Chapter 1

Introduction

1.1 Motivation

Many statistical studies contain data structures with partially observed data. The sources of missingness of some of the variables may be diverse and may vary from the totally randomness to the strong dependence on the true values of the variables. A general and challenging goal in the presence of missing data is the following: *How could we make correct inferences based on the partially observed information, if we will never know the true behaviour of the unobserved data?* Our present work studies this problem, focused in the field of survival analysis, and provides answers to some of the numerous questions that can be formulated on this topic. Furthermore, based on a real dataset corresponding to an epidemiological study, we approach the problem of estimating the stratified survival function from a right censored sample with partially observed covariates.

Survival analysis, also known as Event History analysis or Reliability analysis, concerns the analysis of data corresponding to the *time to an event*. In these studies, data are the elapsed time between a defined origin until the observation of an event of interest. In epidemiology, this event of interest could be the diagnosis of a disease, the death of an individual, the response to a treatment, the occurrence of a symptom,... When performing the analysis, for instance, at the end of the study, it is common not to have observed the event of interest in all the individuals in the sample. This means that for some of the individuals of the sample we have observed the true survival time, while for the others we have only observed the *time in the study*, that is, a lower bound for the true survival time. This type of incompleteness is called *right censoring* and it is well known in the literature. Although there are other types of censoring like *left censoring, interval censoring* or *double censoring*, we will focus this thesis on the right censoring case.

Many methodologies exist for dealing with right censored samples in order to incorporate the censoring mechanism in the analysis, in particular when the censoring mechanism and the survival time are independent. From the nonparametric point of view the most popular technique is the Kaplan–Meier estimator (KM) (Kaplan and Meier, 1958). Its large-sample properties such as its convergence in distribution to a Gaussian process have been established by Breslow and Crowley (1974) and, using counting processes and martingale theory, by Fleming and Harrington (1991). From the parametric perspective, after assuming a distribution for the survival time and for the censoring mechanism, we can derive the likelihood function and apply the maximum likelihood theory to estimate the vector of parameters. If a vector of covariates is available for each individual, both methodologies can also be used to estimate the conditional survival for each category in the vector of covariates. For this estimation, if we could assume a proportional hazard between categories, we could fit a Cox's proportional hazard model (Cox, 1972) as a semiparametric method for which it is not necessary to specify a baseline distribution. However, if part of the vector of the covariates is missing none of the previous methodologies can be straightforwardly applied.

We will start introducing some basic definitions and reviewing, briefly, some of the existing methodologies to approach the missing data problem.

1.2 State of the art on missing data

A historical review of the literature on missing observations, basically related with multivariate analysis, can be found in Afifi and Elashoff (1966). Nevertheless, one of the first systematic studies on missing data can be found in the paper by Rubin (1976) and a decade later the well-known reference "*Statistical Analysis with Missing Data*" by Little and Rubin (1987). Regression with missing covariates is reviewed by Little (1992). Following Little and Rubin's terminology, a non-response pattern in a vector of data $\boldsymbol{Y} = (Y_1, \ldots, Y_p)'$ with possibly missing components is said to be *monotone* if the no observation of a component implies the no observation of the subsequent ones. This type of missingness is sometimes produced by design when, for instance, after a first set of variables have been collected, only a subsample is the target for subsequent studies. It is also encountered in longitudinal studies in which all the patients would comply with the program of visits except those that would be lost of follow-up. This is the easiest scheme of missingness and specific methodologies have been developed to deal with it. However, in many practical situations it is not realistic, either because the no observation of a component cannot imply the no observation of the others (*e.g.,* when patients only miss some visits in a longitudinal study) or because there is not a natural order within the components of $\boldsymbol{Y}$ (see Section 4.2 for more details on this definition).

A non-response pattern is said to be *missing completely at random* (MCAR) if the probabilities of observing some components and unobserving the others do not depend neither on the observed data nor on the unobserved data. If these probabilities only depend on the observed data, then the non-response pattern is said to be *missing at random* (MAR). Little and Rubin refered to these two types of non-response pattern as *ignorable* in the sense that, in a maximum likelihood approach, the estimates for the parameters of the distribution of $\boldsymbol{Y}$ do not depend on the model for the non-response pattern and therefore it can be ignored. In other words, the observed subsample is a good representation of the sample and henceforth the estimates will be unbiased –although less efficient–. However, if those probabilities depend on the unobserved data, then the non-response pattern is said to be *non-ignorable* and inferences will be wrong if this fact is not taken into account. Some authors refer to non-ignorable non-response patterns as *missing not at random.* One crucial issue is that, in general, there is no way to discard the non-ignorability of a non-response pattern. Gill and Robins (1997) and Gill, Van der Laan and Robins (1997) proved that for a given observed empirical distribution, for each plausible full-data distribution, there exist a non-response pattern that marginalize the observed data.

1.2.1 Analysis of surveys. Multiple imputation

One of the most usual scenarios with missing data is the *analysis of surveys*, and the statistical tool the most used to deal with the non-respondents is the *multiple imputation* (Rubin, 1987). Multiple imputation is the technique that replaces each missing value with a pointer to a vector of m values. The m values come from, as much, m possible scenarios or imputation procedures based either on the observed information or on historical or posterior follow-up registers. The analysis is integrated by m complete-data analyses and the result can be pooled to a summary analysis. On one hand, multiple imputation has the advantage of using complete-data methodologies for the analysis and the ability to incorporate the data collector's knowledge. On the other hand, multiple imputation allows to reproduce the uncertainty due to the sampling variability assuming that the reasons for non-response are known as well as the variability due to the uncertainty about the reasons for non-response. A examination of the sensitivity of inferences to models for the non-response in a particular survey is presented in Heitjan and Rubin (1990).

Rubin (1987) derived expressions for the estimation of the variance of the estimators after a multiple imputation. Rubin and Schenker (1986) report that multiple imputation interval estimates tend to have at least the nominal coverage in a variety of scenarios even for m as small as 2. However, many of the methodologies have been only studied under the assumption of an ignorable non-response pattern. In order to perform a multiple imputation procedure under a non-ignorable non-response, Rubin (1987) suggests some ideas like transforming ignorable imputed values to create non-ignorable imputed values. For example, after drawing m' values to impute from an ignorable non-response mechanism he proposes to distort the probability of drawing from these values using a function of themselves. Obviously, this kind of adjustments require the specification of multiple distributions and, usually, there is not enough information in the sample to validate them. Other authors have been working as well in this direction (Rubin and Schenker, 1986; Meng and Rubin, 1992; Conaway, 1993; Glynn et al., 1993; Efron, 1994; Skinner and Coker, 1996; Cook, 1997). In all these cases only small departures from ignorable non-response pattern have been considered.

The increasing power of the computational resources in the last ten years makes easier the development of software to simultaneously manage large number of variables and many non-response patterns. In these sense, recently, Van Buuren and Oudshoorn (2000) implemented a S-PLUS library called MICE (*Multivariate Imputation by Chained Equations*). In MICE, for each missing variable, a conditional distribution for the missing data given the other data can be specified, and the imputation follows after iterating over these conditional densities by means of Gibbs sampling. Elementary imputation methods in MICE include Bayesian linear regression, predictive mean matching, unconditional mean imputation, logistic and polytomous regressions as well as linear discriminant analysis and proportional odds model.

1.2.2 Parametric modeling. The EM algorithm

When the full-data model and the ignorability assumption are correct, all relevant information about the parameters is contained in the observed-data likelihood. However, the expression for this likelihood can be very complicated and special computations tools are required. Dempster, Laird and Rubin (1977) proposed the well-known Expectation-Maximization (EM) algorithm. Two improvements of the EM algorithm were developed by Louis (1982) and Meilijson (1989). More information on the EM algorithm can be found in Rubin (1991) and Laird (1993). Dempster, Laird and Rubin (1977) and Wu (1983) provide regularity conditions under which the sequence of estimates reliably converges to a stationary point. In well-behaved problems, this point is a global maximum for the observed-data likelihood, however sometimes the EM does not converge to a global maximum. Some interesting examples of these situations can be found in Schafer (1997). Based on the second derivatives of the observed-data log-likelihood, Meng and Rubin (1991b) proposed the Supplemented EM (SEM) algorithm to obtain the asymptotic variance-covariance matrix of an EM-based estimator. An extension of the EM algorithm for situation where the M-step can result hard of performing is the Expectation-Conditional Maximization (ECM) algorithm (Meng and Rubin, 1993). ECM replaces a complicated M-step with a sequence of simpler conditional maximizations. ECM preserves the good convergence properties of EM and simplifies the required computations.

1.2.3 Tables of contingency

Deep work on partially classified *tables of contingency* has been done by Fuchs (1982), assuming ignorability, or by Baker and Laird (1988) and Baker (1994a) dealing with non-ignorable non-response patterns. These approaches use the EM algorithm to obtain the estimates, although in many situations it can be replaced by a faster Iterative Proportional Fitting method (IPF) (Bishop et al., 1975). Closed-form estimates for missing counts in two-way contingency tables using log-linear models were derived by Baker, Rosenberger and DerSimonian (1992). A general approach to incomplete categorical data analysis, with special attention to the problems of identifiability and redundancy of the parameters (Catchpole and Morgan, 1997), can be found in Molenberghs and Goetghebeur (1995) and Molenberghs *et al.* (1999). An interesting and illustrative example of analysis of survey data and tables of contingency is the case of the Slovenian plebiscite in 1991 which was first analyzed by Rubin, Stern and Vehovar (1995), betting for a ignorable approach, and recently by Molenberghs, Kenward and Goetghebeur (2001), allowing for a non-ignorable non-response pattern in a more general sensitivity analysis perspective.

1.2.4 Longitudinal data analysis. Semiparametric modeling

Another field where missing data are often encountered is in *longitudinal data studies*. A basic reference on longitudinal data analysis is the book by Diggle, Liang and Zeger (1994). Diggle (1989) developed a method of testing the MAR hypothesis within experimental treatment groups and Diggle and Kenward (1994) proposed a model for continuous longitudinal data with non-ignorable or informative drop-out. This model was adapted to ordinal longitudinal data by Molenberghs, Kenward and Lesaffre (1997). An extensive review of *parametric modeling* for incomplete continuous and categorical longitudinal data, with special interest in non-ignorable non-response patterns and sensitivity analysis philosophy, can be found in Kenward and Molenberghs (1999). Liang and Zeger (1986) introduced a class of Generalized Estimating Equations (GEE) as a semiparametric extension of the generalized linear models. Robins, Rotnitzky and Zhao (1994, 1995), Robins and Rotnitzky (1995) and Rotnitzky and Robins (1995a,b) extended the GEE methodology to the Inverse Probability Weighted GEE (IPWGEE) to estimate a conditional mean when the data are MAR. Rotnitzky and Robins (1997) and Rotnitzky, Robins and Sharfstein (1998) generalized the IPWGEE methodology to repeated outcomes subject to non-ignorable non-response. In Chapter 5 *semiparametric methods* for longitudinal data with non-ignorable non-response will be considered in detail.

1.2.5 Survival analysis

In the area of the *survival analysis* with incomplete categorical covariates, log-linear models, assuming that the missing data and censoring mechanism are ignorable, were developed by Schluchter and Jackson (1989) and extended by Baker (1994b) in order to incorporate non-ignorable patterns. Lipsitz and Ibrahim (1996a,b) used the EM algorithm for survival data, also under a MAR assumption. The EM algorithm has also been used in a context of quasi-likelihood based regression (Cox, 1975), including Cox's proportional hazards model fitting, with missing covariates under a MAR non-response pattern (Paik, 1995; Paik, 1996; Paik and Tsai, 1997). In this sense Paik and Tsai (1997) suggested to impute the conditional expectation of any statistic in the partial likelihood equations involving missing covariates given the available information. As an example of semiparametric model Robins and Rotnitzky (1992), Robins (1993) and Robins, Rotnitzky and Zhao (1994) used the Inverse Probability Weighted GEE methodology to estimate the survival function under the Cox model, they described the influence functions of the resulting estimators and its efficiency. More details and references can be found in the state of the art in Chapter 5.

1.2.6 Selection models and Pattern-mixture models

One important question when dealing with missing data, specifically with non-ignorable non-response, is the modeling strategy. Little and Rubin (1987) introduced two different modeling approaches. If $\boldsymbol{R} = (R_1, \dots, R_p)'$ denotes the response indicator vector associated to $\boldsymbol{Y}$, the joint density between the full-data and the response indicator, $f(\boldsymbol{Y}, \boldsymbol{R})$, can be modeled as a *selection model*:

$$f(\boldsymbol{Y}, \boldsymbol{R}; \boldsymbol{\theta}, \boldsymbol{\phi}) = f(\boldsymbol{Y}; \boldsymbol{\theta}) \cdot f(\boldsymbol{R}|\boldsymbol{Y}; \boldsymbol{\phi}),$$

i.e., a model for the full-data and a model for the missing data mechanism, or as *pattern-mixture model*:

$$f(\boldsymbol{Y}, \boldsymbol{R}; \boldsymbol{\psi}, \boldsymbol{\gamma}) = f(\boldsymbol{R}; \boldsymbol{\psi}) \cdot f(\boldsymbol{Y}|\boldsymbol{R}; \boldsymbol{\gamma}),$$

i.e., a distribution of probability for the missing data patterns and a model for the data within each pattern. Little (1993) introduced a more general approach: the *pattern-set mixture model*. Pattern-set mixture models divide data into sets of non-response patterns and combine a mixture model across sets and a selection model within sets.

Selection models seem to be of more natural substantive formulation when the inference concerns the entire population, whereas pattern-mixture models are more natural when the interest is in population strata defined by missing data patterns. When the non-response is non-ignorable, drawbacks appear with both strategies. One on hand, selection models are sensitive to the specification of the missing data mechanism and, on the other, pattern-mixture models need to control the non-ignorability of the non-response imposing restrictions on the parameters. In all the cases parameters of missing data are often unidentified or weakly-identified from the data, thus a *sensitivity analysis* to explore the consequences of the non-ignorability is needed. The sensitivity analysis has to be based on the range of all the plausible values for the weakly identified parameters.

Both strategies have been deeply studied and discussed. An interesting discussion on both methodologies is given by Glynn, Laird and Rubin (1986). Baker and Laird (1988) applied selection models to the regression analysis of categorical variables with outcome subject to non-ignorable non-response. Robins, Rotnitzky and Zhao (1994) used a selection perspective for the conditional expectation model in a semiparametric approach. Other references are Robins (1993), Robins and Rotnitzky (1995), Rotnitzky and Robins (1995a), Robins, Rotnitzky and Zhao (1995), Zhao, Lipsitz and Lew (1996), Rotnitzky and Robins (1997), Nielsen (1998), Rotnitzky, Robins and Sharfstein (1998) and Lipsitz, Ibrahim and Zhao (1999). A general class of selection models for the analysis of non-monotone missing data have been proposed by Robins and Gill (1997) for the ignorable non-response hypothesis and for Robins (1997) for the non-ignorable case. Kenward (1998) explored the sensitivity to the selection model specification in longitudinal studies with non-random dropout. In order to avoid the impact of the parametric missing data specification in a selection model perspective, Sharfstein, Rotnitzky and Robins (1999) proposed a semiparametric approach for the missing-data mechanism. On the other hand, pattern-mixture models have been studied and applied, among others, by Little (1993, 1994, 1995), Hogan and Laird (1997), Ekholm and Skinner (1998), Molenberghs *et al.* (1998) and Michiels, Molenberghs and Lipsitz (1999a). Similarities and differences of both types of modeling have been discussed by Kenward and Molenberghs (1999) and Michiels, Molenberghs and Lipsitz (1999b).

Since our inference will be based on the population strata defined by the covariates and not by the missing data patterns, we use a selection model perspective and analyze the role of the parameters in the missing data mechanism.

1.3 About the subsequent chapters

The chapters of this thesis are organized according to the chronological order of our research. That is, from the initial analysis of an epidemiological dataset subject to right censoring and partially observed covariates, through nonparametric and parametric approaches, to the development of a new semiparametric method to estimate the stratified survival under a non-ignorable non-response pattern in the vector of covariates.

In Chapter 2 we introduce the HIV+PTB cohort which is integrated by 494 HIV-infected patients with pulmonary tuberculosis. All of them started treatment against tuberculosis and the epidemiological goal is to make inferences on the survival time from the beginning of treatment until death. The challenging and methodological problem arises when the main predictor covariates, that is the T CD4 lymphocyte counts (or percentages) and the tuberculin skin test, present, respectively, a 38.8% and a 50.4% of missingness.

Chapter 3 develops an imputation strategy following the bilinear imputation model and a bootstrap technique proposed by Efron (1994). The study includes the complete case analysis as well as the imputation scheme under the assumption of a missing at random non-response pattern. Results from this chapter have been published in Serrat and Gómez (1995) and Serrat *et al.* (1998) . Chapter 3 corresponds to the above mentioned paper "*CD4+ lymphocytes and tuberculin skin test* *as survival predictors in pulmonary tuberculosis HIV-infected patients*", where slight corrections and modifications have been made to adapt it to its present structure.

In Chapter 4 we explore the difficulties concerning the parametric approach. On one hand, we present a test to check the MAR condition under a monotone non-response pattern and, on the other, we derive the expression of the likelihood function in terms of the density function for the full-data and the distribution of the non-response pattern. This parametric approach is applied to a subsample of the HIV+PTB cohort for which other covariates, possibly surrogates for the covariates of interest, are completely observed. The chapter ends with a sensitivity analysis of the estimation of the parameters and a comparative analysis after fitting parametric models with different sets of surrogate covariates. In the discussion we analyze some drawbacks that restrict its applicability. The most relevant one concerns the dependency of the method 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 of this possible misspecification, the estimators might be seriously biased and strongly assumption-dependent. The paper by Gómez and Serrat (1999) "*Estudios de supervivencia con datos no observados. Dificultades inherentes al enfoque paramétrico*" (in spanish), that appeared in Qüestiió, has been translated to english and, after having been adapted to the PhD thesis format, it constitutes the contents of Chapter 4.

After these first two approaches we focus our interest in developing a method for survival analysis when we have a non-ignorable non-response pattern, using a semiparametric perspective. In Chapter 5 we briefly review the state of the art and the basic definitions and concepts on semiparametric theory. We introduce the generalized method of moments (GMM) (Newey and McFadden, 1994) as a general framework for the class of the Generalized Estimating Equations (GEE) (Liang and Zeger, 1986) and we define the inverse probability of being observed weighted GEE (IPWGEE) class of estimators (Robins et al., 1994; Rotnitzky et al., 1998).

Our semiparametric proposal is developed in Chapter 6. First, for right censored samples with completely observed covariates, we propose the Grouped Kaplan–Meier estimator (GKM) as an alternative to the standard KM estimator when we are interested in the survival at a finite number of fixed times of interest. However, when the covariates are partially observed, neither the stratified GKM estimator, nor the stratified KM estimator can be directly computed from the sample, because the probability of being at risk at each time in each category may not be available. Henceforth, we propose a class of estimating equations to obtain semiparametric estimates for these probabilities and then we substitute these estimates in the stratified GKM estimator. We refer to this new estimation procedure *Estimated Grouped Kaplan–Meier* estimator (EGKM). We prove that the GKM and EGKM estimators are $\sqrt{n}$-consistent and asymptotically normal distributed, and a consistent estimator for their limiting variances is derived. The advantage of the EGKM estimator is that provides asymptotically unbiased estimates for the survival under a flexible selection model for the non-response probability pattern. We illustrate the method with the HIV+PTB cohort introduced in Chapter 2. At the end of the application, a sensitivity analysis that includes all types of non-response pattern, from MCAR to non-ignorable, allows the epidemiologist to draw conclusions after analyzing all the plausible scenarios.

We close the semiparametric approach by exploring the behavior of the EGKM estimator for finite samples. In order to do that, a simulation study is carried out in Chapter 7. Simulations performed under scenarios taking into account different levels of censoring, non-response probability patterns and sample sizes show the good properties of the proposed estimator. For instance, the empirical coverage probabilities tend to the nominal ones when the non-response pattern used in the analysis is close to the true non-response pattern that generated the data. In particular, it is specially efficient in the less informative scenarios (*e.g.,* around a 80% of censoring and a 50% of missing data).

A final discussion on the different approaches and the results obtained in this work, as well as considerations on further areas of research conclude this PhD thesis in Chapter 8.

We finish this introduction mentioning some computational aspects. The imputation methodology of Chapter 3 has been developed in Pascal language and carried out on a Sun SPARC work station in a UNIX operating system. The parametric and semiparametric approaches, as well as the simulation study, have been implemented in S-PLUS and run in a PC-Pentium based computer in a Windows environment. Main sources of the programs and functions and the results in detail of the simulations are listed in the Appendices, before the Bibliography.