of realistic and difficult data structures and models commonly encountered in biostatistical and epidemiologic practice. It is our belief that a reader who has mastered the techniques described in this book will be ready to attack the many possible variations of the examples covered in the book.

1.2 Tour through the General Estimation Problem.

In both observational and experimantal studies, the full (equivalently, complete) data structure X that one would wish to collect is often incompletely observed on some, possibly all, subjects. In such cases, we say that the study data is subject to censoring or missingness. As an example, in a study of a cohort of HIV infected subjects, the full data X on a subject might consist of the time from seroconversion to the development of AIDS, the time from seroconversion to death, and the time-dependent covariate processes encoding a subjects CD4 lymphocyte count, viral load, and antiviral treatment history from seroconversion to death. Due to the finite duration of the study, to limitations of funds and resources, and/or to the logistical impossibility of performing hourly or even daily laboratory tests, X is only partially observed. We denote a unit or subject's full data structure with random variable X which may be incompletely observed. Rather we observe the random variable

$$Y = \Phi(X, C) \text{ for a known many-to-one-mapping } \Phi, \tag{1.1}$$

where Φ is a known function and C is the censoring or missingness variable that determines what part of X is observed. The following examples should help clarify the notation.

Example 1.1 (Repeated measures data with missing covariate) Let Z be a p-dimensional vector of outcomes and let E be a vector of accurately measured exposures based on blood tests. Let V be a vector of variables that one wants to adjust for in a regression model (such as confounding factors for the causal effect of E on Z). Our goal is to estimate the regression parameters $\alpha = (\alpha_1, \ldots, \alpha_k)$ in a model for the conditional mean of Z, given $X^* = (E, V)$,

$$Z = g(X^* \mid \alpha) + \epsilon,\ E(\epsilon \mid X^*) = 0, \tag{1.2}$$

where ϵ is a p-dimensional vector of residuals and $g(X^* \mid \alpha) = (g_1(X^* \mid \alpha), \ldots, g_p(X^* \mid \alpha))$ is a known function and α is an unknown parameter to be estimated.

Let E^* be a vector of poorly measured surrogates for E obtained on each subject from questionnaire responses. In general one would not wish to adjust for these surrogates E^* in the regression model (1.2) because of the possibility of differential misclassification (i.e., the possibility that E^* and Z may be conditionally dependent given E and V). However if it is very expensive to measure E it may be feasible to obtain data on E only on a subset of the study subjects. In that case, data on on E^* may be useful either to explain informative missingness and/or to recover information from the censored observations (i.e., from the observations lacking data on E) Let Δ be the indicator that E is observed; Then $C = \Delta$ and $Y = \Phi(X, C) = (C, CX + (1 - C)W)$, where $W = (Z, V, E^*)$.

Example 1.2 (Repeated measures data with right-censoring) Consider a longitudinal study in which each subject is supposed to be monitored at time points $0, \ldots, p$, but some subjects drop out before they reach the endpoint p. Let $X = \{X(t) : t = 0, \ldots, p\}$ represent the full data structure on a subject, where $X(t)$ is typically a multivariate vector. Let $\bar{X}(t) = (X(0), \ldots, X(t))$ denote the history through time t. We assume that the measurements $X(t)$ can be divided in outcomes $Z(t)$, covariates $X^*(t)$ that one wants to adjust for in a regression model, and extraneous covariates $V^*(t)$. Let $Z = (Z(0), \ldots, Z(p))$, $X^* = (X^*(0), \ldots, X^*(p))$, and $V^* = (V^*(0), \ldots, V^*(p))$. Consider a regression model

$$Z = g(X^* \mid \alpha) + \epsilon,\ E(\epsilon \mid X^*) = 0, \tag{1.3}$$

where $g(X^* \mid \alpha) = (g_0(X^* \mid \alpha), \ldots, g_p(X^* \mid \alpha))$ and $g_j(X^* \mid \alpha)$ only depends on the history $(X^*(0), \ldots, X^*(j))$ of X^* up to point j, $j = 0, \ldots, p$. For example, if X^* is univariate, we might have

$$g_t(X^* \mid \alpha) = \alpha_0 + \alpha_1 t + \alpha_2 X^*(0) + \alpha_3 (X^*(t) - X^*(0)). \tag{1.4}$$

For other longitudinal data models, we refer to Diggle, Liang and Zeger (1994). Our goal is to estimate the regression parameters $\alpha = (\alpha_1, \ldots, \alpha_k)$.

Let C be the discrete drop-out time with values in $\{0, \ldots, p\}$. The observed data structure is given by $Y = \Phi(X, C) = (C, \bar{X}(C) = (X(0), \ldots, X(C)))$. In other words, if $C = j$, then the subject was followed up to (and including) visit j. □

We assume throughout that we have n study units (or subjects) and observe n identically and independently distributed observations (copies) $Y_1, \ldots, Y_n$ of the random variable Y. We will suppose that the full data structure distribution F_X of X is known to be an element of a specified full data structure model $\mathcal{M}^F$ and that there is a Euclidean parameter $\mu = \mu(F_X) \in \mathbb{R}^k$ of interest. For instance, in both Examples 1.1 and 1.2, μ is the regression parameter.

1.2.1 Estimation in a high-dimensional full data model

Our estimating function methodology for estimating the k-dimensional parameter μ based on the observed data $Y_1, \ldots, Y_n$ requires that we can find a class of k-dimensional estimating functions whose components, when evaluated at any $F_X \in \mathcal{M}^F$, are elements of the orthogonal complement $T^{F,\perp}_{nuis}(F_X)$ of the nuisance tangent space $T^F_{nuis}(F_X)$ in model $\mathcal{M}^F$ at F_X.