{ij}} = \begin{pmatrix} X_{ij1} \\ \vdots \\ X_{ijp} \end{pmatrix}$$

and we write

$$\underset{n_i\times p}{X_i} = \begin{pmatrix} X_{i1}^T \\ \vdots \\ X_{in_i}^T \end{pmatrix}.$$

In the design matrix X_i, the rows correspond to the covariates associated with the different times of measurement, and the columns correspond to the different variables. Thus each subject has a vector of outcomes, Y_i, and a matrix of covariates, X_i. In the setting where j indexes occasions of measurements, the covariate X_{ij} may include functions of explanatory variables measured at or prior to time j. There are two main classifications for the type of covariates found in the analysis of correlated data.

a. Classification according to whether the values taken by a covariate associated with measurements on the same unit remain constant or differ.

   a.1. We say that the kth covariate, $1 \leq k \leq p$, is between-cluster or time-invariant if for all $i = 1, \ldots, N$,

   $$X_{i1k} = X_{i2k} = \cdots = X_{in_ik}$$

   Examples include sex and race in a longitudinal study where individuals are the basic sampling unit, and fixed experimental conditions such as treatment assignment in a longitudinal clinical trial. In the cluster sampling setting, an example would be cluster level covariates, e.g., family income, or disease status of the proband in family studies.

   a.2. We say that the kth covariate, $1 \leq k \leq p$, is within-cluster or time-varying if for some $i = 1, \ldots, N$, $X_{ijk} \neq X_{ij'k}$ for at least one pair j, j' and $j \neq j'$. Examples include time since baseline, experimental condition in crossover or repeated measure designs, current smoking status, smoking history, current height in a longitudinal study, or individual characteristics in a clustered sample survey. In some cases (pure repeated measures designs, or longitudinal studies with fixed time points),

these covariates vary systematically in the same way for each subject so that $X_{ijk} = X_{i'jk}$ for all i and j, for fixed k.

b. Classification according to whether the variable is fixed by design or stochastic

b.1. Covariates which vary systematically over the course of the study but which are fixed by design include treatment group indicators in a crossover design, time since baseline, or individual relationships in a family study.

b.2. Covariates which vary over time and are stochastic include height, current smoking status, or pollution exposure in a longitudinal survey.

Model Formulation

We shall make the following assumptions in formulating a model for relating the response Y_i to the covariates X_i:

(a). $(Y_1, X_1), \ldots, (Y_N, X_N)$ are independently distributed.

(b). Given X_i,

$$E\underset{n_i \times 1}{(Y_i)} = \underset{n_i \times p}{X_i}\ \underset{p \times 1}{\beta}.$$

(c). Given X_i,

$$\text{cov}\,(Y_i) = \underset{n_i \times n_i}{\Sigma_i},$$

where Σ_i is some known function of the covariates X_i.

Without loss of generality we consider X_i random. Thus, when the jth covariate is fixed by design, X_{ij} takes a fixed value x_{ij} with probability 1. In (b) and (c) and throughout this monograph, in a slight abuse of notation, when X_i is stochastic we eliminate the conditioning covariate X_i in the notation for the conditional mean and covariance of Y_i given X_i. Similarly, when X_i is fixed by design, $E(Y_i)$ and $\text{cov}(Y_i)$ refer to the mean and covariance of Y_i taken with respect to the law $f_i(y)$ of Y_i. Assumption (a) says that the sample consists of N independently selected units. Assumption (b) says that the conditional mean of the jth outcome of unit i given $X_{i1}, \ldots, X_{in_i}$ is a linear function of X_{ij} only, i.e.,

$$\begin{aligned} E(Y_{ij} \mid X_i) &= X_{ij}^T \beta \\ &= \beta_1 X_{ij_1} + \cdots + \beta_p X_{ij_p}. \end{aligned}$$

Assumption (c), cov $(Y_i) = \Sigma_i$, allows for dependencies among measurements on the same unit. The covariance may vary with covariates, e.g., across groups (sex, race), or entries may be functions of time. It may also depend upon i only through n_i. In later sections we will be more explicit about different forms for Σ_i; for now, we leave it very general.

There are important and often overlooked limitations related to the interpretability of the regression model assumed in (b) when covariates are stochastic and time-varying. Regression models for the marginal mean of Y_{ij} as assumed in (b) may be used to answer public-policy related questions for the dependence of the outcome Y_{ij} on X_{ij}. The X_{ij} are defined very generally, and may be functions of explanatory variables $Z_{i1}, \ldots, Z_{ij}$ measured at or prior to time j. When the Z_{ij} are stochastic, and the objective is to infer a causal relationship between changes in Z_{ij} and the outcome Y_{ij} there are important issues regarding the validity and interpretability of the model that we now briefly discuss.

1. Validity: Assumption (b) implies that the conditional expectation of Y_{ij} given the entire covariate process $X_{i1}, \ldots, X_{in_i}$ observed on the ith unit is a function only of the covariate X_{ij}. This assumption may be violated in settings where, conditional on X_{ij}, the current value of the outcome Y_{ij} predicts the subsequent value of the covariate $X_{i(j+1)}$. This is so since the conditional dependence of Y_{ij} and $X_{i(j+1)}$ given X_{ij} may violate the condition

$$E\left(Y_{ij}|X_{ij}, X_{i(j+1)}\right) = E\left(Y_{ij}|X_{ij}\right) \tag{1.1}$$

implied by assumption (b). Equation (1.1) might be violated, for example, in longitudinal studies designed to evaluate the effect of cumulative fat intake at time j on cholesterol level at time j. Letting Z_{ij} denote fat intake of subject i at time j, define $X_{ij} = (Z_{i1}, \ldots, Z_{ij})$, and let Y_{ij} denote the cholesterol level of subject i at time j. With these definitions equation (1.1) might be violated. For example, suppose subjects with high mean cholesterol level at time j who report a high level of fat intake at time j, tend to subsequently reduced their fat intake. If at the same time, subjects with the same level of fat intake, but lower cholesterol at time j tend to report continuously high levels of fat intake, then assumption (b) does not hold.

2. Interpretability: Even if assumption (b) holds, the β parameters may not quantify the parameters of interest in applications. For example, model (b) may correctly specify the dependence of the conditional mean of the cholesterol level measured at the last occasion on the covariates, since at the last occasion, X_{in_i} is a function of the entire fat intake history. Yet, the parameter β may fail to quantify the covariate effect

of practical interest when the covariates are stochastic. To simplify our exposition, suppose that all subjects are measured at two time points so that $n_i = 2$, that Z_{ij} is a dichotomous indicator of a high level of fat intake at time j, $j = 1, 2$ and that at baseline, the study population is homogeneous with respect to simultaneous predictors of subsequent fat intake and cholesterol level. One model satisfying assumption (b) might be

$$E(Y_{i2}|X_{i2}) = \beta_1 + \beta_2 Z_{i2} + \beta_3 Z_{i1} \tag{1.2}$$

where $X_{i2} = (1, Z_{i2}, Z_{i1})$ and $j = 1, 2$. According to model (1.2) the effect on cholesterol at time two of a continuously high fat diet versus a continuously low fat diet is given by

$$E(Y_{i2}|Z_{i1} = 1, Z_{i2} = 1) - E(Y_{i2}|Z_{i1} = 0, Z_{i2} = 0) = \beta_2 + \beta_3.$$

The interpretation of $\beta_2 + \beta_3$, however, depends upon implicit model assumptions. If subjects are randomized to high and low fat diets (and adhere to their specified diets), this determines Z_{i1} and Z_{i2}. Hence Z_{i2} is independent of $Y_{i1}|Z_{i1}$, or in general Z_{ij} is independent of

$$(Y_{i1}, \ldots, Y_{i(j-1)}|Z_{i1}, \ldots, Z_{i(j-1)}).$$

In this setting one can assign the desired causal interpretation to $\beta_2 + \beta_3$. In the setting where the predictors are stochastic, one still needs to make independence assumptions to interpret the results causally.

Specifically, $\beta_2 + \beta_3$ measures the causal effect of Z_{i1} and Z_{i2} on Y_{i2} only when *either* the observed cholesterol level at time 1, Y_{i1}, is not a predictor of the observed cholesterol level at time 2, conditional on fat intake history (Z_{i1}, Z_{i2}), i.e.,

$$Y_{i2} \amalg Y_{i1}|Z_{i1}, Z_{i2}, \tag{1.3}$$

or when the observed cholesterol level at time 1 is not a predictor of subsequent fat intake conditional on fat intake at time 1, i.e.,

$$Z_{i2} \amalg Y_{i1}|Z_{i1}. \tag{1.4}$$

Here $X \amalg Y|W$ is used to indicate that X and Y are conditionally independent given W. When (1.4) holds, or in general when

$$Z_{ij} \amalg \left(Y_{i1}, Y_{i2}, \ldots, Y_{i(j-1)}\right)|Z_{i1}, Z_{i2}, .., Z_{i(j-1)} \tag{1.5}$$

is true, the time dependent covariate process Z_{ij} is called an external covariate process (Kalbfleisch and Prentice, 1980). Since (1.3) is not in

general a reasonable assumption in a longitudinal setting, assumption (1.5) is the most important in the context of a longitudinal study. An example of such a stochastic predictor might be air pollution levels in a study of lung function growth.

When both (1.3) and (1.4) are false, the parameter β in model (1.2) does not have a causal interpretation even if model (1.2) correctly specifies the dependency of the *actually* observed cholesterol level at time 2 on past fat intake history. This is so because cholesterol level at time 1 (Y_{i1}) is simultaneously a predictor of fat intake at time 2 (Z_{i2}) and an independent risk factor for cholesterol level at time 2 (Y_{i2}). When risk factors for current cholesterol level, such as past cholesterol level, reduce subsequent high fat intake, fat intake specific means of cholesterol level tend to underestimate the true effect of fat intake. For example, the observed cholesterol mean of subjects with an observed continuously high fat intake will be an underestimate of the overall cholesterol mean when all subjects follow a continuously high fat diet, if subjects with high cholesterol level at time 1 tend to reduce their fat intake level more than those with normal cholesterol levels, and if within levels of fat intake, cholesterol levels at time 1 and 2 are correlated.

Study designs

In the context of longitudinal studies we distinguish between *balanced* and *unbalanced* study designs. A design is balanced when all N individuals are to be measured at the same n occasions and it is unbalanced otherwise. Thus in balanced designs, $n_i = n$ so that each Y_i is a $n \times 1$ vector and Σ_i defined in Section 1.2, (c), is a $n \times n$ matrix. We say that a study with a balanced design is *complete* when all n measurements are actually observed on each study participant, and the study is incomplete otherwise. Notice that in an incomplete study with a balanced design the actual number of measurements observed on each subject may be less than or equal n. However, at least conceptually, we could have observed n measurements, $Y_{i1}, \ldots, Y_{in}$, on each subject. In the model formulated in Section 1.2, Y_i is the vector comprised of these, potentially unobserved, outcomes and the matrix Σ_i refers to the covariance of Y_i. If this covariance does not depend on covariates, then it is the same for all the subjects. In this case we eliminate the index i and simply denote it with Σ. We say that the model for the covariance matrix Σ is *unstructured* when its elements are restricted by the condition that Σ be positive definite and symmetric but are otherwise unconstrained. It is often the case that in studies that are initially designed as balanced, measurements are

not made at the same exact n occasions, for example due to misstimed events. In this case, even if the study is complete, i.e., if n measurements are actually observed on each subject, a single unstructured covariance matrix is often inappropriate. For example, if measurements can be assumed to have constant variance, a single unstructured covariance would indicate that the correlation between the tth and $(t+1)$th measurement is the same regardless of the time elapsed between the measurements.

In familial aggregation studies or in cluster sampling designs, where the number of subjects in the cluster varies for each cluster, Σ depends on i because its dimension $n_i \times n_i$ is a function of i. Furthermore, it is often the case that Σ also depends on i because the covariance of the measurements made on each cluster varies with covariates. Often, parsimonious covariance models are postulated making the elements of Σ_i depend on a small number of cluster-specific covariates and subject specific covariates, e.g., parent or sibling. In such cases we say that Σ_i is *structured.* We will later consider analysis strategies for both structured and unstructured covariance matrices.

Finally, note that the linear model can be written as

$$\underset{M\times 1}{E\begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix}} = \underset{M\times p}{\begin{pmatrix} X_1 \\ \vdots \\ X_N \end{pmatrix}} \underset{p\times 1}{\beta}$$

or

$$\underset{M\times 1}{E\,(Y)} = \underset{M\times p}{X} \quad \underset{p\times 1}{\beta}$$

and

$$\underset{M\times M}{\operatorname{cov}\,(Y)} = \begin{pmatrix} \underset{n_1\times n_1}{\Sigma_1} & 0 & \dots & 0 \\ & \underset{n_2\times n_2}{\Sigma_2} & & 0 \\ & & \ddots & \vdots \\ & & & \Sigma_N \end{pmatrix}, \qquad \text{(Block diagonal matrix)}$$

where $M = \Sigma n_i$. In Chapters 3 through 5 we will additionally assume that, given X_i, Y_i has a multivariate normal distribution. We now consider the formulation of X_i for some specific cases.

1.2 Examples and Special Cases of the Linear Model for Correlated Data

In this section we will show that many "classical" multivariate models are special cases of the LMCD for appropriate choices of X_i, β and Σ_i. We also discuss application in some nonstandard settings.

1. One sample repeated measures.

N subjects are measured repeatedly under n different experimental conditions. The goal is to quantify differences in experimental conditions. The model assumes

$$E\underset{n\times 1}{(Y_i)} = \underset{n\times 1}{\mu} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ & & \ddots & \\ & & & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix}.$$

Here $X_i = I$ (where I denotes the identity matrix) and $\beta = \mu$. In this case, contrasts among the μ_j's are the differences in the outcome means for the different experimental conditions, and as such are of primary interest. Often Σ is left unstructured.

A popular alternative to leaving Σ unstructured is to assume compound symmetry. Under the additional assumption of normal errors, or randomization to the experimental conditions the data can then be analyzed using simple univariate ANOVA methods. We say that Σ has a compound symmetry structure if it can be written as

$$\begin{aligned} \Sigma &= \sigma_\epsilon^2 I + \sigma_\gamma^2 \, 1 \, 1^T \\ &= \begin{pmatrix} \sigma_\epsilon^2 + \sigma_\gamma^2 & \sigma_\gamma^2 & \cdots & \sigma_\gamma^2 \\ & \ddots & & \\ & & \ddots & \\ & & & \sigma_\epsilon^2 + \sigma_\gamma^2 \end{pmatrix}. \end{aligned} \tag{1.6}$$

where 1 denotes an $n \times 1$ vector of one's. Compound symmetry arises by considering the model

$$Y_{ij} = \mu_j + \gamma_i + \epsilon_{ij} \qquad \begin{matrix} j = 1, \ldots, n \\ i = 1, \ldots, N \end{matrix}$$

where the γ_i's and the ϵ_{ij}'s are independent of each other, with var $(\gamma_i) = \sigma_\gamma^2$ and var $(\epsilon_{ij}) = \sigma_\epsilon^2$. This implies that with $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{in})^T$

$$Y_i = \mu + \gamma_i \, 1 + \epsilon_i$$

and hence $\text{cov}(Y_i)$ satisfies (1.6).

2. One way multivariate ANOVA (MANOVA)

We assume G treatment groups, and n measurements are obtained for each of N_g subjects in treatment group g, $g = 1, \ldots, G$. Our goal is to test if the mean vector is the same for all G groups. Letting

$$\underset{n\times 1}{E(Y_{gi})} = \underset{n\times 1}{\mu_g} \qquad \begin{matrix} g = 1, \ldots, G \\ i = 1, \ldots, N_g \end{matrix},$$

we can write

$$E(Y_{gi}) = \underbrace{\underset{n\times n}{[0, \ldots, 0, I, 0, \ldots, 0]}}_{n\times(nG)} \underbrace{\begin{bmatrix} \mu_1 \\ \vdots \\ \mu_g \\ \vdots \\ \mu_G \end{bmatrix}}_{(nG)\times 1}$$

(gth block)

Notice that this model can also be written as

$$E(Y_{gi}) = (0, \ldots, 1, 0, \ldots, 0) \otimes I \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_G \end{bmatrix}$$

$$= a_{gi}^T \otimes I \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_G \end{bmatrix}$$

where a_{gi} is the indicator vector for the gth group and $\otimes$ is the Kronecker or direct matrix product (Harville, 1999). Here the number of free parameters under the null hypothesis

$$H_0 : \mu_1 = \cdots = \mu_G$$

is equal to n. Under the alternative hypothesis H_1 that at least two groups have unequal mean vectors, i.e.,

$$H_1 : \text{there exist } g_1 \text{ and } g_2 \text{ such that } \mu_{g_1} \neq \mu_{g_2}$$

the number of free parameters p satisfies $n < p \leq nG$. The usual MANOVA assumes $\Sigma_i = \Sigma$ is unstructured and normality of the error terms.

3. One group polynomial growth curve model.

N subjects from the same cohort are observed at the same times denoted by $t_1, \ldots, t_n$; for example, a group of children is observed yearly at ages $6, 7, 8, \ldots, 12$. The linear model for the mean response is expressed as a polynomial in t_j. We might assume a quadratic polynomial

$$E(Y_{ij}) = \beta_0 + \beta_1 t_j + \beta_2 t_j^2,$$

and

$$\underset{n\times 1}{E(Y_i)} = \begin{pmatrix} 1 & t_1 & t_1^2 \\ 1 & t_2 & t_2^2 \\ \vdots & \vdots & \vdots \\ 1 & t_n & t_n^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = X\beta.$$

Hence $X_i = X$ and is in general an $n \times q$ matrix of the form

$$X = \begin{pmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^{q-1} \\ 1 & t_2 & t_2^2 & \cdots & t_2^{q-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & t_n & t_n^2 & \cdots & t_n^{q-1} \end{pmatrix}.$$

One model for Σ_i arises from assuming that each subject has his or her own growth curve with parameters β_i:

$$E(Y_i|\beta_i) = X\,\beta_i \text{ and } \operatorname{var}(Y_i|\beta_i) = \sigma^2 I.$$

Thus each subject has their own vector of growth curve parameters $\beta_i = (\beta_{i0}, \beta_{i1}, \beta_{i2})$, where the β_i are themselves considered random with $E(\beta_i) = \beta$ and var $(\beta_i) = D$. Averaging over the β_i's, we have

$$E(Y_i) \;=\; X\beta$$

as before, but now Σ has the variance component type structure

$$\operatorname{var}(Y_i) = \operatorname{var}\,[E(Y_i|\beta_i)] + E\,[\operatorname{var}(Y_i|\beta_i)] = X\,D\,X^T + \sigma^2 I.$$

4. Growth curve with covariates.

As in Example 3, suppose that N subjects are observed at the same n occasions. The model is motivated as follows. Suppose that each subject follows his/her own growth curve indexed by a $q \times 1$ parameter vector β_i. Specifically, for subject i we assume that

$$\underset{n\times 1}{E(Y_i)} = Z\beta_i$$

where the $n \times q$ matrix Z gives the polynomial design on time, e.g., Z is X in the previous example, and the expectation is taken with respect to the law of Y_i, $f_i(y)$. Furthermore, we assume that the covariance of Y_i is the same for all subjects but it has an arbitrary structure, i.e.,

$$\operatorname{cov}(Y_i) = \underset{n \times n}{G}$$

where G is unstructured.

Suppose initially that β_i is a deterministic function of a $q \times p$ time invariant covariate matrix A_i. Specifically, suppose that β_i varies with A_i according to the relationship

$$\beta_i = A_i \beta$$

where β is a $p \times 1$ unknown parameter vector, so that

$$\underset{n \times 1}{E(Y_i)} = ZA_i\beta = X_i\beta.$$

The jth row of A_i gives the design for the regression of the jth component of β_i on the covariates. For example, suppose that in a study of growth in children we assume that height (the outcome) is a linear function of age ($g = 2$) and further that the intercept is gender specific but the slope is the same for both genders. In this case there is a single time invariant covariate, sex, the row dimension of A_i is equal to 2, and $p = 3$ (two intercepts, but only one slope), so that

$$A_i = \begin{bmatrix} 1 & S_i & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

where $S_i = 1$ for boys and $S_i = 0$ for girls, and $\beta = (\beta_1, \beta_2, \beta_3)$. Under this formulation, β_1 represents the girl's intercept, $\beta_1 + \beta_2$ represents the boys intercept and β_3 is the common slope. Letting $X_i = ZA_i$, we can write

$$E(Y_i) = X_i\beta$$

and $\operatorname{cov}(Y_i) = G$ is unstructured and unknown.

The growth curve model has a special form in the case where each element of β_i has a regression on the same $k \times 1$ vector of covariates a_i. That is, $\beta_{i1} = a_i^T\beta^{(1)}$, $\beta_{i2} = a_i^T\beta^{(2)}, \ldots, \beta_{iq} = a_i^T\beta^{(q)}$, where the $\beta^{(l)}$ are the regression parameters for the lth coefficient. Thus $\beta = (\beta^{(1)T}, \beta^{(2)T}, \ldots, \beta^{(q)T})^T$ and $p = kq$. In this case we may write $A_i = a_i^T \otimes I$ where I is $q \times q$ so that $\beta_i = A_i\beta$, and $E(Y_i) = Z(a_i^T \otimes I)\beta$. This

is a very special form because $X_i = ZA_i = Z(a_i^T \otimes I) = a_i^T \otimes Z$. We may also write

$$E(Y_i) = X_i\beta = Z\Delta a_i,$$

where $\Delta(q \times k)$ is a matrix of parameters satisfying $\beta = \text{vec}(\Delta^T)$ and $\text{vec}(\Delta^T)$ is a matrix operator which makes a column vector from a matrix by stacking the columns. The result that $X_i\beta$ is equal to $Z\Delta a_i$ follows from the identity $\text{vec}(ABC) = (C^T \otimes A)\text{vec}(B)$ for conformable matrices A, B, and C. The analysis of this highly structured growth model has been discussed by Grizzle and Allen (1969).

The formulation of the growth curve model with covariates is completed by replacing the assumption that β_i is a deterministic function of A_i with the assumption that β_i is random, it fluctuates around $A_i\beta$ and it has covariance that does not depend on A_i. Specifically, we assume that β_i, $i = 1, \ldots, N$, are independent and satisfy

$$E(\beta_i) = A_i\beta$$

and

$$\text{var}(\beta_i) = \underset{q\times q}{D}.$$

where, if A_i is stochastic, the expectation and covariance are taken with respect to the conditional law of β_i given A_i, and if A_i is fixed by design, they are taken with respect to the law $f_{\beta i}(b)$ of β_i. Thus, with X_i and β defined as above we have that

$$E(Y_i) = X_i\beta$$

but now

$$\begin{aligned} \text{var}(Y_i) &= E\{\text{var}(Y_i|\beta_i)\} + \text{var}\{E(Y_i|\beta_i)\} \\ &= G + ZDZ^T \end{aligned}$$

Often the model is formulated by additionally imposing that $G = \sigma^2 I$, because leaving $G_{n\times n}$ unstructured, does not permit also estimating an additional $q \times q$ matrix of covariance parameter D.

Suppose now that the study design is unbalanced, and in addition that outcomes are not completely observed on all units in the sample. We can still assume the same growth curve, but with subject-specific design matrices which allow each Y_{ij} to be measured at a unique t_{ij}:

$$E\underset{n_i\times 1}{(Y_i)} = \underset{n_i\times q}{Z_i}\ \underset{q\times 1}{\beta_i}.$$

Here the design Z_i depends on occasions of measurement for the ith subject, but the β_i vector remains the same as in the balanced and complete setting.

As before, we let

$$E(\beta_i) = \underset{q\times p}{A_i}\ \underset{p\times 1}{\beta}, \quad \text{var}\,(Y_i|\beta_i) = \sigma^2\ \underset{n_i\times n_i}{I}, \quad \text{and var}\,\beta_i = D, \quad \text{then}$$

$$E(Y_i) = Z_i A_i \beta = X_i \beta$$

and

$$\text{var}\,(Y_i) = Z_i D Z_i^T + \sigma^2 I.$$

Here Σ_i depends on occasions of measurement through Z_i.

5. Epidemiological survey of lung function development in children.

Studies have shown that growth in lung function can be expressed as a linear function of current age and height (Hopper *et al.*, 1991; Laird *et al.*, 1992). Letting response be repeated measures of log forced expiratory volume in one second (log FEV_1), covariates include current age and current height. The mean of log FEV_1 is approximately linear in age and height, implying that we may write

$$\underset{n_i\times 1}{E(Y_i)} = \begin{bmatrix} 1 & a_{i1} & h_{i1} \\ \vdots & \vdots & \vdots \\ 1 & a_{in_i} & h_{in_i} \end{bmatrix} \beta,$$

where $(a_{ij},\ h_{ij})$ are age and height for subject i at occasion j. The issue here is that not all subjects enter the study at the same time; hence we have both "longitudinal" and "cross-sectional" information about lung function dependence on age and height. This will be discussed in Section 1.4. The model for Σ_i is complicated if $\text{var}(Y_{ij})$ depends on both age and height.

6. Family studies: Hypertension in families.

Many family studies are designed to quantify correlation among relatives of a measured factor, such as blood pressure. When the goal is to quantify the inherited genetic component of the correlation, residual correlations are computed which have been adjusted for known or suspected correlates. For example, suppose we have N nuclear families, where each family member is measured on blood pressure and a variety of covariates including age, sex, weight/height, physical activity, diet (various measures), smoking, alcohol, etc. Here i indexes family and j indexes family member. We include all known risk factors in the mean model and then

study residual correlation. The specification of Σ_i is complex for specific genetic models and depends on the gene sharing among relatives. As a simple example consider an "association" study restricted to families with both parents and three children. If we assume that: (a) the covariance is independent of the covariates, (b) blood pressure variability is similar in males and females but differs between children and adults, (c) the blood pressure correlation between children and adults is the same for all family members, and (d) the correlation between siblings in the same family is the same, then we have the following simple structure for the covariance matrix

$$\begin{array}{c} \\ M \\ F \\ C_1 \\ C_2 \\ C_3 \end{array} \begin{array}{c} \begin{array}{ccccc} M & F & C_1 & C_2 & C_3 \end{array} \\ \begin{bmatrix} \sigma_A^2 & & & & \\ \sigma_{MF} & \sigma_A^2 & & & \\ \sigma_{CP} & \sigma_{CP} & \sigma_C^2 & & \\ \sigma_{CP} & \sigma_{CP} & \sigma_{CC} & \sigma_C^2 & \\ \sigma_{CP} & \sigma_{CP} & \sigma_{CC} & \sigma_{CC} & \sigma_C^2 \end{bmatrix} \end{array},$$

where σ_A^2 and σ_C^2 are adult and child blood pressure variance; σ_{CP} is covariance between parent and child, etc. The components of Σ_i are of primary interest, especially the correlations between two siblings and a parent-child combination. In the absence of environmental causes for correlation, σ_{MF} should be approximately zero.

1.3 Models for the Variance/Covariance Matrix

In many cases, the primary focus of a study is change in mean response, which is modeled by $X_i\beta$; the parameters indexing Σ_i are nuisance, or possibly of secondary interest. In this setting, if all subjects are measured at the same n occasions and the covariance is assumed not to depend on covariates, $\Sigma_i = \Sigma$ may be taken as an arbitrary $n \times n$ positive definite matrix, and, provided n is small relative to N, the entries of Σ can be estimated with reasonable precision. If n is large relative to N, or if the design is inherently unbalanced, i.e., clustered or with subjects measured at arbitrary time points, then it may be desirable to impose some parametric model on Σ_i.

Structure is usually built into covariance matrices in one of two general ways: using serial correlation models, and using random effects models. The models considered in Examples 3 and 4 of Section 1.2 were particular random effects models. In Chapter 5 we consider random effects

models in some detail. Here we consider structures for Σ based on serial correlation.

1. *Banded.* Here

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{n-1} \\ & 1 & \rho_1 & \cdots & \rho_{n-2} \\ & & & \ddots & \vdots \\ & & & & 1 \end{bmatrix}.$$

This formulation is often appropriate when measurements are equally spaced, so that

$$\text{corr}\,(Y_{ij}, Y_{ij+k}) = \rho_k \qquad \text{for all } j \text{ and } k.$$

Notice that the formulation also implies constant variance. A special case of a banded covariance is the autoregressive covariance.

2. *Autoregressive.* Here

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ & 1 & \rho & \cdots & \rho^{n-2} \\ & & & \ddots & \vdots \\ & & & & 1 \end{bmatrix}.$$

This formulation also assumes equal spacing. The autoregressive model can be generalized to accommodate unequal spacing and number of observations as follows: let $(t_{i1}, \ldots, t_{in_i})$ denote the observation times for the ith subject and let Σ_i be the $n_i \times n_i$ covariance matrix. Let var (Y_{ij}) be constant, and assume that

$$\text{corr}\,(Y_{ij}, Y_{ik}) = h\,(|t_{ij} - t_{ik}|),$$

where $h\,(x)$ is a decreasing function of x that takes values between -1 and 1. Since repeated measurements in longitudinal studies are assumed to be positively correlated, a convenient form for $h\,(\cdot)$ is

$$h\,(d_{ijk}) = \exp\{-\alpha d_{ijk}^b\} \tag{1.7}$$

where $d_{ijk} = |t_{ij} - t_{ik}|$, $b = 1$ or 2, and $\alpha > 0$. This presumes that the correlation is one if measurements are made repeatedly at the same time, and goes to zero rapidly if α is large. Neither of these features is especially realistic for the most common attributes of interest that are measured repeatedly in human subjects longitudinal studies.

Diggle (1988) proposed a model which combines the autoregressive with compound symmetry:

$$\Sigma_i = \sigma^2 I + \delta^2 J + \tau^2 \Omega_i$$

where J is an $n_i \times n_i$ matrix of ones, and Ω_i is a correlation matrix with the jkth element given by $h(d_{ijk})$ as in (1.7), so that

$$\text{var}\,(Y_{ij}) = \sigma^2 + \delta^2 + \tau^2$$

and

$$\rho_{ijk} = \text{corr}\,(Y_{ij}, Y_{ik}) = \frac{\delta^2 + \tau^2\, h(d_{ijk})}{\sigma^2 + \delta^2 + \tau^2}.$$

Notice that this implies that

$$\rho_{ijk} \to \frac{\delta^2 + \tau^2}{\sigma^2 + \delta^2 + \tau^2} \text{ as } d_{ijk} \to 0,$$

and

$$\rho_{ijk} \to \frac{\delta^2}{\sigma^2 + \delta^2 + \tau^2} \text{ as } d_{ijk} \to \infty.$$

Here σ^2 can be thought of as sampling variability or measurement error; the variance has three components: sampling, subject to subject, and serial. Repeated measurements made at the same occasions are not perfectly correlated unless $\sigma^2 = 0$ (no sampling error), and provided $\delta^2 > 0$ (subject variability), observations on same subject never go to zero, even if widely separated in time.

1.4 Cross-sectional versus Longitudinal Effects

The simple growth curve model of Example 1.2.4 assumes that individuals are measured at the same occasions in time and have the same age at baseline. This is rarely the case in observational studies. Often, subjects enter the study at different ages and have their measurements taken at different points in time. These unbalanced designs provide the opportunity to obtain information on differences due to aging as well as differences in the response variable across cohorts. However, care must be taken in specifying models and interpreting the results of the analysis because the parameters of the model may not reflect the aging effect as intended.

As a simple example of this situation consider a study in which measurements of an outcome of interest are taken at prespecified, equally

spaced, time points $t_1, t_2 = t_1 + t, \ldots, t_n = t_1 + (n-1)t$, on each of N independent subjects. Suppose first that a single age-cohort is followed-up so that age of entry to the study is the same, say a, for all subjects. The data then consists of independent vectors $(Y_{i1}, \ldots, Y_{in})$, $i = 1, \ldots, N$, where Y_{ij}, $j = 1, \ldots, n$, is the outcome of the ith subject at age $x_j = a + (j-1)t$. Suppose that it is assumed that the mean increases linearly with age, or equivalently with x_j, so that we may write

$$E(Y_{ij}) = \mu + \Delta x_j, \qquad j = 1, \ldots, n.$$

Then we have the LMCD model

$$E\begin{bmatrix} Y_{i1} \\ Y_{i2} \\ \cdot \\ \cdot \\ Y_{in} \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & x_n \end{bmatrix} \begin{pmatrix} \mu \\ \Delta \end{pmatrix} \tag{1.8}$$

Clearly, this model implies

$$E\left(Y_{ij} - Y_{i(j-1)}\right) = \Delta, \qquad \text{for } j = 2, \ldots, n. \tag{1.9}$$

Thus, Δ is the mean outcome change between contiguous time points. It is therefore a measure of the effect of aging; the so called "*longitudinal or intra-subject effect.*"

Suppose now that n age-cohorts are followed according to the following design. As before, measurements are taken only at the prespecified time points $t_1, \ldots, t_n$, and age of entry to the study is the same for all subjects. However, in contrast to the previous design, there is now staggered entry, so that age-cohort 1 enters the study at time t_1, age-cohort 2 enters at time t_2, etc. Even though subjects are the same age at entry to the study, they have different ages at times $t_1, \ldots, t_n$ and differing numbers of measurements. Let Y_{ij} and x_j be defined as before. On subject i we then observe only measurements of $Y_{i1}, Y_{i2}, \ldots, Y_{i(n-C_i+1)}$ where C_i denotes the age-cohort number to which subject i belongs. That is, if $C_i = 1$ we observe $Y_{i1}, \ldots, Y_{in}$ while if $C_i = n$ we observe only Y_{i1}. Note that the index j on Y_{ij} now indicates age of measurement on the ith subject and not occasion of measurement. With only one cohort, age of measurement and occasion of measurement are equivalent, but not with staggered cohorts.

Suppose that for analyzing this unbalanced design we fit model (1.8) where now we omit row j if $j > n - C_i + 1$, i.e., if the observation Y_{ij} is not available for subject i. We now show that the model actually fitted

may not appropriately reflect the actual data generating process and, as such, the parameter Δ may no longer have the intended interpretation as quantifying the within-subject effect of aging.

Because Y_{ij} is observed only if $j \leq n - C_i + 1$, then when we fit model (1.8) to the observed data, the actual fitted model is

$$E\left(Y_{ij} | C_i \leq n - j + 1\right) = \mu + \Delta x_j, \tag{1.10}$$

The conditional expectation is the mean response at age $a + x_j t$ for all subjects in cohorts $1, 2, \ldots, n - j + 1$. Note that the conditioning means the model is true for all subjects in cohorts $C_i = 1, \ldots, n-j+1$. Equation (1.10) implies the following interpretation for Δ:

$$\begin{aligned} \Delta = & E\left(Y_{ij} | C_i \leq n - j + 1\right) \\ & - E\left(Y_{i(j-1)} | C_i \leq n - j + 2\right) \qquad \text{for all } j, \end{aligned} \tag{1.11}$$

which can be re-expressed as

$$\begin{aligned} \Delta = & E\left(Y_{ij} - Y_{i(j-1)} | C_i \leq n - j + 1\right) \\ & + \left\{E\left(Y_{i(j-1)} | C_i \leq n - j + 1\right)\right. \\ & \left. - E\left(Y_{i(j-1)} | C_i = n - j + 2\right)\right\} \end{aligned} \tag{1.12}$$

This last equation illustrates the likely misspecification of model (1.10). The first term in the right hand side, $E(Y_{ij} - Y_{i(j-1)} | C_i \leq n - j + 1)$, does measure the mean intra-subject change between ages x_{j-1} and x_j in cohorts $1, 2, \ldots, n - j + 1$. However, the second term,

$$E\left(Y_{i(j-1)} | C_i \leq n - j + 1\right) - E\left(Y_{i(j-1)} | C_i = n - j + 2\right),$$

is a contrast in the mean response at the single age x_{j-1} between cohorts $1, 2, \ldots, n - j + 1$ and cohort $n - j + 2$. This latter measures "secular" or cross-sectional change because it compares individuals in different cohorts. Because Δ does not change with j, model (1.10) implies that this linear combination remains the same for all j. This will only be reasonable when there are no cohort effects. Thus, unless the mean response at entry age a is the same in all cohorts, the parameter Δ cannot be interpreted as a measure of pure longitudinal change.

The point of our example is to illustrate that using naive extensions of models for estimating pure longitudinal change with balanced designs, such as in our problem, a model obtained by simply omitting the rows of the design matrix corresponding to unobserved outcomes, may result in: a) misspecified models or, b) correctly specified models whose parameters no longer retain their interpretation under balanced designs.

Researchers have approached the problem of isolating longitudinal effects in various ways. In the example described above one strategy would be to restrict inferences to subjects measured on all occasions that is, to base inference on data from a single cohort. This strategy will not induce loss of information about longitudinal change when $n = 2$, but will clearly be inefficient otherwise as subjects with incomplete (but more than one) measurements carry information about intra-individual effects. In addition, there is the obvious problem that by restricting attention to a single cohort no inference about cross-sectional effects can be made.

A second general strategy applicable also to continuous outcomes and more complex designs than the one considered in our example is to incorporate subject specific effects into the model. Specifically, assume that the outcomes Y_{ij}, $j = 1, \ldots, n_i$, $i = 1, \ldots, N$ are independently normally distributed with common (unknown) variance σ^2 and mean

$$\mu_{ij} = b_i + \Delta (X_{ij} - X_{i1}) \tag{1.13}$$

where the b_i's are fixed unknown constants and X_{ij} is the age of subject i when Y_{ij} is measured. Here we allow for the following possibilities:

1. age and calendar time at which the initial outcome measurement Y_{i1} is obtained vary across subjects, so that subjects belong to different age-cohorts;

2. the number of measurements taken on each individual varies, i.e., n_i changes with i;

3. the outcome measurements are taken at unequally spaced time points.

We use capital letters for X_{ij} to stress its stochastic nature under these unbalanced designs. We can extend model (1.13) to allow for non-linear age trends by considering

$$\mu_{ij} = b_i + h (X_{ij} - X_{i1}; \Delta) \tag{1.14}$$

where h is a known function of time from baseline $X_{ij} - X_{i1}$ and an unknown parameter Δ satisfying the constraint of being equal to 0 when $X_{ij} = X_{i1}$. In models (1.13) and (1.14) the values of the b_i's vary with subjects. If the aging process is the same across cohorts then the b_i's absorb the cohorts effects and the parameter Δ reflects pure longitudinal

trend. A difficulty with model (1.14) is that the number of nuisance parameters increases with N, hence maximum likelihood gives inconsistent results for inferences (Neyman and Scott, 1951).

Inference about a parameter of interest (in our case Δ) when the number of nuisance parameters (in our case the b_is) grows at the same rate as the sample size can sometimes be resolved by a factorization of the likelihood. If two factors can be found such that one carries most or all of the information about the parameter of interest and is exactly or approximately free of the nuisance parameters. In such case, inference is based on the informative component of the likelihood. In the special case in which μ_{ij} is linear in X_{ij}, and $f(Y_{ij})$ depends upon b_i, X_{ij} and Δ only via μ_{ij}, reparameterization via orthogonalization of the covariate space yields a satisfactory factorization for inference about Δ. Specifically, letting $\tau_i = b_i + (\overline{X}_i - X_{i1})\Delta$ and $Z_{ij} = X_{ij} - \overline{X}_i$, where $\overline{X}_i = \sum_j X_{ij}/n_i$, the mean model is re-expressed as $\mu_{ij} = \tau_i + Z_{ij}\Delta$. It can be easily checked that the distribution of $\widehat{\tau}_i = \sum_j Y_{ij}/n_i = \overline{Y}_i$ does not depend on Δ and is a sufficient statistic for τ_i when Δ and σ^2 take arbitrary fixed values. Further, the conditional distribution of $Y = \{Y_{ij}; i = 1, .., N, j = 1, \ldots, n_i\}$ given $\overline{Y} = (\overline{Y}_1, \ldots, \overline{Y}_N)$, $f(Y|\overline{Y}; \Delta, \sigma^2)$ say, is free of of the τ_i's and carries all the available information about Δ and σ^2. This approach is indeed identical to conducting inference based on the differences from the individual means $W_{ij} = Y_{ij} - \overline{Y}_i, j = 1, \ldots, n_i - 1$.

In the more general non-linear setting (1.14) an analogous strategy would be to base inference about Δ on the conditional distribution of Y given a sufficient statistic for b_i, say $S_i, i = 1, \ldots, N$. Besides yielding valid asymptotic inference about Δ, this approach has the advantage of not requiring estimation of the b_i's. However, it can be quite inefficient if S_i is informative about Δ. As a final remark, notice that the method does not use the data of subjects with a single observation, since for them, the conditional distribution of Y_i given S_i is constant.

There are two alternative approaches to conduct inference in the presence of a large number of nuisance parameters that we now mention briefly.

One possibility is to regard the b_i's as realizations from a random variable following a parametric distribution. The large number of nuisance parameters is then replaced by the finite number of unknown parameters indexing the distribution of the b_i's. This formulation falls into the class of random effects models. Inference under these models is discussed in detail in Chapter 5.

A second possibility is to assume that the b_i's follow a deterministic

function of cohort. For example, if A_i denotes the age of subject i at a fixed time point t_0, then we may assume that $b_i = \beta_0 + \beta_C A_i$. Under the design with staggered entry previously described this assumption implies that

$$E(Y_{ij}|A_i) = \beta_0 + \beta_C A_i + \Delta(X_{ij} - X_{i1}), \tag{1.15}$$

where we write the conditioning on the random variable A_i to emphasize that the model describes the dependence on time since baseline of the outcome mean of subjects belonging to the same cohort. Further, we can use equation (1.15) to model designs in which the baseline measurement is taken at the same calendar time for all subjects but age of entry to the study varies with subjects (and possibly the subsequent times of measurements also differ across subjects). In this case, X_{i1} is equal to A_i and the model becomes

$$E(Y_{ij}|X_i) = \beta_0 + \beta_C X_{i1} + \Delta(X_{ij} - X_{i1}). \tag{1.16}$$

Equation (1.16) defines an LMCD model with covariates X_{i1} and $X_{ij} - X_{i1}$ associated with the Y_{ij}. Notice that the model implies that

$$E(Y_{i1}|X_i) = \beta_0 + \beta_C X_{i1}$$

and

$$E(Y_{ij} - Y_{i1}|X_i) = \Delta(X_{ij} - X_{i1})$$

Thus, β_C reflects the cross-sectional effect of age since it relates how the mean of the baseline measurement changes with age at baseline, and Δ reflects longitudinal effects, since it relates individual changes in age with changes in the outcome.

As an illustration of this method we consider the study by Ware *et al.* (1990) comparing longitudinal and cross-sectional estimates of decline in pulmonary function. Non smoking adults, ages 25–74 on entry, were examined three times at three year intervals between 1974 and 1983. There were 2,454 white subjects with valid measurements at baseline; 1,973 remained in the study at the first follow-up and 1,713 at the final follow-up. Figure 1.1 represents the qualitative results found.

The cross-sectional curve was obtained by applying ordinary univariate regression methods to the first measurement only, Y_{i1}. The longitudinal curve was obtained by fitting a correlated regression model with outcomes equal to the differences $Y_{i3} - Y_{i2}$ and $Y_{i2} - Y_{i1}$. The cross-sectional curve shows a slower rate of decline. The possible reasons for these results include the following:

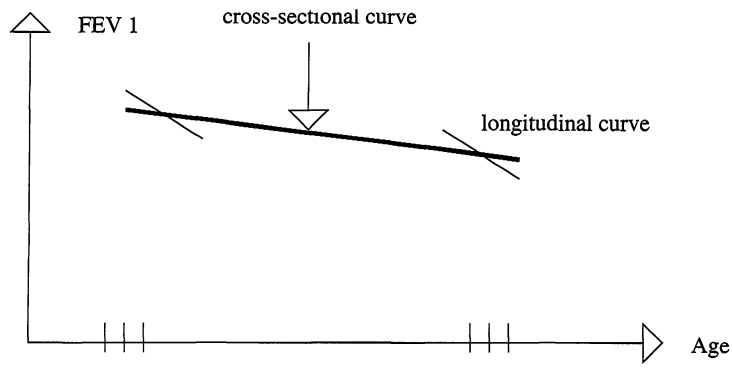

FIGURE 1.1.

- **Cohort effects**:

  Younger cohorts are more prone to respiratory disease because, for example, of higher levels of pollution exposure.

- **Attrition**:

  There might be selection bias, in that at older ages we may only get to see healthier subjects (more subjects with respiratory disease may be dead by age 70 than by age 50).

1.5 Missing Data Issues

The analysis of studies with longitudinal or clustered designs that suffer from non-response poses important methodological challenges. First, we need methods that do not discard units with incomplete measurements but that, instead, efficiently incorporate the observed data on these units. Second, with missing data any analysis method requires important assumptions about the missing data mechanism (MDM) and the usefulness of different methods depends critically on the validity of these assumptions on each specific application. Rubin (1976) and Little and Rubin (1987) provide a classification of the missing data mechanism that is useful for understanding the validity of analysis methods with missing data. To motivate this classification we now provide some examples in the context of the analysis of incomplete correlated outcomes.

An observational longitudinal study with drop-out. The Six Cities Study discussed in Section 1.1 was a longitudinal study of the effects of air pollution on respiratory health. As part of this study, children were followed from age 6 to 18. Lung function was measured with forced expiratory volume (FEV) measurements obtained annually in children in

examinations conducted at the schools. Moving in or out of the school district was a predominant reason for late entry or attrition.

An observational longitudinal study with intermittent non-response. As part of the Muscatine Risk Factor Study (Lauer, 1975) weight and height measurements of five age groups of school children in 1977 (ages 5–7, 7–9, 9–11, 11–13, 13–15) were obtained at three examinations conducted every two years. One of the goals of the study was to determine the prevalence, incidence and risk factors for obesity in this cohort. Parental consent prior to each study wave was a requisite for child participation. The study suffered from substantial non-response. For example, of the 471 children age 11-13 in 1977, only 182 participated in the three study waves (Baker, 1995). Other age groups had similarly high non-response rates (Woolson and Clarke, 1984). Some children had intermittent non-response, i.e., they missed an examination but attended a later one. The two main reasons for non-response were: (1) no parental consent was received by the teachers and, (2) the children were absent from school the day of the examination.

A randomized repeated measures study with intermittent non-response. The International Breast Cancer Study Group (IBCSG) trial VI assessed adjuvant chemotherapy in 1461 patients with node-positive breast cancer. Patients were asked to complete a quality of life (QL) questionnaire. Single-item self-assessment scales measured physical well-being, mood, appetite and perceive adjustment/coping. QL was assessed at the beginning of treatment, 2 months after the start of treatment, every three months, and at 1 and 6 months after recurrence. Hurny *et al.* (1996) reported results of the analysis of QL data for the first 18 months. Patients were excluded from the analysis if they could not be classified in any of the nine culture/language categories. In the first 18 months, 1158 (95%) of the 1221 evaluable patients completed at least one questionnaire, 716 (59%) completed at least six, and 415 (34%) completed all seven.

A randomized repeated measures study with drop out and intermittent non-response. The AIDS Clinical Trial Group Protocol 002 (Fischl *et al.* (1990) compared high vs low dose AZT treatment on the health status of AIDS patients. Measurements of CD4 were to be obtained on 520 subjects at baseline and at clinic visits on weeks $t = 8$, 16, 24 and 32. One objective was to compare the evolution in time of mean CD4 counts in the two treatment arms. The study suffered from drop outs and intermittent non-response. Of the 261 patients in the high dose AZT arm, 135 were drop-outs and 22 had intermittent clinic visits.

For the 259 subjects in the low dose AZT arm, these figures were 107 and 25 respectively.

Missing Data Mechanisms. We will now introduce some notation useful for describing the classification of different missing data mechanisms in the context of studies with correlated outcomes. We will assume that for each of N independent units i, $i = 1, \ldots, N$, outcomes Y_{ij}, and a vector of covariates X_{ij}, $j = 1, \ldots, n$, are to be measured. We will let $Y_i = (Y_{i1}, \ldots, Y_{in})^T$ be the complete-data outcome vector, which may not be fully observed and we will let X_i be the $n \times p$ matrix with rows X_{ij}^T. Throughout we will assume that X_i is always observed and that there is no additional information relevant to predicting the outcome Y_i or non-response. Finally, we will let $R_i = (R_{i1}, \ldots, R_{in})$ where $R_{ij} = 1$ if Y_{ij} is observed and $R_{ij} = 0$ otherwise. We will let $Y_{(r)i}$ denote the subvector of Y_i that is observed when $R_i = r$. Thus, $Y_{(R_i)i}$ is the outcome vector actually observed on the ith unit. We will write $f(Y_i|X_i, \beta, \theta)$ for the density of the $n \times 1$ vector Y_i given X_i, where β is the parameter of interest and θ is a nuisance parameter (finite or infinite dimensional) indexing other components of the joint distribution of Y_i given X_i. We will let $P(R_i|Y_i, X_i, \psi)$ denote the conditional probability function of R_i given (Y_i, X_i) which is indexed by the parameter vector ψ. We are now ready to describe the classification of missing data mechanisms (MDM) as originally introduced by Rubin (1976).

Missing completely at random The outcomes are said to be missing completely at random (MCAR) if

$$P(R_i|Y_i, X_i, \psi) = P(R_i|X_i, \psi); \tag{1.17}$$

i.e., the probability of response is conditionally independent of the full outcome vector (observed and unobserved) given the covariates. As an example, in the Six Cities Study the missing lung function measurements would be MCAR if the reasons for missing a lung function test are unrelated to the respiratory health of the child, as would be the case if the child moved out of the school district because of job relocation of the parents. In contrast, the lung function data would not be MCAR if a family decides to relocate into an area with better air quality because of respiratory problems of the child.

Condition (1.17) is equivalent to the conditional independence of the outcome vector Y_i and the response indicator vector R_i given the covariate vector X_i, i.e.,

$$f(Y_i|R_i = r, X_i; \beta, \theta) = f(Y_i|X_i; \beta, \theta) \tag{1.18}$$

Thus, under MCAR the distribution of Y_i is the same in each subpopulation defined by a specific non-response pattern and in the entire target population consisting of the aggregate of units with complete and incomplete responses. Two important practical implications of this fact are the following.

1. If in (1.18) we take r equal to the vector of ones we obtain that the conditional distribution of the outcome vector is the same in the population of units with complete responses as in the target population. Thus, any analysis method that yields valid inference about β in the absence of missing data will also yield valid inference about β if the outcomes are MCAR and the analysis is conducted based on units with complete responses. The latter is usually refer to as a complete case analysis.

2. Equation (1.18) implies that

$$f\left(Y_{(r)i}|R_i = r, X_i; \beta, \theta\right) = f\left(Y_{(r)i}|X_i; \beta, \theta\right) \qquad (1.19)$$

Thus, the conditional distribution of the observed components of Y_i among units with any non-response pattern coincides with the distribution of the same components of Y_i in the target population. In particular, the first and second moments of $Y_{(R_i)i}$ given the covariates are preserved under the MCAR mechanism. That is, if

$$E\left(Y_i|X_i\right) = \underset{n\times p}{X_i}\ \underset{p\times 1}{\beta} \text{ and } \operatorname{cov}\left(Y_i|X_i\right) = \underset{n\times n}{\Sigma_i},$$

then we may write $Y_{(r)i} = I_i Y_i$, where I_i denotes an $n_i \times n$ matrix obtained by removing from the $n \times n$ identity matrix the jth row if $r_{ij} = 0$, $j = 1, \ldots, n$. Hence

$$E\left(Y_{(r)i}|X_i, R_i\right) = E\left(I_i Y_i|X_i\right) = I_i\, E(Y_i|X_i) = \underset{n_i\times nn\times pp\times 1}{I_i\ X_i\ \beta} = X_{(r)i}\beta$$

and

$$\operatorname{cov}\left(Y_{(r)i}|X_i, R_i\right) = \underset{n_i\times nn\times nn\times n_i}{I_i\ \Sigma_i\ I_i^T} = \Sigma_{(r)i},$$

where $X_{(r)i}$ and $\Sigma_{(r)i}$ denote respectively $I_i X_i$ and $I_i \Sigma_i I_i^T$. Thus, with MCAR outcome data, the appropriate design matrix for the ith subject is simply $X_{(r)i}$ obtained by removing rows of the full data design matrix corresponding to the missing observations. In practice, this has the important implication that standard weighted least squares analyses conducted based on all the available observations of the sampled units yields valid inference about the regression parameter β.

Missing at random The outcomes are said to be missing at random (MAR) if the probability of response is conditionally independent of the unobserved responses given the covariates and the observed responses. That is,

$$P\left(R_i = r|Y_i, X_i; \psi\right) = P\left(R_i = r|Y_{(r)i}, X_i; \psi\right). \tag{1.20}$$

The MAR assumption is attractive because it is less stringent, i.e., it imposes less restrictions on the model for the conditional probability of response given outcomes and covariates, than the MCAR condition. However, an important consequence of the relaxation of assumptions on the model for the missing data mechanism is that the identity (1.19) is no longer guaranteed to be true. This is so because by the Bayes rule,

$$\begin{aligned} &f\left(Y_{(r)i}|R_i = r, X_i; \beta, \theta\right) \\ &\quad = \frac{P\left(R_i = r|Y_{(r)i}, X_i; \psi\right)}{P\left(R_i = r|X_i; \psi\right)} f\left(Y_{(r)i}|X_i; \beta, \theta\right) \end{aligned} \tag{1.21}$$

and the MAR condition does not imply that $P(R_i = r|Y_{(r)i}, X_i; \psi) = P(R_i = r|X_i; \psi)$. This has the important practical implication that complete case analyses as well as least squares analyses based on all available observations no longer yield valid inference about β. However, equation (1.21) implies that the likelihood contribution of unit i can be factored as

$$\begin{aligned} &\prod_r \left\{P\left(R_i = r|X_i; \psi\right) f\left(Y_{(r)i}|R_i = r, X_i; \beta, \theta\right)\right\}^{I(R_i=r)} \\ &\quad = \prod_r \left\{P\left(R_i = r|Y_{(r)i}, X_i; \psi\right)\right\}^{I(R_i=r)} \\ &\qquad \times \prod_r \left\{f\left(Y_{(r)i}|X_i; \beta, \theta\right)\right\}^{I(R_i=r)} \end{aligned} \tag{1.22}$$

where the product ranges over all values taken by r. Thus, if ψ and (β, θ) are variation independent (i.e., the (ψ, β, θ) −parameter space is the Cartesian product of a ψ−parameter space and a (β, θ) −parameter space) and the outcomes are MAR, the likelihood for β is proportional to the likelihood obtained by ignoring the missing data mechanism. When ψ and (β, θ) are variation independent and the outcomes are MAR, the MDM is called *ignorable* (Rubin, 1976). We conclude that under an ignorable MDM, likelihood-based inference does not require the specification of a model for the response probabilities.

The MAR assumption has an easy interpretation when the patterns of non-response are monotone. We say that the non-response patterns

are monotone when

$$R_{ij} = 0 \text{ implies } R_{i(j+1)} = 0 \qquad \text{for any } j = 1, \ldots, n-1. \tag{1.23}$$

Monotone patterns arise, for example, in the context of longitudinal studies as a result of data missing solely due to a drop-out process. Under (1.23) it can be easily shown that the MAR condition (1.20) is equivalent to

$$\begin{aligned} &P\left(R_{ij} = 1 | R_{i(j-1)} = 1, Y_{i1}, \ldots, Y_{in}, X_i\right) \\ &\quad = P\left(R_{ij} = 1 | R_{i(j-1)} = 1, Y_{i1}, \ldots, Y_{i(j-1)}, X_i\right) \\ &\qquad\qquad \text{for } j = 1, \ldots, n. \end{aligned} \tag{1.24}$$

Thus, in the context of longitudinal studies with monotone patterns of non-response, MAR means that the probability of dropping out of the study at each time j is conditionally independent of current and future outcomes given the covariates and the observed history of the outcomes up to but not including time j. For example, in the Six Cities Study described previously the outcomes would be missing at random if the decision to move out of the school district was based only on: (a) the air quality (X_i) of the area of residence, (b) the respiratory health history of the child as measured solely by the past recorded FEV test results and (c) other factors unrelated, i.e., (conditionally) statistically independent of, current and future FEV measurements, such as, for example, parental job relocation. In randomized follow-up clinical studies, such as the ACTG trial described previously, the MAR assumption would hold for example if doctors decide to remove patients from the study based on the recorded history of the health outcome variable of interest. Notice, however, that the MAR assumption would not hold if doctors make their decisions based on other health variables predictive of the outcome of interest that are not available for data analysis.

When the longitudinal study also suffers from intermittent nonresponse, i.e., when the monotonicity condition (1.23) does not hold, an MAR process similar in spirit to (1.24) is given by an assumption that non-response depends only on previously observed outcomes, covariates and non-response pattern, i.e.,

$$\begin{aligned} &P\left(R_{ij} = 1 | R_{i1}, \ldots, R_{i(j-1)}, X_i, Y_i\right) \\ &\quad = P\left(R_{ij} = 1 | R_{i1}, \ldots, R_{i(j-1)}, X_i, R_{i1}Y_{i1}, \ldots, R_{i(j-1)}Y_{i(j-1)}\right) \end{aligned} \tag{1.25}$$

When the covariates X_i are successfully obtained from sources other than the study participants, for example from air pollution measurement stations in the communities where the study participants reside as in the

Six Cities Study, Equation (1.25) says that the decision to return to or miss the next study cycle is based solely on the (a) covariates, (b) the actual cycles previously missed and, (c) the outcomes measured at the non-missed prior study cycles. For example, in the context of a randomized clinical follow-up study, assumption (1.25) would hold if X_i recorded treatment arm assignment and baseline patient characteristics and the decision of doctors to temporarily remove patients from the study or to ask patients to return to the study was based solely on measurements of the health outcome of interest that were recorded while the patient was on the study.

Robins and Gill (1997) showed that in contrast to the monotone data case, (1.25) implies MAR but MAR does not imply (1.25). These authors further show that the condition (1.25) imposes restrictions on the law of the observables and it is therefore subject to empirical verification. Unfortunately, as we illustrate below, MAR processes that do not satisfy (1.25) have the unattractive feature that they do not marginalize, that is, unless (1.25) holds,

$$P\left(R_i = r|Y_i\right) = P\left(R_i = r|Y_{(r)i}\right)$$

does not imply

$$P\left(\widetilde{R}_i = \widetilde{r}|\widetilde{Y}_i\right) = P\left(\widetilde{R}_i = \widetilde{r}|\widetilde{Y}_{(\widetilde{r})i}\right)$$

where $\widetilde{R}$, $\widetilde{r}$ and $\widetilde{Y}$ denote the $n-1$ dimensional subvectors of R, r and Y obtained by removal of an arbitrary component. The important practical message of the lack of marginalization is the limitation of the MAR assumption with intermittent non-response when (1.25) does not hold. This is so because any reasonable assumption about the missing data mechanism should not depend on the design, i.e., on the number of observations that are to be potentially taken on each unit. This discussion indicates that when the study suffers from intermittent non-response, practitioners prepared to assume that the data are MAR should routinely test the null hypothesis that condition (1.25) holds and if rejected, they should discount the possibility that the data are MAR.

We now provide a simple example to show that the assumption of MAR does not marginalize when (1.25) is not true. Suppose that $Y_i = (Y_{1i}, Y_{2i})$ is a bivariate vector of binary outcomes and for simplicity suppose that covariates X_i are not relevant. Suppose that $P(R_{1i} = r_1, R_{2i} = r_2|Y_{1i} = y_1, Y_{2i} = y_2)$ is given by the corresponding entries of the follow-

ing table:

		$(r_1, r_2) = (0,0)$	$(r_1, r_2) = (0,1)$	$(r_1, r_2) = (1,0)$	$(r_1, r_2) = (1,1)$
(y_1, y_2)	$(0,0)$	0.2	0.1	0.3	0.4
	$(0,1)$	0.2	0.2	0.3	0.3
	$(1,0)$	0.2	0.1	0.5	0.2
	$(1,1)$	0.2	0.2	0.5	0.1

Then it can be easily checked that

$$P\left(R_{i1} = 1, R_{i2} = 0 | Y_{1i}, Y_{2i}\right) = P\left(R_{i1} = 1, R_{i2} = 0 | Y_{1i}\right), \tag{1.26}$$

$$P\left(R_{i1} = 0, R_{i2} = 1 | Y_{1i}, Y_{2i}\right) = P\left(R_{i1} = 0, R_{i2} = 1 | Y_{2i}\right), \tag{1.27}$$

and

$$P\left(R_{i1} = 0, R_{i2} = 0 | Y_{1i}, Y_{2i}\right) = P\left(R_{i1} = 0, R_{i2} = 0\right). \tag{1.28}$$

Thus, the data are MAR. However, it follows from the table that

$$P\left(R_{i1} = 0 | Y_{1i} = 0\right) = 0.3P\left(Y_{2i} = 0 | Y_{1i} = 0\right) + 0.4P\left(Y_{2i} = 1 | Y_{1i} = 0\right)$$

and

$$P\left(R_{i1} = 0 | Y_{1i} = 1\right) = 0.3P\left(Y_{2i} = 0 | Y_{1i} = 1\right) + 0.4P\left(Y_{2i} = 1 | Y_{1i} = 1\right),$$

so that $P(R_{i1} = 0)$ depends upon Y_{i1} unless Y_{i2} and Y_{i1} are independent. Hence, (1.26), (1.27) and (1.28) in general do not imply that MAR holds marginally, i.e., in general,

$$P\left(R_{i1} = 0 | Y_{1i}\right) \neq P(R_{i1} = 0),$$

even though $P(R_{i1} = 0, R_{i2} | Y_{i1}, Y_{i2})$ depends only on Y_{i2}.

Non-ignorable. The MDM is non-ignorable when the conditional probability of response depends both on observed and unobserved outcomes i.e., when (1.20) is not true. For example, in the Muscatine Risk Factor Study described previously, the MDM may be non-ignorable because of the possibility of non-response being associated with (unovserved) obesity. In younger children, parents of obese children may be more likely than parents of non-obese children to give consent for participation because of their concern with the health risks associated with obesity. On the other hand, obese preadolescents may be more embarrassed than the non-obese over the examination and more likely to be absent from school on the examination day. In the IBCSG quality of life study, non-ignorable non-response is suspected because patients with worse quality of life may be less likely to want or to be able to fill in a questionnaire. In the ACTG trial 002 non-ignorable non-response

is suspected if the possibility exists that doctors' decision to remove patients from study is based not only on the recorded health outcomes from past clinic visits but also on other patient characteristics predictive of prognosis and not available in the data base. Non-response would also be non-ignorable if subjects whose health had markedly deteriorated between visits j and $j+1$ are more likely than others to either miss the clinic visit at week $j+1$ or to never return for a visit after week j.

Important technical and philosophical considerations arise when the MAR assumption is in doubt. Technically, if we are no longer prepared to assume that the outcomes are MAR, we can no longer factorize the individual likelihood contributions as in (1.22). This has the important implication that likelihood-based inference (and indeed any valid inferential procedure) about β under non-ignorable non-response requires the specification of a model for the conditional probability of response. However, as we shall argue in Chapter 9, identification problems arise when fitting richly parameterized models for the non-response probabilities. Philosophically, the values of the parameters indexing the non-response probabilities determine the process of self selection of the non-respondents. Without external information, the data on the respondents only cannot provide any evidence about the self selection process. The identifiability difficulties encountered when fitting non-ignorable non-response models are the mathematical reflection of this paradigm. We will return to these issues in Chapter 9.

In our discussion so far we have assumed that the study design calls for the collection only of data on Y_i and X_i. However, it is often the case that additional time dependent variables V_{ij}, $j = 1, \ldots, n$, not included in the regression model of scientific interest are also recorded and available for data analysis. For example, in the ACTG trial 002 described previously measurements were obtained at each clinic visit not only on CD4 counts (the outcome of interest) but also on other health variables such as white blood cell counts and the occurrence of pneumonia episodes between study cycles. Notice that the regression model of interest does not include as covariates the variables V_i because these variables are in the causal pathway between X_i (treatment) and the outcomes $\mathrm{CD4}_{ij}$ and hence their inclusion would lead to over-adjustment. On occasions, we may be prepared to assume that conditional on the covariates and on past measurements of Y_{ij} and V_{ij}, the probability of response at each time j is independent of current and future values of the outcomes. That

is,

$$P\left(R_{ij}=1|R_{i1},\ldots,R_{i(j-1)},X_i,Y_i,V_i\right)$$
$$=P\left(R_{ij}=1|R_{i1},\ldots,R_{i(j-1)},X_i,R_{i1}Y_{i1},R_{i1}V_{i1},\ldots,R_{i(j-1)}Y_{i(j-1)},R_{i(j-1)}V_{i(j-1)}\right). \quad (1.29)$$

For example, this would be the case in the ACTG trial 002 if a patient misses a study cycle only by indication of the doctor and the doctor's indication is determined solely by the recorded history of Y_{ij} and V_{ij} on the patient.

When, as in the ACTG trial 002, the scientific goal remains the estimation of the parameter β indexing the conditional mean of Y_i on X_i, the analysis presents important methodological challenges if one is not prepared to assume that the more stringent condition (1.25) holds in addition to (1.29). Specifically, as argued in Chapter 9, likelihood based inference about β (and in particular tests comparing the CD4 count means in the two treatment arms) requires the specification of a parametric model for the conditional distribution of the additional variables V_i given the outcomes Y_i and the covariates X_i and can be non-robust to misspecification of this model. In Chapter 9 we discuss likelihood based inference in this setting in more detail.

The methodological difficulties arising when (1.29) holds but (1.25) is not true can be overcome to some extend if the scientific model of interest can be modified to include as covariates the V_i's, because then the outcomes are MAR and likelihood based inference depends only on the correct specification of the model for the conditional law $f(Y_i|X_i,V_i)$. Modification of the scientific model of interest to include the V_is would perhaps be reasonable in certain studies in which the purpose of the analysis is entirely descriptive. For example, consider a survey of lung function in adolescents. We are interested in relating lung function to age and sex to construct norms. But suppose smokers are less likely to participate in the survey. Because smoking is known to be predictive of lung function in adolescents even after stratification on age and sex, an analysis that does not adjust for the differential response rates between smokers and non-smokers would yield biased estimates of the age-sex-specific distribution of lung function in the targeted study population. However, if non-response is independent of lung function after stratifying on age, sex and past smoking status, then (as argued in our discussion of the MAR mechanism) the age-sex-smoking history-specific lung function distribution among respondents is the same as the corresponding strata specific distribution in the targeted population. Thus, in particular, standard regression analyses based on all available observations will

result in valid inference about the age-sex-smoking history-specific, but not about the age-sex-specific, lung function means.