Resum

Molts estudis estadístics contenen patrons de dades amb valors no observats. Els motius per a la no observació d'alguna de les variables poden ser molt diversos i poden anar des de la total aleatorietat fins a una forta dependència dels valors reals de les variables. Un objectiu general, i alhora un repte en aquest tipus de situació, és el següent: *Com podem fer inferències correctes basant-nos només en la informació parcialment observada, si mai no sabrem el comportament real de les dades no observades?* El nostre treball estudia aquest problema, centrat en el terreny de l'anàlisi de la supervivència, i dóna respostes a algunes de les moltes preguntes que es poden plantejar. A més a més, a partir d'un conjunt real de dades que pertanyen a un estudi epidemiològic, estudiem el problema de l'estimació de la funció de supervivència estratificada a partir d'una mostra censurada per la dreta i amb covariants parcialment observades.

L'anàlisi de la supervivència tracta l'anàlisi de dades que corresponen a un *temps fins a un esdeveniment*. En aquests estudis, les dades són el temps transcorregut des d'un origen fins a l'observació de l'esdeveniment d'interès. En epidemiologia, aquest esdeveniment d'interès pot ser el diagnòstic d'una malaltia, la mort d'un individu, ... A l'hora de fer l'anàlisi, sovint no s'ha observat l'esdeveniment d'interès en tots els individus de la mostra. Això vol dir que per a alguns dels individus de la mostra hem observat el temps de surpervivència vertader, mentre que per als altres hem observat només el *temps en l'estudi*, és a dir, una fita inferior del temps de supervivència vertader. Aquest tipus d'incompletesa es coneix amb el nom de *censura per la dreta* i és ben conegut en la literatura. Hi ha molts mètodes per tractar mostres amb censura per la dreta per tal d'incorporar el mecanisme de censurament en l'anàlisi, més encara quan el mecanisme de censurament i el temps de supervivència són independents. De tota manera, quan part del vector de covariants no ha estat observat cap de les estratègies existents es pot aplicar de manera directa.

Un dels primers estudis sistemàtics sobre dades no observades és l'article de Rubin (1976). Deu anys més tard, conjuntament amb Little, varen publicar la coneguda referència "*Statistical Analysis with Missing Data*" (1987). Un patró de no resposta es diu que és *completament aleatori* (MCAR) si les probabilitats d'observació d'algunes components i de no observació d'altres no depenen ni de les dades obser-

vades ni de les no observades. Si aquestes probabilitats només depenen de les dades observades, aleshores el patró de no resposta es diu que és *aleatori* (MAR). Little i Rubin es refereixen a aquests dos tipus de patrons de no resposta com a *ignorables* en el sentit que, en un context de màxima versemblança, l'estimació dels paràmetres d'una distribució no depèn del model de no resposta i, per tant, el podem ignorar. En altres paraules, la submostra observada és una bona representació de tota la mostra i aleshores els estimadors seran no esbiaixats, encara que no tan eficients. De tota manera, si les esmentades probabilitats depenen de les dades no observades, aleshores el patró de no resposta es diu que és *no ignorable* i les inferències seran incorrectes si aquest fet no es té en compte. El punt clau és que, en general, no hi ha cap test que permeti descartar la hipòtesi de no ignorabilitat del patró de no resposta.

En el Capítol 1 fem una revisió de les tècniques desenvolupades per a tractar el problema de les dades no observades –imputació múltiple (Rubin, 1987), l'algorisme EM (Dempster, Laird and Rubin, 1977), la modelització semiparamètrica (Diggle, Liang i Zeger, 1994; Robins, Rotnitzky i Zhao, 1994)– així com dels diferents tipus d'anàlisis on les dades no observades poden aparèixer –anàlisi d'enquestes, taules de contingència, anàlisi de dades longitudinals, anàlisi de supervivència–. A l'hora d'estudiar el problema de les dades no observades, particularment quan el mecanisme de resposta és no ignorable, un punt molt important a considerar és el del tipus de modelització que es fa servir. Little i Rubin (1987) van introduir dues estratègies de modelització: l'anomenada *"selection model"*, que consisteix en considerar un model per al conjunt de les dades i un model de selecció com a mecanisme de no observació de les dades, i l'anomenada *"pattern-mixture model"* a on es selecciona una distribució de probabilitat per al patró de no resposta i es considera un model de distribució de les dades a cada patró. Com que les nostres inferències es basaran en les categories dins de la població definides per les covariants, més que no pas pels models de no resposta, utilitzarem un model de selecció i analitzarem el paper dels paràmetres en el mecanisme de no resposta.

En el Capítol 2 presentem la cohort HIV+PTB que està formada per 494 pacients seropositius i amb tuberculosi pulmonar. L'objectiu epidemiològic és el de fer inferències sobre el temps de supervivència entre el començament del tractament fins a la mort. El repte metodològic consisteix en què a les principals covariants

predictores de la supervivència, que són el comptador de cél.lules T CD4 (o bé el percentatge) i el resultat de la prova de la tuberculina, presenten un 38.8% i un 50.4% de dades no observades, respectivament.

El Capítol 3 desenvolupa una estratègia d'imputació seguint el model bilineal d'imputació i una tècnica bootstrap proposada per Efron (1994). A l'estudi s'inclou l'anàlisi basat en les dades completament observades així com l'esquema d'imputació a partir de la hipòtesi que el mecanisme de no resposta és aleatori (MAR).

En el Capítol 4 explorem les dificultats relatives a l'enfoc paramètric. Concretament, obtenim una expressió per a la funció de versemblança en termes de la funció de densitat del conjunt de les dades i de la distribució del patró de no resposta. Aquest enfoc paramètric l'apliquem també a una submostra de la cohort HIV+PTB. A la discussió analitzem alguns inconvenients que restringeixen l'aplicabilitat de la metodologia. El punt més important rau en el fet que la metodologia depèn en gran manera de les modelitzacions que es considerin que, per altra banda, poden ser arbitràries i mai es poden validar a partir de les dades observades.

Després d'aquests dos primers enfocs, ens centrem en desenvolupar un mètode per a l'anàlisi de supervivència en un situació d'un patró no ignorable de no resposta, i des d'una perspectiva semiparamètrica. Primer, en el Capítol 5 fem un breu repàs a l'estat de l'art i als conceptes i definicions bàsics sobre teoria semiparamètrica. Partim del mètode dels moments generalitzat (GMM) (Newey and McFadden, 1994) com a enfoc general per a la classe de les equacions generals d'estimació (GEE) (Liang and Zeger, 1986) i definim la classe d'estimadors de les GEE inversament ponderades per la probabilitat de ser observat (IPWGEE) (Robins et al., 1994; Rotnitzky et al., 1998).

La nostra aportació semiparamètrica està desenvolupada en el Capítol 6. Primer, per a una mostra censurada per la dreta però amb covariants completament observades, proposem l'estimador Kaplan–Meier agrupat (GKM) com una alternativa a l'estimador KM estàndard per als casos en què ens interessem per la supervivència en un nombre finit de temps prèviament prefixats. De tota manera, quan les covariants estan parcialment observades no podem calcular cap dels dos estimadors a partir de la mostra, perquè la probabilitat d'estar a risc en un cert moment en un categoria pot no ser disponible. És a partir d'aquí que proposem una classe d'equacions d'estimació per tal d'obtenir estimacions semiparamètriques d'aquestes

probabilitats per tal de després substituir-les en l'estimador GKM. Hem anomenat estimador *Kaplan–Meier agrupat i estimat* (EGKM) al nou estimador que proposem. Demostrem que els estimadors GKM i EGKM són $\sqrt{n}$-consistents i que asimptòticament tenen una distribució normal, i obtenim un estimador consistent per a la variància límit. L'avantatge de l'estimador EGKM és que dóna estimacions asimptòticament no esbiaixades per a la supervivència a partir d'un model de selecció flexible per al patró de no resposta. Novament, il.lustrem el mètode amb la cohort HIV+PTB introduïda al Capítol 2. L'aplicació acaba amb una anàlisis de sensibilitat que contempla tots els tipus de patró de no resposta, des del MCAR al no ignorable, i que permet a l'epidemiòleg obtenir conclusions a partir d'analitzar tots els escenaris plausibles de considerar.

Acabem l'enfoc semiparamètric estudiant el comportament de l'estimador EGKM per a mostres finites. Per fer-ho, duem a terme un estudi de simulació en el Capítol 7. Les simulacions fetes en escenaris que tenen en compte diferents nivells de censurament, de models de no resposta i de grandàries mostrals, proven les bones propietats per a mostres finites de l'estimador proposat. Per exemple, les probabilitats de cobertura empíriques tendeixen a la nominal quan el model de no resposta que fem servir en l'anàlisi és proper al vertader patró de no resposta que ha generat les dades. Concretament, l'estimador és particularment eficient en els escenaris menys informatius (*e.g.,* quan hi ha un 80% de censura i un 50% de dades no observades).

Un capítol dedicat a les conclusions, discussió i recerca futura a desenvolupar, tanca la tesi.

Resumen

Muchos estudios estadísticos contienen patrones de datos con valores no observados. Los motivos para la no observación de alguna de las variables pueden ser muy diversos y pueden ir desde la total aleatoriedad hasta una fuerte dependencia de los valores reales de las variables. Un objetivo general y al mismo tiempo un reto en este tipo de situación es el siguiente: *¿Cómo podemos hacer inferencias correctas basándonos solamente en la información parcialmente observada, si nunca sabremos el comportamiento real de los datos no observados?* Nuestro trabajo estudia este problema, centrado en el área del análisis de la supervivencia y da respuestas a algunas de las muchas preguntas que se pueden plantear. Además, a partir de un conjunto real de datos que pertenecen a un estudio epidemiológico, estudiamos el problema de la estimación de la función de supervivencia estratificada a partir de una muestra censurada por la derecha y con covariantes parcialmente observados.

El análisis de la supervivencia trata el análisis de datos que corresponden a un *tiempo hasta un suceso*. En estos estudios, los datos son el tiempo transcurrido desde un orígen hasta la observación del suceso de interés. En epidemiología, este suceso de interés puede ser el diagnóstico de una enfermedad, la muerte de un individuo, ... En el momento de hacer el análisis, a menudo no se ha observado el suceso de interés en todos los individuos de la muestra. Esto quiere decir que para algunos de los individuos de la muestra hemos observado el tiempo de supervivencia verdadero, mientras que para los otros hemos observado solamente el *tiempo en el estudio*, es decir, una cota inferior del tiempo de supervivencia verdadero. Este tipo de incompletitud se conoce con el mombre de *censura por la derecha* y es bien conocida en la literatura. Hay muchos métodos para tratar muestras con censura por la derecha que incorporen el mecanismo de censura en el análisis, más aún cuando el mecanismo de censura y el tiempo de supervivencia son independientes. De todas formas, cuando parte del vector de covariantes no ha sido observado, ninguna de las estrategias existentes puede aplicarse de manera directa.

Uno de los primeros estudios sistemáticos sobre datos no observados lo tenemos en el artículo de Rubin (1976). Diez años más tarde, conjuntamente con Little, publican la conocida referencia "*Statistical Analysis with Mising Data*" (1987). Un patrón de no respuesta se dice que es *completamente aleatorio* (MAR) si las probabilidades de observación de algunas componentes y de no observación de otras no

depende ni de los datos observados ni de los no observados. Si estas probabilidades solamente dependen de los datos observados, entonces el patrón de no respuesta se dice que es *aleatorio* (MAR). Little y Rubin se refieren a estos dos tipos de patrones de no respuesta como *ignorables* en el sentido de que, en un contexto de máxima verosimilitud, la estimación de los parámetros de una distribución no depende del modelo de no respuesta y por tanto, lo podemos ignorar. En otras palabras, la submuestra observada es una buena representación de toda la muestra y entonces los estimadores serán no sesgados, aunque no tan eficientes. De cualquier forma, si las mencionadas probabilidades dependen de los datos no observados, entonces el patrón de no respuesta se dice que es *no ignorable* y las inferencias serán incorrectas si no se tiene en cuenta este hecho. El punto clave es que, en general, no hay ningúna prueba de hipótesis que permita descartar la no ignorabilidad del patrón de no respuesta.

En el Capítulo 1 hacemos una revisión de las técnicas desarrolladas para tratar el problema de los datos no observados –imputación múltiple (Rubin, 1987), el algoritmo EM (Dempster, Laird y Rubin, 1977), la modelización semiparamétrica (Diggle, Liang y Zeger, 1994; Robins, Rotnitzky y Zhao, 1994)– así como de los diferentes tipos de análisis donde los datos no observados pueden aparecer –análisis de encuestas, tablas de contingencia, análisis de datos longitudinales, análisis de supervivencia–. Cuando se estudia el problema de los datos no observados, particularmente cuando el mecanismo de respuesta es no ignorable, un punto muy importante a considerar es el del tipo de modelización que se emplea. Little y Rubin (1987) introdujeron dos estrategias de modelización: la llamada *"selection model"*, que consiste en considerar un modelo para el conjunto de los datos y un modelo de selección como mecanismo de no observación de los datos, y la llamada *"pattern-mixture model"*, en donde se elige una distribución de probabilidad para el patrón de no respuesta y un modelo de distribución de los datos en cada patrón. Como nuestras inferencias se basarán en las categorías dentro de la población o definidas por los covariantes, más que por los modelos de no respuesta, utilizaremos un modelo de selección y analizaremos el papel de los parámetros en el mecanismo de no respuesta.

En el Capítulo 2 presentamos la cohorte HIV+PTB que está formada por 494 pacientes seropositivos y con tuberculosis pulmonar. El objetivo epidemiológico es

hacer inferencias sobre el tiempo de supervivencia entre el inicio del tratamiento y la muerte. El reto metodológico consiste en que los principales covariantes predictores de la supervivencia, que son el contador de células T CD4 (o bien el porcentaje) y el resultado de la prueba de la tuberculina, presentan un 38.8% y un 50.4% de datos no observados, respectivamente.

El Capítulo 3 desarrolla una estrategia de imputación siguiendo el modelo bilineal de imputación y una técnica bootstrap propuesta por Efron (1994). En el estudio se incluye el análisis basado en los datos completamente observados, así como el esquema de imputación a partir de la hipótesis que el mecanismo de no respuesta es aleatorio (MAR).

En el Capítulo 4 exploramos las dificultades relativas al enfoque paramétrico. Concretamente, obtenemos una expresión para la función de verosimilitud en términos de la función de densidad del conjunto de los datos y de la distribución del patrón de no respuesta. Este enfoque paramétrico lo aplicamos también a una submuestra de la cohorte HIV+PTB. En la discusión analizamos algunos de los inconvenientes que limitan la aplicabilidad de la metodología. El punto más importante radica en el hecho de que la metodología depende en gran manera de las modelizaciones que se consideren que, por otro lado, pueden ser arbitrarias y nunca se pueden validar a partir de los datos observados.

Después de estas dos aproximaciones concentramos nuestro interés en desarrollar un método para el análisis de la supervivencia cuando tenemos un patrón de no respuesta no ignorable, utilizando una perspectiva semiparamétrica.

Primero, en el Capítulo 5 revisamos brevemente el estado del arte y las definiciones y conceptos básicos sobre teoría semiparamétrica. Partimos del método de los momentos generalizado (GMM) (Newey y McFadden, 1994) como enfoque general para la clase de las ecuaciones generales de estimación (GEE) (Liang y Zeger, 1986) y definimos la clase de estimadores de las GEE inversamente ponderadas por la probabilidad de ser observado (IPWGEE) (Robins et al., 1994; Rotnitzky et al., 1998).

Nuestra aportación semiparamétrica está desarrollada en el Capítulo 6. Primero, para una muestra censurada por la derecha pero con covariantes completamente observados, proponemos el estimador Kaplan-Meier agrupado (GKM) como una

alternativa al estimador KM estándar para los casos en que nos interesemos por la supervivencia en un número finito de tiempos previamente prefijados. De todas formas, cuando los covariantes están parcialmente observados no podemos calcular ninguno de los dos estimadores a partir de la muestra, porqué la probabilidad de estar a riesgo en un cierto momento en una categoría puede no estar disponible. Es a partir de aquí que proponemos una clase de ecuaciones de estimación para obtener estimaciones semiparamétricas de estas probabilidades para después sustituirlas en el estimador GKM. Hemos nombrado estimador *Kaplan-Meier agrupado y estimado* (EGKM) al nuevo estimador que proponemos. Demostramos que los estimadores GKM y EGKM son $\sqrt{n}$-consistentes y que asimptoticamente tienen una distribución normal y obtenemos un estimador consistente para la varianza límite. La ventaja del estimador EGKM es que proporciona estimaciones asintóticamente no sesgadas para la supervivencia a partir de un modelo de selección flexible para el patrón de no respuesta. Nuevamente, ilustramos el método con la cohorte HIV+PTB introducida en el Capítulo 2. La aplicación termina con un análisis de sensibilidad que contempla todos los tipos de patrón de no respuesta, desde el MCAR al no ignorable y que permite al epidemiólogo obtener conclusiones a partir de analizar todos los escenarios plausibles.

Terminamos el enfoque semiparamétrico estudiando el comportamiento del estimador EGKM para muestras finitas. Para ello, llevamos a cabo un estudio de simulación en el Capítulo 7. Las simulaciones hechas en escenarios que tienen en cuenta diferentes niveles de censura, de modelos de no respuesta y de tamaño muestral, prueban las buenas propiedades del estimador propuesto, para muestras finitas. Por ejemplo, las probabilidades de cobertura empíricas tienden a la nominal cuando el modelo de no respuesta que utilizamos en el análisis está próximo al verdadero patrón de no respuesta que ha generado los datos. Concretamente, el estimador es particularmente eficiente en los escenarios menos informativos (e.g., cuando hay un 80% de censura y un 50% de datos no observados).

Un capítulo dedicado a las conclusiones, discusión e investigación futura a desarrollar, pone punto final a la tesis.

Abstract

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 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,... When performing the analysis, 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. 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. However, when part of the vector of the covariates is missing none of the existing approaches can be straightforwardly applied.

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). 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 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. One crucial issue is that, in general, there is no way to discard the non-ignorability of a non-response pattern.

In Chapter 1 we review the existing techniques to approach the missing data problem –multiple imputation (Rubin, 1987), the EM algorithm (Dempster, Laird and Rubin, 1977), semiparametric modeling (Diggle, Liang and Zeger, 1994; Robins, Rotnitzky and Zhao, 1994)– as well as the different contexts where the missing data appears –analysis of surveys, tables of contingency, longitudinal data analysis, survival analysis–. 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: *selection model*, that is, a model for the full-data and a model for the missing data mechanism, and *pattern-mixture model*, that is, a distribution of probability for the missing data patterns and a model for the data within each pattern. 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.

In Chapter 2 we introduce the HIV+PTB cohort which is integrated by 494 HIV-infected patients with pulmonary tuberculosis. 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.

In Chapter 4 we explore the difficulties concerning the parametric approach. In particular, 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. 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.

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. First, 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 contribution 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 discussion and future research chapter closes this thesis.