Contents

1 Introduction **1**

1.1 Motivation 1

1.2 State of the art on missing data 2

1.2.1 Analysis of surveys. Multiple imputation 4

1.2.2 Parametric modeling. The EM algorithm 5

1.2.3 Tables of contingency 6

1.2.4 Longitudinal data analysis. Semiparametric modeling 6

1.2.5 Survival analysis 7

1.2.6 Selection models and Pattern-mixture models 7

1.3 About the subsequent chapters 9

2 The Motivating HIV+PTB Cohort Example **13**

2.1 Introduction 13

2.2 The HIV+PTB dataset 13

2.3 The missing data problem 19

2.4 A naive pointwise lower/upper bound for survival estimates 23

3 Imputation and Bootstrap Methodologies 27

3.1 Introduction . . . 27

3.2 Methods . . . 28

3.2.1 Definitions . . . 28

3.2.2 Data . . . 29

3.2.3 Study variables . . . 29

3.2.4 Missing data problem . . . 30

3.2.5 Statistical analysis . . . 30

3.3 Results . . . 32

3.3.1 Complete data analysis . . . 32

3.3.2 Missing data analysis . . . 38

3.4 Discussion . . . 43

4 Parametric Approach 47

4.1 Introduction . . . 47

4.2 Notation and definitions . . . 49

4.3 Testing the non-response model . . . 50

4.3.1 Introduction . . . 50

4.3.2 MCAR validation . . . 50

4.3.3 Parametric approach of the problem . . . 55

4.4 Illustration with the HIV+PTB cohort . . . 58

4.4.1 Introduction . . . 58

4.4.2 Dataset and methods . . . 58

4.4.3 Validation of the MCAR assumption . . . 60

4.4.4 Parametric approach of the problem . . . 61

4.4.5 Results . . . 64

4.5 Discussion 69

5 Preliminaries on Semiparametric Theory and Missing Data Problem 71

5.1 Introduction 71

5.2 State of the art 72

5.3 From a parametric to a semiparametric point of view 74

5.4 GMM class of estimators 80

5.5 IPWGEE class of estimators 82

5.6 Efficiency 86

5.7 Sensitivity analysis 87

6 Semiparametric Approach 89

6.1 Introduction and notation 89

6.2 Grouped Kaplan–Meier (GKM) estimator 90

6.2.1 Definition 90

6.2.2 Asymptotic behavior 92

6.2.3 Asymptotic bias 96

6.2.4 Stratified Grouped Kaplan–Meier estimator $\widehat{\boldsymbol{S_x}}$ 100

6.3 Estimated (stratified) Grouped Kaplan–Meier (EGKM) estimator $\widetilde{\boldsymbol{S_x}}$ 102

6.3.1 An introductory example 103

6.3.2 Semiparametric estimation of $\boldsymbol{p_x^*}$ 105

6.3.3 Asymptotic properties of $\widetilde{\boldsymbol{p_x}}$ and $\widetilde{\widetilde{\boldsymbol{p_x}}}$ 114

6.3.4 Asymptotic properties of $\widetilde{\boldsymbol{S_x}}$ 125

6.4 Returning to the HIV+PTB cohort example 126

6.4.1 Design of the estimation 126

6.4.2 Sensitivity analysis 128

7 Simulation Study **135**

7.1 Introduction 135

7.2 Design of the simulation 135

7.3 Implementation of the simulation 143

7.4 Results 144

7.5 Discussion 154

8 Discussion and Future Research **157**

Appendices **159**

I. HIV+PTB dataset 159

II. Pascal program for the bilinear imputation 175

III. S-PLUS functions for the parametric approach 187

IV. S-PLUS functions for the semiparametric approach 193

V. Simulations Results 215

Notation 215

A. Scenarios with $T_{max} = 3$ years and $p = P(X = 1) = 0.3$ 217

B. Scenarios with $T_{max} = 3$ years and $p = P(X = 1) = 0.5$ 233

C. Scenarios with $T_{max} = 10$ years and $p = P(X = 1) = 0.3$ 249

D. Scenarios with $T_{max} = 10$ years and $p = P(X = 1) = 0.5$ 265

E. Standard errors and coverage probabilities 281

Bibliography **285**

List of Tables

2.1 *Names and description of the variables in the HIV+PTB dataset. Missing values are coded as NA*
[1] *IVDU: Intravenous drug user* 15

2.2 *Categorical covariates in the HIV+PTB dataset for 10 arbitrary cases* 16

2.3 *Overall percentages for the values of the categorical covariates in the HIV+PTB dataset*
[1] *Summarized as 0 = Non recovered, 1 = Recovered, 2 = Other*
[2] *Summarized as 0 = Other, 1 = Exclusively IVDU* 16

2.4 *Continuous variables in the HIV+PTB dataset for 10 arbitrary cases (δ and PPD binary variables are included for completeness)* 17

2.5 *Descriptive statistics for the continuous variables in the HIV+PTB dataset (δ and PPD binary variables are included for completeness)* . 17

2.6 *Table of contingency for the values in the dichotomized CD4 % and PPD. Percentages in parentheses (overall/by rows/by columns)* 19

2.7 *Lower-upper bounds for the estimation of the stratified survival for the covariates $CD4$ and PPD based on the re-allocation, at each time, of the individuals with missing covariates to the worst-best option. Results shown every three months* 24

3.1 *Estimated relative hazards in Pulmonary TB HIV-infected patients, Barcelona (1992-1994)* 34

3.2 *p-values on fitting a multivariate Cox proportional hazards model (n=157) to the Pulmonary TB HIV-infected patients, Barcelona (1992-1994)* . 36

3.3 *Estimated relative hazards and parameters estimates on fitting a Weibull model to the Pulmonary TB HIV-infected patients, Barcelona (1992-1994)* 37

3.4 *Estimated percentiles (in days) of the distributions of survival times on fitting a Weibull model to the Pulmonary TB HIV-infected patients, Barcelona (1992-1994)* 38

3.5 *Parameters estimate and estimated relative percentiles of the distributions of survival times on fitting a Weibull model to the Pulmonary TB HIV-infected patients, Barcelona (1992-1994)* 41

3.6 *Comparative study between complete data analysis and the bootstrap & imputation new methodology in Pulmonary TB HIV-infected patients, Barcelona (1992-1994)* 46

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)* 65

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)* 67

6.1 *Data example to illustrate the Estimated Grouped Kaplan–Meier estimator. $n = 10$, $\{\tau_1, \tau_2\}$ such that $t_i \leq \tau_1$ for $i = 1, \ldots, 7$ and $\tau_1 < t_i \leq \tau_2$ for $i = 8, 9, 10$* 103

6.2 *Complete case life table and stratified Grouped Kaplan–Meier estimator for categories $X = 0$ and $X = 1$ for the data in Table 6.1. $n_x, x = 0, 1$, number of individuals belonging to the category $X = x$. n_{ef}, effective sample size* 104

6.3 *Estimated life table and stratified Grouped Kaplan–Meier estimator for categories $X = 0$ and $X = 1$ for the data in Table 6.1, under the MCAR and MAR hypotheses. $n_x, x = 0, 1$, estimated number of individuals belonging to the category $X = x$. n_{ef}, effective sample size* 105

6.4 *Estimates for the survival at 1 year for categories in $CD4$ and PPD covariates (standard error, in parentheses) resulting from the complete case analysis and the semiparametric methodology for different values of τ and grid in weeks* 129

7.1 *True survival at different times (in years) for the reference distributions in each category of the covariate X* 136

7.2 *Proportion of censoring for different values of $P(X = 1)$ and different observation windows $(0, T_{max}]$* 137

7.3 *Setup of parameters $\alpha_0, \alpha_1, \alpha_2$ and τ for each non-response pattern model* 138

7.4 *Proportion of missing data for different values of $P(X = 1)$ and different observation windows $(0, T_{max}]$, for each non-response pattern* 139

7.5 *Non-response patterns used in the analysis of the simulated data* . . . 140

7.6 *Generating vs analyzing non-response pattern used in the simulation study* 140

7.7 *Configuration of the scenarios for the simulation study* 141

7.8 *Approximate upper bound (in percentage) for the relative bias for the Grouped Kaplan-Meier estimator for the two reference distributions in the Monte Carlo simulations as a function of the observation window $(0, T_{max}]$ and the grid size* 142

7.9 *Time in seconds for computing one iteration in each scenario for the analysis of a simulated data set with a non-ignorable generating and analyzing non-response pattern* 145

7.10 *Monte Carlo proportion of missing data for $T_{max} = 3$ years, different values of $P(X = 1)$ and different sample size n, for each non-response pattern* 146

7.11 *Monte Carlo relative effective sample size of the proposed methodology versus the complete case analysis, for* $T_{max} = 3$ *years and 10 years and* $P(X = 1) = 0.3$*, as a function of the grid, the non-response pattern and the sample size* (n)*.*
† *Results for* $T_{max} = 10$ *years and grid in weeks are not available* . . . 147

7.12 *Monte Carlo estimated effective proportion of individuals with* $X = 1$*, for* T_{max} *=3 / 10 years, grid in months and sample size n = 500 / 1000, as a function of the non-response patterns we use and the true values of* $P(X = 1)$ 149

7.13 *Monte Carlo mean of the estimated survivals in the simulation at 1 year and 2 years (in parentheses the standard error of the estimates) for each category and for the* $T_{max} = 3$ *years, grid in months, sample size* $n = 500$ *and* $P(X = 1) = 0.3$ *scenarios.*
Boldface: the least mean squared error estimate, Italic: the least biased estimate (if different from the least mean squared error estimate) 150

7.14 *Shortest half location parameter for the estimated standard errors (lse), coverage probability of the nominal* 95% *confidence intervals (cp) and simulated standard error (sse) for each category at 1 year for the* $T_{max} = 3$ *years, grid in months, sample size* $n = 500$ *and* $P(X = 1) = 0.3$ *scenarios* 152

7.15 *Asymptotic Relative Efficiency of the different methodologies used in the simulation at 1 year and 2 years and for each category.* ARE_1 *takes the CC methodology as the reference and* ARE_2 *uses the generating non-response pattern as analyzing pattern and reference. The scenarios correspond to* $T_{max} = 3$ *years, grid in months, sample size* $n = 500$ *and* $P(X = 1) = 0.3$*.*
Boldface: the most efficient estimate 153

List of Figures

2.1 *Boxplot of the covariate $CD4\%$ stratified by the result of the tuberculin skin test (PPD)* 20

2.2 *Kaplan–Meier estimates of the survival function for all the sample (solid line) and for the observed subsample (dotted line) for the HIV+PTB cohort* 21

2.3 *Weibull hazard plots for a) $CD4 \leq 14$, b) $CD4 > 14$, c) $CD4 \leq 14$ stratified by PPD and d) $CD4 > 14$ stratified by PPD* 22

2.4 *Lower-upper bounds for the estimation of the stratified survival for the covariates $CD4$ and PPD based on the allocation of missing values to the worst-best case, at each death-time* 25

3.1 *Kaplan–Meier estimates of survival function for Pulmonary TB HIV-infected patients, Barcelona (1992-1994), for which CD4+ lymphocytes % and tuberculin test are available (n=157), stratified by: a) CD4+ lymphocytes %, and b) tuberculin test result. m/n, the proportion of deaths* 33

3.2 *Estimated survival functions for Pulmonary TB HIV-infected patients, Barcelona (1992-1994), on fitting a Weibull model to: a) all the sample (n=494), b) cases for whom CD4+ % and tuberculin test are available (n=157), c) cases with CD4+ % $\leq$ 14 and negative tuberculin (n=80), d) cases with CD4+ % $\leq$ 14 and positive tuberculin (n=18), and e) cases with CD4+ % $>$ 14 (n=59)* 39

3.3 *Kaplan–Meier estimates of survival function for Pulmonary TB HIV-infected patients, Barcelona (1992-1994), stratified by: a) whether or not CD4+ lymphocytes % is available, and b) tuberculin test result. m/n, the proportion of deaths* 40

3.4 *90 % confidence intervals for the relative percentiles estimates for a positive tuberculin patient with respect to a negative one in Pulmonary TB HIV-infected patients, Barcelona (1992-1994): a) negative tuberculin group (reference group), b) cases with CD4+ % ≤ 14 and positive tuberculin, and c) cases with CD4+ % > 14 and positive tuberculin. Relative percentiles estimates were obtained by multiple imputation using a Weibull regression model for each subsample in each imputated bootstrap replica* 42

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* 68

6.1 *Histograms of the survival times whether the $CD4$ covariate has been observed or not, and the interval class is in months or weeks* 127

6.2 *Contour lines for $p_1 = 0.1, \ldots, 0.9$ as a function of p_0 and τ* 128

6.3 *Estimates and 95% confidence bands for the stratified survival at 1 year, for the covariates CD4 and PPD, as a function of the non-ignorability parameter τ and when the grid is in weeks* 130

6.4 *Estimates and 95% confidence bands for the stratified survival at 1 year, for the covariates CD4 and PPD, as a function of the non-ignorability parameter τ and when the grid is in months* 132

6.5 *Estimated survival functions for the covariates CD4 and PPD for four different analyzing strategies: complete case, MAR and non-ignorable with $\tau = -2$ and $\tau = 2$. The grid for the semiparametric approach is setup in weeks. Vertical line corresponds to 365 days. In parentheses, the estimated number of individuals in each category and the effective sample size* 134

7.1 *Reference survival functions for the simulation* 136