Survival Data Analysis with Heavy-Censoring and Long-Term Survivors

Doctoral Dissertation by

Lucas López Segovia

Advised by

Guadalupe Gómez Melis

Universitat Politècnica de Catalunya

and

Anna Espinal Berenguer

Universitat Autònoma de Barcelona

Universitat Politècnica de Catalunya

Barcelona, February 2014

Devoted to

All women who have trusted and supported me in one form or another in my professional development. For those hard-working women, who always encouraged me to continue on my way, my biggest acknowledgement. In special to Doña Conchita Segovia, my mother.

Acknowledgements

Some years ago, I began the most important challenge of my life, the realization of a PhD degree in statistics. During the tour I met different people who contributed in one way or another in my research. Now that I've reached the end of my studies, I would like to thank to these people who have had the biggest influence on this project.

First of all I want to express my gratitude and appreciation to my advisors, Guadalupe Gómez Melis and Anna Espinal Berenguer. I am indebted to them for welcoming me to the Research Group in Statistical Analysis of Survival (GRASS: Grup de Recerca en Anàlisi eStadística de la Supervivència) as well as the guidance that they gave me during my doctoral research. Not only they were readily available for me, but they also always answered my doubts and my enquiries. I deeply appreciate their support, trust and encouragement with this project. Both are two excellent advisors and fantastic people. To them: Moltíssimes gràcies.

Thanks a lot to the members of the GRASS group of the Department of Statistics and Operations Research at the Universitat Politècnica de Catalunya. From the first moment, I have enjoyed the exchange of ideas in the several seminars and meetings that we have had throughout these years. Thanks for all the spent time and for their friendship.

My sincere thanks to all the graduate students, professors and staff of the Department of Statistics and Operations Research at the Universitat Politècnica de Catalunya, all of them were essential to the conduct of my research.

I express gratitude to all the institutions that have supported my doctoral studies: To the Secretaria de Educación Publica (SEP) en México by the PROMEP program. To the Universidad Juárez Autónoma de Tabasco (UJAT), who proposed me and represented in the SEP. In special to PhD Dora Maria Frias Márquez secretary of academic services and institutional representative in the UJAT by PROMEP, because at supporting to the conclusion of my thesis. To the Universitat Politècnica de Catalunya (UPC), which gave me a research grant during the first half of 2010.

Finally, I make a special acknowledgement of gratitude towards my family who were patient during the long time daily, to my little children Iván and Pamela, and my wife Denisse, for all the time I spent away from them, Thanks. Similarly, I want to thank the understanding and support of my sisters, brothers and friends, especially my mother, Doña Conchita Segovia, my gratitude is immeasurable. I cannot dedicate this thesis to anyone but her, for all the love and care she has given me throughout my life, but over all, for teaching the first concepts in Mathematics and Statistics.

Abstract

Survival analysis with standard methods such as proportional hazards models or parametric models are commonly used to analyze time to event data. A standard assumption in survival analysis is that all individuals will have the event of interest provided the follow-up period is large enough. However, common models might be inappropriate when data contain too much right-censoring. In these cases one might think that the follow-up is not enough, maybe because the data contains long-term survivors or both. In this scenario, standard models are extended to a more general class of models that take into account long-term survivors. These individuals are mixed with those censored individuals due to the termination of the study, yielding a large proportion of censored data for which the subsample of long-term survivors cannot be distinguished from the others. One might assume that this type of data arises from the mixture of two populations, the immune individuals and the non-immune individuals and, in this case, the standard use of the Kaplan-Meier estimator could prove to be wrong since we are dealing with an improper survival function. An standard approach for these data are the cure models. However, before carrying out an analysis with a mixture cure model, one has to ensure, whether follow-up was sufficient. In addition to the problem of heavy right-censoring, sometimes we have to face with the situation in which the interest event can or not occur within a finite interval of time, for example, data coming from veterinary studies: when the event is abortion, the interest interval for the study is determined from conception to before birth, or when the event is death to weaning, the interest interval is birth-weaning. In these cases a model with bounded hazard function could be the most appropriate, or any member of the class of nonlinear transformation models introduced by Tsodikov (2003).

In this dissertation we propose recommendations for use of the Cox model in presence of heavy censoring, we implement nonparametric tests for assessment of sufficient follow-up, and we show applications of the no standard survival models. The research developed in this thesis has been motivated by two datasets, which are introduced in Chapter 2, one concerning the mortality of calves from birth to weaning while the other refers to survival of patients diagnosed with melanoma. In both cases the percentage of censoring is high, it is very likely to have immune individuals and proper analysis accounting for the possibility of a not negligible proportion of cured individuals has to be performed. Cure models are introduced in Chapter 3 together with the available software to perform the analysis, such as SAS, R and STATA, among others. We investigate the effect that heavy censoring could have on the estimation of the regression coefficients in the Cox model via a simulation study which considers several scenarios given by different sample sizes and censoring levels, results presented in Chapter 4. An application of a mixture cure model, which includes a Cox model for the survival part and a logistic model for the cure part of patients with melanoma, is described in Chapter 5. In addition, discussions about test for sufficient follow-up and censoring levels are also presented for this data. The data analysis is carried out using the macro in SAS: PSPMCM. The results show that patients with Sentinel Lymph Node (SLN): negative status to biopsy, Clark's level of invasion I-III, Histopathological of Malignant Melanoma subtype: Superficial Spreading Melanoma (SSM), younger than 46 years, and female, are more likely to be cured, whereas patients with melanoma in head and neck, Breslow's micrometric depth $\geq$ 4mm and ulceration presents, are patients with increased risk of relapse. In particular, patients with Breslow's micrometric depth $\geq$ 4mm are at higher risk for death. Furthermore, since mixture cure models do not have the property of proportional hazards for the entire population, they can be extended to non-mixture cure models by means of nonlinear transformation models as defined in Tsodikov (2003). An application of the extended hazard models is presented for the mortality of calves in Chapter 6. The methodology allows to get estimates for the cure rate as well as for genetic and environmental effects for each herd. A relevant feature of the non-mixture cure models is that they model, separately, factors which could affect survival from those affecting the cure model, making the interpretation of these models

relatively easy. Results are shown in section 6.3.1, and were obtained using the library NLTM of the statistical package R. The short (mortality) and long term (survivors) effects are determined for each factors, as well as its statistical significance in each herd. For example in the herd 1, we find that calving month and difficulty at birth is the set of statistically significant factors for the nonsusceptible (long-term survivors) proportion. Calves born in the period march-august have lower probability of survive than those born in september-february; and the probability of survive is much lower for those that have difficulties at calving for herd 1. For herd 7 the effect of difficulty at calving is different as for herd 1, here only is significative the category strongly assisted. Calves that born from strongly assisted calving have lower probability of survive that calves from without assistance calving. Regarding short-term (mortality) effects, we only find statistically significant predictors in herd 7 where the risk of death of calves born from older mothers, hence with a longer reproductive life, is twice the risk of death of calves born from younger mothers. The obtained results have been compared with those coming from standard survival models. It is also included, a discussion about the likely erroneous conclusions that may yield from standard models, without taking into account the cure.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The idea for this thesis came from the seminaries of the GRASS group (Grup de Recerca en Anàlisi eStadística de la Supervivència) on survival analysis, held at the UPC (Universitat Politècnica de Catalunya). The initial motivation was a dataset of the mortality of calves (Tarrés et al. (2005)). The particularity of this dataset was that the event of interest *mortality* could or could not occur within a period of finite time: the birth-weaning period. Furthermore due to the fact that the majority of the calves survived the study, the data collected presented heavy right censoring. Furthermore the calves can be classified as immune or non-immune (equivalent to die or not within the birth-weaning period). A second motivation was given by a sample of diagnosed patients with skin cancer, being the relapse (or death) due to cancer the event of interest. As patients receive a treatment, most of them do not experience a relapse (long-term survival), resulting survival data with high level of right censoring. These data types can be analyzed via survival cure models. The cure model was first proposed as a scheme where the event of interest can or cannot happen in a period of sufficiently long time (Boag (1949), Berkson and Gage (1952)), thus, it is considered as an interval of infinite time. The cure model has been widely studied by many researchers, among them Maller and Zhou (1996). In their work, they have summarized the majority of the contributions on the subject until 1996, and this book will be our basic point of reference.

The main objective in this thesis will be to extend the cure models to a situation with survival data subject to a heavy right-censoring level. Our purpose is to investigate the effects due to the heavy right-censoring on the regression coefficients in the Cox model when the assumption of this model is true. Determine the appropriate sample size to ensure acceptable results using the Cox model in a scenario of heavy censoring. Identify the causes of heavy censoring and the most appropriate models for the analysis in such scenario. Show the advantages of the cure models with respect to the standard models. Implement statistical tests to determine whether the follow-up is or not enough, when the existence of immune and susceptible individuals is suspected in the population. Identify the cure models that do not have the property of proportional hazard, study the properties of their estimated parameters. Review the useful statistical software to perform an analysis using a survival model that takes into account long term survivors. Apply the extended hazard models to the data of calves, and compare the results with those obtained using a standard model.

This chapter aims to make an introduction about the mixture cure models in survival analysis. In section 1.1 we give the basic concepts and notation needed for the subsequent chapters. In section 1.2 we present the notation for right-censored data. Section 1.3 introduces the formulation of mixture cure models. Finally, we give an outline of the subsequent chapters in Section 1.4.

1.1 Notation and definitions

In this thesis the random variables are denoted by capital letters, for example T, C, U, Y. Covariates or covariate vectors are denoted by small letters, say x or z.

Let T be a positive random variable representing the survival time, defined as the time until an event of interest occurs. F denotes the distribution function of T and S the survival function defined by $S(t) = 1 - F(t)$. $S_P(t)$ denotes the survival function of T in a specific population.

Let C be a positive random variable to represent the time to censoring, with distribution function G.

***Definition 1*:** A survival function $S(t)$ is said to be *improper*, if $S(\infty) > 0$.

***Definition 2*:** The right extreme of the distribution function $F(t)$ is defined as $\tau_F = \inf\{t \geq 0 : F(t) = 1\}$.

***Definition 3*:** An individual who does not present the event of interest during a long enough follow-up, is defined as a *long-term survivor, immune, cured or nonsusceptible.* While an individual who presents the event is defined as *susceptible or uncured.*

***Definition 4*:** Let $\varphi(u)$ be a nonnegative strictly monotonically decreasing function defined for all $u \geq 0$, such that its first derivative φ' is continuous and $\varphi(0) = -\varphi'(0) = 1$. The φ-hazard rate $r(t)$ for the survival function $S(t)$ is defined by the relation, $r(t) = \frac{d}{dt}\varphi^{-1}(S(t))$, where φ^{-1} is the inverse function for φ. It is clear that the function $r(t)$ reduces to the traditional hazard rate $h(t)$ if one chooses φ of the form $\varphi(u) = \exp(-u)$.

***Definition 5*:** Let $r(t)$ be the φ-hazard rate for the survival function $S(t)$. Then $S(t)$ has increasing φ-hazard rate average, if $\frac{1}{t}\int_0^t r(u)du$ is increasing in $t > 0$. When $r(t)$ is the traditional hazard rate $h(t)$, then it is said that the survival function $S(t)$ has increasing hazard rate average.

1.2 Right censoring

In survival analysis, the survival time is the time from a well-defined, possibly random, starting point until some event E occurs. Some examples of survival times are: lifetime of an organism, time to re-offense of a criminal, remission times for cancer. An observation is censored, or more specifically, right-censored at time C if the survival time T is unknown

except for the fact that T is greater than C. In this thesis, we suppose that censoring is random, and that censoring times are independent of survival times.

The general problem of right censoring in survival analysis was first treated by Lagakos (1979), who represented the life process of a subject using random variables (T, U, δ), where T is the time to event E, with distribution function F, U is the observed portion of T with $U \leq T$ and, δ=1 if $U = T$ and δ=0 if $U < T$.

Nevertheless, in independent censoring models, the survival time of a subject can be represented by random variables (T, C, U, δ) where T is the time until event E with distribution F and C represents the time until censoring, with distribution G, T and C are stochastically independent. Within this formulation the observed pair (U, δ) are structurally represented by $U = \min\{T, C\}$ and, δ=1 if $U = T$ or δ=0 if $U = C$ (Williams and Lagakos (1977)). The distribution function of U (which we denote by L) is given by the relation

$$L(u) = 1 - [1 - F(u)][1 - G(u)]. \tag{1.1}$$

Furthermore,

$$P(U \leq u, \delta = 0) = P(U \leq u, T > C) = \int_0^u [1 - F(s)]dG(s). \tag{1.2}$$

1.3 Mixture survival models

Mixture survival models are widely applied to carry out survival analysis when the population being studied is heterogeneous and it is not possible to distinguish between individuals of different types. Mixture survival models are so called because they are formed from a combination of different survival functions, each one corresponding to a group of individuals of the entire population.

$$S_p(u) = \pi_1 S_1(u) + \pi_2 S_2(u) + \cdots + \pi_k S_k(u), \tag{1.3}$$

where $\sum_{i=1}^{k} \pi_i = 1$. The properties of a mixture model depend on the properties of the components involved in the mixture. Models with more than two components are rarely

used, and as explained by Lawless (2003), they can be difficult to estimate.

Mixture models with two population components were originally proposed by Boag (1949) and Berkson and Gage (1952). These models must be used together with the proportion of cures and the survival time of a disease. In this particular case, mixture models are called mixture cure fraction models or mixture cure models. By cure it is meant that an individual will have little or no risk of experiencing the event of interest again (e.g. return of disease). In this scenario, the majority of the patients survive the disease and only a small percentage of mortality can be observed, resulting in a mixture of two populations, immune patients and not immune (Maller and Zhou (1996)). All this could be complicated by the presence of a censoring random variable, and combined with the additional problem of patients cured, resulting in the problem of mixture cure models with heavy censoring.

Mixture models have been widely applied in criminology (Schmidt and Witte (1989)), finances (Cole and Gunther (1995)) and, as Farewell (1986) discusses, they should not be used indiscriminately. In order to use mixture cure models a good empirical or biological evidence of a nonsusceptible population is required as well as large sample and long-follow up combined with a not excessive censoring during the period when events can occur (Sy and Taylor (2000)).

1.4 Outline of the subsequent chapters

In this section we describe the structure and composition of the thesis. The chapters are presented thematically in relation to each other, but they can also be read and understood separately without much difficulty.

Chapter 2, contains a presentation of the datasets that have motivated the development of this thesis, the first deals with data from a clinical trial of patients with melanoma, the second deals with a calf mortality study. A preliminary analysis was done separately, showing the limitations of the standard survival model. Some questions that arise from

the analysis are made, and the subsequent chapters are developed to provide answers to these questions.

Chapter 3 presents the state of the art about cure models, from its origins to the present, including applications in various fields, as well as the development of the mixture cure models and the non-mixture cure models approaches. At the end of this chapter a review of existing software to analyse survival data with cure models is provided together with a discussion about availability.

In Chapter 4, we propose new recommendations for survival analysis specially designed for heavy censoring. We examine how the level of right censoring affects the properties of parameter estimators in survival models. First we will confront the problem of simulating survival data with a controlled level of right censoring. We must, however, obtain the theoretical results to control the censoring, and establish the bases for simulation. We begin with a simulation study to investigate the effects of heavy censoring in estimates of the regression coefficients in the Cox model. For these simulations we consider different scenarios for various sample sizes and censoring levels. An analysis of the bias, variance, relative bias and coverage of Cox´s regression coefficient estimator will be carried out, with simulated data with heavy right censoring. All the methodology for the simulation study was implemented as functions in the statistical package R (see Appendix A.2.1).

Chapter 5 presents the application of the mixture cure models to the dataset of melanoma. First a discussion about testing for sufficient follow-up is presented, second a formulation about the estimation process for a semiparametric mixture cure model is formulated via maximum likelihood and EM algorithm. The nonparametric tests for sufficient follow-up and the estimation for cure rate, were implemented with the statistical package R (see Appendix A.2.2). The results to estimate the regression coefficients in the mixture cure models were obtained using the SAS program.

Chapter 6 presents an extension of the cure models through the nonlinear transformation

models, resulting in a large family of extended hazard models. These models were recently developed, and this chapter presents an application to calves mortality data. Results presented are obtained using the statistical package R.

Finally, chapter 7 summarizes the results in this thesis and addresses those aspects which remain to be completed. When we compare the results obtained with the proposed cure models with those results calculated from standard survival models, the later mistakenly assess the effects of factors and may lead to erroneous conclusions. Moreover, our cure models are easy to interpret since the effects of the factors, both the cure and survival, are modelled separately.

Chapter 2

Motivation data

2.1 Introduction

This work is motivated by population studies from two areas of application, from oncology studies in human populations and from veterinary studies in animal populations. The first is related to a group of skin cancer patients and the second is a dataset about mortality of calves, both of which were collected in Catalonia, Spain. The objective of this chapter is to give an introduction to these studies and to make a descriptive presentation of these two datasets, which are amply discussed in the development of this thesis.

2.2 Melanoma dataset

Cancer is one of the most important diseases in the developed world for its incidence, prevalence and mortality. The cancer derived from melanocytes (cutaneous melanoma) represents 2-3% of all malignancies and is responsible for 80% of deaths from skin cancer.

As in many other neoplasms, prognostic depends on the extension of the disease. Early localized disease can be curable, but once the melanoma cells migrate to lymphatic nodes or distant sites, the disease free survival (DFS) and overall survival (OS) dramatically decays. Revisions to the melanoma staging system were published in the 7th edition of

the American Joint Committee on Cancer (AJCC) in 2009 and implemented January, 2010 (see Balch et al. (2010)). The melanoma staging system, known as the Tumour, Node, Metastasis (TNM) system is based on depth of tumor invasion, the number of regional lymph nodes involved, and the presence of metastasis (see Table 2.1). The AJCC classification stratifies patients in three main categories according to the disease extension and their prognostic. These categories are recognized: Stages I and II (limited to the

Table 2.1: TNM system for Cutaneous Melanoma

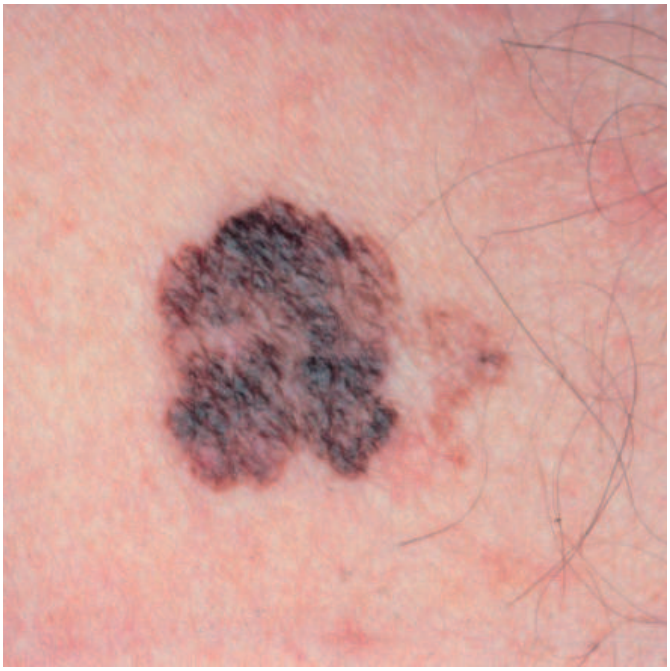
T classification	Thickness	Ulceration Status/Mitoses
Tis	N/A	N/A
T1	≤ 1.0 mm	a: w/o ulceration and mitosis <1/mm^2
		b: with ulceration or mitoses ≥ 1/mm^2
T2	1.01 - 2.0 mm	a: w/o ulceration
		b: with ulceration
T3	2.01 - 4.0 mm	a: w/o ulceration
		b: with ulceration
T4	> 4.0 mm	a: w/o ulceration
		b: with ulceration
N classification	**# of Metastatic Nodes**	**Nodal Metastatic Mass**
N0	0 nodes	N/A
N1	1 node	a: micrometastasis*
		b: macrometastasis**
N2	2-3 nodes	a: micrometastasis*
		b: macrometastasis**
		c: in-transit met(s)/satellite(s) without metastatic nodes
N3	4 or more metastatic nodes, or matted nodes, or in-transit met(s)/satellite(s) with metastatic node(s)	
M classification	**Site**	**Serum LDH**
M0	0 sites	N/A
M1a	Distant skin, subcutaneous, or nodal mets	Normal
M1b	Lung metastases	Normal
M1c	All other visceral metastases	Normal
	Any distant metastasis	Elevated

*Micrometastases are diagnosed after sentinel lymph node biopsy and completion lymphadenectomy (if performed).

**Macrometastases are defined as clinically detectable nodal metastases confirmed by therapeutic lymphadenectomy or when nodal metastasis exhibits gross extracapsular extension.

primary site), Local regional stage III disease (once the melanoma cells migrate to regional nodal basins through lymphatic vessels) and disseminated stage IV disease (once melanoma cells had spread to distant organs). In localized melanoma (see Figure 2.1), the prognostic is related to the tumor burden, and currently, the diagnostic and therapeutic protocol for this type of cancer includes local wide excision and, in tumors with Breslow thickness deeper than 1.0mm (or thinner ones but with high mitotic rate or ulceration), to explore regional nodal involvement. Complete Lymph Node Dissection (CLND) is only performed in those patients with evidence of nodal involvement. Sentinel Lymph Node (SLN) biopsy is a technique that allows the identification and analysis of the lymph node(s) in the regional basin, that receives a direct afferent drainage from a solid tumor, and therefore is at greatest risk of harboring regional metastases.

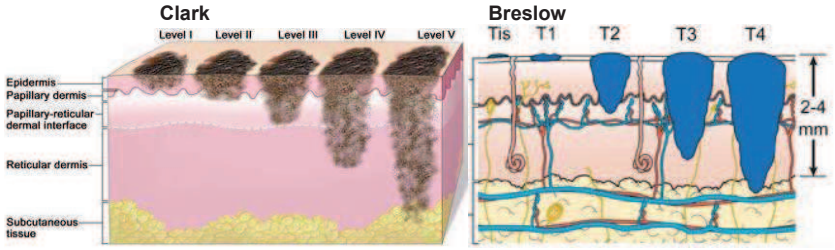
Figure 2.1: Melanoma



The two main measures to estimate the primary tumor burden are: Clark's level of invasion (a discrete measure that defines the deepest skin layer invaded by melanoma); and Breslow's micrometric depth. The Clark's classification measures the level of invasion

into the different layers of the skin. *Level I*: lesions involving only the epidermis (*in situ* melanoma); not an invasive lesion. *Level II*: tumor invades the papillary dermis, but it does not reach the papillary-reticular dermal interface. *Level III*: invasion fills and expands the papillary dermis, but the tumor does not penetrate the reticular dermis. *Level IV*: invasion into the reticular dermis but not into the subcutaneous tissue. *Level V*: invasion through the reticular dermis into the subcutaneous tissue. Breslow's micrometric depth of invasion measures the vertical thickness of the melanoma from the granular endermic layer. *Thickness 1*: 0.75 mm or less. *Thickness 2*: 0.76 mm to 1.50 mm. *Thickness 3*: 1.51 mm to 4.0 mm. *Thickness 4*: 4.0 mm or greater (see Figure 2.2).

Figure 2.2: Clark levels and Breslow thickness



2.2.1 Data description

The data corresponds to an observational study of all patients diagnosed with primary cutaneous melanoma treated at the Hospital Universitari Germans Trias i Pujol (HUGTIP) from Badalona, between September 1998 and January 2008. The dataset is formed by a group of 400 patients with melanoma who underwent SLN biopsy. The result of the biopsy was positive in 80 (20%) patients (see Table 2.2), and all of them underwent lymphadenectomy.

Variables under study

For each patient (positive, negative) different characteristics of their melanoma were evaluated, and are summarized in Table 2.2. The categorizations for each of the variables or characteristics of melanoma were given by the hospital staff responsible of the study. In this table we

Table 2.2: Characteristics of patients

Variables	Categories	Count(%)	Total
Age	≤ 45	137(34.42)	
	$46-60$	107(26.88)	
	$61-70$	84(21.10)	
	>70	70(17.58)	398
Gender	*F*	233(58.25)	
	M	167(41.75)	400
HMM subtype	*SSM*	206(55.83)	
	ALM	26(7.05)	
	LMM	4(1.08)	
	NM	133(36.04)	369
SLN Status	*Negative*	320(80.00)	
	Positive	80(20.00)	400
Localization	*Extremities*	199(49.75)	
	Head and Neck	39(9.75)	
	Trunk	162(40.50)	400
Breslow level	<1	106(26.97)	
	$[1,2)$	123(31.30)	
	$[2,4)$	101(25.70)	
	≥ 4	63(16.03)	393
Ulceration	*No*	254(70.36)	
	Yes	107(29.64)	361
Clark level	*I-III*	122(31.28)	
	IV-V	268(68.72)	390
Mitotic index	≤ 1	102(40.32)	
	>1	151(59.68)	253

Histopathological of Malignant Melanoma(HMM) subtype: Superficial Spreading Melanoma (SSM), Acral Lentiginous Melanoma (ALM), Lentigo Maligna Melanoma (LMM), Nodular Melanoma (NM).

see that the melanoma is mainly located in Extremities (49.75%) and Trunk (40.5%); Breslow's depth of melanoma is more frequent in the class [1-2)mm (31.30%), followed by the class $< 1mm$ (26.97%) and [2-4)mm (25.70%); 29.64% of patients had Ulceration; the levels of anatomical invasion of the melanoma in the skin described by Clark's level and classified into two categories were, I-III in 31.28% of the cases and IV-V in the remanning 68.72%; Histopathological of Malignant Melanoma(HMM) subtype (see Swetter et al. (2005) for melanoma subtype) more frequent were Superficial Spreading Melanoma (SSM) (55%) and Nodular Melanoma (NM) (36.04%); Mitotic index classified into two categories were, ≤ 1 mitosis/mm^2 in 40.32% of the cases and > 1 mitosis/mm^2 in the remaining 59.68%. In the same Table 2.2, we observe that 34.42% of the patients are less than 45 years old and 58.25% are women. Furthermore, characteristics of melanoma such as Breslow, Ulceration, Clark, HMM subtype, Mitotic index and Age, were not collected for all patients.

Missing data

The information for some of the variables under study is missing. In what follows we describe the characteristics of those patients with missing information. Concerning Breslow's depth of melanoma we have 7 patients for which this variable is missing. Among these, one has a positive result of the biopsy, two relapsed and one died. There are ten patients (including the previous seven) for which Clark's level is not reported. Missing information for Ulceration and HMM subtype is around 10% and 8%, respectively, and are summarized in Table 2.3. The information about Ulceration was missing for 39 patients (26 women and 13 men, most of them younger than 60 years old), among these 31 had a negative SLN, 19 of them had the melanoma Localized in the Extremities, 21 had Breslow's depth between 1 and 2mm and Clark's levels in one of the first 3 categories. The information about HMM subtype was missing for 31 patients (18 women and 13 men, most of them younger than 60 years old), among these 28 had a negative SLN, 16 of them had the melanoma Localized in the Trunk, 11 had Breslow's depth less than 1mm and 12 between 1 and 2mm, 18 had Clark's levels between IV and V categories. Mitotic index is only reported for 253 (63%) of the patients (see Table 2.2), and due to this high percentage of missing data (37%) this variable will be excluded for the next analysis.

Patients relapsed

In the group of 400 patients with a melanoma diagnosis, each patient is followed from the di-

Table 2.3: Description of missing data

		Ulceration			HMM subtype		
		SLN status			SLN status		
Variables	Characteristic	Negative	Positive	Total (%)	Negative	Positive	Total (%)
Age	*≤ 45*	13	3	16 (41.03)	13	1	14 (45.16)
	46-60	8	3	11 (28.21)	9	1	10 (32.26)
	61-70	6	1	7 (17.94)	5	0	5 (16.13)
	> 70	4	1	5 (12.82)	1	1	2 (6.45)
Gender	*F*	22	4	26 (66.67)	17	1	18 (58.06)
	M	9	4	13 (33.33)	11	2	13 (41.94)
HMM subtype	*SSM*	17	3	20 (51.28)			
	ALM	1	1	2 (5.13)			
	LMM	1	0	1 (2.56)			
	NM	7	4	11(28.21)			
	Missing	5	0	5 (12.82)	28	3	31 (100)
Localization	*Extremities*	14	5	19 (48.72)	10	0	10 (32.29)
	Head and neck	4	2	6 (15.39)	4	1	5 (16.13)
	Trunk	13	1	14 (35.89)	14	2	16 (51.58)
Breslow	*<1*	7	0	7 (17.95)	11	0	11 (35.48)
	[1-2)	17	4	21 (53.84)	11	1	12 (38.71)
	[2-4)	5	1	6 (15.39)	3	1	4 (12.90)
	≥4	0	2	2 (5.13)	2	1	3 (9.68)
	Missing	2	1	3 (7.69)	1	0	1 (3.23)
Ulceration	*No*				21	2	23 (74.19)
	Yes				2	1	3 (9.68)
	Missing	31	8	39 (100)	5	0	5 (16.13)
Clark	*I-III*	14	5	19 (48.72)	12	0	12 (38.71)
	IV-V	14	2	16 (41.02)	15	3	18 (58.06)
	Missing	3	1	4 (10.26)	1	0	1 (3.23)
Dead	*No*	30	6	36 (92.30)	28	3	31 (100.0)
	Yes	1	2	3 (7.70)	0	0	0
Relapse	*No*	29	5	34 (87.18)	26	3	29 (93.55)
	Yes	2	3	5 (12.82)	2	0	2 (6.45)

Histopathological of Malignant Melanoma(HMM) subtype: Superficial Spreading Melanoma (SSM), Acral Lentiginous Melanoma (ALM), Lentigo Maligna Melanoma (LMM), Nodular Melanoma (NM).

agnosis date until the date of the last visit at hospital (Data was collected until January 2008). During follow-up some patients relapsed or died due to melanoma (or other causes). In this dataset we observed 63 patients (15.75%) with a melanoma relapse while the remaining of the patients (84.25%) stay relapse-free.

Table 2.4 describes the characteristics of 61 (15.25%) patients who relapsed (two patients were not included in this table due to missing information in the relapse type). Among these, 39.34% of patients are older than 70 years, 24.59% are aged between 61-70, and the same percentage for 46-60 years. Among the patients who relapsed, most of them had a Visceral type (28 patients) followed by 16 with Satellitosis or in transit type, 14 with Nodal type and only 3 had a Local type. The melanoma was located in the Extremities for 29 patients, in the Trunk for 19 and the remainder 13 had melanoma in the Head and Neck. Approximately half of the patients (55%) had a Negative biopsy result, and half (55%) suffered an Ulceration. Moreover, 29 patients died due to melanoma and 30 patients were still Alive at the end of follow-up.
Two events are of great importance for research in this group of patients: relapse and death. We define two survival times of interest. First is the *disease-free survival time*, defined as the difference between the date of the relapse and the date of diagnosis. The second is the *overall survival time*, defined as the difference between the date of death and the date of the diagnosis.

2.2.2 Disease-free survival time

To analyze the disease-free survival time, we define censored patients as all patients alive without relapse at the end of the study. We have a total of 337 patients censored, the maximum censoring time was 4136 days. The maximum time until relapse was 3065 days, and is attained when the study is closed and 81 patients, still at-risk, have not relapsed. The estimated probability of survivors beyond approximately 8 years is given by $\widehat{S}_{KM}(3065) = 0.797$[1] and 95% confidence interval of [0.750, 0.847]. Observe that this probability is slightly smaller than the probability of not relapsing given by the proportion $\frac{337}{400} = 0.8425$.

Table 2.5 describes the characteristics of relapsed patients. In this table, we observe that patients with a positive SLN have a percentage of relapse higher than the patients with negative

[1]Kaplan-Meier estimator

Table 2.4: Characteristics of patients who relapsed

Variables	Categories	Nodal	(%)	Local	(%)	Satellitosis or in transit	(%)	Visceral	(%)	Total (%)
Age	≤45	3	(42.86)	0	(0.00)	1	(14.29)	3	(42.86)	7 (11.47)
	46-60	3	(20.00)	0	(0.00)	5	(33.33)	7	(46.67)	15 (24.59)
	61-70	4	(26.67)	2	(13.33)	3	(20.00)	6	(40.00)	15 (24.59)
	>70	4	(16.67)	1	(4.17)	7	(29.17)	12	(50.00)	24 (39.34)
Gender	F	2	(7.41)	1	(3.70)	10	(37.04)	14	(51.85)	27 (44.26)
	M	12	(35.29)	2	(5.88)	6	(17.65)	14	(41.18)	34 (55.73)
HMM subtype	SSM	8	(40.00)	1	(5.00)	4	(20.00)	7	(35.00)	20 (33.89)
	ALM	1	(16.67)	0	(0.00)	2	(33.33)	3	(50.00)	6 (10.16)
	LMM	0	(0.00)	0	(0.00)	0	(0.00)	1	(100.0)	1 (1.69)
	NM	5	(15.62)	2	(6.25)	9	(28.12)	16	(50.00)	32 (54.23)
SLN status	Negative	13	(38.24)	1	(2.94)	9	(26.47)	11	(32.35)	34 (55.73)
	Positive	1	(3.70)	2	(7.41)	7	(25.93)	17	(62.96)	27 (44.26)
Localization	Extremities	7	(24.14)	0	(0.00)	11	(37.93)	11	(37.93)	29 (47.54)
	Head and neck	1	(7.69)	1	(7.69)	3	(23.08)	8	(61.54)	13 (21.31)
	Trunk	6	(31.58)	2	(10.53)	2	(10.53)	9	(47.37)	19 (31.14)
Breslow	<1	2	(40.00)	0	(0.00)	1	(20.00)	2	(40.00)	5 (8.47)
	[1-2)	3	(30.00)	0	(0.00)	3	(30.00)	4	(40.00)	10 (16.94)
	[2-4)	6	(31.58)	0	(0.00)	7	(36.84)	6	(31.58)	19 (32.20)
	≥4	3	(12.00)	3	(12.00)	4	(16.00)	15	(60.00)	25 (42.37)
Ulceration	No	6	(24.00)	2	(8.00)	8	(32.00)	9	(36.00)	25 (44.64)
	Yes	8	(25.81)	1	(3.23)	7	(22.58)	15	(48.39)	31 (55.35)
Clark	I-III	1	(14.29)	0	(0.00)	3	(42.86)	3	(42.86)	7 (11.86)
	IV-V	13	(25.00)	3	(5.77)	12	(23.08)	24	(46.15)	52 (88.13)
Cexit	Dead_Melanoma	7	(24.14)	0	(0.00)	5	(17.24)	17	(58.62)	29 (47.54)
	Dead_Others	0	(0.00)	0	(0.00)	2	(100.0)	0	(0.00)	2 (3.27)
	Alive	7	(23.33)	3	(10.00)	9	(30.00)	11	(36.66)	30(49.18)

Histopathological of Malignant Melanoma(HMM) subtype: Superficial Spreading Melanoma (SSM), Acral Lentiginous Melanoma (ALM), Lentigo Maligna Melanoma (LMM), Nodular Melanoma (NM); Censoring type (Cexit).

result. We observe that the percentage of relapse due to melanoma increases when patients get older. Men are more susceptible to relapse than women. Regarding the location of melanoma, patients with melanoma in Head and Neck have a higher percentage of relapse than patients with melanoma in Extremities or Trunk. In addition, we observe that the higher the tumor volume (Breslow level) is at higher risk of relapse. Likewise patients who present ulceration have a percentage of relapse more higher than patients with non-ulcerated tumors. The group of patients with Clark I-III has a lower percentage of relapse than patients with Clark IV-V. On the other hand, the Histopathological of Malignant Melanoma (HMM) for Acral Lentiginous Melanoma (ALM), Lentigo Maligna Melanoma (LMM) and Nodular Melanoma (NM) types have a percentage of relapse higher than Superficial Spreading Melanoma (SSM) type.

It follows from Table 2.5 that patients with primary tumor ulceration or positive result of the SLN biopsy are at higher risk of a relapse compared to patients that have non ulcerated tumor or negative result in the biopsy. Similarly, it is observed that the risk of relapse increases when patients get older and when Breslow's thickness increases. These and other evidences were analyzed using the log-rank test and of Peto-Peto test, the latter to detect early differences.

Figure 2.3 shows plots of the disease-free survival curves by groups defined from categorial variables, with similar percentage of relapse (where one might suspect equality of survival curves) into variable. Table 2.6 presents log-rank and Peto-Peto statistics[2] value (Harrington and Fleming (1982)) together with the p-value for these groups. Not significant differences were found between levels. Figure 2.3 plots the survival curve according to the melanoma location (Figure 2.3 a), Breslow's categories (Figure 2.3 b), Histopathological of Malignant Melanoma (HMM) subtype (Figure 2.3 c) and Age (Figure 2.3 d). We observe that, for instance, the survival of patients with melanoma in Extremities and Trunk is very similar. Concerning HMM subtype we observe a much higher disease-free survival for Superficial Spreading Melanoma (SSM) than the rest. These considerations suggest a new categorization for these 4 variables, before considering such factors in the survival model.

[2]Nonparametric test for comparing two or more survival curves where some of the observations may be censored.

Table 2.5: Characteristics of relapsed patients

Variables	Categories	Relapsed(%)	Total
`Age`	≤ 45	7(5.10)	137
	$46-60$	15(14.01)	107
	$61-70$	17(20.23)	84
	>70	24(34.28)	70
`Gender`	*F*	27(11.58)	233
	M	36(21.55)	167
`HMM subtype`	*SSM*	20(9.70)	206
	ALM	6(23.07)	26
	LMM	1(25.00)	4
	NM	34(25.56)	133
`SLN Status`	*Negative*	36(11.25)	320
	Positive	27(33.75)	80
`Localization`	*Extremities*	30(15.07)	199
	Head and Neck	13(33.33)	39
	Trunk	20(12.34)	162
`Breslow level`	<1	5(4.71)	106
	$[1,2)$	10(8.13)	123
	$[2,4)$	21(20.79)	101
	≥ 4	25(39.68)	63
`Ulceration`	*No*	25(9.84)	254
	Yes	33(30.84)	107
`Clark level`	*I-III*	7(5.73)	122
	IV-V	54(20.14)	268

Histopathological of Malignant Melanoma(HMM) subtype: Superficial Spreading Melanoma (SSM), Acral Lentiginous Melanoma (ALM), Lentigo Maligna Melanoma (LMM), Nodular Melanoma (NM).

Cox model for the disease-free survival time

Based on the results obtained in the previous section, we present, in Table 2.7, the new recoding variables. We are now using only two categories for location type (Extremities-Trunk versus

Figure 2.3: Disease-free survival time by categories of the studied variables

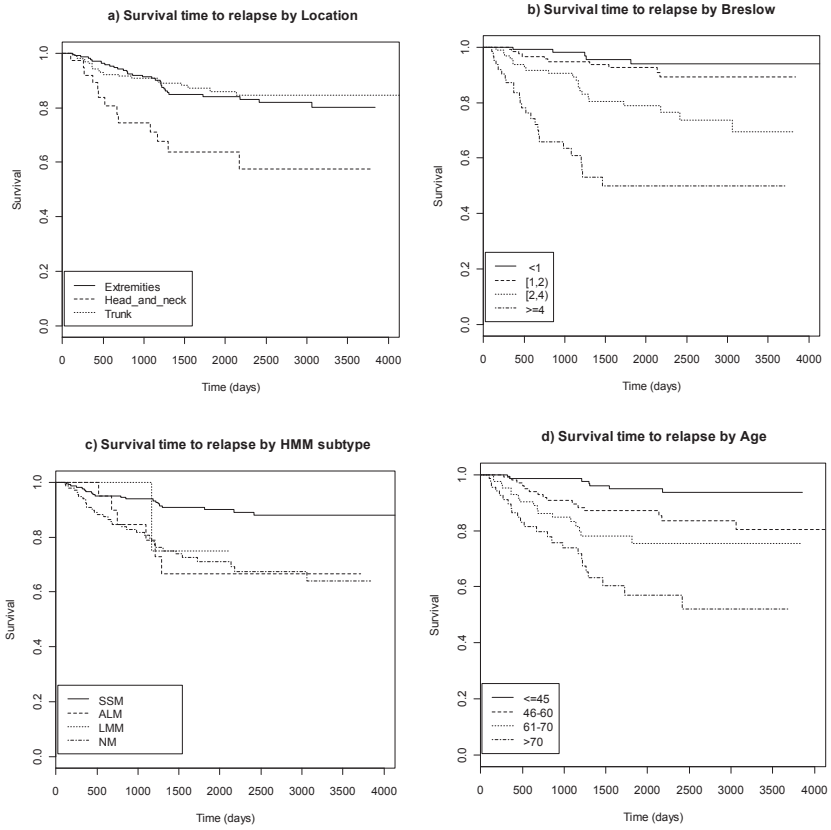


Table 2.6: Tests for equality of survival curves

Tests	Log-rank			Peto-Peto		
Groups	χ^2	*df*	*p-value*	χ^2	*df*	*p-value*
`Location`: *Extremities, Trunk*	0.3	1	0.615	0.2	1	0.674
`Breslow` : *<1, [1, 2)*	1.1	1	0.286	1.2	1	0.283
`HMM` : *ALM , LMM, NM*	0.0	2	0.976	0.1	2	0.949
`Age` : *46-60, 61-70*	1.9	1	0.170	2.1	1	0.149

Head-Neck), three categories for Breslow level ($< 2, [2, 4)$ and ≥ 4), two categories for HMM subtype (SSM versus ALM-LMM-NM) and three categories for Age (≤ 45, 46-70, and > 70). In this table we can see the distribution of the total of relapses (and percentage) by categories for each of the studied variables.

Table 2.7: Genetic factors for relapse due to melanoma

Factor	Categories	Relapsed(%)	Total
SLN status	*Negative*	36(11.25)	320
	Positive	27(33.75)	80
Localization	*Extremities-Trunk*	50(13.85)	361
	Head and Neck	13(33.33)	39
Breslow level	< 2	15(6.55)	229
	$[2, 4)$	21(20.79)	101
	≥ 4	25(39.68)	63
Ulceration	*No*	25(9.84)	254
	Yes	33(30.84)	107
Clark level	*I-III*	7(5.73)	122
	IV-V	54(20.14)	268
HMM subtype	*SSM*	20(9.70)	206
	ALM-LMM-NM	41(25.15)	163
Age	≤ 45	7(5.10)	137
	$46 - 70$	32(16.75)	191
	> 70	24(34.28)	70
Gender	*F*	27(11.58)	233
	M	36(21.55)	167

Histopathological of Malignant Melanoma(HMM) subtype: Superficial Spreading Melanoma (SSM), Acral Lentiginous Melanoma (ALM), Lentigo Maligna Melanoma (LMM), Nodular Melanoma (NM).

The results obtained with the Cox model, are summarized in Table 2.8. Highly statistically significant factors for survival are SLN status, Breslow, Ulceration, Clark, and Age. This table shows that, patients with nodal metastasis in the SLN have twice the risk of relapse than patients with negative SLN; patients with Breslow ≥ 4 have three times more risk of relapse than those with Breslow < 2; patients with Ulceration have two times more risk of relapse than those without Ulceration. Another important factor is Clark's level, patients with Clark in categories IV-V are three times more at risk of relapse than those with Clark I-III.

Table 2.8: Statistically significant factors for disease free survival using a Cox model. Reference group: SLN status= negative, Breslow level $< 2mm$, Ulceration= no, Clark level= I-III and Age$\leq$ 45 years.

Predictors	β	se(β)	p	e^{β}	$L_{.95}$	$U_{.95}$
`SLN Status`						
positive	0.718	0.288	0.012	2.050	1.164	3.610
`Breslow level`						
[2,4)mm	0.611	0.381	0.108	1.842	0.872	3.889
$\geq 4mm$	1.191	0.427	0.005	3.289	1.423	7.605
`Ulceration`						
Yes	0.758	0.311	0.014	2.135	1.160	3.928
`Clark level`						
IV-V	1.083	0.493	0.027	2.955	1.124	7.764
`Age`						
46-70	1.339	0.452	0.003	3.815	1.572	9.256
> 70	2.183	0.466	$<.001$	8.871	3.556	22.133

n=352 (48 missing observations)

The risk of relapse one increases when increasing the patient's Age: patients aged between 46 and 70 years are more than three times at risk of relapse than those younger than 45 years, and patients older than 70 years are around nine times more at risk of relapse than those younger than 45 years. These results indicate that patients with Ulcerated melanoma, thick (high Breslow and Clark) and extended to SLN have a higher risk of relapse and this risk increases if the patient is older than 45 years.

In this Thesis, testing for the proportional hazards assumption in the Cox regression model is carried out using the *cox.zph* function in R. Table 2.9 summarizes the results for the validation of the assumption of proportionality. The column *rho* is the Pearson's product-moment correlation between the scaled Schoenfeld residual and $g(t)$, where $g(.)$ is the Kaplan-Meier transformation

$(1 - S(t))$. The significance level p is for testing whether the proportional hazards assumption is true or not. As we can see all the p-values are larger than 0.13, excluding a linear dependence between the coefficient of the variable and time. Therefore, we conclude that the proportional hazards condition is satisfied.

Table 2.9: Test for proportional hazards.

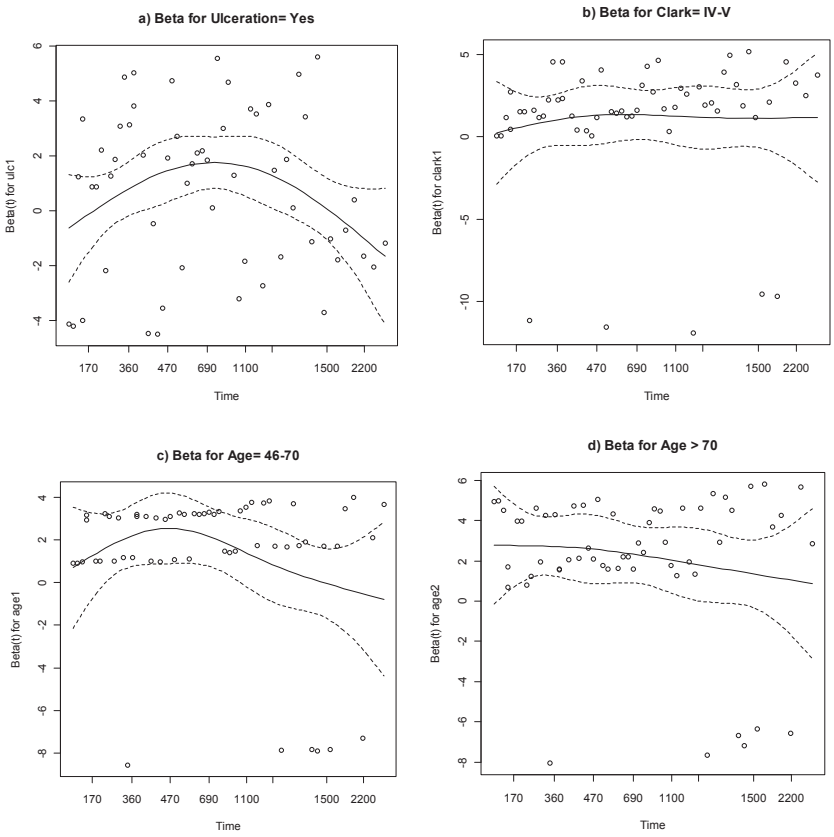
Predictors	rho	chisq	p
`SLN Status`			
positive	-0.0146	0.015	0.904
`Breslow level`			
[2,4)mm	-0.0276	0.059	0.807
$\geq 4mm$	-0.0914	0.778	0.378
`Ulceration`			
Yes	-0.0433	0.160	0.689
`Clark level`			
IV-V	-0.0403	0.102	0.749
`Age`			
46-70	-0.2054	2.266	0.132
> 70	-0.1617	1.460	0.227

Schoenfeld residuals (Schoenfeld (1982)), useful, for checking the proportional hazards assumption in Cox model, are shown in Figure 2.4 for Ulceration=Yes (Ulc1), Clark=IV-V (clark1), Age=46-70 (age1) and Age>70 (age2). Although the proportional hazards assumption is satisfied, we observe residual outliers, in Figure 2.4 b), c) and d). In the remaining variables we did not observe any outlier.

2.2.3 Overall survival time

To analyze the overall survival time, we defined censored patients as all patients who are still alive at the end of the study. We have a total of 369 patients censored, the maximum censoring time was 4136 days (the same value for disease-free time). The maximum time until death was 3358

Figure 2.4: Schoenfeld residuals



days, and is attained when 59 patients are still alive at the last visit and hence they are censored because the study was completed. The estimated probability of survivors beyond approximately 9 years is given by $\widehat{S}_{KM}(3358) = 0.872$ and 95% confidence interval of (0.824, 0.923). Notice that their value is bit smaller than the level of censoring at end the follow-up, $\frac{369}{400} = 0.9225$.

Table 2.10 describes the percentage of dead patients by categories of the studied variables. In this table, we observe that patients with a positive SLN have a percentage of death higher than the patients with a negative one. We observe that the percentage of death due to melanoma increases when patients get older. The patients with Breslow ≥ 4 have a death proportion higher than the rest of the patients. Likewise the patients who present an ulcerated tumor have a percentage of death higher than the patients who do not have. The group of patients with Clark I-III has

Table 2.10: Characteristics of dead patients

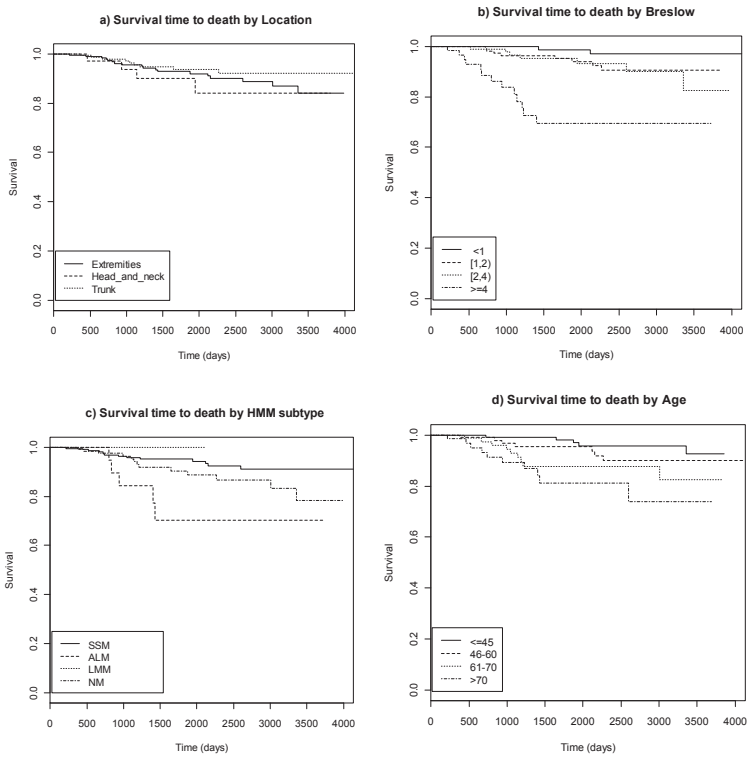
Variables	Categories	Dead(%)	Total
`Age`	≤ 45	5(3.64)	137
	$46-60$	7(6.54)	107
	$61-70$	9(10.71)	84
	>70	10(14.28)	70
`Gender`	*F*	17(7.29)	233
	M	14(8.38)	167
`HMM subtype`	*SSM*	13(6.31)	206
	ALM	5(19.23)	26
	LMM	0(0.00)	4
	NM	13(9.77)	133
`SLN status`	*Negative*	16(5.00)	320
	Positive	15(18.75)	80
`Localization`	*Extremities*	18(9.04)	199
	Head and Neck	4(10.25)	39
	Trunk	9(5.55)	162
`Breslow level`	<1	2(1.88)	106
	$[1,2)$	8(6.50)	123
	$[2,4)$	7(6.93)	101
	≥ 4	13(20.63)	63
`Ulceration`	*No*	12(4.72)	254
	Yes	16(14.95)	107
`Clark level`	*I-III*	3(2.45)	122
	IV-V	27(10.07)	268

Histopathological of Malignant Melanoma(HMM) subtype: Superficial Spreading Melanoma (SSM), Acral Lentiginous Melanoma (ALM), Lentigo Maligna Melanoma (LMM), Nodular Melanoma (NM).

a percentage of death lower than the patients with Clark IV-V. On the other hand, the HMM subtype for Acral Lentiginous type have a percentage of death higher than the rest of the patients.

From Table 2.10, similarly as in Table 2.5 for relapse, it is concluded that patients with an ulcerated melanoma or metastasis in the SLM, are at higher risk of death compared to patients that don't have ulceration or a negative SLM. Similarly, it is observed that the risk of death increases with age. For those variables where it is unclear whether there is or not a difference between categories in the proportion of deaths, we proceed to investigate its significance via hypothesis testing. Figure 2.5 and Table 2.11 show the survival curves and the test statistic to

Figure 2.5: Overall survival time by categories of the studied variables



compare them, respectively, by Localizations, Breslow index, HMM subtype and Age. As it is observed the plots are quite similar among the categories. The p-value in the Table 2.11 for log-rank and Peto-Peto statistics confirm that the equality of survival curves cannot be rejected.

These results suggest redefining the characteristics of Location, Breslow and HMM subtype of melanoma, and Age of the patient, before considering such factors in the survival model.

Table 2.11: Tests for equality of survival curves

Tests		Log-rank			Peto-Peto		
Groups		χ^2	*df*	*p-value*	χ^2	*df*	*p-value*
`Location`	*Extremities, Head and Neck, Trunk*	1.9	2	0.390	1.8	2	0.398
`Breslow`	*<1, [1, 2), [2, 4)*	4.2	2	0.121	4.2	2	0.122
`HMM`	*SSM, LMM, NM*	3.1	2	0.217	2.9	2	0.233
`Age`	*≤45, 46-60*	1.3	1	0.250	1.3	1	0.247
`Age`	*61-70, >70*	1.2	1	0.270	1.2	1	0.270

Cox model for the survival time

Taking into account the results obtained in the previous section, in Table 2.12 there are the new recoded variables. We are now using only two categories for Breslow level (< 4 versus ≥ 4), two categories for HMM subtype (SSM-LMM-NM versus ALM) and two categories for Age (less than 60 years old and higher than 60). In this table we can see the distribution of the total number of deaths in each category, such as will enter the survival model.

The results obtained with the Cox model, are summarized in Table 2.13. Highly statistically significant factors for survival are SLN Status, Ulceration, Clark and Age. This table shows that, patients with positive SLN have three times more risk of dead than patients with negative status; patients with an ulcerated melanomas are three more times at risk of dead than those without ulceration; patients with Clark's level IV-V are five more times at risk of dead than those with Clark I-III; and patients older than 60 years is three times higher the risk of dead that younger.

These results indicate that patients with ulcerated, thick and advanced Clark melanoma, which has extended to the nodes are at high risk of dead, and this risk increases if the patient is older than 60 years.

Table 2.14 summarizes the results for the validation of the assumption of proportionality. In this table we see that in all cases, the test is not significant, and conclude that the proportional hazards condition is satisfied.

Moreover, Schoenfeld residuals shown in Figure 2.6, confirms the proportional hazards assump-

Table 2.12: Genetic factors for death due to melanoma

Factor	Categories	Dead(%)	Total
`SLN Status`	*Negative*	16(5.00)	320
	Positive	15(18.75)	80
`Breslow level`	< 4	17(5.15)	230
	≥ 4	13(20.63)	63
`Ulceration`	*No*	12(4.72)	254
	Yes	16(14.95)	107
`Clark level`	*I-III*	3(2.45)	122
	IV-V	27(10.07)	268
`HMM subtype`	*SSM-LMM-NM*	26(7.58)	343
	ALM	5(19.23)	26
`Age`	≤ 60	12(4.91)	244
	> 60	19(12.33)	154
`Gender`	*F*	17(7.29)	233
	M	14(8.38)	167

Histopathological of Malignant Melanoma(HMM) subtype: Superficial Spreading Melanoma (SSM), Acral Lentiginous Melanoma (ALM), Lentigo Maligna Melanoma (LMM), Nodular Melanoma (NM).

tion in Cox model. However, we observe two residuals with values atypical in Figure 2.6 c).

Table 2.13: Statistical significant factors for survival to death using a Cox model. Reference group: SLN Status= negative, Ulceration= no, Clark level= I-III and Age≤ 60 years.

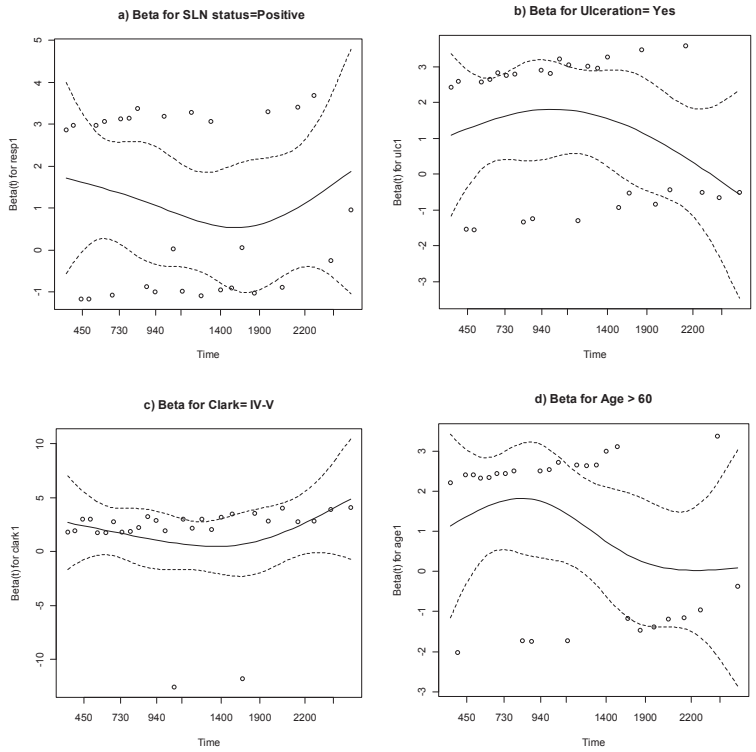
Predictors	β	se(β)	z	p	e^{β}	$L_{.95}$	$U_{.95}$
`SLN Status`							
presence of metastasis	1.111	0.390	2.848	0.004	3.038	1.414	6.528
`Ulceration`							
Yes	1.268	0.388	3.265	0.001	3.556	1.661	7.614
`Clark level`							
IV-V	1.647	0.744	2.212	0.027	5.190	1.206	22.332
`Age`							
> 60	1.053	0.394	2.674	0.007	2.866	1.325	6.203

n=353 (47 missing observations)

Table 2.14: Test for proportional hazards.

Predictors	rho	chisq	p
`SLN Status`			
presence of metastasis	-0.078	0.157	0.692
`Ulceration`			
Yes	-0.211	1.124	0.289
`Clark level`			
IV-V	0.054	0.079	0.779
`Age`			
> 60	-0.287	2.060	0.151

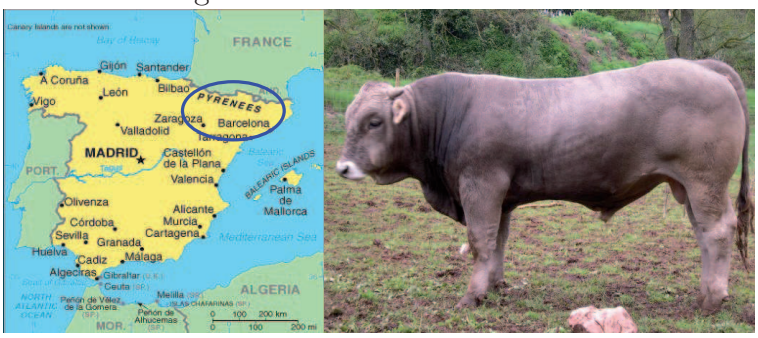
Figure 2.6: Schoenfeld residuals



2.3 Dataset mortality up to weaning of calves

This data corresponds to the cow pedigree Old Brown Swiss, called Bruna dels Pirineus, reported in Tarrés et al. (2005). The reproduction system of the Bruna dels Pirineus race conforms to the seasons, passing the winters in the valleys close to the villages and summers in the praries of the high mountains (port), accompanied normally by their calves. Many individuals of this race pass the winter in the open air. The herds are located in the Pyrenean mountains areas of Catalonia (Spain), see Figure 2.7. Although classified as a special protected race, the Bruna dels Pirineus constitute 80% of the bovine meat in Catalonia. During the time period from birth until the moment of weaning (birth-weaning period), approximately the first 180 days of the calf´s life, a certain percentage of calves die due to natural causes. Even though this percentage is not very high, it reduces cattle farm incomes and adds significantly to cattle productions costs (see Goyache et al. (2003), for a review). For example, the effect of spontaneous abortion on the dairy industry is substantial, costing the industry around $200 million per year in California

Figure 2.7: Bruna dels Pirineus



alone (Hanson et al. (2003)).

2.3.1 Data description

A study was designed with the objective of identifying calf survival traits so as to reduce the mortality of calves from birth to weaning. The data was recorded between 1994 and 2002 in three breeding herds, Tarrés et al. (2005). Table 2.15 presents the distribution of calves during the nine years of research, the first row represents the number of deaths and the second the total number of births. In this table we observe that in the initial and final year of the study no deaths occurred. The data base contains information on 2,504 calves, 68 of which died during the period of the first 180 days after birth. Mortality is displayed in Figure 2.8 where the highest percentage of deaths was 4.07% registered in 1998.

Table 2.15: Distribution of calves by year

Year	1994	1995	1996	1997	1998	1999	2000	2001	2002	Total
dead	0	7	5	13	11	13	13	6	0	68
Total	80	308	334	346	270	355	413	310	88	2504

Table 2.16 shows the distribution of the calves by month, the first row represents the number of deaths and the second row represents the total births. In this table we can see that most of the calving (73%) occurred between January and April.

Figure 2.8: Percentage of dead calves by year

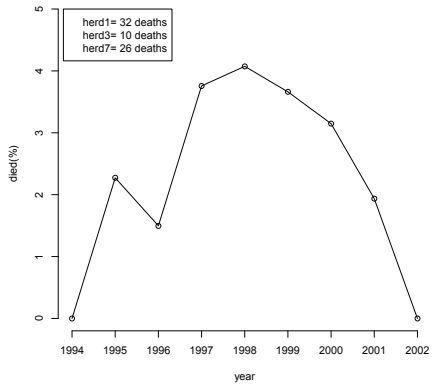


Table 2.16: Distribution of calves by month

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dic
died	6	11	12	22	6	2	2	2	1	0	2	2
Total	451	599	482	295	175	73	55	55	63	46	47	163

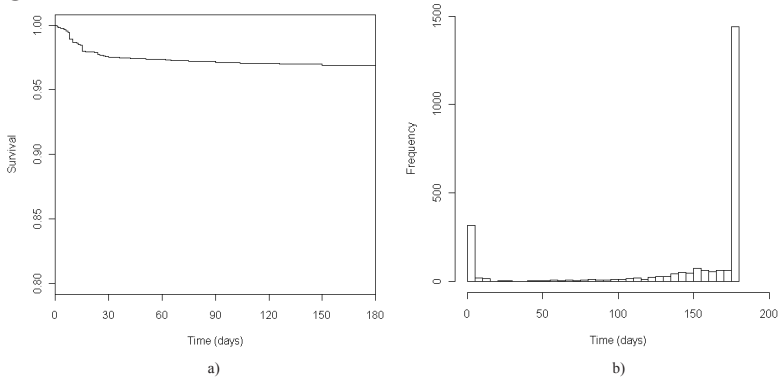
In May, the cows and calves migrate to the grazing pastures of the mountains until September when they return to the valleys below where they stay for the winter. When they return to the valley in September, the calves are weaned. Consequently, a calf born in January is weaned at eight months and a calf born in April is weaned at four months. This is a reason for which a calf born in the month of April has the greatest risk of death, with a probability of 0.074.

For some of the calves born in 2002, the date of weaning was unknown and was replaced with the date of the end of the study, and these were considered as censored. The birth-weaning period for cattle begins at birth and lasts for the first 180 days of life. The survival time was estimated as the difference between the date of death and the date of birth. The results of the study are as follows: 68 complete records (2.7% dead calves) and 2,436 censored records (97.3% censored calves). The total of censored calves is distributed into two groups: 1384 (55.3%) censored because they survived to 180 days and 1052 (42%) censored due to loss during the follow-up study

or because the study finished.

The Kaplan-Meier estimator of the survival time curve of the calves is shown in Figure 2.9 a), and the corresponding histogram in Figure 2.9 b). We observe that the survival curve in a), is asymptotically improper (not converging to zero). The altitude of the curve is given by the Kaplan-Meier estimator as $\widehat{S}_{KM}(t_n) = 0.968$, where t_n is the greatest time of failure observed. Note that this value is similar to the level of censoring 0.973 end of the period, contained in the data. In the histogram of Figure 2.9 b, the highest bar represents the group formed by the 1,384 censored calves which survived until the end of the follow-up.

Figure 2.9: Kaplan-Meier estimator and survival time histogram



The group of calves which entered the study belong to the herds: herd1, herd3 and herd7. Table 2.17 shows the distribution of the calves considering their status at the end of the follow-up and the herd to which they belong. In this table we observe that around 50% of the calves belong to herd7. In contrast, we also observe that 95% of the calves from herd1 completed the follow-up and survived. However in herd3, there are approximately 80% of the calves did not complete the follow-up and are censored before 180 days. In comparison, in herd7 there are approximately 98% censoring, 45% due to loss during follow-up and 53% due to survival beyond 180 days. Additionally we want to emphasize that in the last year of the study all of the calves that entered the study were from herd3, constituting a total of 88 calves, see Table 2.15.

Until now we have given a description about distribution of calves into the herds and a temporal

Table 2.17: Follow-up of calves from 1994 to 2002

Status	herd1 (%)	herd3 (%)	herd7 (%)
dead	32 (4.89)	10 (1.53)	26 (2.16)
lost during follow-up	0 (0.00)	518 (79.69)	534 (44.50)
alive at 180 days	622 (95.10)	122 (18.76)	640 (53.33)
Total	654 (26.11)	650 (25.95)	1200 (47.92)

description about distribution of calves by year and month. In the following section we present genetic and environmental factors, which could influence the survival of the calves.

2.3.2 Survival factors

The data set collected includes variables at the time of calving such as: cow's longitude of productive life at calving; calf birth-weight; gender; month and year of birth; difficulties at calving; as well as the herd to which the cow belongs. All these factors could contribute to the mortality and the survival time of calves, and were categorized based on the information provided by the veterinarians and farmers (see Tarrés et al. (2005)).

Table 2.18 summarizes the descriptive statistics for each factor, such as the productive life of the cow, denoted by `Lpl`, dichotomized into groups < 1300 *days* and > 1300 *days*; month of birth, denoted by `Month`, dichotomized into groups *September to February* and *March to August*); `Gender` (*female*, *male*); Difficulties at calving, denoted by `Difficulty`, categorized into *without assistance*, *slightly assisted by the farmer* and *strongly assisted by the farmer or the veterinary practitioner*; and the weight at birth, denoted by `Weight`, categorized into *small (<42.9kg)*, *median-large (>42.9kg)* and *missing*. In this Table we have excluded a total of 427 records: 168 records, corresponding to births in 1994 and 2002, were excluded because an irregular distribution of calves in the herds (see Table 2.15), and 259 records were excluded due to an insufficient follow-up (censored at $t = 1$ day) (see Figure 2.9 b). We analyze the remainder 2077 calves born between 1995 and 2001 from three different herds, with a total of 68 uncensored observations (3.27% dead calves). In this table and in the future, we exclude the `Year`, due to time-dependent of the same way as does the `Month` of birth.

Table 2.18: Characteristics of Calves

Factors	herd1 (%)	herd3 (%)	herd7 (%)
`Lpl`[3]			
$< 1300d$	360 (58.72)	263 (63.52)	498 (47.42)
$> 1300d$	253 (41.27)	151 (36.47)	552 (52.57)
`Month`			
sep-feb	204 (33.27)	319 (77.05)	650 (61.90)
mar-aug	409 (66.72)	95 (22.94)	400 (38.09)
`Gender`			
female	308 (50.24)	216 (52.17)	544 (51.80)
male	305 (49.75)	198 (47.82)	506 (48.19)
`Difficulty`			
without assistance	299 (48.77)	379 (91.54)	916 (87.23)
slightly assisted	38 (6.19)	25 (6.03)	50 (4.76)
strongly assisted	3 (0.48)	9 (2.17)	83 (7.90)
missing	273 (44.53)	1 (0.24)	1 (0.09)
`Weight`			
small (<42.9kg)	101 (16.47)	123 (29.71)	259 (24.66)
median-large (>42.9kg)	12 (1.95)	284 (68.59)	506 (48.19)
missing	500 (81.56)	7 (1.69)	285 (27.14)
Total	613 (29.51)	414 (19.93)	1050 (50.55)

[3] Length of productive life of the cow

In Table 2.18 we see that in the herd1 and herd3 there is a higher percentage of young cows than in herd7. A highest number of births occurred in the period from September to February in the herd3 and herd7, whereas than in the herd1 occurred into the period from March to August. There is a high percentage of missing information for difficulty at calving in the herd1. Similarly, we observe that a high percentage of weights were not recorded in the herd1 and herd7. These

observations in addition to those obtained from Table 2.17, are clear evidence that these herds are heterogeneous among themselves.

Standard survival analysis

Given the heterogeneity between herds, the analysis of calves' survival time up to weaning is presented by herd. Taking account that all the calves that live up to 180 days were defined them as censored. We have then a censored percentage of 94.78% for herd1, 97.59% for herd3, and 97.52% for herd7, see Table 2.19 similarly as in Table 2.17.

Table 2.19: Follow-up of calves from 1995 to 2001

Status	herd1 (%)	herd3 (%)	herd7 (%)
dead	32 (5.22)	10 (2.41)	26 (2.48)
lost during follow-up	0 (0.00)	334 (80.68)	391 (37.24)
alive at 180 days	581 (94.78)	70 (16.91)	633 (60.28)
Total	613 (29.52)	414 (19.93)	1050 (50.55)

Herd 1: The estimated probability of survivors beyond 180 days is given by $\widehat{S}_{KM}(180) = 0.948$ (is the estimated probability of being censored) and 95% confidence interval of [0.930, 0.966]. In this herd, the weight factor (calf's weight at birth) was not considered in the analysis, because 81.56% of its data is lost, see Table 2.18. The gender has not been considered as a survival factor because there are not significant differences between females and males, the log-rank test (p-value = 0.479) and Peto-Peto (p-value = 0.47) were not significant.

Table 2.20 summarized the characteristics of 273 (44.53%) calves that were excluded from the analysis because the difficulty at calving of these calves was not recorded. This table shows that 61.53% of their births correspond to young cows, 64.47% were born between March-August and their weights were not recorded.

The results obtained with the Cox model with reference group given by *Month= Sep-Feb* and *Difficulty= without assistance*, are summarized in Table 2.21. Highly statistically significant

Table 2.20: Description of missing data for Difficulty in the herd1

Factor	Categories	Total(%)
`Lpl`	$< 1300d$	168(61.53)
	$> 1300d$	105(38.47)
`Month`	*sep-feb*	97(35.53)
	mar-aug	176(64.47)
`Gender`	*female*	150(54.94)
	male	123(45.06)
`Difficulty`	*without assistance*	
	slightly assisted	
	strongly assisted	
	missing	273(100)
`Weight`	*small* ($< 42.9kg$)	
	median-large($> 42.9kg$)	
	missing	273(100)

factors for survival were *Month* and *Difficulty*. Calves that were born between March-August are seven more times at risk of dead within their 180 days of life than those were born between September-February, as shown on the Table. Calves that come from slightly or strongly assisted calving are six and eight more times at risk of dead before 180 days of life than those coming from calving without assistance.

Table 2.22 summarizes the results for the validation of the assumption of proportionality. In this table we see that in all cases, the test was not significant, and conclude that the proportional hazards condition was satisfied.

Moreover, Schoenfeld residuals shown in Figure 2.10, confirms the proportional hazards assumption in the Cox model for herd1.

Herd 3: In this herd the estimated probability of survivors beyond 180 days is given by $\widehat{S}_{KM}(180) = 0.976$ and 95% confidence interval of [0.961, 0.991], similarly to the estimated

Table 2.21: Statistical significant factors for survival up to weaning using a Cox model for herd1. Reference group: month of birth= sep-feb and difficulty at calving= without assistance.

Predictors	β	se(β)	z	p	e^{β}	$L_{.95}$	$U_{.95}$
`Month`							
mar-aug	2.032	1.029	1.974	0.048	7.630	1.015	57.37
`Difficulty`							
slightly assisted	1.853	0.475	3.901	$< .001$	6.382	2.515	16.19
strongly assisted	2.134	1.051	2.030	0.042	8.449	1.076	66.35

n=340 (273 missing observations)

Table 2.22: Test for proportional hazards for herd1.

Predictors	rho	chisq	p
`Month`			
month2= mar-aug	-0.381	2.707	0.0999
`Difficulty`			
difficulty2= slightly assisted	0.170	0.547	0.4594
difficulty3= strongly assisted	-0.136	0.346	0.5563

probability of being censored (0.975).

When comparing the survival curves between levels for each factor, we did not find statistically significant differences. Table 2.23 summarizes the obtained results with the log-rank test and Peto-Peto. These results suggest that in this herd there is not difference between the categories of each factor, and thereby there are not statistically significant factors for survival.

Figure 2.10: Schoenfeld residuals for herd 1

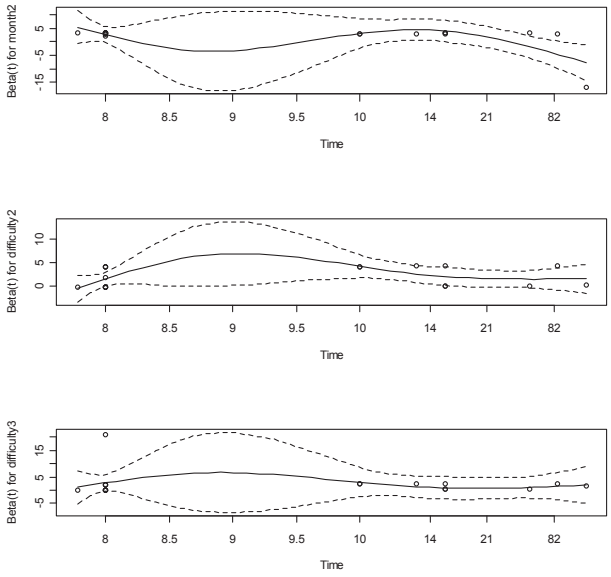


Table 2.23: Tests for equality of survival curves in the herd3

Tests		Log-rank			Peto-Peto		
Groups		χ^2	*df*	*p-value*	χ^2	*df*	*p-value*
`Lpl`	: *<1300 d, >1300 d*	0.8	1	0.372	0.8	1	0.376
`Month`	: *sep-feb, mar-aug*	0.3	1	0.565	0.3	1	0.568
`Gender`	: *female ,male*	0.0	1	0.897	0.0	1	0.900
`Difficulty`	: *without assistance, slightly assisted, strongly assisted*	3.6	2	0.167	3.6	2	0.162
`Weight`	: *small (< 42.9kg), median-large (> 42.9kg)*	1.9	1	0.171	1.9	1	0.172

Herd 7: The estimated probability of survivors beyond 180 days is given by $\widehat{S}_{KM}(180) = 0.975$ and 95% confidence interval of [0.966, 0.985], similarly to the estimated probability of being censored.

Table 2.24 summarized the characteristics of 285 (27.14%) calves whose weight at calving were not recorded. This table shows that 53.33% of their births coming from young cows, 64.56% were born between September-February and 87.72% coming from calving without assistance.

When comparing the survival curves between levels for each factor, we did not find statistically significant differences for Lpl, Month, Gender, and Weight. We also found no significant differ-

Table 2.24: Description of missing data for Weight in the herd7

Factor	Categories	Total(%)
`Lpl`	$< 1300d$	152(53.33)
	$> 1300d$	133(46.67)
`Month`	*sep-feb*	184(64.56)
	mar-aug	101(35.44)
`Gender`	*female*	158(55.44)
	male	127(44.56)
`Difficulty`	*without assistance*	250(87.72)
	slightly assisted	4(1.40)
	strongly assisted	30(10.53)
	missing	1(0.35)
`Weight`	*small* ($< 42.9kg$)	
	median-large($> 42.9kg$)	
	missing	285(100)

ences between without assistance calving and slightly assisted calving in the difficulty at calving. Table 2.25 summarizes the obtained results with the log-rank test and Peto-Peto. These results suggest that difficulty at calving is an important factor for the survival of calves, and may be recoded in without assistance- slightly assisted versus strongly assisted.

Table 2.25: Tests for equality of survival curves in the herd7

Tests		Log-rank			Peto-Peto		
Groups		χ^2	*df*	*p-value*	χ^2	*df*	*p-value*
`Lpl`	: *<1300 d, >1300 d*	3.4	1	0.064	3.4	1	0.066
`Month`	: *sep-feb, mar-aug*	1.6	1	0.202	1.6	1	0.201
`Gender`	: *female, male*	0.0	1	0.841	0.0	1	0.851
`Difficulty`	: *without assistance, slightly assisted*	0.0	1	0.991	0.0	1	0.995
`Weight`	: *small (< 42.9kg), median-large (> 42.9kg)*	1.1	1	0.293	1.1	1	0.294

The results obtained with the Cox model with reference group given by *Difficulty= without*

assistance or slightly assisted, are summarized in Table 2.26. Highly statistically significant for survival is *Difficulty*. Calves that come from strongly assisted calving are three more times at risk of dead before 180 days of life than those coming from without assistance or slightly assisted calving.

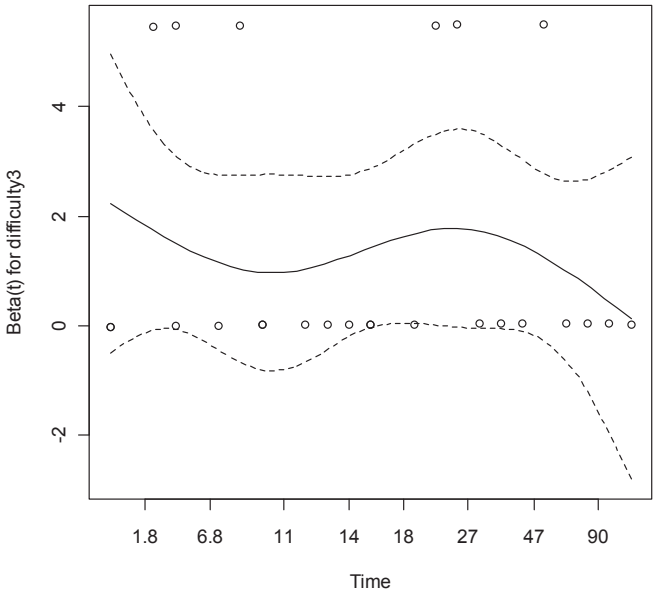
Table 2.26: Statistical significant factors for survival up to weaning using a Cox model for herd7. Reference group: difficulty at calving= without assistance or slightly assisted.

Predictors	β	se(β)	z	p	e^{β}	$L_{.95}$	$U_{.95}$
`Difficulty`							
strongly assisted	1.332	0.468	2.844	0.004	3.788	1.513	9.486

n=1049 (1 missing observations)

The test for the validation of the assumption of proportionality is not significant (*difficulty3= strongly assisted*), whit rho= -0.109, chisq= 0.295 and p= 0.587. Conclude that the proportional hazards condition is satisfied. Moreover, Schoenfeld residuals shown in Figure 2.11, confirms the proportional hazards assumption in the Cox model.

Figure 2.11: Schoenfeld residuals for herd 7



2.4 Limitations of the standard survival techniques

The databases presented here, concerning the evaluation of survival time until the occurrence of an event of interest, including those features that could be considered as survival factors, have one characteristic in common: they are subject to a large (heavy) censoring level. This characteristic of heavy censoring leads to two questions which have to be investigated. First, if the assumption of a proportional hazards model is satisfied in these data, then which are the effects of the heavy censoring in the inference about regression coefficients? and what is the appropriate sample size?. Second, the Kaplan and Meier curve for these data suggests a survival model that does not converge to zero as time goes on indefinitely, then we seek a non-standard survival model that takes into account the improperness of the curve. Which are these models and what advantages have over standard models? The answers to these questions are investigated in the following chapters of this thesis, and nonstandard analyses are presented using these databases. Concerning limitations of standard models, a PH model could be adequate in a setting where $S(\infty|x) = \pi > 0$, since $S(t|x)$ approaching a positive limit is not excluded. The proportion of immune $S(\infty|x)$ in a censored sample, is often estimated by Kaplan-Meier estimator, $\hat{\pi} = \hat{S}_{KM}(t_{(n)}|x)$, where $t_{(n)}$ is the last observed survival time. Properties of consistency under heavy censoring scenario are presented in Zukang (1997) (see also Maller and Zhou (1992)). Although in a PH model we have $S(\infty|x) = [S_0(\infty)]^{\psi(x;\beta)}$, it lacks flexibility since β determines both the hazard ratio for persons who are *susceptible*, $\frac{h(t|x_2)}{h(t|x_1)} = \frac{\psi(x_2;\beta)}{\psi(x_1;\beta)}$ and the *nonsusceptible* fraction ratios, $\frac{\pi_2}{\pi_1} = [S_0(\infty)]^{\psi(x_2;\beta)-\psi(x_1;\beta)}$, where $\psi(x;\beta) = \exp(-\beta' x)$ and $S_0(t)$ is the baseline survival function.

Chapter 3

Review on Cure Models

3.1 Introduction

In many clinical studies (especially in cancer research), there might be a certain percentage of patients who respond favourably to treatment. After a sufficient period of follow-up they appear to be risk free or even *cured* of the disease. Only a proportion of the population is susceptible to the target illness within the survival-time of the data while others remain immune. Empirical evidence to confirm this population trend appears in a graph as a long, stable plateau which, contains heavy censoring at the end of the Kaplan-Meier survival curve. *Cure models* or *Cure fraction models* also referred to as *Cure rate models* were specifically introduced for the purpose of modeling time-to-event data and incorporating a cure fraction. These types of models are becoming increasingly useful in clinical trials, especially in oncology studies. More recently, they have been applied to a wide range of fields of research such as psychology, criminology, economics, education, etc.

Within survival analysis with cure, there are two major approaches which take the cure fraction $S(\infty) > 0$ into account, when dealing with the modeling of survival times in a population with survival function S. The first type of cure model introduced in statistical literature is called the *mixture cure model* (Boag (1949)) where it is assumed that a proportion $S(\infty) = \pi$ of patients are cured and are no longer at risk of experiencing the disease. However, remains a proportion (1-π), which is still uncured. Consequently, these people will eventually experience the illness

and thus the survival function $S_1(t)$, will tend to zero for these subjects. With this approach in mind, a *mixture cure model* is defined as

$$S(t) = \pi + (1 - \pi)S_1(t). \tag{3.1}$$

The second type of cure model is the *non-mixture cure model* (Tsodikov (1998)), which is defined by a bounded cumulative hazard function, $H(\infty) = \theta$ as follows

$$S(t) = \exp\{-\theta F(t)\}, \tag{3.2}$$

where $H(t) = \theta F(t)$ and $F(t)$ is any distribution function and $H(t)$ represents a standardized cumulative hazard function. Within this representation the cure fraction is given by $S(\infty) = \exp(-\theta)$.

This chapter reviews cure survival models, from their origin up until the writing of this thesis. The presentation is in chronological order and distinguishes between the various approaches related to the topic. At the end of this chapter there is a description of the software available to carry out an analysis with a cure model.

3.2 Mixture cure models

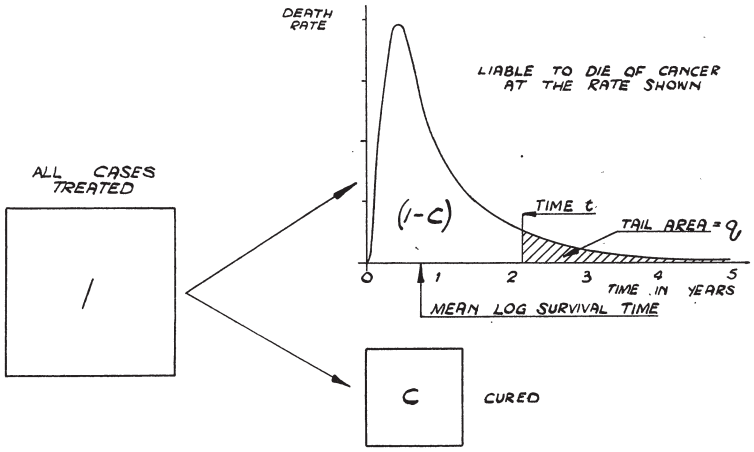
In 1949, John W. Boag published an article in the *Journal of the Royal Statistical Society SERIES B* under the title *Maximum likelihood estimates of the proportion of patients cured by cancer therapy*. In this article he discusses the modeling of the survival times of a group of cancer patients. After undergoing the treatment the results vary; some patients respond well and were cured, however, others continue to have the illness despite undergoing the same treatment. Because of this (and based on the schema of Figure 3.1 taken from his article), Boag introduced a mixture cure model for the modeling of both the survival function and cure fraction in a population as follows

$$S(t) = S_0(t)\{c + (1 - c)S_1(t)\}. \tag{3.3}$$

In this model $S_0(t)$ indicates the probability of a patient surviving to time t when all causes of death except the original cancer are considered. Meanwhile $S_1(t)$ indicates the probability of a patient, who has not been permanently cured, surviving to time t when only the cause of cancer

is considered. This formulation assumes that the specific causes of cancer $S_1(t)$ are independent of other causes $S_0(t)$. In Figure 3.1, c denotes the proportion of cured patients and $S_1(t)$ denotes a lognormal survival function. By applying the process of maximum likelihood to estimate $S_1(t)$

Figure 3.1: Statistical model of a clinical experiment, Boag (1949).



and c, it becomes possible to ignore the factors that contain $S_0(t)$.

The most popular cure model is the *mixture cure model* introduced by Berkson and Gage (1952). Based on empirical evidence from cancer studies, a simple function, in terms of two physically meaningful parameters, has evolved, which fits such survival data very well. These two parameters can be used to compare simultaneously the mortality of two groups, which differ in type of treatment, type of cancer, or other characteristics. Berkson and Gage assume that patients with a specific cancer are, all before treatment, subject to the effect of two mortality forces, C_s representing the cancer in question (cause-specific) and C_o representing all other diseases (other-causes), and that these act independently and simultaneously. After treatment, a fraction c of the population is *cured* and they are only subject to the mortality forces C_o, while the remainder (1-c) are subject to two forces, C_o and C_s, the value of C_s being not necessarily equal to zero, and presumably less than one, before treatment.

When we think about the two hypothesized cohorts of the population separately, then the probability of survival to time t of the cured (*nonsusceptible*) cohort is given by $P[T > t|$ only C_o acts],

and for the uncured (*susceptible*) cohort it is given by $P[T > t|$ both forces of mortality act$]$. The probability of survival to time t for the total population $S(t) = P[T > t]$ is given as

$$S(t) = cP[T > t|\text{ only } C_o \text{ acts}] + (1-c)P[T > t|\text{ both forces of mortality act}]. \tag{3.4}$$

If $S_0(t) = P[T > t|$ only C_o acts$]$ and $S_1(t) = P[T > t|$ only C_s acts$]$, under the additional supposition of independence of C_o and C_s, the survival to time t in the total population can be rewritten as (3.3). $S_0(t)$ represents the survival function for a population subject to other causes, which can be obtained from standard life tables, while $S_1(t)$ represents the survival function for a population subject only to the specified cancer. Berkson and Gage (1952) argued that $S_0(t)$ can be a constant value, estimated from general life tables which are considered applicable to the population at hand. Hence, the survival model due to the specific causes of cancer in a population free of other factors which can cause death is given by

$$S(t) = c + (1-c)S_1(t). \tag{3.5}$$

Berkson and Gage (1952) assumed (3.5) a constant excess mortality rate $(1-c)$ for the susceptible group and an exponential distribution for the time of incidence

$$S_1(t) = \exp(-\lambda t), \tag{3.6}$$

and values of c and λ maybe estimated by a least-squares procedure. Farewell (1977) introduced a version of the previous model in which he assumed that a binary variable Y could specify the incidence of a particular disease with $Y = 1$ or a lifetime free of the disease with $Y = 0$, and denoting by $\pi = P[Y = 1]$. The new formulation results in,

$$S(t) = P[T > t, Y] = [1 - \pi] + \pi S_1(t), \tag{3.7}$$

where $P[T > t|Y = 0] = 1$ and $S(\infty|x) = 1 - \pi(x)$ corresponds to the rate of the subjects free of the disease. He studies the efficiency of the model by modelling the probability π as a function of a vector of covariates by means of a logistic model given by (3.8)

$$\pi(x) = P[Y = 1|x] = \frac{\exp(bx)}{1 + \exp(bx)}, \tag{3.8}$$

and $S_1(t)$ is as well Exponential as in (3.6). Later, Farewell (1982) reanalyzed a toxicological experiment analyzed by Pierce et al. (1979), using model (3.8) and a Weibull regression model for the survival time of susceptible individuals

$$S_1(t|x) = \exp(-(\lambda t)^{\delta}), \tag{3.9}$$

where δ is a shape parameter and λ is related to x by $\lambda = \exp(-\gamma_0 - \gamma' x)$. The vector γ represents unknown regression coefficients. Notice that, in terms of hazard functions, (3.9) can be rewritten as $h_1(t|x) = \exp(\beta x)[\alpha\delta(\alpha t)^{\delta-1}]$, where $\beta = -\delta\gamma'$ and $\alpha = \exp(-\gamma_0)$. And more generally,

$$h_1(t|x) = \exp(\beta x) h_0(t), \tag{3.10}$$

where the conditional baseline hazard $h_0(t) = h(t|Y = 1, x = 0)$ is Weibull with parameters δ and α. Yamaguchi (1992) assume (3.8) in combination with a general class of accelerated failure time model for (3.9), namely, the extended family of generalized Gamma models; which includes exponential, Weibull, reciprocal Weibull, log-normal, and Gamma as its special cases. And shows an application to the analysis of permanent employment in Japan. Kuk and Chen (1992) proposed a semiparametric generalization of Farewell's model using a Cox proportional hazards model in the susceptible group. Their model is also based on (3.8) and (3.10) but $h_0(t)$ can be any arbitrary unspecified hazard function not necessarily in the Weibull family. In terms of survivor functions, the assumption (3.10) of proportional hazard implies $S_1(t|x) = S_0(t)^{\exp(\beta x)}$, where $S_0(t) = S(t|x = 0)$. For the purpose of estimating the regression coefficients b in (3.8) and β in (3.10), they proposed a marginal likelihood approach where $S_0(t)$ is a nuisance baseline function. To simplify the calculations, they suggest a Monte Carlo approximation of the marginal likelihood which can be maximized using existing software. Moreover, Ghitany et al. (1994) provide sufficient conditions for the existence, consistency, and asymptotic normality of maximum likelihood estimators for the parameters in the model (3.7) using (3.8) and (3.9). In Maller and Zhou (1996) there is an extense discussion of the model proposed by Farewell (1982).

Gieser et al. (1998) apply a mixture cure model like (3.5), using (3.8) for c and a Gompertz model for $S_1(t|x)$ to analyze data of acute lymphocytic leukaemia in children. Peng et al. (1998) considered the model (3.5), using (3.8) and a generalized F for the susceptible group distribution, i.e.,

$$S_1(t|x) = \int_0^k \frac{u^{s_2-1}(1-u)^{s_1-1}}{B(s_2, s_1)} du,$$

where $k = s_2[s_2 + s_1 \exp(w)]^{-1}$, $w = [\log(t) - \mu]/\sigma$ and $\mu = \beta x$. The generalized F distribution includes many commonly used distributions as special cases, such as the log-normal, Weibull, gamma, and log-logistic distributions. Calculation problems with the model, model and covariate selection methods are discussed. They compared maximum likelihood estimates with those

obtained from mixture models under other distributions. Other different methods of estimation were given, among others, by Anscombe (1961), Meeker (1987), De Angelis et al. (1999), Peng and Dear (2000), Sy and Taylor (2000), Betensky and Schoenfeld (2001), Aljawadi et al. (2011), Dikta (2014).

3.2.1 Split population models

The cure models defined out from a binary variable Y given by Farewell (1977), have a wide applications in various areas, such as criminology, reliability, marketing, education (see Maller and Zhou (1996)). In some of these areas of application, the *mixture cure models* have been known as a *split population models*, in the sense that the population is separated by the categorical variable Y. We briefly describe below, some of the references which have applied these methodologies.

Schmidt and Witte (1989) analyzed data on a cohort of releasees from the North Carolina prison system, where the survival time, T, is defined as time to recidivism and $Y = 1$ when the individual is recidivist. They analyzed this data using a mixture cure model of the form (3.7), with $P(Y = 1)$ like (3.8) and several distributions for $S_1(t)$, such as Exponential, Weibull, lognormal and loglogistic, all with scale parameters dependent of covariates. Cole and Gunther (1995) studied the factors influencing bank failures using a model of the form (3.7), in their formulation, $Y = 1$ if the bank fails and T is the time until failure. In the present context, they used a log-logistic distribution for $S_1(t)$ and a logistic model for $P[Y = 1]$ like (3.8).

And more recently, Mavromaras and Orme (2004) considered a split population model for the duration of temporary layoffs in the German labour market (see also Yamaguchi (1992) for the analysis of permanent employment in Japan); the population being split according to whether a layoff is temporary or permanent. In this case the binary random variable (unobserved) Y, partitions (splits) the population into temporary ($Y = 1$) or permanent ($Y = 0$) layoffs, and T denotes the duration of a layoff conditional on $Y = 1$, that is, T denotes the duration of a temporary layoff. They defined a survival model $S_1(t)$ by means of a flexible piecewise constant hazard and the probability $\pi(x) = P[Y = 1]$ was modeled using a probit link $\pi(x) = \Phi(x'\beta)$, where $\Phi(.)$ denotes the standard normal distribution function.

3.3 Non-Mixture Cure Models

Although model (3.1) is intuitively attractive and widely used, it does not have a proportional hazards structure for $S(t)$, which is desirable in carrying out survival analysis with covariates. An alternative cure model, with a proportional hazards structure for the population, is the so called *non-mixture cure model* or *non-mixture cure fraction model* given by (3.2) and was introduced by Tsodikov (1998). In the formulation of the survival model, Tsodikov takes into account that the cumulative hazard function can be written as $H(t) = -\log(S(t))$ and according to (3.1), $H(t) \leq \theta$ and $H(\infty) = \theta$. A convenient way to adjust for the above property is to consider $H(t) = \theta F(t)$, where $F(t)$ is the distribution function of a nonnegative random variable. In conclusion, the survival function in the general population can be rewritten as (3.2). In addition, if the observed covariates are related to θ by means

$$\theta(x) = \exp(x\beta), \tag{3.11}$$

then we get a PH model with a cure faction $S(\infty) = \exp(-\theta)$. Tsodikov (1998) proposes an estimation method for θ via the profile likelihood, when F is completely unknown.

The motivation behind non-mixture cure models (3.2), is due to the function F, since F can be defined as the distribution function of a unobserved or latent variable, computationally very attractive from a *Bayesian approach*. For this approach, a basic reference is given by Ibrahim et al. (2001), as we mention in chapter 5.

Chen and Ibrahim (2001) proposed maximum likelihood estimation methods via EM algorithm for parameters in the models proposed by Tsodikov (1998), when $F(t)$ is defined as a piecewise exponential and $\theta(x)$ as in (3.11). Broët et al. (2001) introduce a new proposal for the two-sample comparison of survival times with long-term survivors, the approach is made by means a semiparametric generalization of the improper Gompertz model. This paper was the starting point for a new line of research on extensions of the *non-mixture cure model*, such as Tsodikov (2002), Sposto (2002), Tsodikov (2003), Kim et al. (2009), among others. Extensions which we discuss in more detail and show some of its applications in Chapter 6 of this thesis.

3.4 Cure rate models unified

The mixture cure model $S(t|z,x) = \pi(z) + [1 - \pi(z)]S_1(t|x)$ and non-mixture cure models $S(t|z,x) = \exp(-\theta(z)F(t|x))$ are the most widely used cure rate models which may be seen as competitors. Recent applications of these models to oncology data are presented by Lambert et al. (2010), Kim et al. (2011), Othus et al. (2012), and Aljawadi et al. (2013), among others. Each model offers its own advantages as well as its disadvantages (Ibrahim et al. (2001)). When covariate z is absent and $F(t|x)$ is unspecified, defined $\pi = \exp(-\theta)$ and $S_1(t|x) = 1 - F(t|x)$, the mixture and non-mixture cure models are equivalent. They are simply different forms of the same model. When z is present and x is absent, the non-mixture model is a proportional hazards model ($h(t|z) = \theta(z)f(t), f(t) = F'(t)$) and the mixture cure models is not, and they are clearly models for different data structures. When both x and z are present, both models are flexible and they can be considered for modeling survival data with a cure fraction.

The first attempt to unify these models in a more general class of models was given by Yin and Ibrahim (2005a) and Yin and Ibrahim (2005b), via the Box-Cox transformation. Recently Peng and Xu (2012) propose a unified cure model, the proposal is based on a review the unified cure model and a novel biological interpretation for the non-mixture cure model given by Hanin et al. (2001).

3.5 Available Software

Before and during the development of this thesis, extensive literature on mixture cure models was reviewed, the main goal was to identify the statistical software available to carry out an analysis with mixture cure models. Among the most important in the literature are those listed below.

`Windows`: CANSURV
A windows program for population-based cancer survival analysis. The program is available at `http://srab.cancer.gov/cansurv`, but it only reads data-bases created by the Surveillance, Epidemiology and End Results (SEER) program of the National Cancer Institute. Therefore, the use of this software is limited. The CANSURV is based on the models proposed by Yu et al. (2005).

SAS: PSPMCM

A macro for parametric and semiparametric mixture cure models (PSPMCM). This macro is available at http://www.isped.u-bordeaux2.fr/recherche/biostats/FR-biostats-accueil.htm#programmes. It does not have any restriction and it can be used freely. This macro is given by Corbière and Joly (2007), and is based on the methodology proposed by Peng and Dear (2000) and Sy and Taylor (2000).

STATA: CUREREGR

It is a macro to fit a Parametric Cure Model (PCM) in either the non-mixture or mixture class. The program works with one or multiple records by observation or subject and time-varying-covariates also can be estimated. This macro is available at http://econpapers.repec.org/scripts/search/search.asp?kw=parame tric+near+cure+near+regression. Despite that this macro is based on the models proposed by Sposto (2002), was not until 2004 when it was created for a Stata version 8.2. However, the methodology was later extended by Lambert et al. (2007), who discusses how the scale and shape parameters in the Weibull distribution can be modeled as a function of covariates, for both the mixture and non-mixture models.

STATA: SPSURV

In the standard survival model, the risk of failure is non-zero for all cases. A split-population (or cure) survival (SPSURV) model relaxes this assumption and allows an (estimable) fraction of cases never to experience the event. This macro is available at http://econpapers.repec.org/scripts/search/search.asp?kw=split-population+near+model

STATA: strsmix and strsnmix

Other important softwares have been developed by Lambert (2007). He describes the strsmix and strsnmix commands written in Stata, which fit the two main types of cure fraction model, namely, the mixture and nonmixture cure fraction models. These models allow incorporation of the expected background mortality rate and thus enable the modeling of relative survival when cure is a possibility.

R: NLTM

A library for the free software R, to fit a non-linear transformation model (NLTM) for analyzing survival data. The class of NLTM are extensions of the non-mixture cure models and includes a proportional hazards model with cure as a particular case. This package is available at `http://cran.r-project.org/web/packages/nltm/index.html`. This package was created in 2009, and the models included in this library were discussed in Tsodikov (2002). This class of models are related to (3.2), and when the distribution function $F(t)$ in (3.2) depends on covariates, the resulting model is called a non-linear transformation model. Tsodikov (2003) studies the properties and self-consistence of semiparametric models using Nonparametric maximum likelihood estimation.

Chapter 4

Heavy right-censoring

We investigate using simulation, the effect that different levels of right-censoring have on the estimation of the relative risk in a proportional hazards model. We suppose a fixed censoring model for the censoring distribution and a Cox model for the survival times. The simulations were done assuming a binary covariate as an explanatory variable and a fixed percentage of censoring, which we called the censoring level. In order to evaluate the effect of the censoring level on the relative risk we have studied the properties of its estimator such as: bias, variance, mean square error, relative bias and coverage.

This chapter is organized as follows: Section 4.1 describes a dataset where there is a time to event variable and heavy right censoring. In section 4.2 we establish the notation and provide the definitions that we will use throughout this chapter. In section 4.3 we formulate the model and obtain relations between the probability of censoring, the probability of success for a binary variable and the relative risk, using a proportional hazards model for T and a fixed censoring model for C. In section 4.4 we display the simulation results: bias, variance, relative bias and coverage for the relative risk in the Cox model, under fixed censoring. This Chapter 4 concludes with a discussion in section 4.5.

4.1 Motivation

In survival studies the individuals are often right-censored due to either the *end of study* or *loss to follow-up*. End of study censoring is, in general, administered by the researcher and it is a particular case of fixed censoring. The level of administrative censoring depends on the window of observation and the population in study. Loss to follow-up is mainly due to some random mechanism that can depend or not on the survival time. In this chapter we will consider some scenarios where administrative censoring due to end of the study is the unique cause of censoring.

Heavy right-censored data often arises in survival analysis due to an *insufficient follow-up*. When the follow-up period is not long enough, the event for a large percentage of individuals is not observed, these individuals without the event are right-censored. A second instance producing heavy right-censored data is the presence of *long-term survivors*. Those individuals for whom the event was not observed during the follow-up period but that in the future will fail, are right-censored at end of study. Finally, one third possibility is to have *immunes or cured* individuals. In this scenario, the immune or cured subpopulation is censored at the end of study. Maller and Zhou (1996) discuss this special case of censoring, usually named as heavy censoring.

The motivation of this chapter is based on Tarrés et al. (2005) and it is described in the Chapter 2 of this thesis. The first analysis of these data, carry out survival analysis using a proportional hazards model for calf mortality data. The survival time was estimated as the difference between the date of death and the date of birth in the first 180 days. The data collected included the survival time of 2504 calves, with 68 complete records (2.7% dead calves) and 2436 censored records (97.3% censored calves). When we fit a Cox's model, and censoring is heavy, many questions arise on the quality of the adjustment and on the estimators: (1) which properties of the estimators remain valid? (2) when is censoring too heavy for modelling a survival time via a proportional hazards model? (3) in which scenarios we would not recommend to analyze data using a Cox's model?

Although the Cox's model has been intensively investigated using simulation studies to get information about bias and efficiency of the estimated regression coefficients for a variety of situations (Tsiatis and Davidian (1998), Bender et al. (2005) among others), much less has been

done under heavily censored data and recommendations on whether or not the Cox model is appropiate are lacking and it can be very useful.

4.2 Fixed right-censoring mechanism

Fixed right-censoring occurs in studies which have a finite duration time, say τ. An observation on a subject is (fixed) right-censored if the subject is still alive at the end of the study and the measurement of interest has not yet been made for the subject. The recorded variable for the subject is the time at the end of the study. Among others we find this type of censoring in toxicology experiments and are discussed by Groggel et al. (1989).

4.2.1 Formulation

Let T be the survival time and x a covariate vector. Assume that $T|x$ follows a Weibull regression model with shape parameter α and scale parameter λ_x. Let $S(t|x)$ represent the survival function of T with covariate x. We have

$$S(t|x) = \exp(-(\lambda_x t)^\alpha).$$

We will link λ_x with x by assuming that $\lambda_x = \exp(-\beta' x)$. This model has hazard function $h(t|x) = \lambda_x^\alpha(\alpha t^{\alpha-1}) = (\exp(-\beta' x))^\alpha(\alpha t^{\alpha-1})$ and the ratio $\frac{h(t|x_1)}{h(t|x_2)} = (\exp(-\beta'(x_1 - x_2)))^\alpha$ is constant in t.

Under a fixed censoring model, the distribution of the censoring variable C concentrates all their probability in some point τ (Breslow (1970)). This is equivalent to define a time window $[0, \tau]$ within which we would observe the event of interest E, i.e., a fixed censoring at time τ. The probability, p, that T is right-censored at τ is given by $p = P(T > \tau)$. Consequently, in this formulation the censoring level on T is given by parameter p, and the scenarios we investigate here are under a *heavy right-censoring* level, defined for $p \geq 0.70$.

4.2.2 Comparing two groups

In survival analysis is common to compare the survival curves of two groups (by gender, treatment), with the aim of investigating which group is more susceptible to the event of interest.

With this idea in mind, we evaluate the effect of the heavy censoring on the estimation of the relative risk between two groups: group 1 ($x = 1$) and group 0 ($x = 0$), where $\kappa = P(x = 1)$ (under a fixed window of observation $[0, \tau]$).

Under these assumptions, the probability of censoring p is given by the marginal law of T, when a window of observation $[0, \tau]$ is fixed, that is,

$$p = [1 - \kappa]S(\tau|x = 0) + \kappa S(\tau|x = 1),$$

and if a Weibull law with shape parameter α and scale parameter λ_x is considered we have

$$p = [1 - \kappa] \exp(-(\lambda_0 \tau)^\alpha) + \kappa \exp(-(\lambda_1 \tau)^\alpha).$$

The choice of a Weibull model with equal shape parameter for both groups yields a proportional hazard relation as follows

$$RR = \frac{h(t|x = 1)}{h(t|x = 0)} = \frac{\lambda_1^\alpha}{\lambda_0^\alpha},$$

and if we take, without loss of generality, $\lambda_0^\alpha = 1$, we have $RR = \lambda_1^\alpha$ and we can write

$$p = [1 - \kappa] \exp(-\tau^\alpha) + \kappa \exp(-RR\tau^\alpha). \tag{4.1}$$

In order to evaluate the effect of the censoring level p on the RR, we will compute the values of τ for fixed values of RR given p, κ and α by means of the relation

$$RR = -\frac{1}{\tau^\alpha} \log\{\frac{p - [1 - \kappa] \exp(-\tau^\alpha)}{\kappa}\}. \tag{4.2}$$

4.3 Simulation design

In the simulation study, T is the survival time of interest within a fixed window of observation $[0, \tau]$ and subject to a censoring level p. We will evaluate the relative risk between the group 1 ($x = 1$) versus group 0 ($x = 0$), given $\kappa = P(x = 1)$, $T|(x = 1)$ as a Weibull distribution with

scale parameter λ_1, $T|(x=0)$ as a Weibull distribution with scale parameter λ_0, and both with the same shape parameter α. The main goal of this study is to estimate the relative risk (4.2) subject to a right-censoring, and evaluate the effect that a heavy right-censoring mechanism has on this estimation.

In designing the simulation, we first need to determine the parameter τ in terms of the censoring level p using (4.1), where the relative risk RR, the probability κ, and the shape parameter of Weibull distribution α are fixed. Secondly, we have to determine the sample size in each group, say n_1 and n_0, given the total sample size n. Observe that given the values α and RR, the scale parameter λ_1 is given by $\lambda_1 = (RR)^{1/\alpha}$. Then the number of observations in the sample from group 1 and group 0 are given as, $n_1 = n\kappa$ and $n_0 = n(1-\kappa)$, respectively.

After calculating ($\tau(p)$, λ_1, $n1$, $n0$) given (α, κ, RR, p, n) as mentioned above, the simulation study is summarized in the following steps:

- Given α, κ, RR, p and n
- For $x = \mathbf{1}_{n_1}$, generate a vector t_{n_1} with n_1 survival times following a Weibull(α, λ_1) distribution function.
- For $x = \mathbf{0}_{n_0}$, generate a vector t_{n_0} with n_0 survival times following a Weibull$(\alpha, 1)$ distribution function.
- Take $t = (t_{n_1}, t_{n_0})$ and $x = (\mathbf{1}_{n_1}, \mathbf{0}_{n_0})$ and generate a sample of survival data $\{u_i, \delta_i, x_i\}_1^n$ with a censoring level p, where $u_i = min\{t_i, \tau(p)\}$, $\delta_i = 1$ if $u_i = t_i$ or $\delta_i = 0$ if $u_i = \tau(p)$, $i = 1:n$.
- Fit a Cox's model to the sample $\{u_i, \delta_i, x_i\}_1^n$ using the function *coxph* from the R software 2.9.2, and compute the estimator $rr = \widehat{RR}$ together with a confidence interval at 95%.

4.4 Evaluation criteria

In the simulation study two important aspects have to be discussed: First, since $\kappa = P(x=1)$, its value controls the proportion between the groups in the sample; Second, since $\lambda_1 = (RR)^{1/\alpha}$,

the values of RR control the level of the risk between the groups, as it is described in Table 4.1

Table 4.1: Evaluation of values for κ and RR

κ	n_0 compared with n_1	RR	group 0 compared with group 1
0.2	unbalanced (>)	0.2	5.00 times more risk
0.5	balanced (=)	0.4	2.50 times more risk
0.8	unbalanced (<)	0.6	1.65 times more risk
		0.8	1.25 times more risk
		1	equal risks

In the following Example 4.1 the main objective is showing the behavior of censored samples and the effect that the level of censoring has in the estimation of the relative risk. Every sample has been simulated under a model of proportional hazards and different censoring levels.

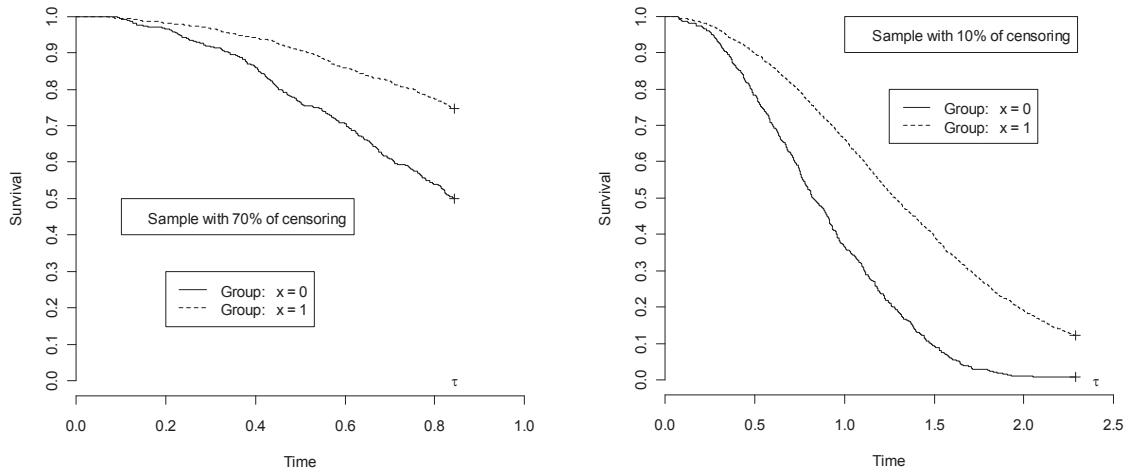
Example 4.1: Two samples of survival times were generated according to the steps summarized in section 4.3, with $n_0 < n_1$, RR=0.4 and each with a sample size of 2500. The sample A with censoring level of 70% and the sample B with 10% of censoring. In sample A we observed 755 events, equivalent to 70% of censoring, distributed in 10% for group 0 and 60% for group 1. In sample B we observed 2249 events, equivalent to 10% of censoring, distributed in 0% for group 0 and 10% for group 1. These results are summarized in Table 4.2. The risk in group 0 is always greater than in group 1 and it satisfies the property of proportionality, as it can be seen in Figure 4.1. The relative risk was estimated, and 95% confidence interval was constructed in each sample after fitting a Cox's model. For the sample A with a censoring level of 70%, the relative risk was 0.42 with an interval of [0.36, 0.48]. For the sample B with a censoring level of 10%, the relative risk was 0.40 and an interval of [0.36, 0.45].

In the previous example we observe an overestimation of the value of the parameter RR (=0.40) when using Sample A containing 70% of censoring, and its confidence interval is slanted to the right. This suggests that if increasing the level of censoring, the bias should increase; such situation is very important for studying the behavior of the bias in terms of the size of sample.

Table 4.2: Simulated data, $n_0 < n_1$ and RR=0.4

	Sample A 70% of censoring		Sample B 10% of censoring	
Groups	records	events	records	events
group 0	500	250	500	496
group 1	2000	505	2000	1753
Total	2500	755	2500	2249

Figure 4.1: Kaplan-Meier estimator, two samples with 70% and 10% of censoring



4.5 Results

We start the study with 45 configurations given by three levels for α (the shape parameter of the Weibull distribution), three levels for $\kappa = P(x = 1)$, and five levels for RR. In each configuration we consider 3 levels of heavy censoring ($p = 0.7, 0.8, 0.9$) and 10 different sample sizes. Table 4.3 summarizes the 1350 scenarios for the simulation, each one replicated 1000 times. All simulations were implemented as functions of the statistical package R, using the R-packages `splines` and `survival` (see Appendix A.2.1). We display the simulation results of the bias, variance, mean square error (mse), relative bias and coverage in the Cox model for the estimator rr of the relative risk RR.

The preliminary results showed that the shape parameter of the Weibull, α, does not have an influence in the results (in terms of bias and mse). We present the results for $\alpha = 1$ and discuss the results for κ, RR, p, and n. Hereby we conducted the simulation for 450 scenarios with three censoring levels, after having eliminated the value of α. The conduction in terms of bias, variance and mean square error are similar for the unbalanced case as for the balanced case. We present and discuss here the situation with $\kappa = 0.8$, hence $n_0 < n_1$.

Table 4.3: Simulation scenarios

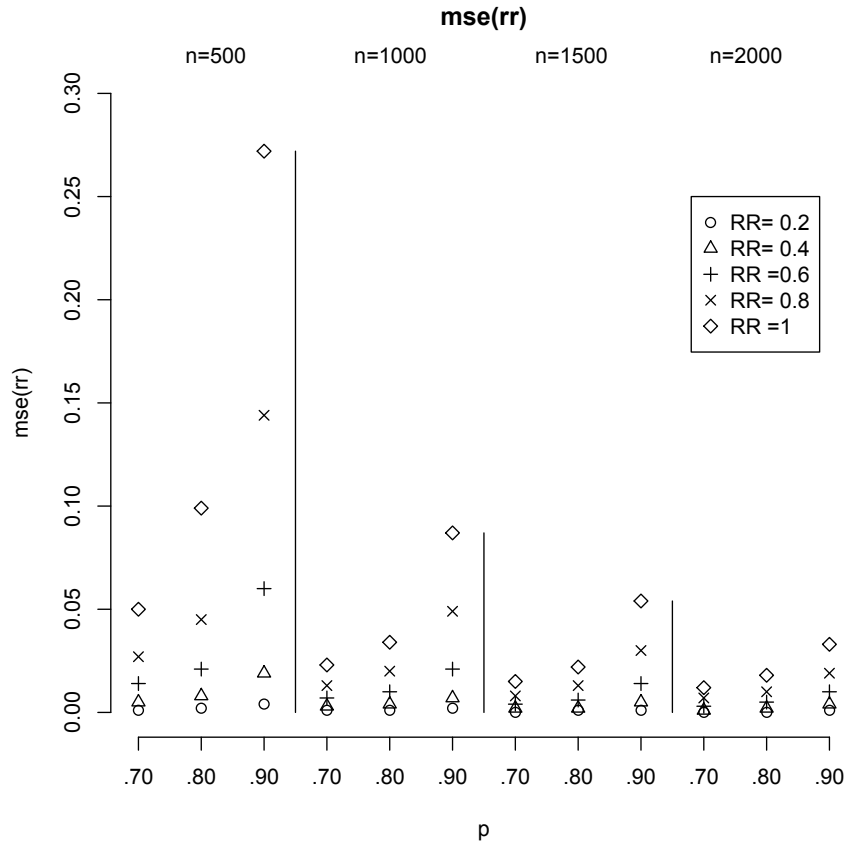
Number of replications	m	1000
Common shape parameter Weibull distributions	α	0.5, 1, 2
Probability of choosing an individual from group 1	κ	0.2, 0.5, 0.8
Relative risk	RR	0.2, 0.4, 0.6, 0.8, 1
Probability of censoring	p	0.7, 0.8, 0.9
Sample size	n	50, 100, 200, 300, 400,
		500, 1000, 1500, 2000, 2500

In Table 4.4 (at the end of this chapter) we present the bias, variance and mean square error for the scenarios defined in the previous Table. When the size of the sample is small, sometimes one of the groups does not present events, and generates problems in the convergence of the process of estimation of the relative risk. These atypical cases were excluded in Table 4.3. We observe that given a value of the relative risk, the bias and variance grow when increasing the censoring level, and both decrease when the sample size increases. If the value of relative risk increases to 1, the bias and variance decrease slowly.

Since the RR is a relative quantity, a small variation in this value can involve big differences between the risks of the groups, as it is shown in Table 4.1. We are ready to accept a variation of the relative bias up to 14%. That is to say, $|(\frac{rr-RR}{RR}) * 100| < 14$, which implies that $0.86RR < rr < 1.14RR$. For example, when $RR = 1$, we accept an estimation if it is contained in the interval (0.86, 1.14). In Table 4.5 (at the end of this chapter), we showed the relative bias in percentage, the coverage and the number of times that the procedure converges and obtains the estimator. For each value of RR, we observed that the relative bias decreases when n increases,

and presents big variations when $n < 500$. Also we observed that there are convergence problems when $n < 500$, mainly when the censoring level is 90%. The coverage is consistent to 95% when $n > 500$. From this table we can deduce that, if the percentage of censoring is 70% or below, the Cox model can be used for the estimation of the relative risk whenever the total sample size in both groups is larger than 100. For an 80% of censoring, a larger than 200 size would be required to get an approximately unbiased and precise estimator for the relative risk. Whenever the percentage of censoring is 90% or larger, a sample size of at least 500 is required, unless the RR is relatively small.

Figure 4.2: Mean square error of the relative risk estimator, $n_0 < n_1$



In Figure 4.2 we presented a graph of the behavior of *mse* under different values from RR with three censoring levels, and $n = 500, 1000, 1500, 2000$. The *mse* decreases when RR decreases, and increases when the censoring level increases. This behavior stays for each level of n. For

each level of RR, the mse is maximum when the censoring level is 90%.

4.6 Conclusion

In the presence of censoring levels from 70% to 90%, the Cox model is always suitable if the sample size is greater or equal than 500. The study has verified, that the behavior of the relative risk, in terms of mse, is better if the sample is balanced. If censoring is too heavy, the Cox model should not be used or used cautiously when sample size is smaller than 500. In this case, there can be problems of convergence or one could have an erroneous estimation of the relative risk, since it tends to overestimate the parameter. The Cox model can be used cautiously for the estimation of the relative risk whenever the total sample size in both groups is larger than 100 for a percentage of censoring of 70% or below, or larger than 200 for an 80% of censoring. If the sample size is larger than 200 and the RR is relatively small with a percentage of censoring of 90%, the Cox model can be used too. Although in this last case it could have convergence problems, these problems will disappear if the groups are balanced.

Table 4.4: Properties of the estimator for relative risk under heavy censoring

$n_0 < n_1$	bias(rr)			var(rr)			mse(rr)		
n\p	0.70	0.80	0.90	0.70	0.80	0.90	0.70	0.80	0.90
RR =0.2									
50	0.035	0.060	0.077	0.025	0.052	0.053	0.027	0.056	0.059
100	0.010	0.020	0.050	0.008	0.015	0.042	0.008	0.015	0.045
200	0.004	0.007	0.019	0.003	0.005	0.022	0.003	0.005	0.023
300	0.003	0.004	0.012	0.002	0.003	0.008	0.002	0.003	0.008
400	0.004	0.004	0.008	0.002	0.002	0.006	0.002	0.002	0.006
500	0.003	0.003	0.004	0.001	0.002	0.004	0.001	0.002	0.004
1000	0.000	0.001	0.002	0.001	0.001	0.002	0.001	0.001	0.002
1500	0.000	0.000	0.001	0.000	0.001	0.001	0.000	0.001	0.001
2000	0.001	0.001	0.002	0.000	0.000	0.001	0.000	0.000	0.001
2500	0.001	0.001	0.002	0.000	0.000	0.001	0.000	0.000	0.001
RR =0.4									
50	0.103	0.128	0.039	0.156	0.190	0.110	0.167	0.206	0.112
100	0.036	0.066	0.107	0.043	0.088	0.168	0.044	0.093	0.180
200	0.012	0.023	0.062	0.014	0.023	0.090	0.014	0.024	0.094
300	0.008	0.013	0.044	0.009	0.014	0.043	0.009	0.014	0.045
400	0.008	0.013	0.025	0.007	0.011	0.027	0.007	0.011	0.028
500	0.005	0.006	0.017	0.005	0.008	0.019	0.005	0.008	0.019
1000	0.001	0.002	0.009	0.003	0.004	0.007	0.003	0.004	0.007
1500	0.000	0.001	0.006	0.002	0.002	0.005	0.002	0.002	0.005
2000	0.002	0.003	0.005	0.001	0.002	0.004	0.001	0.002	0.004
2500	0.002	0.003	0.006	0.001	0.001	0.003	0.001	0.002	0.003
RR=0.6									
50	0.179	0.175	-0.035	0.373	0.379	0.150	0.405	0.410	0.151
100	0.075	0.135	0.148	0.122	0.270	0.322	0.127	0.288	0.344
200	0.024	0.056	0.123	0.036	0.107	0.234	0.037	0.110	0.249
300	0.017	0.030	0.095	0.025	0.040	0.161	0.025	0.040	0.170
400	0.018	0.026	0.052	0.019	0.032	0.072	0.020	0.033	0.075
500	0.010	0.013	0.047	0.014	0.021	0.058	0.014	0.021	0.060
1000	0.003	0.005	0.022	0.007	0.010	0.021	0.007	0.010	0.021
1500	0.001	0.003	0.013	0.004	0.006	0.014	0.004	0.006	0.014
2000	0.004	0.006	0.010	0.003	0.005	0.009	0.003	0.005	0.010
2500	0.005	0.005	0.012	0.002	0.004	0.008	0.002	0.004	0.008
RR=0.8									
50	0.258	0.145	-0.145	0.695	0.486	0.172	0.761	0.507	0.193
100	0.135	0.227	0.155	0.321	0.585	0.460	0.339	0.636	0.484
200	0.046	0.091	0.216	0.080	0.220	0.559	0.083	0.229	0.606
300	0.026	0.063	0.174	0.050	0.087	0.397	0.050	0.091	0.427
400	0.027	0.046	0.093	0.040	0.072	0.203	0.041	0.074	0.211
500	0.015	0.029	0.090	0.027	0.045	0.136	0.027	0.045	0.144
1000	0.006	0.013	0.041	0.013	0.020	0.047	0.013	0.020	0.049
1500	0.003	0.008	0.022	0.008	0.013	0.030	0.008	0.013	0.030
2000	0.006	0.011	0.014	0.007	0.010	0.019	0.007	0.010	0.019
2500	0.006	0.008	0.021	0.005	0.008	0.018	0.005	0.008	0.018
R R = 1									
50	0.303	0.087	-0.278	0.964	0.541	0.193	1.055	0.549	0.271
100	0.224	0.275	0.117	0.652	0.888	0.573	0.702	0.963	0.586
200	0.079	0.156	0.279	0.162	0.523	0.801	0.168	0.547	0.879
300	0.041	0.095	0.226	0.089	0.187	0.679	0.090	0.196	0.730
400	0.038	0.061	0.155	0.072	0.118	0.446	0.073	0.121	0.470
500	0.023	0.046	0.136	0.049	0.097	0.254	0.050	0.099	0.272
1000	0.009	0.019	0.059	0.023	0.034	0.084	0.023	0.034	0.087
1500	0.006	0.011	0.033	0.015	0.022	0.053	0.015	0.022	0.054
2000	0.010	0.016	0.020	0.012	0.018	0.033	0.012	0.018	0.033
2500	0.009	0.011	0.029	0.009	0.013	0.031	0.009	0.013	0.032

Table 4.5: Consistency of the estimator for relative risk under heavy censoring

$n_0 < n_1$	(bias(rr)/RR)x100%			coverage			convergence		
n\p	0.70	0.80	0.90	0.70	0.80	0.90	0.70	0.80	0.90
RR =0.2									
50	17.358	29.773	38.335	0.951	0.970	0.992	999	992	849
100	4.952	9.958	25.135	0.947	0.953	0.976	1000	1000	993
200	1.806	3.525	9.713	0.956	0.957	0.957	1000	1000	1000
300	1.572	2.141	6.231	0.959	0.949	0.952	1000	1000	1000
400	2.205	2.051	4.114	0.950	0.941	0.943	1000	1000	1000
500	1.463	1.410	1.926	0.954	0.960	0.949	1000	1000	1000
1000	0.049	0.324	0.812	0.944	0.944	0.962	1000	1000	1000
1500	0.003	0.199	0.512	0.947	0.944	0.940	1000	1000	1000
2000	0.409	0.530	0.869	0.948	0.953	0.949	1000	1000	1000
2500	0.318	0.708	1.067	0.952	0.950	0.946	1000	1000	1000
RR =0.4									
50	25.656	31.898	9.626	0.964	0.971	0.981	996	977	807
100	9.078	16.517	26.740	0.948	0.965	0.975	1000	1000	971
200	2.953	5.648	15.471	0.961	0.952	0.960	1000	1000	999
300	2.028	3.345	10.898	0.950	0.943	0.952	1000	1000	1000
400	1.926	3.227	6.178	0.942	0.942	0.949	1000	1000	1000
500	1.366	1.621	4.282	0.951	0.947	0.955	1000	1000	1000
1000	0.183	0.599	2.326	0.941	0.945	0.964	1000	1000	1000
1500	0.117	0.286	1.512	0.948	0.942	0.943	1000	1000	1000
2000	0.599	0.674	1.333	0.945	0.950	0.952	1000	1000	1000
2500	0.624	0.661	1.400	0.952	0.946	0.944	1000	1000	1000
RR=0.6									
50	29.770	29.203	-5.786	0.966	0.963	0.972	989	944	737
100	12.563	22.504	24.640	0.957	0.964	0.969	1000	999	939
200	4.026	9.309	20.567	0.958	0.956	0.970	1000	1000	999
300	2.816	4.931	15.778	0.947	0.953	0.959	1000	1000	1000
400	2.943	4.336	8.684	0.939	0.948	0.944	1000	1000	1000
500	1.641	2.132	7.768	0.944	0.955	0.953	1000	1000	1000
1000	0.561	0.848	3.730	0.944	0.947	0.958	1000	1000	1000
1500	0.223	0.552	2.143	0.936	0.946	0.955	1000	1000	1000
2000	0.742	0.929	1.667	0.940	0.948	0.954	1000	1000	1000
2500	0.783	0.851	2.000	0.954	0.941	0.943	1000	1000	1000
RR=0.8									
50	32.243	18.157	-18.127	0.962	0.962	0.961	978	899	666
100	16.815	28.326	19.397	0.964	0.969	0.967	1000	995	897
200	5.773	11.425	27.051	0.953	0.947	0.975	1000	1000	993
300	3.301	7.856	21.768	0.947	0.956	0.957	1000	1000	1000
400	3.321	5.766	11.606	0.943	0.947	0.953	1000	1000	1000
500	1.924	3.650	11.235	0.948	0.948	0.949	1000	1000	1000
1000	0.702	1.599	5.167	0.944	0.949	0.948	1000	1000	1000
1500	0.378	0.999	2.718	0.941	0.943	0.947	1000	1000	1000
2000	0.775	1.388	1.738	0.946	0.941	0.952	1000	1000	1000
2500	0.773	1.028	2.664	0.939	0.950	0.946	1000	1000	1000
R R = 1									
50	30.264	8.734	-27.828	0.959	0.953	0.947	956	851	606
100	22.350	27.457	11.702	0.965	0.951	0.959	1000	978	852
200	7.877	15.558	27.935	0.952	0.954	0.973	1000	1000	985
300	4.132	9.535	22.555	0.954	0.953	0.960	1000	1000	996
400	3.828	6.124	15.465	0.950	0.950	0.956	1000	1000	999
500	2.339	4.600	13.649	0.954	0.954	0.950	1000	1000	1000
1000	0.890	1.874	5.897	0.940	0.956	0.958	1000	1000	1000
1500	0.558	1.120	3.320	0.940	0.953	0.946	1000	1000	1000
2000	0.999	1.618	2.004	0.944	0.946	0.960	1000	1000	1000
2500	0.900	1.099	2.865	0.943	0.954	0.945	1000	1000	1000

Chapter 5

Analysis of the melanoma data via mixture cure models. Assessment of sufficient follow-up

5.1 Introduction

In this Chapter we return to the discussion of the melanoma data described in Chapter 2, section 2.1. In the development of this chapter, we allow the possibility that immune or cured individuals are present in the population. An analysis via a mixture cure model, defined in Chapter 3, is then more appropriate than a standard survival model, since it takes into account the immune proportion and the survival function of the nonimmune individuals. However in a scenario of immune and nonimmune individuals in the population, three aspects must be taken into account before proceeding to use the mixture cure model. First, there must be some empirical evidence to suppose the presence of immune individuals in the population. Second, the presence of immune individuals in addition to those susceptible who are censored before the end of the study could result heavy censoring percentage. Third, it has to be checked whether the follow-up was enough to make sure that individuals in the population are really immune.

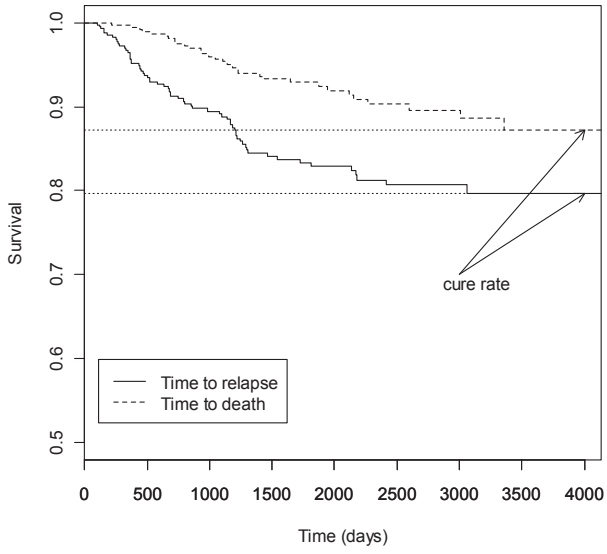
This Chapter is composed as follows, Section 5.2 presents a description of the cure proportion,

level of censoring and the sufficient follow-up: nonparametric tests are introduced in 5.2.1 and 5.2.2, both implemented as functions of R (see Appendix A.2) and applied to the melanoma dataset to evaluate the follow-up, results of these tests are summarized in 5.2.3. Subsequently an analysis via a mixture cure model is presented in Section 5.3 using the SAS macro PSPMCM: the model formulation and discussion of the estimation procedure are presented in 5.3.1, discussion of software for data analysis are presented in 5.3.2, and 5.3.3 summarizes the results. Finally, Section 5.4 contains an overall conclusion from the results.

5.2 Sufficient follow-up in the case of melanoma data

We begin this section by examining the Kaplan and Meier curves for the melanoma data described in Chapter 2. Figure 5.1 shows the K-M curves for the *disease-free time* (time to relapse) and *overall time* (time to death), we see that both have the property of an improper survival (both curves are around 0.8), which is an empirical evidence which suggests the presence of immunes or cured individuals.

Figure 5.1: Kaplan and Meier curves for *disease-free time* (time to relapse) and *overall time* (time to death), both survival curves are improper, $S(\infty) > 0$.



The K-M estimator in Figure 5.1, represents the estimated curve for survival function of the

population $S(t)$, as stated in equation (3.7).

$$S(t) = [1 - \pi] + \pi S_1(t). \tag{5.1}$$

This approach leaves too an estimator for the cure rate $S(\infty)$, given by $\widehat{S}_{KM}(u_{(n)})$, where $u_{(n)}$ is the maximum survival time observed in the sample of size n (see Figure 5.1). These estimates are presented in Table 5.1 for both, the *disease-free time* and the *overall time*, along with its the standard error and an 95% confidence interval.

Table 5.1: Estimation of the cure rate

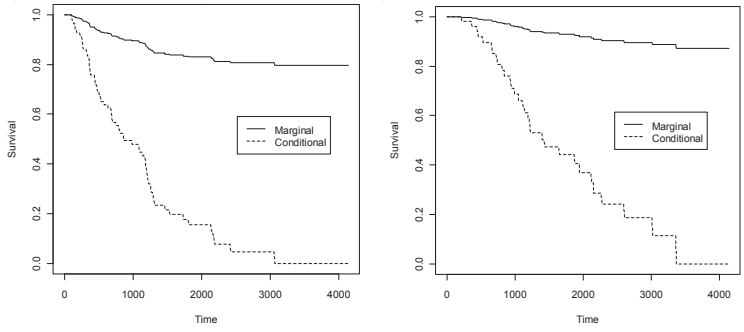
Period	cure	se(cure)	low.95	upper.95
Disease-free time	0.797	0.025	0.750	0.847
Overall time	0.872	0.025	0.823	0.923

Similarly, the estimated survival function of nonimmune individuals $S_1(t)$ may be obtained from (3.7) via Kaplan-Meier

$$\widehat{S}_1(t) = \frac{\widehat{S}_{KM}(t) - \widehat{S}_{KM}(u_{(n)})}{1 - \widehat{S}_{KM}(u_{(n)})}, \qquad t > 0. \tag{5.2}$$

The estimated curves for the marginal and conditional survival function are shown in Figure 5.2. Sometimes S_1 is called *conditional survival*, while S is called the *marginal survival.* To

Figure 5.2: Kaplan and Meier curves for conditional and marginal survival: a) *disease-free time* (time to relapse), b) *overall time* (time to death).



a) b)

distinguish between censored due to lost to follow-up and end of follow-up, we have taken the maximum observed time to event as the reference point. The maximum observed *time to relapse* was $t_{(n)} = 3065$ days, while that, the maximum observed *time to death* was $t_{(m)} = 3358$ days (see Figure 5.3).

Figure 5.3: Censoring scheme: $t_{(n)}$ is reference point for *disease-free time* (time to relapse) and $t_{(m)}$ is reference point for *overall time* (time to death).

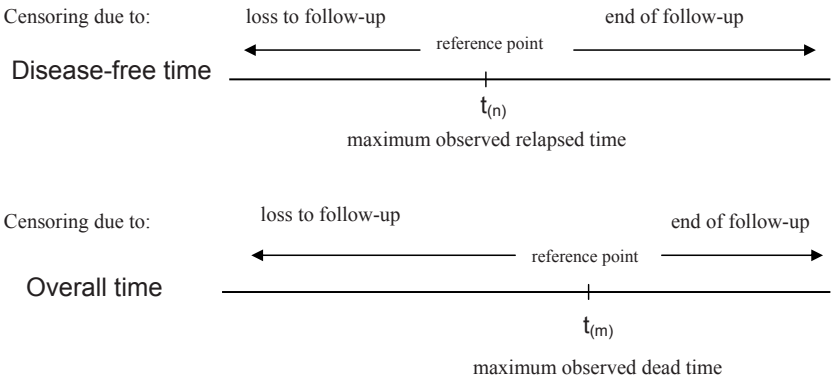


Table 5.2 summarizes the level of censoring for each case that are illustrated in the scheme of the Figure 5.3. In this table we see a relatively small percentage of individuals censored due to the end of follow-up for *overall time*, such individuals are contributing to the cure rate. We also found that 6.5% censored patients due to end of follow-up in the period *disease-free time* (censored in interval $(t_{(n)}, t_{(m)})$), contribute to censoring due to lost of follow-up for *overall time*.

Table 5.2: Censoring levels

	Censoring due to		
Period	loss to follow-up	end of follow-up	total
Disease-free time	64.00%	20.25%	84.25%
Overall time	77.50%	7.75%	85.25%

Although in Figure 5.1 and Table 5.1 show strong evidences to suppose the existence of immune

individuals in the population, it isn't clear whether follow-up was sufficient. For example, for *overall* survival we see that the level of censoring due to the end of follow-up is low (see Table 5.2), the conditional survival curve falls slowly to zero (see Figure 5.2 b). Could indicate that the proportion of censoring due to end of follow-up not due to the presence of immune individuals but rather a insufficient follow-up.

In order to prove whether or not there is sufficient follow-up in this data, two non-parametric tests are reviewed and are introduced below. Both tests have been implemented as functions in the statistical package R and are applied to the dataset, a discussion about application is presented in the last subsection and Appendix A.2.

5.2.1 α-test

Let F and F_1 the distribution functions defined by $F = 1-S$ and $F_1 = 1-S_1$, G is the distribution function of the censoring and $\tau_{_F}$, $\tau_{_{F_1}}$, $\tau_{_G}$ are the right extremes, respectively, defined in section 1.1. When $0 < \pi < 1$, $\widehat{\pi}_n = 1 - \widehat{S}_{KM}(u_{(n)})$ is consistent if and only if

$$\tau_{_{F_1}} \leq \tau_{_G},$$

where $\tau_{_{F_1}} \leq \tau_{_G}$ means a sufficient follow-up (Maller and Zhou (1992)). The consistency properties for $\pi = 1$ under a scenario of heavy censoring can see in Zukang (1997), Wellner (1985), among others.

Maller and Zhou (1992, 1994, 1996) proposed a nonparametric statistic, called the α_n-test, to test the hypothesis

$$H_0 : \tau_{_{F_1}} > \tau_{_G} \quad \text{versus} \quad H_1 : \tau_{_{F_1}} \leq \tau_{_G}. \tag{5.3}$$

The test rejects H_0 if $\alpha_n < \alpha$, where $\alpha_n = (1 - \frac{N_n}{n})^n$, N_n is the number of failure times in the interval $(2t_{(n)} - u_{(n)}, t_{(n)}]$, $t_{(n)}$ is the maximum observed failure time and $u_{(n)}$ is the maximum observed failure or censored time in a sample of size n. Observe that if w_n and δ_n are the values observed of $T_{(n)}$ and $U_{(n)} - T_{(n)}$ in the sample, then $w_n - \delta_n = t_{(n)} - (u_{(n)} - t_{(n)}) = 2t_{(n)} - u_{(n)}$, where δ_n is the length of the interval $(t_{(n)}, u_{(n)}]$.

5.2.2 Δ-test

Due to the instability of $\widehat{\pi}_n$ when follow-up is large, Klebanov and Yakovlev (2007) proposed a procedure to test the hypothesis (equivalent to (5.3))

$$H_0 : S(T_0) = S_1(T_0) \quad \text{versus} \quad H_1 : S(T_0) > S_1(T_0), \tag{5.4}$$

where T_0 is the duration of follow-up and $S_1(t)$ has non-decreasing φ-hazard rate average defined in section 1.1. Under a random censoring scheme with $T_0 < \min\{\tau_{F_1}, \tau_G\}$, the test is based on the Kolmogorov goodness-of-fit statistic

$$\Delta_n = \widehat{S}_n(T_0) - \varphi(\frac{T_0}{t_0}\varphi^{-1}(\widehat{S}_n(t_0))) - [1 + \frac{T_0}{t_0}]\frac{D_\alpha}{\sqrt{n}}\widehat{S}_n(t_0)(1 + A_n(t_0)),$$

where $A_n(t_0) = n\sum_{(i:t_i<t_0)} \frac{\delta_i}{(n-i)(n-i+1}$, δ_i is the censoring indicator and D_α is the $(1-\alpha)$th percentile of the asymptotic Kolmogorov distribution. The test rejects H_0 at a significance level of less that α, if $0 < \Delta_n$. In addition a consistent estimator $\widetilde{\pi}_n$ for π is given by equation (26) of Klebanov and Yakovlev (2007), and by Maller and Zhou (1992) it is concluded that follow-up was sufficient. These two tests are not available yet in statistical software. We have implemented them in R (see Appendix A.2).

5.2.3 Results

For the implementation of the Δ-test, t_0 and $\widetilde{\pi}_n$ were obtained using equation (26) of Klebanov and Yakovlev (2007). The statistic Δ_n has been computed using that $D_\alpha \simeq \frac{z_\alpha}{\sqrt{n}}$, see Gibbons (1985). The Δ-test can be used since as widely discussed by Klebanov and Yakovlev (2007), $S_1(t)$ has non-decreasing φ-hazard rate overage, especially in applications in cancer data.

The results obtained of the α-test and the Δ-test using the melanoma data are summarized in Table 5.3. Observe that, both tests are at the limit of the rejection region, contrasting with the slow convergence to zero in the conditional survival function (Figure 5.2 b)) and explained by the excessive censoring due to loss to follow-up (Table 5.2). We can see that with both tests we reject H_0 for each case, so we can assume that there is sufficient follow-up, for both, *disease-free time* (with $\widetilde{\pi}_n = 0.536$) and *Overall time* (with $\widetilde{\pi}_n = 0.319$). However, regarding the estimation of the cure rate, we must take into account the censoring level (see Table 5.2) due to excessive loss of patients during follow-up (Sy and Taylor (2000)), and utilize $\widetilde{\pi}_n$ as a nonparametric lower

bound and $\widehat{\pi}_n$ as a nonparametric upper bound for cure rate π (Klebanov and Yakovlev (2007)). In this way, we can infer an interval for the cure rate obtained from the Δ-test: cure $[0.536, 0.797]$ for *disease-free time* and $[0.319, 0.872]$ for *overall time*.

Table 5.3: Test for sufficient follow-up

Test-Statistic	conclusion	implication
	Disease-free time	
$\alpha_n = 0.006$	reject $H_0 : \tau_{F_1} > \tau_G$	sufficient follow-up
$\Delta_n = 0.204$	reject $H_0 : S(T) = S_1(T)$	$\widetilde{\pi}_n$ consistent
	Overall-time	
$\alpha_n = 0.049$	reject $H_0 : \tau_{F_1} > \tau_G$	sufficient follow-up
$\Delta_n = 0.045$	reject $H_0 : S(T) = S_1(T)$	$\widetilde{\pi}_n$ consistent

5.3 Analysis of the melanoma data

From Table 5.2 in previous section we have seen that the proportion of censoring due to loss to follow-up is 64% for *disease-free time* and about 77% for *overall time*, and the rest of the censored patients can be defined as *cured* patients. In the first case *cured* patients are those who are not expected relapse or die (20.25% of patients), whereas in the second case, are those that may relapse but not die (7.75% of patients). From the heavy censoring analysis in the chapter 4, we found that for a censoring percentage of 80% need a sample size larger than 200; in the melanoma dataset there are 400 patients. Moreover, in Section 5.2 we found that in both periods *disease-free* and *overall time*, the follow-up is sufficient.

The next subsection summarizes the inference procedure for a mixture cure model introduced by Farewell (1977). This methodology was implemented as a SAS macro by Corbière and Joly (2007), and is used in this thesis to analyze the data melanoma, the results are presented in the last section.

5.3.1 Likelihood function

Let T be the survival time for a patients with cancer and C be the censoring time, T and C independent, T with distribution function F and C with distribution function G. Suppose that T is right censored by C, and let U be observed survival time and δ the censoring indicator, defined in section 1.2. Let Y be a binary variable that specifies that an individual is nonimmune to cancer (susceptible) with Y=1 or immune (cured) with Y=0, and $\pi = P[Y = 1]$, Farewell (1977). Let $\{(u_i, \delta_i, y_i, x_i, z_i) : i = 1, 2, \ldots, n\}$ be survival data observed in a sample n patients with cancer, where the *i-th* patient has an observed survival time u_i with censoring indicator δ_i, cure indicator y_i (partially observed) and a vector of covariates (x_i, z_i). The full likelihood function observed under an independent, noninformative, random censoring model is given by

$$L(b, \beta) = \prod_{i=1}^{n} \{\pi_i(z_i|b) f(u_i|\beta, x_i)\}^{\delta_i} \{[1 - \pi_i(z_i|b)] + \pi_i(z_i|b) S(u_i|\beta, x_i)\}^{1-\delta_i}, \tag{5.5}$$

where $\pi(z_i|b) = P[Y = 1|z_i]$ is the probability of being susceptible given a covariate vector z_i, $S(u_i|\beta, x_i) = S(u_i|Y = 1, x_i)$ is the survival function for susceptible individuals given a covariate vector x_i and $f(u_i|\beta, x) = -\frac{d}{du_i} S(u_i|\beta, x_i)$. Furthermore x is a vector of covariates explaining the survival time of susceptible, z is a covariates vector including the intercept explaining the proportion of susceptible, which may include the same covariate as x.

Maximization procedure

For a parametric mixture cure model, β can be estimated specifying a distribution function to the survival time of susceptible patients, $S(u|Y = 1)$. Discussions of parametric mixture cure models can be found in Farewell (1977), Farewell (1982), Ghitany et al. (1994), Peng et al. (1998), De Angelis et al. (1999), Yu et al. (2005). Main problem with the parametric mixture cure models is that it is difficult to verify the distributional assumptions. An alternative to these models are semiparametric mixture cure models, such as Cox's proportional hazard models.

For a Cox's proportional hazard mixture cure models, the conditional distribution of the susceptible population is defined by

$$S(u|Y = 1, x) = S_0(u|Y = 1)^{\exp(\beta' x)}, \tag{5.6}$$

where $S_0(u|Y=1)$ is left arbitrary. By replacing $S(u|Y=1,x)$ of the (5.6) in (5.3) can obtain estimates for b and β maximizing $L(b,\beta,S_0(u|Y=1))$ via EM algorithm (see Appendix A.1). Discussions of nonparametric mixture cure models can be found in Kuk and Chen (1992), Taylor (1995), Sy and Taylor (2000), Peng and Dear (2000) and Corbière and Joly (2007).

5.3.2 The macro PSPMCM

Several mixture cure models using parametric and nonparametric methods have been programmed in different softwares, see section 3.4. Here we use the SAS macro called PSPCM given by Corbière and Joly (2007), to estimate the model introduced by Farewell (1982), Sy and Taylor (2000) and Peng and Dear (2000).

For the melanoma's cancer we have established a Logistic-Cox model given by

$$S(u|z,x) = [1-\pi(z|b)] + \pi(z|b)S(u|\beta,x), \tag{5.7}$$

where the effects of z and x are modeled via

$$\pi(z|b) = \frac{\exp(b'z)}{1+\exp(b'z)} \quad \text{and} \quad S(u|\beta,x) = S_0(u)^{\exp(\beta'x)}.$$

Being $\pi(z|b) = P[Y=1|z]$ is the probability of relapse (probability of incidence) to cancer and $S(u|\beta,x) = P[U>u|Y=1,x]$ the conditional survival of the time to relapse (conditional distribution of latency). The main goal is to identify factors that would increase the probability of incidence and which are the factors that would accelerate the occurrence when this can occur.

5.3.3 Results for the melanoma data

In the analysis all factors were introduced to the mixture cure model, in the cure part and survival. For *disease-free time* we use the Location type (Extremities-Trunk versus Head-Neck), Breslow level ($<2, [2-4)$ and ≥ 4), Histopathological of Malignant Melanoma subtype (SSM versus ALM-LMM-NM) and Age (≤ 45, 46-70, and >70) (see Table 2.7, Chapter 2). For *overall time* we use Breslow level (<4 versus ≥ 4), Histopathological of Malignant Melanoma subtype (SSM-LMM-NM versus ALM) and age (less than 60 years old and higher than 60) (see Table 2.12, Chapter 2).

Significant covariates for the probability of the event (relapse/death) and conditional survival of the time to event using model (5.7) are presented in the Table 5.4 and Table 5.5.

Table 5.4: Statistical significant factors for probability of relapse and the time to relapse using a Logistic-Cox model. Reference group: SLN status= Negative, Localization= Extremites-Trunk, Bres< $2mm$, Ulceration= No, Clark level= I-III, HMM subtype= SSM, Age $\leq$ 45 years and Gender= Female.

Predictors	β	e^{β}	se(β)	p	$L_{.95}$	$U_{.95}$
LOGIT PART						
`SLN status`						
Positive	0.762	2.142	0.358	0.033	1.061	4.324
`Clark level`						
IV-V	1.414	4.111	0.482	0.003	1.599	10.571
`Age`						
46-70 years	1.238	3.448	0.432	0.004	1.479	8.036
>70 *years*	2.647	14.117	0.490	0.000	5.407	36.859
`Gender`						
Male	0.899	2.458	0.313	0.004	1.331	4.541
`HMM subtype`						
ALM-LMM-NM	0.744	2.105	0.354	0.035	1.052	4.212
SURVIVAL PART						
`Localization`						
Head-Neck	1.013	2.754	0.426	0.017	1.195	6.349
`Breslow level`						
[2,4)mm	0.028	1.028	0.398	0.944	0.471	2.244
$\geq 4mm$	0.992	2.696	0.446	0.026	1.124	6.465
`Ulceration`						
Yes	1.479	4.390	0.382	0.000	2.078	9.275

We observe in Table 5.4 that the characteristics of the tumor, such as *Sentinel Lymph Node status*, *Clark level*, *Histopathological of Malignant Melanoma subtype*, as well as *Age* and *Gender* of the patient are highly significant factors for the incidence of cancer. For example the odds ratio to *Age*>70 *years* on reference group *Age*≤ 45 *years* is given as $OR(Age > 70 years | Age \leq 45 years) = \exp(2.647) = 14.117$. That is, the odds of relapse of patients older than 70 years are 14 times the odds of relapse of patients younger than 45 years; the odds of relapse of patients between 46-70 years are three times more than the odds of relapse of patients younger than 45 years. The odds of relapse of patients with positive Sentinel Lymph Node status are twice the odds of relapse of patients with negative status. The odds of relapse of patients with level of invasion into the skin from IV-V are four times the odds of relapse of patients with level I-III. The odds of relapse of patients with Histopathological of Malignant Melanoma ALM-LMM-NM are twice the odds of relapse of patients with SSM subtype. There is twice more the odds of relapse for males than for females.

In the same Table 5.4, we observe that the *Location*, *Breslow* and *Ulceration* of the cancer are significant factors for survival in susceptible patients. These factors accelerate the relapse or development of cancer. Patients with ulceration are four times more risk to relapse than patients without ulceration. Patients with level of Breslow$\geq 4mm$ have two times more risk to relapse than patients with Breslow$< 2mm$. There are two times more risk of relapse among patients with head and neck cancer that with patients with cancer in extremities and trunk.

We observe in Table 5.5 that the *Ulceration* and *Clark* are characteristics of the tumor highly significant for death due to cancer. The odds of death of patients with Clark level between IV-V are seven times more than the odds of death of patients with Clark level between I-III. The odds of death of patients with ulceration are twice the odds of death of patients without ulceration. Observe that the *Ulceration* is a factor that accelerating the time to relapse (Table 5.4), increasing the probability of death (Table 5.5). Moreover, in Table 5.5 we see that the *Breslow* is a highly significant factor for the survival time of patients. Patients with a Breslow level$\geq 4mm$ have about five times more risk to die than patients with a Breslow level$< 4mm$.

Table 5.5: Statistical significant factors for probability of death and the time to death using a Logistic-Cox model. Reference group: SLN status= negative, Bres< $4mm$, Ulceration= no, Clark level= I-III.

Predictors	β	e^{β}	se(β)	p	$L_{.95}$	$U_{.95}$
LOGIT PART						
`SLN status`						
Positive	0.687	1.989	0.362	0.057	0.977	4.046
`Ulceration`						
Yes	0.737	2.089	0.349	0.034	1.054	4.141
`Clark level`						
IV-V	2.015	7.500	0.667	0.002	2.029	27.719
SURVIVAL PART						
`Breslow level`						
$\geq 4mm$	1.582	4.869	0.567	0.005	1.603	14.791

5.4 Conclusion

In the analysis for melanoma data described in Section 5.2, we obtain evidence that individuals cured (and not cured) can be in the population, increasing the assumption that a survival model that takes into account both populations may be more appropriate, than a standard proportional hazards model described in Chapter 2, Section 2.2. However, before applying a mixture cure model, we describe the percentage of censoring that arises due to loss to follow-up and end of follow-up, then evaluate whether the follow-up has been sufficient to ensure the presence of immune individuals in the population, and discard any effect only of the censoring mechanism due to follow-up. The information in Table 5.2 suggests the follow-up assessment, for this purpose, we apply the nonparametric tests described in section 5.2. The results of these tests, Table 5.3, show that in both, *disease-free time* and *overall time* the follow-up is sufficient. Therefore, it is valid to assume that immune individuals are actually present in the population, and a mixture cure model is more appropriate to analyze of this data. The advantage of this analysis over the

standard model of survival, is the separate modeling, effects of factors on the nonsusceptible proportion $[1 - \pi(z|b)]$ (or cure rate) and effects of factors on time survival of susceptible $S(u|\beta, x)$ (or not cured) individuals, like in Table 5.4 and 5.5. In the analysis for *disease-free time*, some of the factors with significant effects on the standard proportional hazards model, such as *Sentinel Lymph Node status*, *Clark level* and *Age* (Table 2.8), are now highly significant for the nonsusceptible proportion (Table 5.4) in the mixture cure model. Similarly, for analysis of *overall time*, *Ulceration* and *Clark level* (Table 2.13), now significant for the nonsusceptible proportion (Table 5.5). In addition, the mixture cure model incorporates other significant factors, such as *Gender*, *Histopathological of Malignant Melanoma subtype* and *Localization* for *disease-free time*, and *Breslow thickness* for *overall time*.

Chapter 6

Extended hazard models

6.1 Introduction

Mortality of calves from birth to weaning (approximately at 180 days) reduces farm's income, and significantly increases cattle production costs (see Goyache et al. (2003) for a review). Thus, it is important to take into account the survival pattern of calves into the overall breeding. Tarrés et al. (2005) used a standard survival analysis to study how genetic and environmental factors influence mortality up to weaning. However, and due to the high proportion of censoring in the data, one could think of the presence of a mixture of two subpopulations of calves: those susceptible to die before weaning and those who don't. A binary mixture model, also known as cure model, (e.g. Farewell, 1982), which takes into account a fraction of *cured* individuals, could be appropriate in this situation.

In this chapter we present an application of the extended hazard models (EHM) proposed by Tsodikov (2002) which is developed to combine both *long-term* and *short-term* effects. EHM models include as a particular case the proportional hazard cure models. We established a proportional hazard-proportional hazard cure (PHPHC) model to fit both genetic and environmental factors and discriminate between mortality of calves effects (*short-term* effects) and survival or cure effects (*long-term* effects).

6.2 Nonlinear transformation model

Let T be a non-negative random variable denoting the failure time of interest, with improper survival function $S_p(t|z)$ and bounded cumulative hazard function $H_p(t|z)$ such that $\pi(z) = S_p(\infty|z) > 0$ and $\theta(z) = H_p(\infty|z) < \infty$ being z a vector of covariates. A model that takes into account the cure fraction $\pi(z)$ can be formulated in two ways (as we have explained in section 3.1):
(i) A mixture cure model (Farewell, 1982) given by

$$S_p(t|z) = \pi(z) + [1 - \pi(z)]S(t|z), \tag{6.1}$$

where $S(t|z)$ is defined as the survival function for the time to failure conditional upon ultimate failure, i.e. $S(t|z) = P[T > t|T < \infty, z]$;
(ii) By specifying a bounded cumulative hazard function $H_p(t|z)$ of the population (Tsodikov (2002)) and taking the survival function of T as

$$S_p(t|z) = \exp\{-\theta(z)F(t|z)\}, \tag{6.2}$$

where $F(t|z) = \frac{H_p(t|z)}{H_p(\infty|z)}$. In terms of the estimation of the cure fraction $\pi(z) = \exp\{-\theta(z)\}$, the two representations (6.1) and (6.2) are equivalent within a nonparametric framework. Model (6.1) does not have the proportional hazard property, however when F does not depend on z, model (6.2) has the proportional hazard property and is referred as the proportional hazard cure model (PHC) (Tsodikov (2003)).

The standardized cumulative hazard function $F(t|z)$, itself a distribution function, might depend on the covariate vector z. Thus, its corresponding survival function, $1-F(t|z)$, can be specified as a parametric transformation of the baseline survival function S_0 (representing a reference group of individuals) in terms of a second predictor $\eta(z)$ (Tsodikov (2003)). In particular, Lehmann alternatives for $1 - F(t|z)$ can be assumed, that is, $1 - F(t|z) = S_0^{\eta(z)}(t)$, yielding a PH model for $1 - F(t|z)$. The combined PH-PHC model is then given by

$$S_p(t|z) = \exp\{-\theta(z)[1 - S_0^{\eta(z)}(t)]\}, \tag{6.3}$$

and it allows separate modelling of the *long-term* and *short-term* effects in terms of the covariate vector z, which may not necessarily be the same set for each predictor. This extension includes

the PHC model when there are not short-term predictors, that is, when $\eta(z) = 1$; and the PH model when there are no long-term predictors, that is, when $\theta(z) = 1$. Due to we are assuming $\eta(z) = \exp(\beta_\eta z)$ and $\theta(z) = \exp(\beta_\theta z + \beta_c)$, parameters β_η and β_θ are the regression coefficients for short-term effects and for long-term effects, respectively, and β_c is an additional regression parameter for the reference category of the cure fraction.

Inference procedures for regression coefficients β_η, β_θ and β_c are based on the generalized log-likelihood for a non linear transformation model. The R-package `nltm` includes the PH-PHC model, among others, and uses restricted Nonparametric Maximum Likelihood Estimation procedure (Tsodikov, 2002 and 2003) to get parameter estimates.

6.3 Mortality and survival up to weaning of beef calves

In this section we return to the data presented in section 2.3. We remind that these data have two main characteristics: first, the herds are heterogeneous among themselves, second; they are severely censored. Moreover this dataset contains a high level of missing data, mainly in the variable weight (weight at birth), see Table 2.17. With this evidence in hand, we propose to use the models described in section 6.2, excluding of the analysis the variable weight. The obtained results are presented below.

6.3.1 Results

A sample of 2077 calves in three different herds has been analyzed (see Chapter 2, section 2.3 for details about the data). Here the covariates included in the model were, the length of productive live of the cow, say `lpl`, dicotomized into groups < 1300 *days* and > 1300 *days*, month of birth, say `month`, dicotomized into groups *September to Febrary* and *March to August*, `gender` (*female, male*) and the type of difficulties at calving, say `difficulty`, categorized into *without assistance*, *slightly assisted by the farmer* and *strongly assisted by the farmer or the veterinary practitioner*.

Due to heterogeneity among the three herds, separate PH-PHC models (as in 6.3) were fitted for each herd. Table 6.1 displays the results for those models. Concerning long-term (cure) effects we find that calving month and difficulty at birth is the set of statistically significant factors

for the nonsusceptible proportion (*long-term effects*) of calves for herd 1, calving difficulty is the only significant factor for herd 7, and there are no significant predictors among this set of covariates for herd 3.

Table 6.1: Statistical significant factors for mortality and cure for each herd using a PH-PHC model. Reference group for herd 1: Month= calves born between September and February and Difficulty= without assistance, for herd 7: Difficulty= calves born without assistance.

Predictors	β	e^{β}	se(β)	p	$L_{.95}$	$U_{.95}$
			herd1			
Long term predictor						
`Month`						
mar-aug	1.96	7.097	0.989	0.047	1.022	49.263
`Difficulty`						
slightly assisted	1.87	6.476	0.474	0.000	2.557	16.401
strongly assisted	2.17	8.716	1.051	0.039	1.110	68.386
			herd7			
Long term predictor						
`Difficulty`						
slightly assisted	0.01	1.007	1.027	0.990	0.134	07.549
strongly assisted	1.33	3.798	0.471	0.004	1.507	09.569
Short term predictor						
`Length productive`						
>1300 days	0.89	2.440	0.466	0.056	0.978	06.080

We point out that the interpretation of the regression parameters for the cure fraction $\pi(z)$ is such that a higher value for e^{β} would represent a lower probability of cure for the corresponding factor. Note that model (3), together with $\eta(z) = \exp(\beta_\eta z)$ and $\theta(z) = \exp(\beta_\theta z + \beta_c)$, implies that $\pi(z) = (\pi(0))^{e^{\beta}}$ where $\pi(0) = \exp(-\exp(\beta_c))$ represents the probability of cure of the refer-

ence group. Observe that if β_1 is associated with $z = 1$ and $\exp(\beta_1) > 1$, means that $\theta(1) > \theta(0)$, then $\pi(1) < \pi(0)$. In particular, calves born in the period march-august have lower probability of cure than those born in september-february; and the probability of cure is much lower for those that have difficulties at calving for herd 1. For herd 7 the effect of difficulty is different as for herd 1, here only is significative the category *strongly assisted*. Calves that born from strongly assisted calving have lower probability of cure that calves from without assistance calving. The last two columns of Table 6.1 contain the lower ($L_{.95}$) and upper ($U_{.95}$) limits of a 95% confidence interval for the ratio $\frac{\log \pi(1)}{\log \pi(0)} = \exp(\beta_1)$.

Regarding short-term (mortality) effects, we only find statistically significant predictors in herd 7 where the risk of death of calves born to older mothers, hence with a longer reproductive life, is twice the risk of death of calves born to younger mothers ($\beta_\eta = 0.89$, $e^{\beta_\eta} = 2.44$, p-value $= 0.056$). The last two columns of Table 6.1 contain the limits of a 95% confidence interval, in this case, for the risk ratio e^{β_η}.

The results obtained by Tarrés et al. (2005), when the three herds were considered as one only with 2504 records, by using a standard Cox model, are different from those presented here. In this case, the calves borns from September-Febrary, had the lowest mortality risk. Calves from cows younger that 1300 days of productive life had a higher risk of mortality. The non-assisted calvings presented the smallest risk of mortality, and it increased up to five times according to the calving became more difficult.

Due to a complete parametrization of the probability of cure, that is survival up to weaning, given by $\pi(z)$, we obtained estimations of it for each of the categories of the significant covariates for the long-term effects given in Table 6.1. In Table 6.2 are estimates and confidence interval for the cured probability of the different groups for each herd, obtained using the relationship $\pi(z) = (\pi_0)^{e^\beta}$. Observe that if $(L_{.95}, U_{.95})$ is a 95% confidence interval for $\pi(0)$, then $(L_{.95}, U_{.95})^{e^{\beta_1}}$ is a 95% confidence interval for $\pi(1)$. We observe lower probabilities of cure for calves born between March and August and for calves born with assistance for herds 1. Whereas for the herd 7, there is low probability of cure for calves born strongly assisted. Furthermore, note that herd 7 is the only herd for which the length of productive live of the cow has an influence on the risk of death of the calves, and this short-term effect is influencing the probability of cure

Table 6.2: Estimates of the Probability of Cure $\pi(z)$ and 95% Semiparametric Likelihood Ratio Confidence Intervals (in parentheses). Reference group for herd 1: Month= calves born between September and February and Difficulty= without assistance, for herd 7: Difficulty= calves born without assistance.

Predictors	herd1 ($L_{.95}$, $U_{.95}$)	herd3 ($L_{.95}$, $U_{.95}$)	herd7 ($L_{.95}$, $U_{.95}$)
Reference Group	.993 (.955, .999)	.975 (.955, .986)	.980 (.968, .987)
`Month`			
mar-aug	.953 (.723, .992)		
`Difficulty`			
slightly assisted	.957 (.743, .993)		.981 (.968, .987)
strongly assisted	.942 (.671, .991)		.930 (.887, .953)

(survival up to weaning) in such a way that the confidence interval for those calves born with strong assistance $(.887, .953)$ is strictly below the confidence interval for calves born without assistance $(.968, .987)$. Thus,the probability of survival up to weaning of calves born without assistance is significatively higher than the probability of survival up to weaning of calves born with strong assistance.

6.4 Conclusion

Concluding, we point out that the PH-PHC model is an alternative to the standard Proportional Hazards model when there is a proportion of nonsusceptible individuals in the population. This model allows us to jointly estimate the proportion of cure (survival up to weaning) and the effect of different set of covariates for short and long-term on individuals in a heterogeneous population. Moreover, we have been able to use the same approach for the three herds, providing a unified method for situations, such as the one described in this chapter, where the initial set of covariates has different short and long effects on each herd.

Chapter 7

Contributions, Future research and Conclusions

In this chapter we present the major contributions we have made in this research, some tasks to develop and the conclusion of this thesis. Section 7.1 summarizes the contributions made in three categories: data analysis, methodological and implementation. In section 7.2 we discuss some conceptions that can be developed as future works, and finally in Section 7.3 the general conclusion is presented

7.1 Contributions

Survival data about oncology and veterinary studies discussed here are complex data. It involves indistinguishable mixtures of two populations, the susceptible and nonsusceptible; with a combination of random and fixed censoring. Besides complex vectors of covariates, which may or not influence in the survival times of the susceptible population and the nonsusceptible proportion. Data of this nature has to be analyzed via cure models, which involve a new methodology, not implemented in the statistical software. Estimation procedure, effect of censoring levels on estimators of the parameters in the model, statistical tests about sufficient follow-up, among others, has to be researched, developed and implemented in some statistical software.

Data analysis

- *Analysis of oncological data.* We begin this thesis with a discussion of an observational study of 400 patients diagnosed with skin cancer, which were followed for a fixed time period. For its medical importance, the time to relapse and time to death due to cancer, were defined as response variables to be investigated, and a set of tumor and personal characteristics as covariates. In a preliminary analysis we observed that some covariates were not completely collected, the Kaplan-Meier estimator of the curves for overall and disease-free time did not converge to zero. We also found that some covariates have to be recategorized, with the aim of getting more interpretable results. These and other results are presented in section 2.2, where we make a description of missing values of covariates, and then a semi-parametric survival analysis and nonparametric one thoroughly discussed for both response variables. However, due to a high percentage of patients do not relapse or survive to cancer, the time of these patients were defined as a censored observation at study end (fixed censoring or administered), along with those who were censored during follow-up (random censoring). The data of this nature give rise to a high level of right censoring, which could skew the results obtained with the Cox model. To understand what occurs in this scenario, a comprehensive analysis via simulation is carried out to investigate these effects, and the results are presented in Chapter 4.

  On the other hand, patients who remain until the end of the study without experiencing the event of interest and that are declared as censored, are called long-term survivors. These patients may be defined as cured (immune or nonsusceptible), if the time to follow-up is long enough. A priori, these patients are indistinguishable from those patients experiencing the event, called uncured (nonimmune or susceptible). A survival curve that does not converge to zero, called improper survival curve, can be an empirical evidence of the presence of immune individuals in the study population. In Chapter 5, we discuss about the validation of sufficient follow-up in the sample of patients with cancer. Estimates of the proportion of cured patients and nonparametric statistical tests to assess the follow-up, and thereby ensure the existence of cured individuals, are thoroughly discussed in section 5.2. Survival analysis using a more general model than the Cox's model is presented in section 5.3. This model composed of a cure fraction and a survival function of uncured patients, is called mixture cure model. The advantages gained by using the mixture cure

model are based on the separation of statistically significant factors for the susceptible and unsusceptible part in the estimation process, better quality of the estimators, consistent results and ease of interpretation of those obtained from the standard models. These results are widely exposed for both, the disease-free and overall survival time.

- *Analysis of veterinary data.* The other main study analyzed in thesis, begins with the discussion made about a sample of 2504 calves, begins with the discussion made about a sample of 2504 calves, which were followed from birth to the first 180 days of life (when weaning takes place). The study was conducted in order to identify the genetic and environmental factors that influence the time to death before weaning. Features of the calf, the cow and the herd at calving, were recorded as factors that might influence survival time. In a preliminary analysis of the data, we found relevant aspects that need to be taken into account in the modeling, such as, heterogeneous herds, invalid data, heavy censoring and missing data. These preliminary results led to a better analysis strategy, helped identify factors that provided little or no information, also allowed us to debug and summarize the characteristics of calves per herd. In Chapter 2, Section 2.3, we begin the discussion with a description of the temporal and spatial mortality, concluding that the herds included in the study, were heterogeneous. Discussions about genetic and environmental factors, as well as missing data in the sample, are fully explained. A survival analysis using the Cox model per herd, was carried out. The results and model validation were thoroughly discussed as well.

  In this study we observed several compositions and various natures of the censoring level (see Table 2.17 and 2.19), mainly due to the characteristics of the cows, as well as of the type of assistance received at calving, and the follow-up type of calves in each herd. This behavior of the censoring, is reflected in the survival curves of Kaplan and Meier, showing an improper type. This was empirical evidence to suppose the presence of calves which are more likely to die within the first 180 days of life than the others. We take into account that the follow-up period is held within a finite time interval, and that beyond this time there is no interest in whether or not the event occurs. In this scenario a survival model defined by specifying a bounded cumulative hazard function is more appropriate for modeling these data. In Chapter 6 we present survival analysis using cure models with an extended risk function, which includes the cure model of proportional hazards. The

analysis is presented by herd, and covariates were included into the model via nonlinear transformation models. The results and discussions are presented, together with estimates of the proportion of survivors calves by herds.

- *Determination of risk factors and the proportion of cure in patients with an SLN biopsy via mixture cure models.* Analysis of the melanoma data via mixture cure models, presented in chapter 5, is carried out using the macro in SAS: PSPMCM. The results show that patients with negative Sentinel Lymph Node status, Clark level I-III, Histopathological of Malignant Melanoma subtype externa-superficial, younger than 46 years, and female, are more likely to be cured, whereas patients with melanoma in head and neck, Breslow level $\geq$ 4mm and ulceration presents, are patients with increased risk of relapse. In particular, patients with Breslow level $\geq$ 4mm are at higher risk for death.

- *Determination of risk factors and the proportion of cure of beef calves up to weaning via extended hazard models.* In Chapter 6, we present a review and general discussion about the extended hazard models (EHM) proposed by Tsodikov (2002). An analysis of the calves data is carried out via extended hazard models by herd. Results are shown in section 6.3.1, and were obtained using the library "NLTM" of the statistical package R. The short and long term effects are determined for each covariates, as well as the immune the proportion of calves per herd. For example in the herd 1, we find that calving month and difficulty at birth is the set of statistically significant factors for the nonsusceptible proportion. Calves born in the period march-august have lower probability of cure than those born in september-february; and the probability of cure is much lower for those that have difficulties at calving for herd 1. For herd 7 the effect of difficulty is different as for herd 1, here only is significative the category strongly assisted. Calves that born from a strongly assisted calving have lower probability of cure that calves from an unassisted calving. Regarding short-term (mortality) effects, we only find statistically significant predictors in herd 7 where the risk of death of calves born to older mothers, hence with a longer reproductive life, is twice the risk of death of calves born to younger mothers.

Methodological

- *Methodological review about the cure models and software.* We present in Chapter 3 a

methodological review about the mixture and nonmixture cure models, from their origin up until the writing of this thesis, which included its formulation, justification, development and applications. The presentation is in chronological order and distinguishes among various approaches related to the topic. We make a thoroughly revised about available software to carry out analysis with a cure model, and at the end of this chapter we presented a description of the more relevant software, which is available.

- *Determination of relative risk under a scenario of heavy right-censoring in the Cox model.* In Chapter 4, section 4.2 we present a methodological proposal to evaluate the effects of the heavy right-censoring on the estimates of relative risk in a proportional hazards model. The proposal assumes a binary covariate as an explanatory variable and a fixed percentage of censoring, which we called the censoring level. The methodology is based on the study of the relationships between the probability of censoring, the probability of success for binary covariate and the relative risk.

- *Recommendations for use of the Cox model in presence of heavy right-censoring.* In chapter 4, section 4.5, the properties and consistency of the estimator of the relative risk are presented in tables, under three censoring levels and different sample sizes. Recommendations for use of the Cox model in presence of heavy censoring are determined. The main recommendation is, if censoring is too heavy, the Cox model should not be used or used cautiously when sample size is smaller than 500.

- *Discussion and application about tests to assessment of sufficient follow-up.* We begin Chapter 5, section 5.2, with a discussion about the justification for using the mixture cure model. First, there must be some empirical evidence to suggest the presence of immune or cured individuals in the population (and estimate the cure rate, possibly in a scenario of heavy censoring). Then it has to be checked whether follow-up was enough to ensure that individuals of the population are actually immune. The discussion includes a review of the statistical tests for this purpose, like its applications to patients with melanoma cancer.

R implementations

- *Assessment of heavy right-censoring in the Cox model by Simulation.* Proposed methodology to evaluate the effect of the censoring level on the estimation of the relative risk in

a proportional hazards model, was implemented as a function in the statistical package R (see Appendix A.2.1), and a study by simulation was performed. The simulation study include 450 scenarios with three censoring levels. The results are presented in Chapter 4, sections 4.3-4.6.

- *Implementation of nonparametric tests for assessment of sufficient follow-up in the case of melanoma data.* Non-parametric tests to determine if follow-up was or was not enough in a sample, were implemented as functions in the statistical package R (see Appendix A.2.2), and applied to the sample of patients with melanoma, the results are described in section 5.2.3.

7.2 Future research

Among the different issues that remain open after the completion of this thesis, we point out six which we plan to develop in the near future.

- In the study of melanoma data some analyses will still be done. With the goal of a better interpretation, the main Medical Doctor of the study, suggested introducing variable Histopathological of Malignant Melanoma (HMM) subtype in three categories: SMM, ALM/LMM and NM. To this end, this recategorized variable, will be included in the models for the disease-free survival time as well as for the overall survival.

- We plan to extend the study and simulations presented in Chapter 4 to two dichotomous variables. In this case the probability of censoring

$$p = [1-\kappa]\exp(-\tau^{\alpha}) + \kappa\exp(-RR\tau^{\alpha}), \tag{7.1}$$

with dichotomous variables x_1 and x_2 had the form

$$p = \exp(-\tau^{\alpha})\kappa_{00} + \exp(-RR_2\tau^{\alpha})\kappa_{01} + \exp(-RR_1\tau^{\alpha})\kappa_{10} + \exp(-RR_1RR_2\tau^{\alpha})\kappa_{11},$$

where $\kappa_{ij} = P[x_1 = i, x_2 = j]$, $RR_1 = \frac{\lambda^{\alpha}_{1x_2}}{\lambda^{\alpha}_{0x_2}}$, $RR_2 = \frac{\lambda^{\alpha}_{x_11}}{\lambda^{\alpha}_{x_10}}$, and $\lambda_{ij} = \exp(-\beta_1(i) - \beta_2(j))$.

- Also we plan to extend the study and simulations presented in Chapter 4 using other distributions with one or two dichotomous variable.

- We plan to study the theoretical properties of the estimator of RR under the scenario described in Chapter 4, for small and large samples, and extend it to the new scenarios proposed above.

- Similar to the simulation study presented in Chapter 4, we plan to extend it, when the survival time T is right-censored by a random variable C within a time interval bounded by τ. In this case, T can be censored by C (random censoring) or τ (fixed censoring) due to the end of the study. Then the probability of censoring p is given by

$$p = P[T > C|T \leq \tau, x] + P[T > \tau|x], \tag{7.2}$$

  where x is a dichotomous variable. Indeed when T follows a Weibull regression model with shape parameter α and scale parameter λ_x and C follows Weibull model with shape parameter α and scale parameter λ, then equation (7.2) has the form

$$p = A + \frac{k\lambda^\alpha + RR * B}{\lambda^\alpha + RR},$$

  where $A = \frac{1-\kappa}{\lambda^\alpha+1}[\lambda^\alpha + \exp(-(\lambda^\alpha + 1)\tau^\alpha)]$, $B = \exp(-(\lambda\tau)^\alpha)[p_\tau - (1-\kappa)\exp(-\tau^\alpha)]$ and $p_\tau = P[T > \tau|x]$ is the same equation (7.1), with $\lambda_0^\alpha = 1$ and $\lambda_1^\alpha = RR$.

- Another extension to the study and simulations presented in Chapter 4 is to simulate scenarios with parameter values close to those encountered in the skin cancer data and in calf mortality data: $RR > 1$ in combination with heavy fixed censoring and moderate random censoring, $RR > 1$ in combination with moderate fixed censoring and heavy random censoring, among others.

- One of the main differences when using the non-mixture model instead of the Cox model is to distinguish the -short or long term-effects that the covariates have on the subpopulations. This flexibility allows an estimating the proportion of the cured population. We plan to

study the convergence properties of the estimators with the Cox model when the true distribution is given by one of the cure rate model.

7.3 Conclusions

This thesis has studied the cure rate model, has reviewed mixture and non-mixture cure models from a classical statistics approach, and has shown the advantages of these models over the standard Cox models. A review about available software to carry out analysis with a cure rate model is presented. A standard assumption in survival analysis is that all individuals will have the event of interest provided the follow-up period is large enough. However, common models might be inappropriate when data contain too much right-censoring. Simulation is used to analyze the effects of heavy right censoring and sample size on the relative risk of the Cox model. Results show that in the presence of censoring levels from 70% to 90%, the Cox model is always suitable if the sample size is larger or equal than 500. The study has verified, that the behavior of the relative risk, in terms of *mse*, is better if the sample is balanced. The thesis has been motivated by two studies that have a high percentage of right-censoring, and where it is likely that there are immune individuals or that the follow-up has not been long enough to see how the entire population or both fail. In this situation, standard methods of survival analysis like the Kaplan-Meier and the Cox model have limitations and can produce biased results.

Mixture cure models are presented, discussed and applied to the melanoma dataset. However, before carrying out an analysis with these models, one has to ensure, whether follow-up was sufficient. To assess whether the follow-up is sufficient or not, nonparametrics tests are presented and implemented in a statistical package R. The main role of these tests is to ensure the presence of immune individuals in the population, and discard any effect only of the censoring mechanism due to follow-up. We apply these nonparametric tests to the data, and the results show that for both times: *disease-free time* and *overall time* the follow-up is sufficient. Therefore, it is valid to assume that immune individuals are actually present in the population, and a mixture cure model is more appropriate to analyze this data. The advantage of this analysis over the standard models of survival analysis, is that the mixture cure model allows the separate modeling, factors effects on the nonsusceptible proportion $[1 - \pi(z|b)]$ (or cure rate) and factors

effects on time survival of susceptible $S(u|\beta, x)$ (or not cured) individuals, like are presented in Table 5.4 and 5.5. In the analysis for *disease-free time*, some of the factors with significant effects on the standard proportional hazards model, such as Sentinel Lymph Node status, Clark and Age (Table 2.7), are now highly significant for the nonsusceptible proportion (Table 5.4) in the mixture cure model. Similarly, for the analysis of *overall time*, Ulceration and Clark (Table 2.12), now statistically significant for the nonsusceptible proportion (Table 5.5). In addition, the mixture cure model incorporates other significant factors, such as Gender, Histopathological of Malignant Melanoma subtype and Localization for *disease-free time*, and Breslow for *overall time*.

In addition to the problem of heavy right-censoring, we face the situation when the interest event can or not occur within a finite time interval, such is the case of veterinary data, where the event is death before ending weaning, where the interval of interest comes from birth to weaning. In these cases a model with a bounded hazard function could be the most appropriate, or any member of the class of nonlinear transformation models like PH-PHC model. This model allows us to jointly estimate the proportion of cure (survival up to weaning) and the effect of different set of covariates for short and long-term on individuals in a heterogeneous population. Then estimations and confidence intervals for the cure rate could be carried out. The results obtained show lower probabilities of survival up to weaning for calves born between March and August and for calves born with assistance for herds 1. Whereas for the herd 7, there is low probability of survival up to weaning for calves born strongly assisted. Furthermore, in the herd 7 the length of productive live of the cow has an influence on the risk of death of the calves, and this short-term effect is influencing the probability of survival up to weaning. Thus, the probability of survival up to weaning of calves born without assistance is significatively higher than the probability of survival up to weaning of calves born with strong assistance.

Bibliography

Aljawadi, B. A. I., M. R. A. Bakar, and N. A. Ibrahim (2011). Non-parametric maximum likelihood estimation of cure fraction for interval survival data. *International Journal of Applied Mathematics and Statistics 21*(J11), 118–130.

Aljawadi, B. A. I., M. R. A. Bakar, N. A. Ibrahim, and M. AlOmari (2013). Parametric maximum likelihood estimation of cure fraction using interval-censored data. *Journal of Advanced Computing 1*, 43–58.

Anscombe, F. J. (1961). Estimating a mixed exponential response law. *Journal of the American Statistical Association 56*(295), 493–502.

Balch, C., J. Mihm, J. Gershenwald, and S.-J. Soong (2010). The revised melanoma staging system and the impact of mitotic rate. *The Melanoma Letter 28*(3), 1–6.

Bender, R., T. Augustin, and M. Blettner (2005). Generating survival times to simulate Cox proportional hazards models. *Statist. Med. 24*, 1713–1723.

Berkson, J. and R. P. Gage (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association 47*(259), 501–515.

Betensky, R. A. and D. A. Schoenfeld (2001). Nonparametric estimation in a cure model with random cure times. *Biometrics 57*(2), 282–286.

Boag, J. M. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal Royal Statistical Society B 11*, 15–53.

Breslow, N. (1970). A generalized Kruskal-Wallis test for comparing k samples subject to unequal patterns of censorship. *Biometrika 57*(3), 579–594.

Broët, P., Y. D. Rycke, P. Tubert-Bitter, J. Lellouch, B. Asselain, and T. Moreau (2001). A semiparametric approach for the two-sample comparison of survival times with long-term survivors. *Biometrics 57*, 844–852.

Chen, M.-H. and J. G. Ibrahim (2001). Maximum likelihood methods for cure rate models with missing covariates. *Biometrics 57*, 43–52.

Cole, R. A. and J. W. Gunther (1995). Separating the likelihood and timing of bank failure. *Journal of Banking & Finance 19*, 1073–1089.

Corbière, F. and P. Joly (2007). A sas macro for parametric and semiparametric mixture cure models. *Computer Methods and Programs in Biomedicine 85*, 173–180.

De Angelis, R., R. Capocaccia, T. Hakulinen, B. Soderman, and A. Verdecchia (1999). Mixture models for cancer survival analysis: Application to population-based data with covariates. *Statist. Med. 18*, 441–454.

Dikta, G. (2014). Asymptotically efficient estimation under semi-parametric random censorship models. *Journal of Multivariate Analysis 124*, 10–24.

Farewell, V. T. (1977). A model for a binary variable with time-censored observations. *Biometrika 64*(1), 43–46.

Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics 38*, 1041–1046.

Farewell, V. T. (1986). Mixture models in survival analysis: Are they worth the risk? *Canadian Journal of Statistics 14*, 257–262.

Ghitany, M. E., R. A. Maller, and S. Zhou (1994). Exponential mixture models with long-term survivors and covariates. *Journal of Multivariate Analysis 49*, 218–241.

Gibbons, J. (1985). *Nonparametric statistical inference*. NY.

Gieser, P. W., M. N. Chang, P. V. Rao, J. J. Shuster, and J. Pullen (1998). Modelling cure rates using the Gompertz model with covariate information. *Statists. in Med. 17*, 831–839.

Goyache, F., J. P. Gutiérrez, I. Alvarez, I. Fernández, L. J. Royo, and E. Gómez (2003). Genetic analysis of calf survival at different preweaning ages in beef cattle. *Livest. Prod. Sci 83*, 13–20.

Groggel, D., R. Schaefer, and J. Skillings (1989). Effects of withdrawals on tests involving censored data in toxicology experiments. *American Society for Testing and Materials: Aquatic Toxicology and Environmental Fate 11*, 321–338.

Hanin, L. G., M. Zaider, and A. Y. Yakovlev (2001). Distribution of the number of clonogens surviving fractionated radiotherapy: a long-standing problem revisited. *Int. J. Radiat. Biol. 77*, 205–213.

Hanson, T., E. J. Bedrick, W. O. Johnson, and M. C. Thurmond (2003). A mixture model for bovine abortion and foetal survival. *Statist. Med. 22*, 1725–1739.

Harrington, D. P. and T. R. Fleming (1982). A class of rank test procedures for censored survival data. *Biometrika 69*, 553–566.

Ibrahim, J. G., M.-H. Chen, and D. Sinha (2001). *Bayesian Survival Analysis*. Springer.

Kim, S., M.-H. Chen, and D. K. Dey (2011). A new threshold regression model for survival data with a cure fraction. *Lifetime Data Analysis 17*, 101–122.

Kim, S., Y. Xi, and M.-H. Chen (2009). A new latent cure rate marker model for survival data. *The Annals of Applied Statistics 3*(3), 1124–1146.

Klebanov, L. B. and A. Y. Yakovlev (2007). A new approach to testing for sufficient follow-up in cure-rate analysis. *Journal of Statistical Planning and Inference 137*(11), 3557–3569.

Kuk, A. Y. C. and C.-H. Chen (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika 79*(3), 531–541.

Lagakos, S. W. (1979). General right censoring and its impact on the analysis of survival data. *Biometrics 35*, 139–156.

Lambert, P. C. (2007). Modeling of the cure fraction in survival studies. *The Stata Journal 7*(3), 1–25.

Lambert, P. C., P. W. Dickman, C. L. Weston, and J. R. Thompson (2010). Estimating the cure fraction in population-based cancer studies by using finite mixture models. *Applied Statistics. 59*(1), 35–55.

Lambert, P. C., J. R. Thompson, C. L. Weston, and P. W. Dickman (2007). Estimating and modeling the cure fraction in population-based cancer survival analysis. *Biostatistics 8*(3), 576–594.

Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data*. Wiley.

Maller, R. and S. Zhou (1992). Estimating the proportion of immunes in a censored sample. *Biometrika 79*(4), 731–739.

Maller, R. and X. Zhou (1996). *Survival Analysis with Long-Term Survivors*. Wiley.

Mavromaras, K. G. and C. D. Orme (2004). Temporary layoffs and split population models. *Journal of Applied Econometrics* (19), 49–67.

Meeker, W. Q. (1987). Limited failure population life tests: Application to integrated circuit reliability. *Technometrics 29*(1), 51–65.

Othus, M., B. Barlogie, M. L. LeBlanc, and J. J. Crowley (2012). Cure models as a useful statistical tool for analyzing survival. *Clinical Cancer Research 18*(4), 1–6.

Peng, Y. and K. B. G. Dear (2000). A nonparametric mixture model for cure rate estimation. *Biometrics 56*, 237–243.

Peng, Y., K. B. G. Dear, and J. W. Denham (1998). A generalized F mixture model for cure rate estimation. *Statist. Med. 17*, 813–830.

Peng, Y. and J. Xu (2012). An extended cure model and model selection. *Lifetime Data Analysis 18*(2), 215–233.

Pierce, D. A., W. H. Stewart, and K. J. Kopecky (1979). Distribution-free regression analysis of grouped survival data. *Biometrics 35*, 785–793.

Schmidt, P. and A. D. Witte (1989). Predicting criminal recidivism using split population survival time models. *Journal of Econometrics 40*, 141–159.

Schoenfeld, D. (1982). Partial residuals for the proportional hazard regression model. *Biometrika 69*, 239–241.

Sposto, R. (2002). Cure model analysis in cancer: an application to data from the children's cancer group. *Statist. Med. 21*, 293–312.

Swetter, S. M., J. C. Boldrick, S. Y. Jung, B. M. Egbert, and J. D. Harvell (2005). Increasing incidence of lentigo maligna melanoma subtypes: Northern california and national trends 1990-2000. *Journal of Investigative Dermatology* (125), 685–691.

Sy, J. P. and J. M. G. Taylor (2000). Estimation in a Cox proportional hazards cure models. *Biometrics 56*, 227–236.

Tarrés, J., J. Casellas, and J. Piedrafita (2005). Genetic and environmental factors influencing mortality up to weaning of Bruna dels pirineus beef calves in mountain areas. a survival analysis. *Animal Science. 83*, 543–551.

Taylor, J. (1995). Semi-parametric estimation in failure time mixture models. *Biometrics 51*, 899–907.

Tsiatis, H. P. and M. Davidian (1998). Estimating the parameters in the Cox model when covariate variables are measured with error. *Biometrics 54*, 1407–1419.

Tsodikov, A. (1998). A proportional hazards model taking account of long-term survivors. *Biometrics 54*(4), 1508–1516.

Tsodikov, A. (2002). Semi-parametric models of long- and short-term survival: an application to the analysis of breast cancer survival in Utah by age and stage. *Statist. Med. 21*, 895–920.

Tsodikov, A. (2003). Semiparametric models: a generalized self-consistency approach. *Journal Royal Statistical Society B 65*(3), 759–774.

Wellner, J. A. (1985). A heavy censoring limit theorem for the product limit estimator. *The Annals of Statistics 13*(1), 150–162.

Williams, J. S. and S. W. Lagakos (1977). Models for censored survival analysis: constant-sum and variable-sum models. *Biometrika 64*(2), 215–224.

Yamaguchi, K. (1992). Accelerated failure time regression model with a regression model of surviving fraction: An aplication to the analysis of permanent employment in japan. *Journal of the American Statistical Association 87*, 284–292.

Yin, G. and J. G. Ibrahim (2005a). Cure rate models: a unified approach. *The Canadian Journal of Statistics 33*(4), 559–570.

Yin, G. and J. G. Ibrahim (2005b). A general class of bayesian survival models with zero and nonzero cure fractions. *Biometrics 61*, 403–412.

Yu, B., R. C. Tiwari, K. A. Cronin, C. McDonald, and E. J. Feuer (2005). Cansurv: A windows program for population-based cancer survival analysis. *Computer Methods and Programs in Biomedicine 80*, 195–203.

Zukang, Z. (1997). On the uniform consistency of the kaplan-meier estimator under heavy censoring. *Appl. Math. 12B*(1), 33–38.

Appendix A

Algorithms and programs

The Appendix contains a brief description of R codes (including libraries and functions) used in the thesis.

A.1 EM Algorithm

The usual process to fill the lacking information is the traditional algorithm EM. This algorithm makes wear of the log-likelihood of the complete data, namely the log-likelihood of the data censored like those are not censored. The algorithm was made in two steps: step E that results on apply the expectation at the log-likelihood and, the step M that results of maximize this expectation.

Example: *Weibull case.*
Supposed that $T|Y=1 \sim Weibull(\alpha,\gamma)$, with survival function $S(t|Y=1)=e^{-\alpha t^{\gamma}}$ and hazard function $h(t|Y=1)=\alpha\gamma t^{\gamma-1}$, and with $p=P[Y=1]$.

The likelihood of the complete data assuming independent, noninformative, random censore model, and independence of the susceptibility, is given by

$$L(p,\alpha,\gamma;\underline{y})=\prod_{i=1}^{n} p^{y_i}[1-p]^{1-y_i}[\alpha\gamma t_i^{\gamma-1}]^{y_i}e^{-\alpha t_i^{\gamma} y_i}.$$

Step E: Let $O=\{y_{i's}$ observadas, $(t_i,\delta_i):i=1:n\}$ and $\theta=(p,\alpha,\gamma)$. Making $Q(\theta^{(m)})=$

$E[\log[L(p,\alpha,\gamma;\underline{y}|\theta^{(m)},O)]]$, we have that $Q(\theta^{(m)}) = Q_1(p) + Q_2(\alpha,\gamma)$ where

$$Q_1(p) = n\ln[1-p] + \ln[\frac{p}{1-p}]\sum_{i=1}^{n} E\left[Y_i|\theta^{(m)},O\right],$$

$$\begin{aligned} Q_2(\alpha,\gamma) &= \ln(\alpha\gamma)\sum_{i=1}^{n} E[Y_i|\theta^{(m)},O] + (\gamma-1)\sum_{i=1}^{n}[\ln(t_i)]\,E[Y_i|\theta^{(m)},O] \\ &\quad -\alpha\sum_{i=1}^{n} t_i^{\gamma} E[Y_i|\theta^{(m)},O]. \end{aligned}$$

$$E[Y_i|\theta^{(m)},O] = \delta_i + (1-\delta_i)\frac{pe^{-\alpha t_i^{\gamma}}}{[1-p]+pe^{-\alpha t_i^{\gamma}}}$$

Step M: We maximize the likelihood considering the lacking dates as if those were observed, and with $g_i^{(m)} = E[Y_i|\theta^{(m)},O]$.

$$\begin{aligned} Q(\theta|g^{(m)}) &= \left[n\ln[1-p] + \ln[\frac{p}{1-p}]\sum_{i=1}^{n} g_i^{(m)}\right] \\ &\quad + \left[\ln(\alpha\gamma)\sum_{i=1}^{n} g_i^{(m)} + (\gamma-1)\sum_{i=1}^{n}[\ln(t_i)]\,g_i^{(m)} - \alpha\sum_{i=1}^{n} t_i^{\gamma} g_i^{(m)}\right] \end{aligned}$$

The maximization method leads to a double maximization :

$$\begin{aligned} \max_{\theta} Q(\theta|g^{(m)}) &= \max_{p}\left[n\ln[1-p] + \ln[\frac{p}{1-p}]\sum_{i=1}^{n} g_i^{(m)}\right] \\ &\quad + \max_{(\alpha,\gamma)}\left[\ln(\alpha\gamma)\sum_{i=1}^{n} g_i^{(m)} + (\gamma-1)\sum_{i=1}^{n}[\ln(t_i)]\,g_i^{(m)} - \alpha\sum_{i=1}^{n} t_i^{\gamma} g_i^{(m)}\right]. \end{aligned}$$

The exponential case results when $\gamma = 1$ in the Weibull distribution. On this case the maximization of the EM algorithm leads a double simple maximization:

$$\begin{aligned} \max_{\theta} Q(\theta|g^{(m)}) &= \max_{p}\left[n\ln[1-p] + \ln[\frac{p}{1-p}]\sum_{i=1}^{n} g_i^{(m)}\right] \\ &\quad + \max_{\alpha}\left[\ln(\alpha)\sum_{i=1}^{n} g_i^{(m)} - \alpha\sum_{i=1}^{n} t_i g_i^{(m)}\right], \end{aligned}$$

where

$$g_i^{(m)} = \delta_i + (1-\delta_i)\frac{pe^{-\alpha t_i}}{[1-p]+pe^{-\alpha t_i}}, \qquad \theta = (p,\alpha).$$

A.2 R Program

A.2.1 Heavy censoring

```
wcf= function(alf,px,tau,rr,m,n)
{
#Bias, variance and MSE of the estimator rr.

#alf is the shape parameter of the distribution of T.
#rr  is the relative risk of Cox's model
#tau is the value of censoring level
# m  is the number of replicas
# n  is the sample size (less than or equal to 2500)
# px is the probability of x = 1.



library(splines)
library(survival)

      lan1= (rr)^(1/alf)
        n1= n*px
        n0= n*(1-px)
         x= c(rep(1,n1),rep(0,n0))

#initialization of results vector
    rep= matrix(numeric(3*m), nrow=m,ncol=3)

#Seed: by if someone wants to redo the study
set.seed(23571317)

       for(j in 1:m)
       {
```

```
       #Simulation of the failure times Weibull(alf, lan1)
        v1= runif(n1,0,1)
        t1= ( (-log(v1))^(1/alf) )/lan1

       #Simulation of the failure times Weibull(alf, lan0= 1)
        v0= runif(n0,0,1)
        t0= (-log(v0))^(1/alf)

         t= c(t1,t0)

       #Survival Data (times, cens)
             cens= 0*t
              obs= which(t<=tau)
        cens[obs]=1
            times= pmin(t,tau)

         fit.cox= coxph(Surv(times,cens)~as.factor(x),method="br")
               k= summary(fit.cox)$conf.int
         rep[j,]= k[c(1,3,4)]
         }

#As in some iterations the procedure does not converge
ojo =which(rep[,3]<1/0)
rep =rep[ojo,]
  m1 =length(ojo)


#Estimation of bias and relative bias of rr
 s.rr= mean(rep[,1])-rr
sr.rr= (s.rr/rr)*100

                   #Variance of rr
```

```
 v.rr= var(rep[,1])


                   #mean square error of rr
  mse.rr= v.rr + ((s.rr)^2)


                   #Coverage of rr
 li= rep[,2]
 ls= rep[,3]
 nc= which(li < rr & rr < ls)

                   #Coverage level for rr
p.rr=length(nc)/m1

tab=c(s.rr, v.rr, mse.rr, sr.rr, p.rr, m1)
}


  be= c(0.2, 0.4, 0.6, 0.8, 1)
   c= length(be)
vtau= c(0.6502,  0.5273,  0.4509,  0.3971,  0.3566,
        0.3923,  0.3255,  0.2809,  0.2482,  0.2231,
        0.1799,  0.1519,  0.1321,  0.1171,  0.1053)


tau= matrix(vtau,nrow=5,ncol=3)

tf=matrix(numeric(50*3*6),nrow=50,ncol=3*6)
```

```
for(h in 1:c)
{
va= tau[h,]
 r= length(va)
vn= c(50,100,200,300,400,500,1000,1500,2000,2500)
 k= length(vn)

tabla=matrix(numeric(6*k*r),nrow=k, ncol=r*6)

       for(j in 1:r)
       {
       ta=matrix(numeric(6*k),nrow=k, ncol=6)
             for(i in 1:k)
             {
              ta[i,]=wcf(1,0.5,va[j],be[h],1000,vn[i])
             }
        tabla[,j]= ta[,1]
      tabla[,j+3]= ta[,2]
      tabla[,j+6]= ta[,3]
      tabla[,j+9]= ta[,4]
     tabla[,j+12]= ta[,5]
     tabla[,j+15]= ta[,6]

      tabla=round(tabla,4)
      }

tf[(1+10*(h-1)):(10*h),]=tabla
tf
}
tf
```

A.2.2 Test for assessment sufficient follow-up

```
                  ----------------------------
                           $\alpha$-test
                  ----------------------------
----------------------------
#Disease free survival time
----------------------------
   m = read.table('melanoma_new.txt', header=T)
times= m$seg_reci
 cens= m$recid

----------------------------
#Overall survival time
----------------------------
   m = read.table('melanoma_new.txt', header=T)
times= m$seg
 cens= m$cens



  obs1= which(cens==1)
timesf= times[obs1]
  Tnes= max(timesf)
    Tn= max(times)
    tf= which(timesf>(2*Tnes-Tn) & timesf<=Tnes)

 k= length(times[tf])
 k
 n= length(times)
alfn=(1-(k/n))^n
alfn
```

```
                    ---------------------------------
                              $\triangle$-test
                    ---------------------------------
----------------------------
#Disease free survival time
----------------------------
    m= read.table('melanoma_new.txt', header=T)
times= m$seg_reci
 cens= m$recid

----------------------------
#Overall survival time
----------------------------
   m = read.table('melanoma_new.txt', header=T)
times= m$seg
 cens= m$cens



   Tn= max(times)
library(splines)
library(survival)

     km= summary(survfit(Surv(times,cens)~1))
     tf= km$time
    stf= km$surv

#m number of partitions of the vector t
an= function(Tn, m, tf, stf)
 {
```

```
xmin= tf[1]+.5
xmax= Tn

 t= seq(xmin, xmax, by = ((xmax-xmin)/(m-1)) )
st= 0*t

 for (i in 1:length(t))
  {
  obs= which( tf<t[i] )
  pos= length(tf[obs])
st[i]= stf[pos]
  }

rnt= -(1/t)*log(st)
zhi= exp( -Tn*rnt )
  n= length(tf)

cotainf= ( 1 -( (1-stf[n])/(1-zhi) )  )

values= cbind(t,cotainf)
 }



  tabt= an(Tn, 1000, tf, stf)
    nr= nrow(tabt)

      for (i in 1:nrow(tabt))
       {
       if (tabt[i,2]<0)
           {
            tabt[i,2]=0
```

```
            }
          else if (tabt[i,2]>=0)
            {
             tabt[i,2]= tabt[i,2]
            }
          }

 max_an= max(tabt[,2])
    obs= which(tabt[,2]==max_an)
  tabt0= tabt[obs,]
plot(tabt[,1],tabt[,2],type='l',xlab='t',ylab='a_n(t)')

     t0= tabt0[1]
   obs1= which(tf< t0)
   snt0= (stf[max(obs1)])



ant= function(times, cens, t0)
 {
    n= length(times)
  obs= which(times<t0)
   di= cens[obs]

  sum= 0
  for (i in 1:length(di))
   {
   if (di[i]==1)
     {
     sum = sum + (   1/( (n-i)*(n-(i-1)) )  )
     }
    else if (di[i]==0)
     {
```

```
      sum = sum
     }
   }

antf= ( n*sum  )

 values= antf

}

ant0= ant(times,cens,t0)

   n= length(times)
Dalf= 1.36/sqrt(n)
ter1= min(stf)
ter2= exp(- (Tn/t0)*(-log(snt0)) )
ter3= ( 1+(Tn/t0) )*( Dalf/sqrt(n) )*snt0*( 1+ant0 )

#reject H0: S(T)=S_0(T) if (ter1-ter2-ter3)>0
rr=ter1-ter2-ter3
rr
```