Chapter 8

Discussion and Future Research

In this work we have approached three different methodologies –nonparametric, parametric and semiparametric– to deal with data patterns with missing values in a survival analysis context. The first two approaches have been developed under the assumption that the investigator has enough information and can assume that the non-response mechanism is MCAR or MAR. In this situation, in Chapter 4 we have analyzed the drawbacks encountered to get correct inferences, as well as, we have proposed some strategies to take into account the information provided by other fully observed covariates.

However, in many situations it is impossible to assume the ignorability of the non-response probabilities. Then, as a third approach, we have proposed, as an alternative, the semiparametric methodology and a final sensitivity analysis. The first main positive feature of this approach is that we have not to model the distribution of interest –which is specially relevant because it is often not feasible to check its validity– and the problem is reduced to specify a model for the non-response probability pattern. The second advantage is that, no matter which is the specification of the non-response mechanism, the resulting estimators are $\sqrt{n}$-consistent and asymptotically normal distributed. Then, when performing a sensitivity analysis, the investigator has a tool to check the impact on the resulting inferences of the non-ignorable assumptions in the non-response mechanism.

In our work we have used a logit link model depending on all the variables of interest for the non-response probabilities. It could seem that, whereas an assumption

of ignorability often can be justified or rejected from knowledge about how the data were generated, a specific non-ignorable missing data mechanism could be almost impossible to justify. However, our election is due to the fact that it is very interesting to be able to understand the role of the non-ignorability parameters in the model (*e.g.,* in terms of log odds-ratio). We have also proposed to complement this choice with other plausible (and preferably understandable) non-response mechanisms.

We have seen that when the proportion of missing data is quite large, it is hard that any strategy becomes successful. As a consequence, more research would be needed in order to get more appropriate and complementary methodologies to solve the problem when large amount of data are missing.

We plan to address our future research in both the theoretical area as well as in the applied one. The following topics, among others, are in our next middle-term outlook.

a) Theoretical area

- To study the efficient choice for the function $\phi(.)$,

- To develop a pattern mixture, or a pattern-set mixture, approach for survival analysis,

- To compare both methodologies,

- To extend the methodology to the general Kaplan–Meier estimator,

- To generalize the findings for time dependent or continuous covariates.

b) Applied analysis and software development area

- To extend the S-PLUS functions and the sensitivity analysis to the multivariate case,

- To update and reanalyze the HIV+PTB dataset,

- To develop a public domain library to perform the semiparametric approach and the sensitivity analysis.