Chapter 4

Parametric Approach

4.1 Introduction

The missing data problem is already a classical problem that has not been yet solved satisfactorily. This problem includes those situations where the dependent variable is completely observed but the information on the covariates could be partially observed. Furthermore, in survival analysis the outcome is a survival time which itself could be censored.

Most of the existing methodologies are based on the assumption that non observed data are missing completely at random (MCAR) or at most missing at random (MAR) (Little and Rubin, 1987). A brief review of the different approaches and their advantages and inconvenients follows.

First, we could base the analysis on those individuals with all the observed covariates. The inference based on the so-called complete cases is biased and not consistent because the observed individuals are not necessarily a good representation of the overall sample. Furthermore, estimators based on the complete sample would be less accurate due to reduction in the sample size. Therefore, although this approach is appealing because can be handled with the existing software, its use has to be avoided.

A second approach consists in the imputation of the non-observed values (Glynn et al., 1993; Efron, 1994; Serrat and Gómez, 1995). The main problem here re-

lies in that the observed values are used to model the non-observed ones and this assumption might not be true. The resulting estimators could then be seriously biased.

A parametric modelization and the solution via maximum likelihood is another possible approach. As it is known this methodology is asymptotically efficient. This approach started in the 70's with the work done by Little and Rubin and until now this is a widely used method (Little and Rubin, 1987; Glynn et al., 1986); however, its implementation is not straightforward and the corresponding estimators rely heavily in a large number of assumptions that cannot be validated. Recently, some authors (Baker, 1994b) propose an hierarchical methodology, based also on the maximum likelihood method, that allows to combine several non-response models and to study the sensitivity of the resulting estimators under those assumptions.

The semiparametric approach is a fourth way of handling the missing data problem that allows to model only what is strictly necessary; in our situation we will have to model the relationship between survival, the covariates and the non-response pattern. Semiparametric estimators are unbiased, consistent and asymptotically normal (Newey, 1990; Rotnitzky and Wypij, 1994; Robins et al., 1994).

In this chapter, and previously to an analysis based on a semiparametric approach, we use a completely parametric point of view. This parametric analysis has two main goals: on one hand to show the drawbacks, practical and philosophical, in the specification of the likelihood function and in its optimization, and on the other to design a methodology that would allow to determine how the resulting estimators depend on the non-response pattern.

Our methodology is motivated by clinical and epidemiological studies. In this type of situations we have a cohort of patients and we are interested in studying their survival and to find the best set of predicting covariates for it. Specifically, since 1994 we are collaborating with the epidemiological unit of the *Institut Municipal de la Salut*, in Barcelona. This collaboration has motivated the necessity of finding a methodology that allows to deal with the non-observed values in the covariates of interest.

The chapter is organized as follows. In the next section we introduce the notation as well as some needed terminology. In Section 4.3, we present the problem and we

solve it parametrically. In Section 4.4 we show, as an illustration, the analysis of the data that motivated this approach. Last Section is devoted to the discussion.

4.2 Notation and definitions

The M-dimensional **potential data vector** $\boldsymbol{L} = (L_1, L_2, \ldots, L_M)'$ for an arbitrary individual is defined as the vector that contains his or her observed and non-observed data. The **response vector** $\boldsymbol{R} = (R_1, R_2, \ldots, R_M)'$ for this individual has m-th ($m = 1, \ldots, M$) component equal to 1 if the m-th variable has been observed and 0 otherwise.

We can split $\boldsymbol{L}$ into the subvectors $\boldsymbol{L}_{(\boldsymbol{R})}$ and $\boldsymbol{L}_{(\bar{\boldsymbol{R}})}$ corresponding to the observed and unobserved data, respectively. The subvector $\boldsymbol{L}_{(\boldsymbol{R})}$, is integrated by those components l of the vector $\boldsymbol{L}$ for whom $R_l = 1$, meanwhile the subvector of the non-observed variables, $\boldsymbol{L}_{(\bar{\boldsymbol{R}})}$, consists of those components l of $\boldsymbol{L}$ with $R_l = 0$. For example, if $M = 5$ and $\boldsymbol{R} = (1, 0, 1, 1, 0)'$ then $\boldsymbol{L}_{(\boldsymbol{R})} = (L_1, L_3, L_4)'$ and $\boldsymbol{L}_{(\bar{\boldsymbol{R}})} = (L_2, L_5)'$.

We will denote by $\boldsymbol{r} = (r_1, r_2, \ldots, r_M)'$, $r_m \in \{0, 1\}$, $m = 1, \ldots, M$, a realization of the response vector for an arbitrary individual. The conditional probability of $\boldsymbol{r}$ given the potential data vector $\boldsymbol{L}$ will be denoted by $\pi_{\boldsymbol{L}}(\boldsymbol{r}) = P(\boldsymbol{R} = \boldsymbol{r}|\boldsymbol{L})$, and the different non-response patterns will depend on the modelization of $\pi_{\boldsymbol{L}}(\boldsymbol{r})$.

The non-response pattern is Missing Completely at Random (MCAR), if and only if, the conditional probability of a realization $\boldsymbol{r}$ is constant, that is $\pi_{\boldsymbol{L}}(\boldsymbol{r})$ is independent of $\boldsymbol{L}$. The pattern is Missing at Random (MAR), if and only if, $\pi_{\boldsymbol{L}}(\boldsymbol{r})$ depends at most of the observed data $\boldsymbol{L}_{(\boldsymbol{r})}$. The non-response pattern is non-ignorable (NI), if and only if, $\pi_{\boldsymbol{L}}(\boldsymbol{r})$ depends on the subvector of non-observed data $\boldsymbol{L}_{(\bar{\boldsymbol{r}})}$.

If we assume that the components of the vector $\boldsymbol{L}$ follow a pre-established order, the non-response pattern is called to be **monotone** if the **no** observation of a variable implies the **no** observation of those that follow; that is, if $R_l = 0$ implies $R_m = 0$ for all $m > l$.

In a survival study the main variable T, is usually the elapsed time between an origin (randomization date in a clinical trial, treatment initiation, etc) and the

realization of an event (death, diagnosis of AIDS, etc). Data in these studies is often right censored and the observed data are $Y = \min\{T, C\}$ and $\delta = \mathbf{1}\{T \leq C\} = \mathbf{1}\{Y = T\}$ where C is the censoring time. For each individual we also have the values of certain covariates. Let $\boldsymbol{X}$ be the covariate vector. If $\boldsymbol{X}_*$ is a subvector of $\boldsymbol{X}$ and $\boldsymbol{V}_*$ is another set of covariates, we will say that $\boldsymbol{V}_*$ is a **surrogate** of $\boldsymbol{X}_*$ if $\boldsymbol{X}_*$ and $\boldsymbol{V}_*$ are strongly correlated. We will denote by $\boldsymbol{V}$ the vector of surrogate covariates (not included in $\boldsymbol{X}$) for the covariates of interest. The goal in a survival study, with missing data in the covariates, is to model T in terms of the covariates vector $\boldsymbol{X}$, using the information provided by $\boldsymbol{X}$ and by the surrogate vector $\boldsymbol{V}_*$. We will assume that the distribution of the censoring time, C, is independent of T, given the vector of covariates $(\boldsymbol{V}', \boldsymbol{X}')'$.

Suppose that we have a sample available of survival data with sample size n; according to the previous notation, for each individual i ($i = 1, \ldots, n$) we denote by $\boldsymbol{L}_i = (Y_i, \delta_i, \boldsymbol{V}_i', \boldsymbol{X}_i')'$ the vector of potential data, by $\boldsymbol{R}_i$ the response vector and by $\boldsymbol{L}_{(\boldsymbol{R}_i)}$ and $\boldsymbol{L}_{(\bar{\boldsymbol{R}}_i)}$ the subvectors of observed data and non-observed data, respectively. In our study, by construction of the vector $\boldsymbol{L}_i$, we have $L_{i1} = Y_i$ and $L_{i2} = \delta_i$, and as a consequence $R_{i1} = R_{i2} = 1$ for all $i = 1, \ldots, n$.

4.3 Testing the non-response model

4.3.1 Introduction

The goal of this section is the study and validation of the non-response pattern. We start testing whether the non-response process is MCAR. Then, under a totally parametric philosophy, we introduce a hierarchical scheme to check the sensitivity of the model parameters under different non-response patterns. In particular, this methodology allows us to elucidate about the non-ignorability of the non-response pattern.

4.3.2 MCAR validation

Given the observed data, $(\boldsymbol{R}_i, \boldsymbol{L}_{(\boldsymbol{R}_i)})_{i=1,2,\ldots,n}$, the test to check whether or not the non-response process is MCAR is based on the comparison of the probabilities

$P(R_{im}=1), i=1,\ldots,n$ and $P(R_{im}=1|\boldsymbol{L}_i), i=1,\ldots,n$ for every $m=1,\ldots,M$. The hypothesis test is formulated as

$$\begin{aligned} H_0 &: \text{MCAR non-response pattern} \\ H_A &: \text{MAR or NI non-response pattern,} \end{aligned}$$

and two different procedures are developed depending on whether or not the non-response pattern is monotone.

If the non-response pattern is monotone the previous comparison becomes reduced to the comparison between probabilities $P(R_{im}=1|R_{i(m-1)}=1), i=1,\ldots,n$ and $P(R_{im}=1|R_{i(m-1)}=1, L_{i1},\ldots,L_{i(m-1)}), i=1,\ldots,n$ for each $m=1,\ldots,M$, as it is shown in the next proposition.

Proposition 4.3.1 *If the non-response pattern is monotone, the MAR condition*

$$P(\boldsymbol{R}_i=\boldsymbol{r}|\boldsymbol{L}_i)=P(\boldsymbol{R}_i=\boldsymbol{r}|\boldsymbol{L}_{(\boldsymbol{r})i}), i=1,\ldots,n \qquad [A]$$

where $\boldsymbol{r}\in\{0,1\}^M$*, is equivalent to the condition*

$$P(R_{im}=1|R_{i(m-1)}=1,\boldsymbol{L}_i)=P(R_{im}=1|R_{i(m-1)}=1,L_{i1},\ldots,L_{i(m-1)}), i=1,\ldots,n \qquad [B]$$

for each $m=1,\ldots,M$.

In particular, the non-response pattern is MCAR if and only if for each $m=1,\ldots,M$ *the probability of observing the* m*-th variable, given the observation of the* $(m-1)$*-th variable and the potential data, is constant, that is*

$$P(R_{im}=1|R_{i(m-1)}=1,\boldsymbol{L}_i)=P(R_{im}=1|R_{i(m-1)}=1), i=1,\ldots,n$$

for each $m=1,\ldots,M$.

Proof: When the non-response pattern is monotone, the sample space of the M-dimensional response vector, $\boldsymbol{R}_i$, is reduced to the set

$$\Omega_{\boldsymbol{R}_i}=\{\boldsymbol{r}_0=(0,\ldots,0)',\boldsymbol{r}_1=(1,0,\ldots,0)',\boldsymbol{r}_2=(1,1,0,\ldots,0)',\ldots,\boldsymbol{r}_M=(1,\ldots,1)'\}$$

and the condition [A] can be rewritten as

$$P(\boldsymbol{R}_i = \boldsymbol{r}_m|\boldsymbol{L}_i) = P(\boldsymbol{R}_i = \boldsymbol{r}_m|L_{i1}, \ldots, L_{im}), i = 1, \ldots, n \qquad [A']$$

for each $m = 0, \ldots, M$.

- To prove that [A] $\Rightarrow$ [B], we will prove by induction on the m index that, under the hypothesis [A′], the probabilities

  $$P(R_{im} = 1|\boldsymbol{L}_i) \text{ and } P(R_{im} = 1|R_{i(m-1)} = 1, \boldsymbol{L}_i)$$

  only depend on the data $L_{i1}, \ldots, L_{i(m-1)}$.

  Indeed, if $m = 1$, we get

  $$P(R_{i1} = 1|\boldsymbol{L}_i) = 1 - P(R_{i1} = 0|\boldsymbol{L}_i) = 1 - P(\boldsymbol{R}_i = \boldsymbol{r}_0|\boldsymbol{L}_i)$$

  that, by the condition [A′], does not depend on the data.

  Assume that the proposition has been proved until the j-th index, $(1 < j < M)$. To prove the result for the $(j + 1)$-th index, it is enough to verify the relationships

  $$P(R_{i(j+1)} = 1|\boldsymbol{L}_i) = P(R_{ij} = 1|\boldsymbol{L}_i) - P(\boldsymbol{R}_i = \boldsymbol{r}_j|\boldsymbol{L}_i) \quad (4.1)$$

  $$P(R_{i(j+1)} = 1|R_{ij} = 1, \boldsymbol{L}_i) = 1 - \frac{P(\boldsymbol{R}_i = \boldsymbol{r}_j|\boldsymbol{L}_i)}{P(R_{ij} = 1|\boldsymbol{L}_i)}. \quad (4.2)$$

  Since the right-hand part of the previous equalities depend at most, by condition [A′] and the induction hypothesis, on the data $L_{i1}, \ldots, L_{ij}$, it is proved that the condition [A] implies the condition [B].

  The expression (4.1) is straightforwardly derived because, due to the monotonicity of the non-response pattern, $\{R_{ij}\}$ is disjoint union of $\{R_{i(j+1)}\}$ and $\{\boldsymbol{R}_i = \boldsymbol{r}_j\}$.

  In a similar way, the equality (4.2) holds from

  $$\begin{aligned} P(R_{i(j+1)} = 1|R_{ij} = 1, \boldsymbol{L}_i) &= P(R_{ij} = 1|R_{ij} = 1, \boldsymbol{L}_i) - P(\boldsymbol{R}_i = \boldsymbol{r}_j|R_{ij} = 1, \boldsymbol{L}_i) = \\ &= 1 - P(\boldsymbol{R}_i = \boldsymbol{r}_j|R_{ij} = 1, \boldsymbol{L}_i) = \\ &= 1 - \frac{P(\boldsymbol{R}_i = \boldsymbol{r}_j, R_{ij} = 1|\boldsymbol{L}_i)}{P(R_{ij} = 1|\boldsymbol{L}_i)} = \\ &= 1 - \frac{P(\boldsymbol{R}_i = \boldsymbol{r}_j|\boldsymbol{L}_i)}{P(R_{ij} = 1|\boldsymbol{L}_i)}. \end{aligned}$$

- To prove that the condition [B] is sufficient to have a MAR non-response pattern, it is enough to verify that condition [A′] holds.

  If $m = 0$, condition [A′] is true because $P(\boldsymbol{R}_i = \boldsymbol{r}_0|\boldsymbol{L}_i) = P(R_{i1} = 0|\boldsymbol{L}_i) = 1 - P(R_{i1} = 1|\boldsymbol{L}_i)$ does not depend on the data, under hypothesis [B].

  For $m = 1, \ldots, M$, and by using that the non-response pattern is monotone, we get

$$\begin{aligned} P(\boldsymbol{R}_i = \boldsymbol{r}_m|\boldsymbol{L}_i) &= P(R_{i1} = 1|\boldsymbol{L}_i) \cdot P(R_{i2} = 1|R_{i1} = 1, \boldsymbol{L}_i) \ldots \\ &\quad \ldots P(R_{im} = 1|R_{i(m-1)} = 1, \boldsymbol{L}_i) \cdot P(R_{i(m+1)} = 0|R_{im} = 1, \boldsymbol{L}_i) \end{aligned}$$

  and, if [B] holds, all the factors in the right-hand term of the previous equality depend on, at most, the data $L_{i1}, \ldots, L_{im}$.

In particular, the condition to have a MCAR non-response pattern is that the probabilities $P(R_{im} = 1|R_{i(m-1)} = 1, \boldsymbol{L}_i)$, $m = 1, \ldots, M$, are constant. If we denote by π_m the conditional probability $P(R_{im} = 1|R_{i(m-1)} = 1, \boldsymbol{L}_i)$ then the probabilities of the sample space $\Omega_{\boldsymbol{R}_i}$ are

$$P(\boldsymbol{R}_i = \boldsymbol{r}_m|\boldsymbol{L}_i) = \begin{cases} 1 - \pi_1 & \text{if } m = 0 \\ \pi_1\pi_2 \cdots \pi_m(1 - \pi_{m+1}) & \text{if } m = 1, \ldots, M-1, \\ \prod_{m=1}^{M} \pi_m & \text{if } m = M \end{cases}$$

that they do not depend on the data $\boldsymbol{L}_i$. □

If the conditional probabilities $P(R_{im} = 1|R_{i(m-1)} = 1, L_{i1}, \ldots, L_{i(m-1)})$ are modeled through a logit link, that is

$$\text{logit } (P(R_{im} = 1|R_{i(m-1)} = 1, L_{i1}, \ldots, L_{i(m-1)})) = \alpha_{m1} + \alpha'_{m2} h_m(L_{i1}, \ldots, L_{i(m-1)}),$$

where $h_m(L_{i1}, \ldots, L_{i(m-1)})$ is a vectorial arbitrary function of the data $L_{i1}, \ldots, L_{i(m-1)}$, the test H_0 versus H_A can be seen as the following M simultaneous tests:

$$\begin{aligned} H_{0m} &: \text{logit } (P(R_{im} = 1|R_{i(m-1)} = 1, L_{i1}, \ldots, L_{i(m-1)})) = \alpha_{m1} \\ H_{Am} &: \text{logit } (P(R_{im} = 1|R_{i(m-1)} = 1, L_{i1}, \ldots, L_{i(m-1)})) = \\ &\quad = \alpha_{m1} + \alpha'_{m2} h_m(L_{i1}, \ldots, L_{i(m-1)}) \qquad m = 1, \ldots, M. \end{aligned}$$

For each m, $m = 1, \ldots, M$, the statistic based on the likelihood ratio of H_{0m} versus H_{Am} give us a p-valor, p_m. Next proposition proves that if the data are MCAR, *i.e.,* all hypotheses $H_{0m}, m = 1, \ldots, M$, are true, then resulting p-values are independent and uniformly distributed in (0,1).

Proposition 4.3.2 *If the non-response pattern is monotone and the missing data mechanism is MCAR, p-values, $p_1, p_2, \ldots, p_M$, resulting of the statistic based on the likelihood ratio test of H_{0m} versus H_{Am} in the M tests*

$$\begin{aligned} H_{0m} &: \textit{logit}\left(P(R_{im}=1|R_{i(m-1)}=1, L_{i1}, \ldots, L_{i(m-1)})\right) = \alpha_{m1} \\ H_{Am} &: \textit{logit}\left(P(R_{im}=1|R_{i(m-1)}=1, L_{i1}, \ldots, L_{i(m-1)})\right) = \\ &= \alpha_{m1} + \alpha'_{m2} h_m(L_{i1}, \ldots, L_{i(m-1)}) \qquad\qquad m = 1, \ldots, M, \end{aligned}$$

are independent and uniformly distributed in (0,1).

Proof: For a value of m, $m = 1, \ldots, M$, since the non-response pattern is MCAR the hypothesis H_{0m} is true and logit $(P(R_{im} = 1|R_{i(m-1)} = 1)) = \alpha_{m1}$, and therefore

$$P(R_{im} = 1|R_{i(m-1)} = 1) = \frac{\exp(\alpha_{1m})}{1 + \exp(\alpha_{1m})} = \pi_m.$$

As a consequence, the random variable resulting from computing the relative frequencies of the event "observation of the m-th variable given the observation of the $(m-1)$-th variable", $\{R_{im} = 1|R_{i(m-1)} = 1\}$, denoted by f_{im}, follows, asymptotically, a random variable $N\left(\pi_m, \sqrt{\frac{\pi_m(1-\pi_m)}{n_m}}\right)$, where n_m is the total number of individuals for whom the variable $L_{i(m-1)}$ has been observed (by construction, n_1 is the sample size).

If p_m denotes the p-value resulting of the likelihood ratio statistic of the test H_{0m} versus H_{Am}, we derive that $p_m = P\left(|f_{im} - \pi_m| > |f_{im,data} - \pi_m|\right)$.

To prove that $p_m \sim U(0,1)$ it is enough to prove that $\forall p \in [0,1]$, $P(p_m \leq p) = p$. In fact, for a value p, $0 \leq p \leq 1$, if we denote by I_{1-p} the interval with probability $1-p$ centered in π_m for the distribution $N\left(\pi_m, \sqrt{\frac{\pi_m(1-\pi_m)}{n_m}}\right)$, we get $P(p_m \leq p) = P\left(f_{im} \notin I_{1-p}\right) = p$, as we wanted to prove.

The M p-values, $p_1, p_2, \ldots, p_M$, are independent because the conditional probabilities π_m do not depend on the data $\boldsymbol{L}_i$. □

For the overall interpretation of the M resulting p-values, $p_1, \ldots, p_M$, we use the combined statistic $S = -2\sum_{m=1}^{M} \log p_m$, that follows, under H_0, a χ^2 distribution with $2M$ degrees of freedom.

If the non-response pattern is not monotone, the comparison H_0 versus H_A has to be solved from comparing, for each $i = 1, \ldots, n$ and for each m, $m = 1, \ldots, M$, the probabilities $P(R_{im} = 1)$ and $P(R_{im} = 1|$ the observed data for $m' \neq m)$. In this case the p-values that we obtain are not independent and we have to use simultaneous inference techniques to design the respective tests (Miller, 1980), for example the Bonferroni's t statistic. In general, these techniques are more conservative and, therefore, the power of the resulting tests will be lower. In Section 4.4.2 we show in detail the application of this methodology.

4.3.3 Parametric approach of the problem

We solve the problem of estimating a model for the survival T from a totally parametric perspective and starting with a potential data vector $\boldsymbol{L}_i = (Y_i, \delta_i, \boldsymbol{V}_i', \boldsymbol{X}_i')'$, $i = 1, 2, \ldots, n$, where $\boldsymbol{X}$ is a vector of partially observed covariates and $\boldsymbol{V}$ is a vector of completely observed covariates. This approach will allow us to introduce the non-response probabilities modeling and, as a consequence, to study the goodness of fit of these models. The estimation is based on the standard maximum likelihood methodology. In order to specify a likelihood function, L, for the sample, let $f_c(l; \boldsymbol{\theta})$ be the density function for the complete data and $P(\boldsymbol{R}_i = \boldsymbol{r}|\boldsymbol{L}_i; \boldsymbol{\psi})$ be the probability of observing certain $\boldsymbol{L}_i$ components. These functions depend on $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ parameters, and we will suppose that both parameters are distinct.

The contribution of the i-th individual to the likelihood function $L(\boldsymbol{\theta}, \boldsymbol{\psi})$ is: $f_c(\boldsymbol{L}_i; \boldsymbol{\theta}) \cdot P(\boldsymbol{R}_i = \boldsymbol{1}|\boldsymbol{L}_i; \boldsymbol{\psi})$ if the individual has been completely observed (*i.e.,* $\boldsymbol{R}_i = \boldsymbol{1}$), and $\int f_c(\boldsymbol{L}_i; \boldsymbol{\theta}) \cdot P(\boldsymbol{R}_i = \boldsymbol{r}|\boldsymbol{L}_i; \boldsymbol{\psi}) d\boldsymbol{L}_{(\bar{\boldsymbol{r}})i}$ if the individual has been partially observed and his or her response vector is $\boldsymbol{R}_i = \boldsymbol{r}$ with $\boldsymbol{r} \neq \boldsymbol{1}$. This second expression corresponds to the marginalization of the first one with respect to then non-observed

data. Therefore, the likelihood function $L(\boldsymbol{\theta}, \boldsymbol{\psi})$ from the observed data is given by

$$\begin{aligned} L(\boldsymbol{\theta}, \boldsymbol{\psi}) &= \prod_{i=1}^{n} \Bigg\{ [f_c(\boldsymbol{L}_i; \boldsymbol{\theta}) \cdot P(\boldsymbol{R}_i = \mathbf{1} | \boldsymbol{L}_i; \boldsymbol{\psi})]^{I(\boldsymbol{R}_i = \mathbf{1})} \\ &\quad \prod_{\boldsymbol{r} \neq \mathbf{1}} \left[\int f_c(\boldsymbol{L}_i; \boldsymbol{\theta}) \cdot P(\boldsymbol{R}_i = \boldsymbol{r} | \boldsymbol{L}_i; \boldsymbol{\psi}) d\boldsymbol{L}_{(\bar{\boldsymbol{r}})i} \right]^{I(\boldsymbol{R}_i = \boldsymbol{r})} \Bigg\}. \end{aligned}$$

Next proposition shows that, when the non-response pattern is MCAR or MAR (*i.e.,* probabilities $P(\boldsymbol{R}_i = \boldsymbol{r} | \boldsymbol{L}_i)$ do not depend on $L_{(\bar{\boldsymbol{r}})i}$) the likelihood function, $L(\boldsymbol{\theta}, \boldsymbol{\psi})$, can be decomposed in such way that the maximum likelihood estimate for the parameter $\boldsymbol{\theta}$ is independent of the non-response pattern considered in the model.

Proposition 4.3.3 *If the non-response pattern is MCAR or MAR, the maximization of the likelihood function does not depend on the non-response probabilities.*

Proof: According to the introduced notation, it is enough to prove that the function $L(\boldsymbol{\theta}, \boldsymbol{\psi})$ can be decomposed as a product of two functions $L_1(\boldsymbol{\theta})$ and $L_2(\boldsymbol{\psi})$.

Indeed, if the non-response probabilities $P(\boldsymbol{R}_i = \boldsymbol{r} | \boldsymbol{L}_i; \boldsymbol{\psi})$ do not depend on the non-observed data $\boldsymbol{L}_{(\bar{\boldsymbol{r}})i}$ then

$$\int f_c(\boldsymbol{L}_i; \boldsymbol{\theta}) \cdot P(\boldsymbol{R}_i = \boldsymbol{r} | \boldsymbol{L}_i; \boldsymbol{\psi}) d\boldsymbol{L}_{(\bar{\boldsymbol{r}})i} = P(\boldsymbol{R}_i = \boldsymbol{r} | \boldsymbol{L}_i; \boldsymbol{\psi}) \cdot \int f_c(\boldsymbol{L}_i; \boldsymbol{\theta}) d\boldsymbol{L}_{(\bar{\boldsymbol{r}})i}$$

and the likelihood function

$$\begin{aligned} L(\boldsymbol{\theta}, \boldsymbol{\psi}) &= \prod_{i=1}^{n} \Bigg\{ [f_c(\boldsymbol{L}_i; \boldsymbol{\theta}) \cdot P(\boldsymbol{R}_i = \mathbf{1} | \boldsymbol{L}_i; \boldsymbol{\psi})]^{I(\boldsymbol{R}_i = \mathbf{1})} \\ &\quad \prod_{\boldsymbol{r} \neq \mathbf{1}} \left[\int f_c(\boldsymbol{L}_i; \boldsymbol{\theta}) \cdot P(\boldsymbol{R}_i = \boldsymbol{r} | \boldsymbol{L}_i; \boldsymbol{\psi}) d\boldsymbol{L}_{(\bar{\boldsymbol{r}})i} \right]^{I(\boldsymbol{R}_i = \boldsymbol{r})} \Bigg\} \end{aligned}$$

admits the factorization $L(\boldsymbol{\theta}, \boldsymbol{\psi}) = L_1(\boldsymbol{\theta}) \cdot L_2(\boldsymbol{\psi})$ where

$$\begin{aligned} L_1(\boldsymbol{\theta}) &= \prod_{i=1}^{n} \left\{ f_c(\boldsymbol{L}_i; \boldsymbol{\theta})^{I(\boldsymbol{R}_i = \mathbf{1})} \cdot \prod_{\boldsymbol{r} \neq \mathbf{1}} \left[\int f_c(\boldsymbol{L}_i; \boldsymbol{\theta}) d\boldsymbol{L}_{(\bar{\boldsymbol{r}})i} \right]^{I(\boldsymbol{R}_i = \boldsymbol{r})} \right\} \\ L_2(\boldsymbol{\psi}) &= \prod_{i=1}^{n} \left\{ \prod_{\boldsymbol{r}} P(\boldsymbol{R}_i = \boldsymbol{r} | \boldsymbol{L}_i; \boldsymbol{\psi})^{I(\boldsymbol{R}_i = \boldsymbol{r})} \right\}. \end{aligned}$$

□

In our study the density function, $f_c(\boldsymbol{L}_i;\boldsymbol{\theta})$, can be written as

$$f_c(\boldsymbol{L}_i;\boldsymbol{\theta}) = f_c((Y_i,\delta_i,\boldsymbol{V}_i',\boldsymbol{X}_i')';\boldsymbol{\theta}) = f_c(\boldsymbol{X}_i;\boldsymbol{\theta})\cdot f_c(Y_i,\delta_i|\boldsymbol{X}_i;\boldsymbol{\theta})\cdot f_c(\boldsymbol{V}_i|Y_i,\delta_i,\boldsymbol{X}_i;\boldsymbol{\theta}).$$

and therefore the parametric approach imposes the correct specification of a) the distribution of the covariates of interest, $\boldsymbol{X}$, b) the conditional distribution of the observed times, Y, given the covariates $\boldsymbol{X}$, c) the conditional distribution of the surrogate covariates, $\boldsymbol{V}$, given Y and $\boldsymbol{X}$, and d) the conditional non-response probabilities given the potential data. It is very important to know that these specifications can become totally arbitrary, due to the fact that none of them can be validated from the observed data (Gill et al., 1997; Gill and Robins, 1997).

In what follows we suppose that the covariate vector for the i-th individual, $\boldsymbol{X}_i$, is formed by p discrete random variables, $X_{i1}, X_{i2}, \ldots, X_{ip}$ and the vector of surrogate covariates $\boldsymbol{V}_i$ has q discrete random variables, $V_{i1}, V_{i2}, \ldots, V_{iq}$.

The specification of the densities $f_c(\boldsymbol{X}_i;\boldsymbol{\theta})$ and $f_c(\boldsymbol{V}_i|Y_i,\delta_i,\boldsymbol{X}_i;\boldsymbol{\theta})$, as well as of the probabilities $P(\boldsymbol{R}_i=\boldsymbol{r}|\boldsymbol{L}_i;\boldsymbol{\psi})$, can be done in terms of the logarithms of the conditional *odds ratio* of each category with respect to the baseline category (or conditional logistic regressions if the covariates are binary), that is, by modeling each one of these expressions

$$\begin{aligned}
\log\frac{P(X_{ij}=k|X_{i1},\ldots,X_{i(j-1)})}{P(X_{ij}=0|X_{i1},\ldots,X_{i(j-1)})} \quad & j=1,\ldots,p \quad k\neq 0,\\
\log\frac{P(V_{ij}=k|V_{i1},\ldots,V_{i(j-1)})}{P(V_{ij}=0|V_{i1},\ldots,V_{i(j-1)})} \quad & j=1,\ldots,q \quad k\neq 0 \quad \text{and}\\
\text{logit}\,(P(R_{ij}=1|R_{i1},\ldots,R_{i(j-1)})) \quad & j=1,\ldots,p.
\end{aligned}$$

As we will see in the illustration of the next section, the use of different models for the probabilities $P(R_{ij}=1|R_{i1},\ldots,R_{i(j-1)})$, $j=1,\ldots,p$, as a function of the data $\boldsymbol{L}_i$, allows us to perform a sensitivity analysis of the estimates to the non-response pattern considered.

To model the survival times and their relationship with the covariates, $\boldsymbol{X}$, it is enough to specify a density function $f_c(Y_i,\delta_i|\boldsymbol{X}_i;\boldsymbol{\theta})$. Previous analysis based on the

completely observed subsample can be useful to achieve this goal; however, once more, these assumptions can not be validated from the observed data.

An extra drawback, to add to the mentioned misspecifications, is the curse of dimensionality of the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$. This fact reduces, in an important way, on one hand the speed of convergence of the corresponding implementations and, on the other hand, the relative sample size. For example, it is straightforward to compute that in a simple scenario where all the covariates would be binary and we would use only linear models without interactions, we would get

$$\begin{aligned} \dim(\boldsymbol{\theta}) &= (2^p - 1) + (p + 1) + (2^q - 1)(p + 3), \\ \dim(\boldsymbol{\psi}) &= (2^p - 1)(p + q + 3), \end{aligned}$$

and for a small setting with $p = q = 3$, $\dim(\boldsymbol{\theta}) = 53$, $\dim(\boldsymbol{\psi}) = 63$ and therefore we would have to estimate a 116-dimensional parameter in the likelihood function.

4.4 Illustration with the HIV+PTB cohort

4.4.1 Introduction

To illustrate the methodology introduced in the previous section, we present here an application to the study of survival time in a cohort of pulmonary tuberculosis HIV-infected patients. The project is part of the collaboration with the epidemiological unit of the *Institut Municipal de la Salut*, in Barcelona, since 1994. Data come from several medical records of patients belonging to the Prevention and Control of the Tuberculosis Program. Some of the epidemiological goals of this Program are: a) the study of the AIDS progression in TB patients and b) finding predictors of survival in TB HIV-infected patients.

4.4.2 Dataset and methods

We have a sample integrated by 418 HIV+ patients with pulmonary tuberculosis. All of them, resident in Barcelona city and diagnosed of tuberculosis in the 1992–1994 period. Data were collected in September 30th 1995. The survival time of interest is the elapsed time between the diagnosis of tuberculosis (and, as a consequence,

the starting of treatment against TB) and death. For each individual of the sample we have, potentially, the following sociological and clinical variables, all of them collected at the beginning of the study: gender, age, district of residence, prison history, treatment against tuberculosis history, belonging to an HIV transmission group, tuberculosis site, radiological pattern, microbiological results, percentage of lymphocyte subsets (T-CD4+ and T-CD8+), tuberculin skin test result, among others. The sample is a subsample of the cohort introduced in Chapter 2 and it is integrated by all the patients for whom the treatment against tuberculosis history, the radiological pattern and the microbiological result covariates are available.

Previous analysis (Caylà et al., 1993a; Serrat and Gómez, 1995) done with complete observed data show that the T-CD4+ percentage (in particular, its categorization in high an low level of immunosuppression) and the result to the tuberculin skin test are the covariates of interest for the estimation of the mentioned survival. We will refer to this variables by $CD4$ and PPD and they will be the components of the vector $\boldsymbol{X}$ introduced in Section 4.2. Patients will be classified as whether or not the $CD4$ covariate is larger than 14%. PPD variable takes value 1 if the result of the tuberculin skin test is positive and 0 otherwise. The methodological problem is motivated by the fact that both variables have a 37.5% and a 50.5% of missing values, respectively. Specifically, both of them are only available in a 31.3% of the sample and there is a 19.1% of the sample for whom none of both variables is available.

The goal of the study is to evaluate the predictive character of the $CD4$ and PPD indicators, using all the information available in the sample; in particular, to study the result to the tuberculin test as a complementary quality measure of the immunosuppression level given by the T-CD4+ counts. To this goal we will use the methodology described in the previous section.

After studying, together with the epidemiological team, which variables could provide qualitative information about the non-response pattern or about $CD4$ and PPD covariates, we chose as a surrogate covariates, $\boldsymbol{V}$, of the covariates of interest, $\boldsymbol{X}$, a) to have been previously treated against tuberculosis (TR: 1 = Yes, 2 = No), b) the radiological pattern (RA: 0 = Normal, 1 = Abnormal with cavitary pattern and 2 = Abnormal without cavitary pattern) and c) the bacteriological result (BA: 0 = Negative, 1 = Positive and 2 = Positive with bacteriological culture). So,

according to the notation introduced in Section 4.2, our vector of potential data is:

$$\boldsymbol{L}_i = (Y_i, \delta_i, TR_i, RA_i, BA_i, \mathbf{1}\{CD4_i > 14\}, PPD_i)'.$$

In order to simplify the notation we will omit the subindex i, except if it is necessary for the comprehension of the text.

Note that, due to the fact that the surrogate covariates are completely observed, *i.e.,* $R_3 = R_4 = R_5 = 1$, the response vector, $\boldsymbol{r} \in \{0,1\}^7$, will only take the following values (1,1,1,1,1,1,1), (1,1,1,1,1,1,0), (1,1,1,1,1,0,1) and (1,1,1,1,1,0,0).

4.4.3 Validation of the MCAR assumption

For the MCAR validation we apply the methodology introduced in Section 4.3.1. Since data are non monotone, we have to compare the probability $P(R_{CD4} = 1)$ with $P(R_{CD4} = 1|Y, \delta, TR, RA, BA, PPD)$, and $P(R_{PPD} = 1)$ with $P(R_{PPD} = 1|Y, \delta, TR, RA, BA, \mathbf{1}\{CD4 > 14\})$.

In the first case, if we model the non-response probability to the $CD4$ covariate as a function of the other variables according to the logistic model

$$\text{logit } (P(R_{CD4} = 1|Y, \delta, TR, RA, BA, PPD)) =$$

$$= \alpha_0 + \alpha_1 Y + \alpha_2 \delta + \alpha_3 TR + \alpha_4 RA + \alpha_5 BA + \alpha_6 PPD,$$

the comparison between probabilities $P(R_{CD4} = 1)$ and $P(R_{CD4} = 1|Y, \delta, TR, RA, BA, PPD)$ is equivalent to the hypothesis test

$$\begin{aligned} H_0 &: \alpha_i = 0 \quad \forall i \in \{1, \ldots, 6\} \\ H_A &: \exists i \in \{1, \ldots, 6\} \mid \alpha_i \neq 0. \end{aligned}$$

If the null hypothesis is true, the probability of responding to the $CD4$ variable will not depend on other observed variables, and its contribution to the MCAR hypothesis test will be in the sense of not refusing this assumption. Otherwise, if the null hypothesis is false, we will have statistical evidence to reject the MCAR non-response hypothesis.

Analogously, we also apply this methodology to the non-response probability to the PPD covariate, by means of the model

$$\text{logit } (P(R_{PPD} = 1|Y, \delta, TR, RA, BA, \mathbf{1}\{CD4 > 14\})) =$$

$$= \beta_0 + \beta_1 Y + \beta_2 \delta + \beta_3 TR + \beta_4 RA + \beta_5 BA + \beta_6 \mathbf{1}\{CD4 > 14\}.$$

To combine both tests we use the Bonferroni's correction. This correction means to use as a signification level in each partial hypothesis test the global signification level divided by the total number of tests we wish to combine. Then, the decision criteria is the next: we reject the null hypothesis when some of the partial hypothesis tests rejects its respective null hypothesis. We can observe that, initially, it seems to be easier to reject the null hypothesis by the fact of being using more than one test; however, in each of the partial test we are using a fraction of the signification level and therefore we need more evidence against the corresponding partial null hypothesis. The combined effect is a test that, in general, it is more conservative.

In our dataset, the above hypothesis tests for the variables R_{CD4} and R_{PPD} have p-value 0.03766 and 0.1733, respectively. If we use a global signification level of 5%, both results are not significant (they are larger than 0.025) and, therefore, in both variables we do not detect evidence against the null hypothesis. As a consequence, the MCAR non-response hypothesis can not be rejected.

4.4.4 Parametric approach of the problem

Following steps in Section 4.3.2, the contribution of an individual to the likelihood function $L(\boldsymbol{\theta}, \boldsymbol{\psi})$ can be computed after these four modelizations.

a) For the distributions of the covariates of interest, $CD4$ and PPD, we use logistic models for the probability of $\mathbf{1}\{CD4 > 14\}$ and for the conditional probabilities of $PPD = 1$ given the $\mathbf{1}\{CD4 > 14\}$ values, that is

$$\begin{aligned} \text{logit}\,(P(\mathbf{1}\{CD4 > 14\} = 1)) &= \alpha_1, \\ \text{logit}\,(P(PPD = 1 | \mathbf{1}\{CD4 > 14\} = 1)) &= \alpha_2 \quad \text{and} \\ \text{logit}\,(P(PPD = 1 | \mathbf{1}\{CD4 > 14\} = 0)) &= \alpha_3. \end{aligned}$$

b) In a preliminary study carried out on the same cohort of patients (Serrat et al., 1998) we observed that the survival time could be satisfactorily modeled through a Weibull distribution depending on $CD4$ and PPD covariates. This model is also used in this application and therefore the density function

$f_c(Y, \delta|\boldsymbol{X}; \boldsymbol{\theta})$ can be written as

$$f_c(Y, \delta|CD4, PPD; \boldsymbol{\theta}) = \left(\frac{1}{\sigma Y}(e^{-\beta}Y)^{\frac{1}{\sigma}}\right)^{\delta} \cdot e^{-(e^{-\beta}Y)^{\frac{1}{\sigma}}},$$

where $\beta = \beta_0 + \beta_1 \mathbf{1}\{CD4 > 14\} + \beta_2 PPD$ and $\sigma = \sigma_0 + \sigma_1 \mathbf{1}\{CD4 > 14\} + \sigma_2 PPD$.

c) The conditional distributions of TR, RA and BA given $Y, \delta, CD4, PPD$ are modeled via the log-odds ratios with respect to the reference group, given the previous variables. If we denote by $\lambda_.$, in a generic way, one of the following odds ratio

$$\begin{aligned}
\lambda_1 &= \frac{P(TR = 1)}{P(TR = 0)}, \\
\lambda_{2ij} &= \frac{P(RA = j|TR = i)}{P(RA = 0|TR = i)} \quad i = 0, 1 \quad j = 1, 2, \\
\lambda_{3ijk} &= \frac{P(BA = k|TR = i, RA = j)}{P(BA = 0|TR = i, RA = j)} \quad i = 0, 1 \quad j = 0, 1, 2 \quad k = 1, 2,
\end{aligned}$$

then its logarithm, $\log(\lambda_.)$, is specified by

$$\log(\lambda_.) = \gamma_{.0} + \gamma_{.1} \log(Y) + \gamma_{.2}\delta + \gamma_{.3} \log(Y)\delta + \gamma_{.4} \mathbf{1}\{CD4 > 14\} + \gamma_{.5} PPD,$$

and it is a function of Y, δ and $\boldsymbol{X}$ variables, of some possible interactions between them, and on the $(\gamma_{.0}, \ldots, \gamma_{.5})$ parameter. Note that, in order to reduce the effect of the extreme values in the observed survival times, we rescale the time to logarithmic scale.

d) We configure the different non-response patterns by modelling the probabilities $P(R_{ij} = 1|R_{i1}, \ldots, R_{i(j-1)})$ in terms of the covariates $\boldsymbol{V}$ and $\boldsymbol{X}$. We define logistic models for the probability of observing the $CD4$ covariate and for the conditional probabilities of observing the PPD covariate, given the $CD4$

values, according to the following scheme:

$$\begin{aligned}
\text{logit } (P(R_{CD4}=1)) &= \alpha_{10}+\alpha_{11}TR+ \\
&+ \alpha_{12}RA1+\alpha_{13}RA2+\alpha_{14}BA1+\alpha_{15}BA2 \\
&+ \alpha_{16}\mathbf{1}\{CD4>14\}+\alpha_{17}PPD, \\
\text{logit } (P(R_{PPD}=1|R_{CD4}=1)) &= \alpha_{20}+\alpha_{21}TR+ \\
&+ \alpha_{22}RA1+\alpha_{23}RA2+\alpha_{24}BA1+\alpha_{25}BA2 \\
&+ \alpha_{26}\mathbf{1}\{CD4>14\}+\alpha_{27}PPD \quad \text{and} \\
\text{logit } (P(R_{PPD}=1|R_{CD4}=0)) &= \alpha_{30}+\alpha_{31}TR+ \\
&+ \alpha_{32}RA1+\alpha_{33}RA2+\alpha_{34}BA1+\alpha_{35}BA2 \\
&+ \alpha_{36}\mathbf{1}\{CD4>14\}+\alpha_{37}PPD.
\end{aligned}$$

where $RAi = \mathbf{1}\{RA = i\}, i = 1, 2$ and $BAi = \mathbf{1}\{BA = i\}, i = 1, 2$ denote binary dummy variables for the effects of the categories in radiology and bacteriology, respectively.

The parameter $\boldsymbol{\theta}$ of the density function $f_c(\boldsymbol{L}_i;\boldsymbol{\theta})$ in the likelihood $L(\boldsymbol{\theta},\boldsymbol{\psi})$ is given by

$$\boldsymbol{\theta} = (\alpha_1, \alpha_2, \alpha_3, \beta_0, \beta_1, \beta_2, \sigma_0, \sigma_1, \sigma_2, \gamma_{.0}, \ldots, \gamma_{.5})$$

and it has dimension equal to 111. The nuisance parameter resulting from the non-response probabilities modelling is

$$\boldsymbol{\psi} = (\alpha_{10}, \ldots, \alpha_{37})$$

and is 24-dimensional. Note that the parameter of interest to be estimated is only made by the components $(\beta_0, \beta_1, \beta_2, \sigma_0, \sigma_1, \sigma_2)$ of the vector $\boldsymbol{\theta}$ and it has dimension 6.

The hierarchical scheme proposed in the previous step d) allows us to simulate, in a nested way, the different non-response patterns defined in Section 4.2. More

precisely, we optimize the likelihood function for the following five scenarios:

$$\begin{aligned}
\alpha_{ij} = 0 \quad i = 1,2,3 \quad j = 1,\ldots,7 \quad &\Rightarrow \quad \text{MCAR} \\
\alpha_{ij} = 0 \quad i = 1,2,3 \quad j = 6,7 \quad &\Rightarrow \quad \text{MAR} \\
\alpha_{ij} = 0 \quad i = 1,2,3 \quad j = 7 \quad &\Rightarrow \quad \text{Non-ignorable (first case) NI1} \\
\alpha_{ij} = 0 \quad i = 1,2,3 \quad j = 6 \quad &\Rightarrow \quad \text{Non-ignorable (second case) NI2} \\
\text{No constrictions on } \alpha_{ij} \text{ values} \quad &\Rightarrow \quad \text{Non-ignorable (third case) NI3}
\end{aligned}$$

Table 4.1 shows the estimated relative quartiles for the positive tuberculin group ($PPD = 1$) with respect to the negative tuberculin ($PPD = 0$), under different assumed sets of surrogate covariates and, for each of them, by using the above mentioned non-response patterns.

The use of other models based on the same set of covariates would allow to perform a more exhaustive sensitivity analysis of the resulting estimates and it would provide a global response to the problem as a function of the underlying non-response pattern.

This methodology has been implemented in S-PLUS and run in a PC-Pentium Pro, 200 Mhz, 32 Mb RAM computer, in a Windows 95 environment.

4.4.5 Results

Concerning the MCAR validation, we have seen that, although there exist efficient tests for its validation under a monotone non-response pattern, the existing tools for the non monotone case are not powerful enough. In other words, if we use these techniques we need more statistical evidence in the data to reject the MCAR non-response hypothesis. In this illustration it would be necessary to use a maximum confidence level of 92.4% to reject the MCAR assumption.

The sensitivity of the estimation of the parameters to the non-response model assumption, as it is shown on Table 4.1, proves that the non-response pattern is not MCAR and it illustrates the low power of that methodology.

Surrogate covariates $\boldsymbol{V}$	Non-response pattern M_i	First quartile	Median	Third quartile	*Deviance* D_{M_i}	Number of parameters n_i
None	MCAR=MAR	2.983	1.630	1.012	4102.055	12
	NI1	2.293	1.384	0.929	4092.145	15
	NI2	2.675	1.479	0.927	4092.993	15
	NI3	1.784	1.125	0.782	4085.622	18
TR	MCAR/MAR	2.674	1.489	0.939	4522.524/4513.648	18/21
	NI1	1.973	1.270	0.897	4504.61	24
	NI2	2.640	1.474	0.931	4504.236	24
	NI3	1.780	1.121	0.779	4496.693	27
RA	MCAR/MAR	2.963	1.592	0.976	4688.37/4683.306	24/30
	NI1	2.251	1.341	0.892	4672.024	33
	NI2	2.652	1.483	0.938	4674.658	33
	NI3	1.178	1.122	0.781	4664.661	36
BA	MCAR/MAR	2.912	1.644	1.048	4962.181/4957.538	24/30
	NI1	2.262	1.391	0.948	4947.412	33
	NI2	2.657	1.484	0.937	4947.574	33
	NI3	1.776	1.124	0.783	4939.764	36
TR, RA	MCAR/MAR	2.751	1.496	0.926	5102.458/5087.667	42/51
	NI1	2.013	1.265	0.877	5077.458	54
	NI2	2.623	1.479	0.942	5077.824	54
	NI3	2.039	1.284	0.892	5070.985	57
TR, BA	MCAR/MAR	2.102	1.264	0.847	5356.891/5343.023	42/51
	NI1	2.164	1.229	0.787	5334.777	54
	NI2	2.182	1.286	0.848	5340.131	54
	NI3	1.245	1.423	1.580	5326.711	57
RA, BA	MCAR/MAR	2.890	1.641	1.050	5513.294/5504.302	60/72
	NI1	2.316	1.426	0.974	5492.744	75
	NI2	2.740	1.524	0.960	5494.363	75
	NI3	2.273	1.415	0.973	5487.814	78
TR, RA, BA	MCAR/MAR	2.739	1.451	0.880	5888.679/5863.155	114/129
	NI1	2.691	1.442	0.881	5853.455	132
	NI2	2.773	1.456	0.877	5854.811	132
	NI3	1.898	1.177	0.807	5838.814	135

Table 4.1: *Estimated relative quartiles for the positive tuberculin group versus the negative tuberculin group, under different assumed set of surrogate covariates ($\boldsymbol{V}$) and non-response patterns(M_i)*

Concerning the parametric estimation and under a fixed set of surrogate covariates, in order to compare results coming from different non-response pattern assumptions, we will use the statistic resulting of the difference between the respective deviances ($-2\log L$).

If we denote by D_{M_i} the deviance obtained under the model M_i, $i = 1, \ldots, 5$, the statistic $\Delta_{M_iM_j} = D_{M_i} - D_{M_j}$ follows a $\chi^2_{n_j-n_i}$ distribution, where n_i and n_j ($n_i < n_j$) are the number of parameters to be estimated under each one of the nested models M_i and M_j. Table 4.2 shows the p-values of the comparisons between the parametric models, under different surrogate covariates ($\boldsymbol{V}$) and non-response pattern models (M_i).

Analyzing Tables 4.1 and 4.2 we can conclude:

1. In our illustration, the MAR model is not significative with respect to the MCAR one. Therefore, the observation of the $CD4$ and PPD covariates does not depend heavily on the observed values of the surrogate covariates.

Specifically, we only obtain significative differences when we use the indicator of having been previously treated against tuberculosis as a surrogate. In this case the estimates for the coefficients of the variable TR are positive; this means, and it seems to be logical, that patients which have treatment against tuberculosis history have a larger probability of being observed the $CD4$ and PPD variables.

2. In all the scenarios, non-ignorable models are significative with respect to the respective MCAR and MAR. This fact allows us to establish that, the probability of observing the $CD4$ and PPD variables depends on the potential values of these variables. Note that this result would not have been possible from hypothesis tests based on the observed data (Section 4.3.1)

3. The positivity in the tuberculin skin test is a better prognosis factor for short-term survival; however, the long-term survival is worse. To illustrate this fact we present in Figure 4.1 the estimation of the survival function stratified by the immunosuppression level and the positive or negative tuberculin skin test. The estimates correspond to the scenario in which the surrogates variables are TR and RA, and the assumed non-response pattern is NI2.

Surrogate covariates $\boldsymbol{V}$	Non-response pattern M_i	Non-response pattern M_j	$\Delta_{M_iM_j}$	Degrees of freedom $n_j - n_i$	p–value
None	MCAR=MAR	NI1	9.91	3	0.019*
	MCAR=MAR	NI2	9.062	3	0.029*
	NI1	NI3	6.523	3	0.089*
	NI2	NI3	7.371	3	0.061*
TR	MCAR	MAR	8.876	3	0.031*
	MAR	NI1	9.038	3	0.029*
	MAR	NI2	9.412	3	0.024*
	NI1	NI3	7.917	3	0.048*
	NI2	NI3	7.543	3	0.057
RA	MCAR	MAR	5.064	6	0.536
	MAR	NI1	11.282	3	0.010*
	MAR	NI2	8.648	3	0.034*
	NI1	NI3	7.363	3	0.061
	NI2	NI3	9.997	3	0.019*
BA	MCAR	MAR	4.643	6	0.590
	MAR	NI1	10.126	3	0.018*
	MAR	NI2	9.964	3	0.019*
	NI1	NI3	7.648	3	0.054
	NI2	NI3	7.81	3	0.050
TR, RA	MCAR	MAR	14.791	9	0.097
	MAR	NI1	10.209	3	0.017*
	MAR	NI2	9.843	3	0.020*
	NI1	NI3	6.473	3	0.091
	NI2	NI3	6.839	3	0.077
TR, BA	MCAR	MAR	13.868	9	0.127
	MAR	NI1	8.246	3	0.041*
	MAR	NI2	2.892	3	0.409
	NI1	NI3	8.066	3	0.045*
	NI2	NI3	13.42	3	0.004**
RA, BA	MCAR	MAR	8.992	12	0.704
	MAR	NI1	11.558	3	0.009**
	MAR	NI2	9.939	3	0.019*
	NI1	NI3	4.93	3	0.177
	NI2	NI3	6.549	3	0.088
TR, RA, BA	MCAR	MAR	25.524	15	0.043*
	MAR	NI1	9.7	3	0.021*
	MAR	NI2	8.344	3	0.039*
	NI1	NI3	14.641	3	0.002**
	NI2	NI3	15.997	3	0.001**

Table 4.2: *Comparative analysis after fitting a parametric model, under several assumed surrogate covariates ($\boldsymbol{V}$) and nested parametric models for the non-response pattern (M_i). $\Delta_{M_iM_j} = (-2\log L_{M_i}) - (-2\log L_{M_j}) = D_{M_i} - D_{M_j}$.*
** 95% significative result, ** 99% significative result*

The resulting estimates for β and σ parameters of the Weibull distribution are $\beta = 6.878 + 1.362 \cdot \mathbf{1}\{CD4 > 14\} + 0.153 \cdot PPD$ and $\sigma = 1.426 + 0.244 \cdot \mathbf{1}\{CD4 > 14\} - 0.651 \cdot PPD$. By using these distributions, we derive the distribution for the quartiles in each category. For the most immunosuppression level: 164, 576 and 1547 days if the tuberculin skin test is negative, and 431, 851 and 1457 days if the tuberculin skin test is positive. Analogously, for the less immunosuppression level, 473, 2055 and 6539 days or 1241, 3040 and 6160 days, according to the negative or positive tuberculin skin test. Indeed, short-term and middle-term survivorship (until the 72th-percentile, approximatively) of the positive tuberculin test is better than the corresponding to the negative tuberculin group, independently of the immunosuppression level.

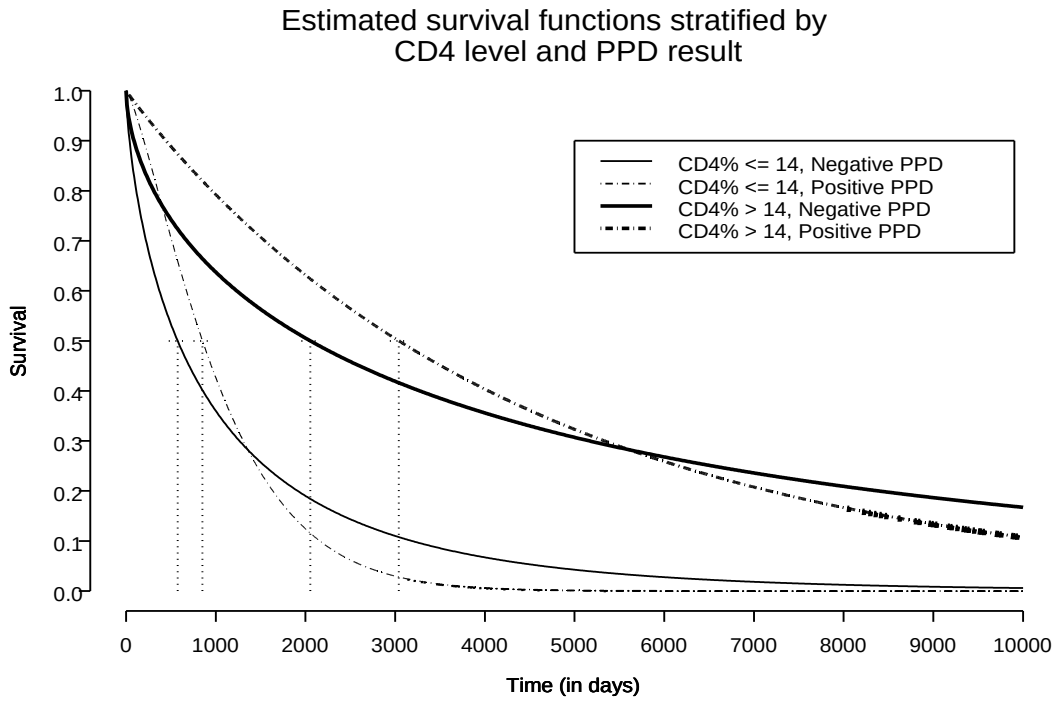

Figure 4.1: *Estimated survival functions for the HIV+PTB cohort according to the immunosuppression level (high* $\equiv CD4\% \leq 14$*, low* $\equiv CD4 > 14\%$*) and the result to the tuberculin skin test (negative* $\equiv PPD = 0$*, positive* $\equiv PPD = 1$*), when we use covariates* TR *and* RA *as surrogates and we assume the NI2 non-ignorable non-response pattern*

4.5 Discussion

The parametric methodology studied in this chapter have some drawbacks that it is necessary to know and that they restrict its applicability. First, the methodology depends on a large number of assumptions on the specification of the models that can be quite arbitrary and cannot be validated from the observed data. As a consequence, the estimators might be biased and strongly assumption-dependent.

Secondly, the choice of the surrogate covariates is crucial in the sense that they have to be based on clinical and epidemiological considerations and the collecting data process itself, in order to capture information about the non-response pattern and/or about the partially observed covariates. For these reasons, it is also necessary to perform a complementary sensitivity analysis that it allows to make reasonable interpretations of the estimates under different assumptions on the non-response pattern.

Thirdly, the geometrical growth of the parameter dimension is another limitation of this approach. As a unique alternative to solve this difficulty we propose to restrict the number of covariates of interest and the number of surrogate covariates. A reduction in the modelization (*e.g.,* introducing functional relationships between the covariates) could also reduce the dimension of the parameters in the optimization algorithm. Once more, these functional relationships will not be able to be validated from the observed data.

Finally, the execution of a complete analysis is extremely computationally costly. Time needed to estimate a model may vary, depending on the complexity of the model, from minutes in the simplest to days in the most complex.

These considerations strongly conditioned the use of the parametric approach and it makes necessary the use of less restrictive methodologies. Summarizing, all the points considered in this chapter evidence the underlying difficulties in the design, implementation and interpretation of the parametric methodology to analyze survival data with missing covariates and they illustrate that alternative ways have to be developed. In this sense, next chapters will deal with the semiparametric approach in order to solve these drawbacks as much as possible.