Hedge funds and Survival analysis

by

Blanche Nadege Nhogue Wabo

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
In partial fulfillment of the requirements
For the M.A.Sc. degree in
Mathematics and Satistics

Department of mathematics
Faculty of Science
University of Ottawa

Abstract

Using data from Hedge Fund Research, Inc. (HFR), this study adapts and expands on existing methods in survival analysis in an attempt to investigate whether hedge funds mortality can be predicted on the basis of certain hedge funds characteristics. The main idea is to determine the characteristics which contribute the most to the survival and failure probabilities of hedge funds and interpret them. We establish hazard models with time-independent covariates, as well as time-varying covariates to interpret the selected hedge funds characteristics. Our results show that size, age, performance, strategy, annual audit, fund offshore and fund denomination are the characteristics that best explain hedge fund failure. We find that 1% increase in performance decreases the hazard by 3.3%, the small size and the less than 5 years old hedge funds are the most likely to die and the event-driven strategy is the best to use as compare to others. The risk of death is 0.668 times lower for funds who indicated that an annual audit is performed as compared to the funds who did not indicated that an annual audit is performed. The risk of death for the offshore hedge funds is 1.059 times higher than the non-offshore hedge funds.

Acknowledgements

I take this opportunity to thank my husband Dr Wilfred Kepseu who often played the role of mother and father to our children, this allowed me to work faster and performed in my thesis. Words like thankfulness or gratefulness fall short in expressing my gratitude for him.

I am deeply indebted to my supervisor Dr Raluca Balan for her invaluable advice, precise organization and patient guidance and my co-supervisor Dr Pierre-Jérôme Bergeron who introduced me in survival analysis and guided me through. This thesis could not have been written without their constant help and support.

Last but not least, I greatly thank the Lord God Almighty for giving me His grace filled because without Him I could not manage all my occupations with success.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Hedge funds are private, unregulated investment pools for wealthy individuals and institutions which have grown significantly in size and influence in recent years. There are now about $8,000$ funds with a total of over $1 trillion under management (see [1, 2]). Additionally, pension funds, endowments, and individuals have invested in hedge funds to diversify their portfolios. Furthermore, the proliferation of multi-strategy funds and funds of hedge funds has allowed investors to diversify within the hedge funds industry. The goal of most hedge funds is to maximize return on investment (see [3]). Hedge funds always take higher risk to get high-return operations and there are always market movements that affect their performance, either directly or indirectly (via the impact on their underlying investments). Therefore, many of them eventually close, or "die off" as has been seen during the different financial crises of the $1990s$ and $2000s$ (see [4]). The estimation of failure rate of hedge funds ranges from 7% to 10% each year (see [5]). Since estimates of the number of hedge funds range from 7,000 to 9,000, this suggests that several hundred funds cease operations each year.

The survival time of hedge funds is of particular interest to both investors and academics (see Ref [6, 7, 8, 9, 10, 11, 12]), as the majority tend to be short-lived. Many studies have investigated the factors impacting hedge funds lifetime. These factors vary between the various hedge funds strategies which are increasing in the hedge funds industry. Some of these factors include the performance obtained, the assets reported, the reason for liquidation and many others characteristics. The aim of this research is to adapt and expand on existing methods in survival analysis to take into account the various features of hedge funds data. The first main issue is that the database that we examined, contained left-truncated data, that is, funds that were created long be-

fore reporting their first performance in the database, and as such are naturally more successful than hedge funds entering the database at inception time and reporting performance. Secondly, we also had to deal with censored data which appears when the fund disappears from the alive database, it enters the graveyard with a death reason, and the last performance date is equivalent to the death date of the fund. This creates a survivorship bias, that needs to be taken into account (see [13]). The main attribute of survival analysis methods is that they account for truncated and censored lifetimes. Thus to reduce possible modeling bias from our data, several models, both parametric and nonparametric are built. As with linear and logistic regression modelling, the goal of survival analysis is to obtain some measure of the effect that describes the relationship between a predictor variable of interest and time to failure, after adjusting for a subset of significant variables selected. In linear regression modelling, the measure of effect is usually the regression coefficient. In logistic regression, the measure of effect is an odd ratio. In survival analysis, the measure of effect is the hazards ratio (HR). So in the regression context, the approach we take used the proportional hazard model for the time independent predictor of failure times of hedge funds, the extended proportional hazard model for the time varying predictor and mixed proportional model that incorporate both fixed and time varying covariates. The mathematical theory to validate these models is also investigated, to assess the validity and appropriateness of the proposed models.

This dissertation is structured as follows; Chapter 2 presents the hedge fund industry in general. It provides detailed insight into the industry and outlines the categorization of hedge funds. This is followed by a concise account of the historical evolution of hedge funds. Chapter 3 introduces the main concepts and survival distributions used in survival analysis, the definition of the Kaplan-Meier survival curves, parametric models, proportional hazard (PH) model, the estimation and assumptions in the Cox PH model. Model checking using residuals is also described. Chapter 4 discusses the data selection, presents features of hedge funds lifetimes and provides some statistical description. Chapter 5 investigates the relationship between hedge funds characteristics and their survival using various survival models. Some closing remarks and discussion points are offered in Chapter 6.

Chapter 2

Hedge Funds

Hedge funds are the most popular private investment associations in global markets nowadays. They constitute a complex and heterogeneous system and thus, for nearly twenty years, issues related to their operation and their impact on markets and more generally on financial systems have become an increasingly important field of study [1, 6, 14, 15].

In this chapter, we will present the hedge fund industry by considering three aspects. First, we will examine different strategies that define them, then we will discuss the specific characteristics of alternative investment hedge funds, and finally we will look at the characteristics of this industry and their key issues.

2.1 Background

2.1.1 Definition

The definition of a hedge fund is rather ambiguous in the literature because various sources (such as Securities and Exchange Commission(SEC), Financial Services Authority(FSA) or Dodd-Frank Act), which are in some way the regulators and supervisors of financial markets, have each their own criteria for defining a hedge fund. According to FSA, hedge funds are collective investment schemes which are not subject to the restrictions usually applied to large public investment funds, including diversification and marketability of financial assets [16]. Therefore they have the opportunity to acquire illiquid and/or complex assets. They seek investment opportunities in all directions, not only traditional financial markets (stocks and bonds) and their derivatives, but also on raw materials (commodities), works of art, finance of films and all kinds of unusual

investments [16]. They are also speculative investment vehicles designed to exploit superior information held by their managers. Information-based trading requires that the information motivating the trade must be kept secret, at least until after all profitable trades are executed. A hedge fund manager will only be willing to reveal the full details of his strategy after he has decided that it is no longer worth pursuing [1].

2.1.2 Origin

The term *hedge fund* was introduced for the first time in 1966 by Loomis [17], and the first investment qualified as a hedge fund was founded by Alfred Winslow Jones in 1949 [16], an Australian-American sociologist and financial magazine editor who decided to try his hand at investing. His unique investment strategy was to buy stocks with half his investors money, and short sell with the other half. Short selling a stock means betting that it will go down in value rather than up. By buying some stocks while simultaneously short selling others, Jones was in effect hedging his bets (hence the name). If the entire market experienced a downturn, Jones could at least recover some of his losses or ideally turn a profit. With hedge funds, the skill and luck of the manager became the only determining factor in how much money could be made, irrespective of the state of the market [18]. However, other funds that could have been described as hedge funds existed long before, but the Jones' fund combined in original ways the structure of private fund, and investment strategy using leverage and short selling, and performance commissions. These main features have therefore allowed to qualify it as a hedge fund. For more details see [19].

Hedge funds have been developed with significant growth over the last ten years, and since 2002 stocks and trading volumes have increased substantially. It is estimated, as of 2012, that the number of hedge funds is 10,000 and their assets are nearly 1740 billion dollars. The majority of the hedge funds in the world are American and British. At the end of 2006, the 200 largest hedge funds accounted for three-quarters of assets under management. The hedge funds began to increase in assets in the early 1990s, contributing to the surge in 1996 and 2000. After the crises of the 2000s, the number of hedge funds declined, with significant sales. The growth of the hedge fund industry is resulting from the sharp decline in equity markets after the bursting of the Internet bubble of the dot-coms (see [20]).

2.2 Functioning of hedge funds

2.2.1 Strategies

Hedge funds are heterogeneous and can be classified into alternative strategies with specific characteristics, in terms of style, assets used, returns, risk profile, liquidity. Each manager has his own way to implement his strategy, and for the same strategy, the type of implementation for two different hedge funds can be completely different. Many managers adjust their investment process to market conditions, without restraining themselves to a single strategy. There is no consensus on a formal system of classifying hedge funds. Hedge funds can be grouped according to whether their investment approach is systematic or discretionary, or according to their historical roots or even to the geographical location of the assets they trade in. For example, some funds may be focused on European assets while others may be limited to emerging markets. A hedge fund that has a mandate to invest in any country is labeled as a global fund [19]. Hedge Fund Research (HFR), one of the main hedge fund databases, lists 30 separate strategies (with some overlap between them, see [21]). Another widely used database, TASS Research, separates hedge funds into 17 strategy types (see [19]). Table 2.1 compares the strategy classification used by four of the largest index providers.

Note that the world "security" used in this section means a financial instrument that represents an ownership position in a publicly-traded corporation (stock), a creditor relationship with governmental body or a corporation (bond), or rights to ownership as represented by an option. A security is a fungible, negotiable financial instrument that represents some type of financial value. The company or entity that issues the security is known as the issuer (see [22]).

Hereinafter, we will resume some specific hedge fund strategies, grouping them under four broad themes as in Table 2.1: long/short, event-driven, tactical trading, and relative value. We will also discuss the fund to fund strategies. They are not mentioned in Table 2.1, but they constitute a widespread method of hedge fund investment. In fact, the strategy classification is more appropriately applied at the individual fund level.

Event-driven strategies

Event-driven strategies concern situations in which the underlying investment opportunity and risk are associated with an event (see [23]). An event-driven investment strategy finds investment opportunities in corporate transactional events such as consol-

Index Provider:		CSFB/ *Tremont*	MSCI	*Standard and Poors*	*Hedge Fund Research*
STRATEGY CLASS	SPECIFIC STRATEGY				
Event-Driven		X	X	X	X
	Event-Driven	X	X		X
	Event-Driven multistrategy	X			
	Merger/Risk arbitrage	X	X		X
	Distressed	X		X	X
	Special situation			X	
Relative value		X	X	X	X
	Arbitrage		X		
	Statistical Arbitrage		X		
	Specialist Trading		X		
	Convertible Arbitrage	X		X	X
	Relative Value Arbitrage				X
Long/Short		X	X	X	X
	Long/Short Equity	X			
	Dedicated Short-sellers	X			
	Equity Market Neutral	X		X	X
	Equity Hedge				X
	Long bias		X		
	No bias		X		
	Short bias		X		
	Variable Bias		X		
Tactical		X		X	X
	Global Macro	X		X	X
	Managed Futures	X		X	
	Equity/Long			X	
Location		X	X		
	Developed Markets		X		
	Emerging Markets	X	X		
	Global		X		

Table 2.1: Classification of hedge funds strategies adapted from [1]

idations, acquisitions, recapitalisation, bankruptcies, and liquidations. Managers employ such a strategy to capitalize on valuation inconsistencies in the market before or after such events, and take a position based on the predicted movement of these security or securities.

Relative value strategies

Relative value arbitrage strategies take advantage of relative discrepancies in price between securities. The price discrepancy can occur due to mispricing of securities compared to related securities, the underlying security or the overall market. Hedge fund managers can use various types of analysis to identify price discrepancies in securities, including mathematical analysis, technical analysis (security analysis methodology for forecasting the direction of prices through the study of past market data, primarily price and volume) or fundamental techniques (analysis of financial statements and health). Relative value is often used as a synonym for market neutral, as strategies in this category typically have very little or no directional market exposure to the market as a whole (see [23]).

Long/Short strategies

Long/short strategies exploit the ability of hedge fund managers to freely short equities, an opportunity not available to most portfolio managers operating under traditional investment mandates. Long/short strategies separate individual stock risk from market risk (see [1]).

Tactical Trading Strategies

Tactical trading strategies attempt to profit by forecasting the overall direction of the market or a market component. The payoffs of hedge funds specializing in this area depend on how well a hedge fund manager can forecast price movements and predict the exact timing of these movements (see [1]).

Fund of Funds Strategies

Fund of funds strategies invest with multiple managers through funds or managed accounts. The strategy designs a diversified portfolio of managers with the objective of significantly lowering the risk (volatility) of investing with an individual manager. The

Fund of funds manager has discretion in choosing which strategies to invest in for the portfolio. A manager may allocate funds to numerous managers within a single strategy, or to numerous managers in multiple strategies. The minimum investment in a Fund of funds may be lower than an investment in an individual hedge fund or managed account. The investor has the advantage of diversification among managers and styles with significantly less capital than investing with separate managers (see [21]).

2.2.2 Organizational Structure of Hedge Fund

The summary below describes briefly a very common method used to structure the hedge fund and its management company. There are many others and just as hedge funds are creative with their investment strategies, they can also be very creative with their organizational structure. The typical hedge fund structure is really a two-tiered organization (see Figure 2.1).

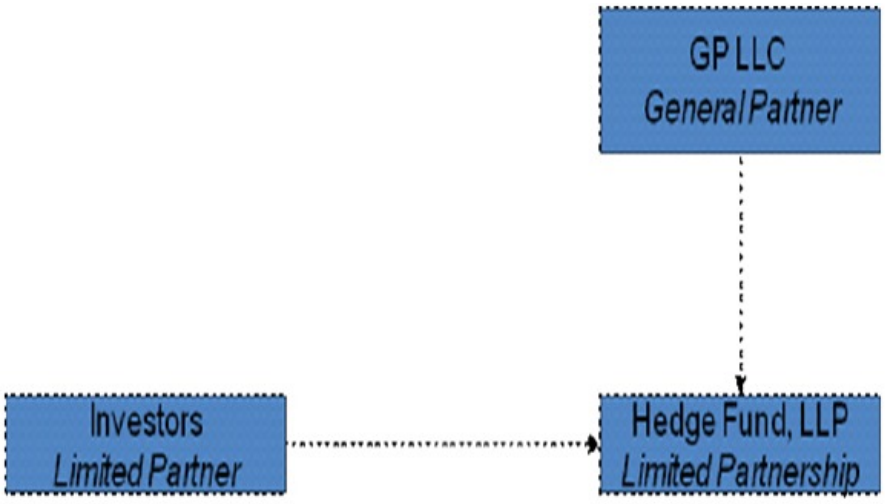

Figure 2.1: General/limited partnership describing the most common hedge funds structure

The most common structure for hedge funds is the *general/limited partnership model* (See Figure 2.1). In this structure, the general partner assumes responsibility for the operations of the fund, while limited partners can make investments into the partnership and are liable only for their paid-in amounts. The second component of the two-tiered organization is the structure of the general partnership. The typical structure used for the general partner is a limited liability company (LLC). The general partner's responsibility

is to market and manage the fund, and perform any functions necessary in the normal course of business, including hiring a fund manager (oftentimes a related company) and managing the administration of the fund's operations (see [22]).

2.2.3 Fee Structure

The hedge fund fee structure is one of the main reason why talented money managers decide to open their own hedge funds to begin with. Not only are the fees paid by investors higher than they are for mutual funds, but they also include some additional fees that mutual funds do not even charge. We have:

- The *management fee* is a periodic payment that is paid by investors in a pooled investment fund to the fund's investment adviser for investment and portfolio management services. The fee pays other people to select which securities a particular investor's money (along with that of the other investors in the fund) is invested into, to do all the paperwork needed and to provide information about the fund's holdings and performance (see [22]).

- The *incentive fee* is usually in place to tie a manager's compensation to their level of performance, more specifically to their level of financial return. However, such fees can sometimes lead to increased levels of risk taking, as managers attempt to increase incentive levels through riskier ventures than outlined in a fund's prospectus (see [22]).

2.2.4 Performance measurement of hedge funds

There are many performance measurement techniques for analyzing hedge funds. The absolute returns, Sharpe Ratio and benchmarks are among the most used techniques:

- *Absolute returns* are the returns that an asset achieves over a certain period of time. A hedge fund must be evaluated based on them, but those returns also need to be consistent with the fund's strategy. There are funds that employ strategies that generate very consistent returns over time with limited volatility (see [22]).

- *Sharpe ratio* is one metric that is widely used in the hedge fund world. It measures the amount of return adjusted for each level of risk taken. It is calculated by subtracting the risk-free rate from annualized returns and dividing the result by the standard deviation of the returns.

- *Benchmark* is a standard measure against which the performance of a security, hedge fund, mutual fund or investment manager can be measured. Generally, broad market, market-segment stock and bond indexes are used for this purpose.

2.2.5 Risk Management

Hedge funds are often mistaken to be very similar in risk to other types of investments, and although they are often measured through the same types of quantitative metrics, hedge funds have qualitative risks that make them unique to evaluate and analyze. The most common risk metrics used in hedge fund analysis are standard deviation, downside capture, drawdown, qualitative factors and leverage. In this work, we will consider only the leverage.

Leverage is a general term for any technique to multiply gains and losses. Common ways to attain leverage are borrowing money, buying fixed assets and using derivatives. *Leverage* is a measure that often gets overlooked, yet it is one of the main reasons why hedge funds incur huge losses. As leverage increases, any negative effect in returns gets magnified and worse, it causes the fund to sell assets at steep discounts to cover margin calls. Leverage has been the primary reason why hedge funds like LTCM and Amaranth have gone out of business. Each of these funds may have had huge losses due to the investments made, but chances are these funds could have survived had it not been for the impact of leverage and the effect it had on the liquidation process. (For more on the possible dangers of leverage, see [22]).

2.2.6 Offshore hedge funds

Investors in offshore hedge funds are comprised of non-U.S. individuals and institutions as well as U.S. tax exempt institutions. Institutional investors include entities such as pension funds, endowments, foundations, and other pools of capital whose assets are managed by a group of fiduciaries. Since offshore funds operate on the assumption that the assets they accept are not subject to taxes in the investors home jurisdiction, they have no obligation to compile the information needed for an investor to calculate and pay taxes. This means that for practical purposes, U.S. taxable investors are prevented from participating directly in offshore funds, because although they would be legally obligated to report and pay tax on any taxable income earned in the fund, their fund manager is unlikely to supply them with the relevant information [24].

2.3 Summary

Hedge funds can be complicated investment vehicles that are often difficult to understand. This is due partly to the complex strategies they use, and partly to the high level of secrecy inherent in trying to prevent others from copying their investment methodology. Hedge funds can generally be characterized as high-risk, high-return operations. Pursuit of risk implies a high failure rate. Various studies have estimated that from 7% to 10% of hedge funds fail each year [5, 10]. This can be explained in part by the financial issues, strategies used, bad performances and so forth.

In the next chapter, we will present the main characteristics of the data to be analyzed and we will introduce the mathematical tools needed for this analysis.

Chapter 3

Survival analysis

3.1 Introduction

The primary purpose of survival analysis is to model and analyze time-to-event data; that is, data that have as a principal endpoint the time when an event occurs. Such events are generally referred to as failures. A failure does not necessarily mean the end of the actual lifetime: it may be the occurrence of an event such as the infection of kidney for dialysis patients [25], the first use of marijuana [26], [27] and others [28, 29, 30, 31]. In the financial field, we study similar times in order to analyze the performance of the global market and predict hedge fund mortality.

In this introduction, we present the main characteristics of the data to be analyzed as well as the technical tools used to describe the distribution of the data.

3.1.1 Illustration of survival data

The survival time measured from a suitable origin has two characteristics. The first is that it is non-negative. The second is structural i.e., for some subjects, the event studied does not occur during the observation period and therefore some data are censored; whereas other subjects enter a study at a particular age (not necessarily the origin for the event of interest) and are followed from this delayed entry time. In fact, data are often only partially collected, due particularly to the process of censorship and truncation.

We will study the survival time of a hedge fund. The event of interest is the death of the hedge fund. All individuals (i.e hedge funds) are followed for 60 weeks after entering the database. In Figure 3.1, we consider four particular hedge funds that will help us illustrate some of the most common survival data.

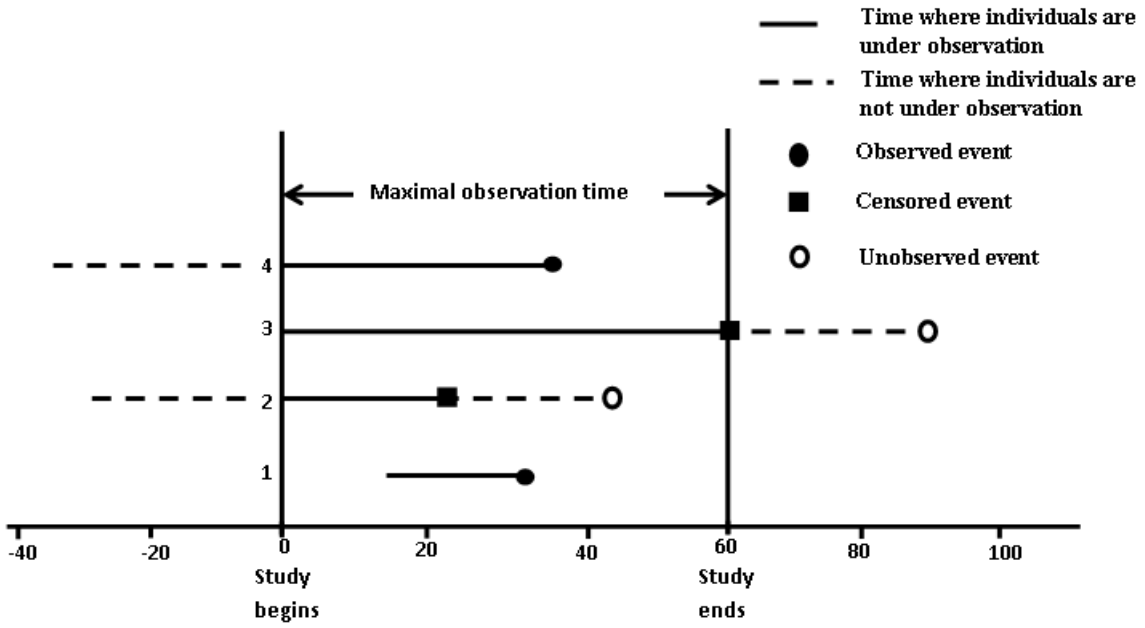

Figure 3.1: Illustration of some of the most common survival data

- Some hedge funds, such as hedge funds 2 and 4, have been created well before entering into database. Such data are truncated. Others are created after the database was created, such as the Hedge Fund 1.
- When the study ends, some hedge funds still have not had the event yet, such as the hedge Fund 3. Such data are censored.
- Other hedge funds drop out or get lost in the middle of the study such as hedge fund 2, and all we know about them is the last time they were still alive. Such data are also censored. Note that this type of right censored is not encountered in our sample data, but it may be encountered in other studies dependenting on how the death is defined.
- Some hedge funds have experienced the event such as hedge funds 4 and 1.

3.1.2 Lifetime distribution function

Let X be the failure time random variable. X is always non-negative. We assume that X has a continuous distribution. There are several equivalent ways to characterize the

distribution of X:

- The *survival function*, which is the probability of an individual surviving beyond time x:

$$S(x) = Pr(X > x) = 1 - Pr(X \leq x) = 1 - F(x) \tag{3.1}$$

  where $F(x)$ is the cumulative distribution function.

- The *hazard function* or *risk function* (also known as the conditional failure rate in reliability theory is):

$$\begin{aligned} h(x) &= \lim_{\triangle x \to 0} \frac{Pr(x \leq X < x + \triangle x \mid X > x)}{\triangle x} \\ &= \lim_{\triangle x \to 0} \frac{Pr(x \leq X < x + \triangle x)}{\triangle x \ Pr(X > x)} = \frac{f(x)}{S(x)} \\ &= -\frac{d \log(S(x))}{dx} \qquad \text{since X is continuous.} \end{aligned} \tag{3.2}$$

  This function is particularly useful in determining the appropriate failure distributions utilizing quantitative information about the mechanism of failure and for describing the way in which the chance of experiencing the event changes with time. There are many general shapes for the hazard rate. Models with increasing hazard rates may arise when there is natural aging or wear. Decreasing hazard functions are much less common but find occasional use when there is a very early likelihood of failure, such as in certain types of electronic devices or in patients experiencing certain types of transplants. Most often, a bathtub-shaped hazard is appropriate in populations followed from birth. Similarly, some manufactured equipments may experience early failure due to faulty parts, followed by a constant hazard rate which, in the later stages of the equipment life, increases. Finally, if the hazard rate is increasing early and eventually begins declining, then, the hazard is termed hump-shaped, see [32] (pages 27-29)

- The *cumulative hazard function* is given by:

$$H(x) = \int_0^x h(u)du = -\log(S(x)). \tag{3.3}$$

  Hence

$$S(x) = \exp(-H(x)) = \exp\left(-\int_0^x h(u)du\right) \tag{3.4}$$

- The *probability density function* (p.d.f.)

$$f(x) = -\frac{dS(x)}{dx} = h(x)\exp\left(-\int_0^x h(u)du\right) \tag{3.5}$$

If we know any one of the four functions defined by relations (3.1), (3.2), (3.3), or (3.5), then the other three can be uniquely determined. For example, the probability density of the *exponential distribution* is:

$$f(x) = \lambda\exp(-\lambda x), \tag{3.6}$$

From equation (3.6), the survival function is easily found to be:

$$S(x) = 1 - F(x) = 1 - (-\exp(-\lambda x) + 1) = \exp(-\lambda x), \tag{3.7}$$

and this distribution is characterized by a constant hazard function:

$$h(x) = \frac{f(x)}{S(x)} = \lambda. \tag{3.8}$$

3.1.3 Quantities derived from the survival distribution

The parameters of interest in survival analysis are:

- *The mean residual life* at time x, which is the mean time to the event of interest, given the event has not occurred at x:

$$mrl(x) = \frac{\int_x^\infty (t-x)f(t)dt}{S(x)} = \frac{\int_x^\infty S(t)dt}{S(x)} \tag{3.9}$$

- *The mean* which represents the average lifetime of an unit in a sample (it is the total area under the survival curve):

$$\mu = E(X) = \int_0^\infty tf(t)dt = \int_0^\infty S(t)dt. \tag{3.10}$$

  Note that the *mean life* $\mu = mrl(0)$

- *The variance* of X is related to the survival function by

$$\begin{aligned} Var(X) &= E[(X - E(X))^2] = E(X^2) - [E(X)^2] \\ &= \int_0^\infty t^2 f(t)dt - [E(X)^2] \\ &= 2\int_0^\infty tS(t)dt - [E(X)^2] \qquad \text{(integration by parts)} \end{aligned} \tag{3.11}$$

- The p^{th} *quantile* (also referred to as the $100p^{th}$ *percentile*) of the distribution of X is the smallest x_p so that

$$S(x_p) \leq 1 - p \;\; i.e. \; x_p = \inf\{t : S(t) \leq 1 - p\}$$

If X is a continuous random variable, then the p^{th} *quantile* is found by solving the equation $S(x_p) = 1 - p$. So, the median lifetime for a continuous random variable X is the value $x_{0.5}$ so that $S(x_{0.5}) = 0.5$. The lifetimes generally have asymmetric distributions to the right, so the mean may be potentially greater than the median, although few individuals have a lifetime superior or equal to the mean. So, an estimate of the distribution median is of more interest in survival analysis. The median is also used as a *measure of location* or *central tendency* and has the advantage that it always exists and is easy to estimate when the data are censored or truncated.

3.2 Some Important Log-Location-Scale Models

There are many statistical distributions used to describe data, but not all of them are used as models for lifetime data. Certain distributions are useful in depicting the lifetime data due to some desirable properties that they possess. Here, we describe some log-location-scale distributions that are used for modelling lifetime data. More details on these distributions can be found in [33] (pages 16-26) and [34].

We recall that X has a density which belongs to a *log-location-scale* if $Y = \log(X)$ has a *location-scale* distribution with probability density function of the form

$$f_Y(y) = \sigma^{-1} f_0\left(\frac{y-\mu}{\sigma}\right), \qquad -\infty < y < +\infty \tag{3.12}$$

where $-\infty < \mu < +\infty$ and $\sigma > 0$ are the location and scale parameters. The standardized random variable $W = (Y - \mu)/\sigma$ has probability density function $f_0(w)$ on $(-\infty, +\infty)$. The distribution function and survival function of Y are $F_0(\frac{y-\mu}{\sigma})$ and $S_0(\frac{y-\mu}{\sigma})$ respectively.

3.2.1 The log-normal distribution

The lifetime X is said to be *log-normally* distributed if $Y = \log(X)$ is normally distributed, say with mean μ, and variance σ^2, i.e. Y has density function:

$$\begin{aligned} f_Y(y) &= \sigma^{-1} f_0\left(\frac{y-\mu}{\sigma}\right) \\ &= \frac{1}{(2\pi)^{1/2}\sigma} \exp\left(-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right), \qquad -\infty < y < +\infty \end{aligned} \tag{3.13}$$

The survival function of Y is:

$$S_Y(y) \;=\; S_0\left(\frac{y-\mu}{\sigma}\right) \;=\; 1 - \Phi\left(\frac{y-\mu}{\sigma}\right), \qquad -\infty < y < +\infty \tag{3.14}$$

where Φ is the standard normal cumulative distribution function of $W = (Y-\mu)/\sigma$. f_0 and S_0 are respectively the density function and survival function of W. W has a standard normal distribution; its density function is given by equation (3.13) with $\mu = 0$ and $\sigma = 1$.

From equation (3.13), the density function of $X = \exp(Y)$ is easily found to be

$$\begin{aligned} f_X(x) &= f_Y[\log(x)] \times \frac{1}{x} \\ &= \frac{1}{(2\pi)^{1/2}\sigma x} \exp\left(-\frac{1}{2}\left(\frac{\log(x)-\mu}{\sigma}\right)^2\right), \qquad x > 0 \end{aligned} \tag{3.15}$$

The log-normal survival function is:

$$S_X(x) = 1 - \Phi\left(\frac{\log(x)-\mu}{\sigma}\right), \tag{3.16}$$

where $\mathbf{\Phi}$ is the standard log-normal distribution function. And the hazard function is given as:

$$h_X(x) = -\frac{d\log(S_x(x))}{dx} \tag{3.17}$$

3.2.2 The log-logistic distribution

The *log-logistic* distribution has a fairly flexible functional form, and it is one of the parametric survival time models in which the hazard rate may be decreasing, increasing, as well as hump-shaped (i.e it initially increases and then decreases). The lifetime X is said to be *log-logistic* distributed with parameters α and β if $Y = \log(X)$ has a

logistic distribution with parameters $\mu = \log(\alpha)$ and $\sigma = \beta^{-1}$. The density function of $Y = \log(X)$ is:

$$\begin{aligned} f_Y(y) &= \sigma^{-1} f_0\left(\frac{y-\mu}{\sigma}\right) \\ &= \frac{\sigma^{-1}\exp[(y-\mu)/\sigma]}{\{1+\exp[(y-\mu)/\sigma]\}^2}, \qquad -\infty < y < +\infty \end{aligned} \tag{3.18}$$

The survival function of Y is:

$$S_Y(y) \;=\; S_0\left(\frac{y-\mu}{\sigma}\right) \;=\; \left(1+\exp\left(\frac{y-\mu}{\sigma}\right)\right)^{-1}, \qquad -\infty < y < +\infty \tag{3.19}$$

where f_0 and S_0 are respectively the density function and survival function of $W = (Y-\mu)/\sigma$. W has a standard logistic distribution; its density is given by equation (3.18) with $\mu = 0$ and $\sigma = 1$.

From equation (3.18), the density function of $X = \exp(Y)$ is easily found to be

$$\begin{aligned} f_X(x) &= f_Y[\log(x)] \times \frac{1}{x} \\ &= \frac{\beta\exp[(\log(x)-\log(\alpha))/\beta^{-1}]}{\{1+\exp[(\log(x)-\log(\alpha))/\beta^{-1}]\}^2} \times \frac{1}{x} \\ &= \frac{\beta/x \times (x/\alpha)^\beta}{[1+(x/\alpha)^\beta]^2} \\ &= \frac{\beta/\alpha(x/\alpha)^{\beta-1}}{[1+(x/\alpha)^\beta]^2}. \end{aligned} \tag{3.20}$$

The survival function of X is:

$$\begin{aligned} S_X(x) &= 1 - F(x) \\ &= 1 - (1+(x/\alpha)^{-\beta})^{-1} \\ &= (1+(x/\alpha)^\beta)^{-1}. \end{aligned} \tag{3.21}$$

The hazard function of X is:

$$h_X(x) = \frac{(\beta/\alpha)(x/\alpha)^{\beta-1}}{[1+(x/\alpha)^\beta]} \qquad , \tag{3.22}$$

where $\alpha > 0$, $\beta > 0$ are parameters and the variable $x > 0$.

We use the notation $Y \;\sim\; Logist\ (\mu, \sigma)$ to indicate that Y has p.d.f. $f_Y(y)$ and $X \;\sim\; LLogist\ (\alpha, \beta)$.

3.2.3 The Weibull distribution and the extreme value distribution

The *Weibull model*, introduced by Waloddi Weibull [35] is an important generalization of the exponential model with two positive parameters. The second parameter in the model allows greater flexibility of the model and different shapes of the hazard function. The convenience of the *Weibull model* for empirical work stems on one hand from this flexibility and on the other hand from the simplicity of the hazard and survival functions. This distribution is the most widely used lifetime data model in both reliability and survival analysis. Extreme Value distributions arise as limiting distributions for maximums or minimums (extreme values) of a sample of independent, identically distributed random variables, as the sample size increases. It models and measures events which occur with very small probability. This implies its usefulness in risk modelling. The class of *Extreme Value Distributions* essentially involves three types of extreme value distributions, types I, II and III. specifically, the Gumbel, Frchet, and Weibull distributions. In this thesis we will use only Gumbel and Weibull distribution.
The lifetime X is said to be *Weibull* (*log-Gumbel*) distributed with parameters γ and λ if $Y = \log(X)$ has an *extreme value* (*Gumbel*) distribution with parameters $\sigma = \gamma^{-1}$ and $\mu = -\log(\lambda)$. The density function of Y is:

$$\begin{aligned} f_Y(y) &= \sigma^{-1} f_0\left(\frac{y-\mu}{\sigma}\right) \\ &= \sigma^{-1} \exp\left(\frac{y-\mu}{\sigma} - \exp\left(\frac{y-\mu}{\sigma}\right)\right), \qquad -\infty < y < +\infty \end{aligned} \tag{3.23}$$

The survival function of Y is:

$$S_Y(y) \;=\; S_0\left(\frac{y-\mu}{\sigma}\right) \;=\; \exp\left(-\exp\left(\frac{y-\mu}{\sigma}\right)\right), \qquad -\infty < y < +\infty \tag{3.24}$$

where f_0 and S_0 are respectively the density function and survival function of $W = (Y - \mu)/\sigma$. W has a standard extreme value distribution; its density is given by equation (3.23) with $\mu = 0$ and $\sigma = 1$. The hazard function of Y is:

$$h_Y(y) = \sigma^{-1} \exp\left(\frac{y-\mu}{\sigma}\right) \tag{3.25}$$

where $\sigma > 0$, $-\infty < \mu < +\infty$ are parameters.

From equation (3.23), the density function of $X = \exp(Y)$ is easily found to be

$$\begin{aligned} f_X(x) &= f_Y[\log(x)] \times \frac{1}{x} \\ &= \exp\left(\frac{\log(x) + \log(\lambda)}{\gamma^{-1}} - \exp\left(\frac{\log(x) + \log(\lambda)}{\gamma^{-1}}\right)\right) \times \frac{\gamma}{x} \\ &= \frac{\gamma}{x} \exp[\log(\lambda x)^{\gamma} - \exp(\log(\lambda x)^{\gamma})] \\ &= \lambda\gamma(\lambda x)^{\gamma-1} \exp(-(\lambda x)^{\gamma}). \end{aligned} \tag{3.26}$$

From equation (3.24), the survival function of $X = \exp(Y)$ is easily found to be

$$S_X(x) = \exp(-(\lambda x)^{\gamma}). \tag{3.27}$$

From equation (3.25), the hazard function of $X = \exp(Y)$ is :

$$h_X(x) = \lambda\gamma(\lambda x)^{\gamma-1}. \tag{3.28}$$

The function $h_X(x)$ is increasing if $\gamma > 1$, decreasing if $\gamma < 1$, and constant for $\gamma = 1$. From equation (3.3), the cumulative hazard function of X is:

$$H_X(x) = -\log[\exp(-(\lambda x)^{\gamma})] \;=\; (\lambda x)^{\gamma}, \tag{3.29}$$

where $\gamma > 0$, $\lambda > 0$ and $x > 0$. Note that when $\gamma = 1$, the Weibull distribution reduces to the exponential distribution with parameter 1.

In place of the Weibull distribution, it is often more convenient to work with the extreme value distribution (which is the logarithm of the Weibull distribution). The main advantage in working with the extreme value distribution is that unlike the Weibull distribution, the extreme value distribution has location and scale parameters.

3.3 Censored data

Censored data arise from incomplete observation of event time, that is, the failure time for some subjects in the sample is not observed directly during the study. This is encountered, for example, when the value of an observation is only partially known. Instead of observing independent and identically distributed variables of duration X, we observe the realization of the variable X subjected to various disturbances, independent or not of the phenomenon. Censorship is commonly encountered when collecting survival data. There are several types of censoring. The three most common forms are right censoring,

left censoring and interval censoring. *Right censoring* arises when a subject leaves the study before an event occurs, or the study ends before the event occurs. *Left censoring* occurs when the event of interest has already happened before the first follow up. This is very rarely encountered. The situation when an event is known to have taken place only between the two-end points of an interval is called *interval censoring*. In this section, we will focus on the right censoring and on the notions of informative and non-informative censoring used in this work. For a detailed account of censored data, one may refer to [32] (pages 64-70).

3.3.1 Right censoring

Let $X_1, X_2, ..., X_n$ be random variables representing failure times. These variables are always non-negative and are assumed to be independent and identically distributed with probability density function $f(x)$ and survival function $S(x)$. Let C_r be a fixed non-negative censoring time (non-random).

Right censoring of type I of survival data occurs when the individual has not experienced the event at the end of the study. It describes the situation where a test ends at a known fixed time. In this case, the censoring time is fixed thus effectively not random. In our work, this is the only type of censoring encountered in our data. We define the censoring (or status) indicator as

$$\delta_i = \begin{cases} 1 & \text{if } X_i \leq C_r \\ 0 & \text{if } X_i > C_r \end{cases} . \tag{3.30}$$

The random variable $T_i = \min(X_i, C_r)$ is called a *censored failure time* random variable. The pairs (T_i, δ_i), $i = 1, ..., n$ constitute a censored survival data set.

3.3.2 Informative and non-informative censoring

Censoring is informative when the censoring cause is related to the lifetimes of the sampling units. On the other hand, when the cause of the censoring is not related to the lifetimes of the units, the censoring is said to be a non-informative censoring, see [36]. Almost all the standard techniques of analyzing censored data are based on the assumption that the censoring is non-informative. For example, when considering Figure 3.1, the hedge fund 3 has been censored at the end of study: this is a non-informative censoring because the censoring cause is not related to its lifetime but due to the end of study.

However, the hedge fund 2 withdraws in the middle of the study perhaps because it is about to go bankrupt: this is an informative censoring because the censoring cause is known and related to its lifetime. To avoid that informative censoring in our sample data, we defined the failure time of a hedge fund as the lack of reporting during 3 consecutive months, therefore, the last performance date is equivalent to the death date of the fund. But if the lack of reporting in the database occurs before the date of bankrupt, there is a backlog bias (see [37]). In fact, some hedge funds stop to report to the database when they do not perform well. However if the period of poor performance is past, those hedge funds update their performance in the database.

3.4 Truncated data

3.4.1 Left truncation

Left truncation of survival data occurs when only individuals who survive a sufficient time are included in the sample (whose event time lies within a certain observational window starting at U). Left truncation is also known by other names: *delayed entry* and *stock sampling with follow-up*. In fact, the subjects enter a study at a particular age (not necessarily the origin for the event of interest) and are followed from this delayed-entry time until either the event occurs or the subject is censored. Let U be a random variable not depending of X. We observe X if only if $X \geq U$.

3.4.2 Right truncation and interval truncation

The right truncation occurs when the values of a random variable can be observed only when they are smaller than an upperbound (the right truncation point); that is, all values greater than this upperbound are not observable (i.e X is right truncated if it is observed only if $X < U$. The truncation variable U is assumed to be independent of X). The interval truncation occurs when the lifetime X is left truncated and right truncated.

3.5 Likelihood of Censored and Truncated Data

The primary interest of parametric statistical inference is to estimate the parameters of a statistical distribution, based on available data. Different methods are employed depending on their necessity. Maximum likelihood estimation is probably the most used

in statistics due to its simplicity and optimality. In constructing a likelihood function for censored or truncated data, we need to consider carefully what information each observation gives us and consider that the lifetimes and censoring times are independent. Therefore, there are several ways to construct the likelihood function depending on the type of censorship or truncation encountered. Below, we will present only the construction of the likelihood function for right censoring of type I and left truncation. For the construction of other types of likelihood function, we refer to [33] (pages 49-71), [32] (pages 64-74) and [34].

3.5.1 The likelihood for a right censoring of type I

For a right-censored observation, all we know is that the event time is larger than the censored time. So a variable is given only by the survival function evaluated at the study time. The likelihood function is based on the probability distribution of (T_i, δ_i), $i = 1, ..., n$. Both T_i and δ_i are random variables and

$$S(C_r) = P(X_i > C_r) = P(T_i = C_r, \delta_i = 0) \tag{3.31}$$

$$f(T_i) = P(T_i = X_i, \delta_i = 1), \qquad T_i \leq C_r \tag{3.32}$$

where C_r is a fixed constant. So the joint p.d.f of T_i and δ_i is

$$f(T_i)^{\delta_i} P(X_i > C_r)^{1-\delta_i}. \tag{3.33}$$

Since the lifetimes $X_1, X_2, ..., X_n$ are independent, the likelihood function is given by:

$$\begin{aligned} L &= \prod_{i=1}^{n} f(T_i)^{\delta_i} P(X_i > C_r)^{1-\delta_i} \\ &= \prod_{i=1}^{n} f(T_i)^{\delta_i} S(C_r)^{1-\delta_i}. \end{aligned} \tag{3.34}$$

3.5.2 Likelihood for left truncated observations

For truncated data these probabilities are replaced by the appropriate *conditional* probabilities. Suppose now that the real survival time X_i is left truncated at U_i. So, a triplet

(U_i, X_i, δ_i), $i = 1, ..., n$ is observed. The *conditional* distribution of X_i given that $X_i \geq U_i$ is:

$$g(x \mid X_i \geq U_i) = \frac{f(x)}{P(X_i \geq U_i)} = \frac{f(x)}{S(U_i)}. \tag{3.35}$$

Therefore, the probability to observe a death at x_d is proportional to

$$g(x_d \mid X_d \geq U_d) = \frac{f(x_d)}{S(U_d)}. \tag{3.36}$$

The probability that the real survival time X_i is right censored at C_r is

$$P(X_i > C_r \mid X_i \geq U_i) = \frac{P(X_i \geq C_r)}{P(X_i \geq U_i)} = \frac{S(C_r)}{S(U_i)} = \frac{S(T_i)}{S(U_i)}, \tag{3.37}$$

because $T_i = \min(X_i, C_r) = C_r$.

Thus, the *conditional likelihood* function is given by

$$\begin{aligned} L &= \left[\prod_{d \in D} f(x_d) \prod_{i \in R} S(T_i)\right] / \prod_{i=1}^{n} S(U_i) \\ &= \left[\prod_{i=1}^{n} f(T_i)^{\delta_i} S(C_r)^{(1-\delta_i)}\right] / \prod_{i=1}^{n} S(U_i) \\ &= \prod_{i=1}^{n} \left[\frac{f(T_i)}{S(U_i)}\right]^{\delta_i} \left[\frac{S(C_r)}{S(U_i)}\right]^{1-\delta_i}, \end{aligned} \tag{3.38}$$

where D is the set of death times and R the set of right-censored observations. Note that the non-truncated observations are like the truncated observations with $U_i = 0$.

3.6 Non-parametric estimation of the survival function and quantiles

The standard estimator of the survival function proposed by Kaplan and Meier [38] is called the *product-limit estimator*. This estimator is used when no assumption has been made on the distribution of survival times. It is extremely popular as it requires only very weak assumptions and yet uses all information content of fully observed right-censored data. It can also be adjusted to take into account left truncation. It is implemented in most statistical packages (such as R) but is simple enough that it can be calculated by hand with small samples. In the present work, we will present only the definition of the Kaplan-Meier estimator without going into the details of the construction. Further

information can be found in [33] (pages 79-124) and [34]. However, we may be interested in the estimation of other functions that characterize the time distribution of events. So, we will briefly present the estimation of the cumulative hazard function, with the Nelson-Aalen estimator, see [39].

3.6.1 The Kaplan-Meier estimator

The *Kaplan-Meier* or *product-limit estimator* of the survival function $S(x)$ associated to a failure time X is defined by:

$$\widehat{S}(x) = \prod_{j:x_j \leq x} \frac{n_j - d_j}{n_j} \tag{3.39}$$

where d_j is the number of deaths at time x_j and n_j the number of individuals who are at risk at time x_j. In fact, n_j is a count of the number of individuals who are alive at x_j or experience the event of interest at x_j.

Concerning the left truncated data, the triplets (U_i, X_i, δ_i), $i = 1, ..., n$ are observed. So, $n_j = \sum_i I(u_i \leq x_j \leq x_i)$ where the individual i is truncated at u_i and is not censored before x_j. Since $n_j = 0$ for $X_j < U_{(1)}$ with $U_{(1)} = \min(U_i)_{i=1,2,...,n}$, there is no information about the survival function. Consequently, we cannot estimate $S(x)$ unless $U_{(1)} = 0$. Therefore, the Kaplan-Meier of the survival function $S(x)$ for the left truncated data is defined as:

$$\widehat{S}_L(x|X > U_{(1)}) = \prod_{j:x_j^* \leq x} \frac{n_j^* - d_j^*}{n_j^*}, \tag{3.40}$$

where $x_1^* < x_2^* < ... < x_k^*$ are the distinct observed failures times, $d_j^* = \sum_i I(x_i = x_j^*, \delta_i = 1)$ and $n_j^* = \sum_i I(u_i \leq x_j^* \leq x_i)$.

Note that without censorship or truncation, for an independent and identically distributed sample of failure times $X_1, X_2, ..., X_n$, a natural estimator of the survival function of X is the *empirical survival* function defined by:

$$\widetilde{S}(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(X_i > x)}. \tag{3.41}$$

The empirical cumulative hazard estimate is:

$$\widetilde{H}(x) = -\log(\widetilde{S}(x)). \tag{3.42}$$

This estimator has good properties in terms of convergence. However, in the case of censored or truncated data, the variable of interest is not the observed variable X. Therefore, estimating the survival function $S(x)$ by the empirical survival function $\widetilde{S}(x)$ provides a biased estimate (the censored data are considered deaths: there is an underestimation of survival).

In order to assess the accuracy of the estimate of $S(x)$, it is useful to estimate its variance. This variance is estimated by the Greenwood's formula:

$$\widehat{\mathrm{Var}}(\widehat{S}(x)) = \widehat{S}(x)^2 \sum_{j:x_j \leq x} \frac{d_j}{n_j(n_j - d_j)} \tag{3.43}$$

and the standard error of cumulative proportion surviving is computed from Greenwood's formula:

$$SE_x = \widehat{S}(x) \sqrt{\sum_{j:x_j \leq x} \frac{d_j}{n_j(n_j - d_j)}} \tag{3.44}$$

The confidence intervals for the survivor function $S(x)$ at a specified value x can be constructed in a variety of ways. The Kaplan-Meier estimate $\widehat{S}(x)$ is asymptotically normally distributed under some conditions. So, if we assume that $\frac{\widehat{S}(x)-S(x)}{\widehat{\sigma}_s(x)} \sim N(0,1)$, then the confidence interval for the survival function $\widehat{S}(x)$ is

$$CI(\alpha) = \left[\widehat{S}(x) \pm z_{\frac{\alpha}{2}} \widehat{\sigma}_s(x)\right] \tag{3.45}$$

where $\widehat{\sigma}_s(x)^2 = \widehat{\mathrm{Var}}[S(x)]$, (see Section 3.2 of [33]).

3.6.2 Standard error estimator of the sample pth-quantile

The standard error of a quantile is calculated from:

$$SE'_x = \frac{SE_{x(p)}}{f[x(p)]} \tag{3.46}$$

where p is a quantile, $SE_{x(p)}$ is the standard error of survival function at $x(p)$,

$$f[x(p)] = \frac{S[u(p)] - S[l(p)]}{l(p) - u(p)} \tag{3.47}$$

and:

$$u(p) = \max\{\ x(j)|\ S[x(j)] \geq 1 - p + \epsilon\} \tag{3.48}$$

$$l(p) = \min\{\ x(j)|\ S[x(j)] \leq 1 - p + \epsilon\} \tag{3.49}$$

ϵ is 0.05, but can be any value between 0 and 1.

3.6.3 The Nelson-Aalen estimator

The Nelson-Aalen estimator of the cumulative hazard function $H(x)$ is given by :

$$\overline{H}(x) = \sum_{j:x_j \leq x} \frac{d_j}{n_j}. \tag{3.50}$$

Its variance can be estimated by:

$$\widehat{\mathrm{Var}}(\overline{H}(x)) = \sum_{j:x_j \leq t} \frac{d_j}{n_j^2}. \tag{3.51}$$

(See Section 3.2 of [33]).

3.7 Regression models

Regression analysis helps us understand how the typical value of the response variable changes when some of the explanatory variables are varied, while the other explanatory variables are held fixed. In survival analysis, it is widely used for prediction and forecasting, and also used to understand which among the explanatory variables are related to the response variable, and to explore the forms of these relationships. Regression analysis of lifetimes involves specifications for the distribution of a lifetime X, given a vector of covariates $\mathbf{z}$. However, it may happen that specifying the distribution is difficult. So we use the non-parametric regression which refers to techniques that allow the regression function to lie in a specified set of functions, which may be infinite-dimensional, see [6]. There are two main approaches to regression modelling for lifetimes. The first approach which uses time transformations, assuming that the effect of covariates is equivalent to altering the rate at which time passes is called the *accelerated failure-time* (AFT). The second approach adopts specifications of the way that the covariates affect the hazard function for X. The most common model for this type is the proportional hazard (PH) regression model, see [33] (pages 269-303). We explain these models in the pages below. Let $\mathbf{z} = (z_1, z_2, ..., z_p)$ be a vector of non-random covariates.

3.7.1 Accelerated failure-time (AFT) regression models

In an AFT model, it is assumed that the survival function of the distribution of Y given $\mathbf{z}$ is given by:

$$S_Y(y \mid \mathbf{z}) = S_0\left(\frac{y-\mu(\mathbf{z})}{\sigma}\right), \qquad -\infty < y < +\infty \tag{3.52}$$

where S_0 is the survival function of a random variable $W = (Y-\mu)/\sigma$ (which does not depend on $\mathbf{z}$). Another way to express this is as $Y = \mu(\mathbf{z}) + \sigma W$.

Weibull AFT model

We have already seen from the *Weibull survival function* equation (3.27) that when we use the transformation $Y = \log(X)$ where X is the lifetime, we get the *extreme value* survival function equation (3.24). From this *extreme value* equation, if the parameter $\mu = -\log(\lambda)$ depends on the covariates $\mathbf{z}$, we get the *Weibull AFT* model with equation:

$$S_Y(y) = \exp\left(-\exp\left(\frac{y-\mu(\mathbf{z})}{\sigma}\right)\right) \qquad -\infty < y < +\infty \tag{3.53}$$

where σ, $-\infty < \mu < +\infty$ are parameters with $\sigma = \gamma^{-1}$ and $\mu(\mathbf{z}) = -\log(\lambda(\mathbf{z}))$

3.7.2 Proportional hazard (PH) regression models

In a *PH model*, the hazard function is of the form:

$$h(x \mid \mathbf{z}) = h_0(x)r(\mathbf{z}) \tag{3.54}$$

where X is the lifetime, $r(\mathbf{z})$ and $h_0(x)$ are positive-valued functions. The function $h_0(x)$ is usually called the baseline hazard function. It is the hazard function for an individual whose covariate vector $\mathbf{z}$ is such that $r(\mathbf{z}) = 1$. The survival function for X given $\mathbf{z}$ is

$$S(x \mid \mathbf{z}) \;=\; (S_0(x))^{r(\mathbf{z})} \tag{3.55}$$

where $S_0(x) = \exp(-H_0(x))$ is a baseline survivor function. A common specification is:

$$r(\mathbf{z}) = \exp(\boldsymbol{\beta}'\mathbf{z})$$

where $\boldsymbol{\beta}' = (\beta_1, \beta_2, ..., \beta_p)$ is a vector of regression coefficient. The name proportional hazards comes from the fact that any two individuals have hazard functions that are constant multiple of one another. The measure of effect is called *hazard ratio*. The estimated *hazard ratio* of two individuals with different covariates z_1 and z_2 is:

$$\widehat{HR} = \frac{h_0(x)\exp(\hat{\boldsymbol{\beta}}'z_1)}{h_0(x)\exp(\hat{\boldsymbol{\beta}}'z_2)} = \exp\left[\hat{\boldsymbol{\beta}}'(z_1 - z_2)\right] \tag{3.56}$$

Since this hazard ratio is time-independent, this model is called a proportional hazard model.

Weibull PH model

Under the *Weibull PH model*, the hazard function of a particular individual with covariates $\mathbf{z} = (z_1, z_2, ..., z_p)$ is given by

$$h(x \mid \mathbf{z}) = h_0(x) \exp(\boldsymbol{\beta}'\mathbf{z}) = \lambda\gamma(\lambda x)^{\gamma-1} \exp(\beta_1 z_1 + \beta_2 z_2 + ... + \beta_p z_p) \tag{3.57}$$

From equation (3.55), the corresponding *Weibull PH* survival function is given by

$$S(x \mid \mathbf{z}) \;=\; \exp(-H_0(x))^{\exp(\boldsymbol{\beta}'\mathbf{z})} = \exp(-(\lambda x)^{\gamma} \exp(\boldsymbol{\beta}'\mathbf{z})), \tag{3.58}$$

where we used equation (3.29) for the second equality above. Due to equation (3.53), the Weibull family is the only set of models that is in both the PH and AFT classes.

From equation (3.27), applying the $\log(-\log(\bullet))$ transformation to the survival function for a *Weibull distribution*, we obtain:

$$\log(-\log(S(x))) \;=\; \gamma \log(\lambda) + \gamma \log(x) \tag{3.59}$$

The plot of $\log(-\log(S(x)))$ versus $\log(x)$ should give approximately a straight line if the *Weibull distribution assumption* is reasonable. The intercept and slope of the line will be rough estimate of $\gamma \log(\lambda)$ and γ respectively. If the two lines for two groups in this plot are essentially parallel, this means that the proportional hazards model is valid. Furthermore, if the straight line has a slope nearly one (i.e $\gamma = 1$), the simpler *exponential distribution* is reasonable. Note that, for an exponential distribution $\log(S(x)) = -\lambda x$. Thus we can consider the graph of $\log(S(x))$ versus x. This should be a line that goes through the origin if the exponential distribution is appropriate. Another approach to assess the suitability of a parametric model is to estimate the hazard function using the non-parametric method. If the hazard function were reasonably constant over time, this would indicate that the exponential distribution might be appropriate. If the hazard function increased or decreased monotonically with increasing survival time, a Weibull distribution might be considered.

3.7.3 Semi-parametric multiplicative hazards models

The *Cox model* is extensively used in the analysis of survival data and it was proposed by Cox in [40]. It is used when the objective is to assess the effect of covariates on lifetime. Let X be a continuous lifetime variable and $\mathbf{z}$ a $p \times 1$ vector of fixed covariates. We consider the PH model defined by equation (3.54) in the case when $r(\mathbf{z}) = \exp(\boldsymbol{\beta}'\mathbf{z})$, so that the hazard function for X given $\mathbf{z}$ takes the form

$$h(x \mid \mathbf{z}) = h_0(x)\exp(\boldsymbol{\beta}'z) = h_0(x)\exp(\beta_1 z_1 + \beta_2 z_2 + ... + \beta_p z_p), \tag{3.60}$$

with $\boldsymbol{\beta}$ a $p \times 1$ vector of regression coefficients. Hence, by equation (3.4):

$$\begin{aligned} S(x \mid \mathbf{z}) &= \exp\left(-\int_0^x h(u \mid \mathbf{z})du\right) \\ &= \left[\exp\left(-\int_0^x h_0(u)du\right)\right]^{\exp(\boldsymbol{\beta}'\mathbf{z})} \\ &= (S_0(x))^{\exp(\boldsymbol{\beta}'\mathbf{z})}. \end{aligned} \tag{3.61}$$

Given a censored random sample of lifetimes (T_i, δ_i), $i = 1, ..., n$ and corresponding covariate vectors $\mathbf{z}_i$, we want to estimate $\boldsymbol{\beta}'$ and $S_0(x)$. Cox introduces an ingenious way of estimating $\boldsymbol{\beta}'$ without considering $S_0(x)$ explicitly, which is now known as the partial likelihood method, see [33] (pages 349-353).

Verifying the Proportional Hazard (PH) assumption

The main assumption of Cox model is the proportional hazards. This means that the hazard function of one individual is proportional to the hazard function of the other individual, i.e. the hazard ratio is constant over time. There are several methods for verifying that a model satisfies the assumption of proportionality [33] (pages 354-358). We will present three common tests for examining this proportionality assumption.

1. **Graphical method**: The plots of $\log[-\log(\widehat{S}_j(x))]$ versus x or $\log(x)$ for two groups (or several groups) should be roughly vertical translations of one another if the PH assumption is reasonable. Here, $\widehat{S}_j(x)$ is the Kaplan-Meier estimate for each group. This method does not work well for continuous predictors or categorical predictors that have many levels because the graph becomes "cluttered". Furthermore, the curves are sparse when there are few time points and it may be difficult to tell how "close to parallel" is close enough.

2. **Test the interaction of covariates with time**: We create time-dependent covariates by creating interactions of the predictors and a function of the survival time, and including them in the model. If the test shows that there exist significant interactions, then the covariates should be time-dependent. This means that the proportionality assumption is violated.

3. **Conduct the Schoenfeld residuals test**: One popular assessment of proportional hazards is based on *Schoenfeld residuals*, which ought to show no association with time if proportionality holds, see [41].

3.7.4 Time-Varying Covariates

A time-dependent variable is defined as any variable whose value for a given subject depends on the time x. In contrast, a time-independent variable is a variable whose value for a given subject remains constant over time. Time-varying covariates can be incorporated into AFT and PH models. The key idea is that the covariates $\mathbf{z}(x)$ effectively alter the rate at which the time x passes. Given a survival analysis situation involving both time-independent and time-dependent predictor variables, we can write the extended PH model that incorporates both types as follow:

$$h(x \mid \mathbf{z}(x)) = h_0(x) \exp\left[\sum_{i=1}^{p_1} \beta_i z_i + \sum_{i=1}^{p_2} \alpha_i z_i(x)\right], \tag{3.62}$$

where $\mathbf{z}(x) = (z_1, z_2, ..., z_{p_1}, z_1(x), z_2(x), ..., z_{p_2}(x))$. The most important feature for this model is that the proportional hazards assumption is not satisfied.

3.8 Model Checking and Goodness-of-Fit Tests

Model validation is possibly the most important step in the model building sequence. It is also one of the most overlooked. There are many statistical tools for model validation, but the primary tool for most process modeling applications is graphical residual analysis. Graphical methods are useful for summarizing information and suggesting possible models. Numerical methods play an important role as confirmatory methods for graphical techniques. Together, they also provide ways to check various assumptions concerning the form of a lifetime distribution and its relationship to covariates. Some procedures that help in formulating and checking the validity of a regression model for lifetime data will be discussed in this section.

3.8.1 Graphical methods

1. **Exponential Distribution**: The plot of $-\log(\widehat{S}(x))$ versus x should yield a straight line which passes through the 0.

2. **Weibull Distribution**: The plot of $\log[-\log(\widehat{S}(x))]$ versus $\log(x)$ should be a straight line.

3. **Log-Normal Distribution**: The plot of $\Phi^{-1}(1-\widehat{S}(x))$ versus $\log(x)$ should be a straight line, where $\Phi(\bullet)$ is the standard normal c.d.f.

4. **Log-Logistic Distribution**: The plot of $\log[(1-\widehat{S}(x))/\widehat{S}(x)]$ versus $\log(x)$ should be a straight line.

5. **Cox-Snell Residuals Plot**: This plot can be applied to any parametric model and Cox PH model. The Cox-Snell residual for the ith individual with observed time x_i is defined as $r_{c_i} = \widehat{H}(x_i|z_i) = -\log(\widehat{S}(x_i|z_i))$. If the fitted model is appropriate, the plot of $\log[-\log(\hat{S}(r_{c_i}))]$ versus $\log(r_{c_i})$ should be a straight line with unit slope through the origin.

3.8.2 Numerical methods

There are three common statistical methods for model comparisons:

1. log-Likelihood (see the equations (3.34) and (3.38))

2. The likelihood ratio test (LRT) can be used to compare the nested models. The LRT is defined as:

$$LRT = -2\log\left[\frac{likelihood\,for\,null\,model}{likelihood\,for\,alternative\,model}\right] \tag{3.63}$$

   The exponential model, the Weibull model and log-normal model are nested within gamma model.

3. The Akaike information criterion (AIC) can be used for comparing models that are not nested. The AIC is defined as:

$$AIC = -2l + 2(k+c), \tag{3.64}$$

   where l is the log-likelihood, k is the number of covariates in the model and c is the number of model-specific ancillary parameters. The term $2(k+c)$ can be thought of as a penalty if non-predictive parameters are added to the model. Lower values of the AIC suggest a better model.

3.9 Summary

In this chapter, we presented some background material from survival analysis to analyze our data. In particular, after describing our data, we introduced some mathematical models and discussed how to estimate the regression parameters.

In the next chapter, we will present our data and explain how we selected the meaningful variables.

Chapter 4

Data selection and methodology

Several recent works have been devoted to the survival analysis of hedge funds. They used different databases to estimate and compare the lifetimes of various hedge funds. These studies have reported that there are many characteristics which play a significant role in the lifetime of hedge funds. In this chapter, we first review some literature concerning the characteristics that have been reported to have a significant effect on the lifetime of a hedge fund. Our data set consisted of $453,728$ data entries, loosely organized and spread over 100 Excel tables. The first step consisted various data management tasks: identifying linking variables, matching tables, selecting and defining variables to prepare the data for in depth analysis of survival times. The second section of this chapter consists in defining and describing the numerous quantities of interest from the data to be used in the analysis. We will take into account the left truncation and right censoring to define the lifetime of hedge funds of our data. Finally, we will use exploratory data analysis techniques to identify the hedge fund characteristics which appear a *priori* to have the strongest relationship with lifetime in our data.

4.1 Literature review

Recent studies focus on the relationship between hedge fund characteristics and the probability of a hedge fund demise. Several of them apply the Cox proportional hazards model to study the effect of explanatory variables on the survival of a hedge fund. In particular, characteristics such as *size*, *age*, *performance* were found to have a potential impact on the fund lifetime.

The *size of the hedge funds* is an important characteristic to analyze before investing.

Several authors have investigated the relation between the fund size and the lifetime of hedge fund and identified a positive relationship between hedge funds size and lifetime (see [6, 7, 8, 9, 10, 11, 12]). Boyson [7], Gregoriou [6] and Liang [8] have found in their work that large hedge funds survive longer than smaller, and that using leverage leads to shortened lifetimes. They also found a strong effect of returns on survival time. Along the same line of research, Baquero et al. [9] showed that funds with a larger size and a higher past return are much more likely to survive, but they did not find any meaningful relationship between incentive fees and survival rates. Getmansky et al. [10] estimated the effects of fund-specific characteristics such as size and age on the likelihood of liquidation for the funds in the Lipper TASS database. They showed that size and age have a significantly negative impact on the liquidation probability. Using several theories of reputation, Boyson [7] investigated how the career concerns of hedge fund managers influences their decisions. In the same order, Gregoriou [11] estimated the survival times of micro-sized hedge funds in $5 million increments, and found the same positive relationship between size and survival even among micro-sized funds. He incorporated both *Live* and *Dead Funds* data into his survival analysis. By including a range of covariates to explain fund's survival lifetime, he found that the fund size has an important impact on survival time, with the funds above the median size having a longer survival time.

Another important characteristic described in the literature as playing a significant role in determining hedge fund survival is the *hedge fund performance*. Hendricks et al. [42] assumed that this performance could improve the survival probability, and also proposed the idea that survivorship bias may induce some patterns in hedge funds survival analysis. Njavro [43] applied the parametric probit regression model and the less restrictive semi-parametric Cox proportional hazard model to investigate the factors that affect the survival and mortality patterns of Asia-focused hedge funds. He found that larger, better performing funds with lower redemption frequency have a higher likelihood of survival. Getmansky [44] used the TASS database from January 1994 until December 2002 to investigate the relationship between industry, fund-specific factors and the probability of hedge fund survival. She found a positive relation between past fund asset size and current fund performance. In fact, when hedge funds which have an optimal size, exceed that size, it has a negative effects on performance, thus on the lifetime. Gregoriou [6] applied survival analysis to the Zurich Capital Markets database from 1990 until 2001 to investigate whether some explanatory variables can predict hedge fund failure. He used several survival models, including the product-limit estimator, the life table method, the

accelerated failure time model, and the semi-parametric Cox proportional hazard model, and found that larger, low leveraged, better performing funds have a higher probability of survival. Furthermore, funds with a higher minimum investment requirement tend to die faster. Liang and Park [45] implemented survival analysis methods to investigate the causes of hedge fund failure. The authors compared the effectiveness of various downside risk measures in predicting hedge fund attrition. They found that funds with high historical performance and high-water mark provisions are less likely to fail. On the other hand, Gregoriou and Rouah [46] investigated the link between the size of a fund and its performance and found that the correlations between fund size and fund performance are statistically insignificant. They showed that the size of the hedge fund has no impact on the fund performance, and they suggested that the investigation of the relationship between hedge fund size and performance be carried out over a longer period.

The *fund age* has been demonstrated to have an effect on the lifetime of the hedge fund. Estimates of the 50% survival time (i.e. time by which half of the hedge funds have died) range from 2.5 years, as obtained by Brown et al. [47] using TASS database, to as high as 10 years, as obtained by Barès et al. [48]. In most studies (see Amin and Kat [49], Gregoriou [6], Securities and Exchange Commission works [50]), the 50% survival time is estimated to be between 5 and 5.5 years. These authors found a higher mortality rate for new hedge funds, i.e. younger hedge funds with low past returns are at a higher risk of failure. They also documented that there is a strong relation between bad performance and consecutive disappearance from the database [51]. In the same order, De Souza and Gokcan [52] found that young funds with poor performance, minimal assets under management, short lock-up periods, short redemption notice periods, and no high watermarks are the funds that are most likely to liquidate. In their study, Amin and Kat [49] used the Cox proportional hazards model applied to an augmented TASS database and observed that hedge funds have a median survival lifetime of 2.5 years and that 40% of funds do not survive past their fifth year of operations. Gregoriou et al. [53] studied the survival of exchange-listed hedge funds. They compared survival times and characteristics of listed and non-listed hedge funds, and found that listed funds are larger and adopt more conservative investment strategies than non-listed funds. They also found that listed funds survive two years longer than non-listed funds, on average. Along the same line of research, Howell [54] found that 7.4% of hedge funds fail in their first year and the chance of failure increases to 20.3% in their second year.

There are some other characteristics mentioned by some studies. Baba and Goko [5] applied a survival analysis to individual hedge funds in the TASS database. They used

several methodologies, including a non-parametric survival analysis and the Cox proportional hazards model with shared frailty, to estimate the effect of fund characteristics and dynamic performance properties on the survival probabilities of hedge funds. They found that funds with a longer redemption notice period and a lower redemption frequency have a higher probability of survival. They also found that large funds with higher returns, recent fund flows, lower volatilities, and higher skewness of returns and assets under management have a lower probability of failure. The impact of other variables, such as age, size, and lock-up provisions, depends on how fund failure is defined. Njavro [43] found that the incentive structure of hedge funds (management fees, incentive fees, and lock-up provisions) does not seem to have an effect on fund survival. He also showed that two models contradict on the impact of leverage on hedge fund survival: the probit model that indicates that higher leverage is beneficial for fund survival, while the Cox model indicates that leverage has no effect on hedge fund survival. Interestingly, Baba and Goko [5] found that leverage does not significantly influence the probability of fund survival. Gregoriou [6] found that funds with low leverage are more likely to live longer than higher leveraged funds, while funds with higher minimum purchases tend to fail faster. Notably, funds with annual redemptions were inclined to have longer survival times. Another interesting finding was that investment in funds-of-hedge funds was a worthwhile strategy due to their higher survival time and their diversification effects. The widely-varying estimates of survivorship bias encountered in the literature has prompted researchers to investigate factors driving hedge fund mortality and liquidation probabilities, in an attempt to understand factors driving survivorship bias. Rouah [13] used a competing risks model to analyze hedge funds survivorship. Hedge funds lifetimes are studied with time-dependent covariates, along with the cause of exit under the assumption of independent risks. He noticed that avoiding to separate liquidation from other kind of withdrawal leads to severe biases (especially over-estimation of the survivorship bias). Amin and Kat [55] explained that this may be due, for instance, to a lack of size or performance. Gregoriou [6] argued that failure to incorporate *Live Funds* into the estimation process would result in a downward bias of their survival lifetime, due to the contribution of the Live Funds to the overall survival lifetime of funds. According to Rouah [13], in most studies survivorship bias is estimated at 2 to 4% per year.

4.2 Data Selection

This section describes the data used in this thesis as well as the biases that are well documented in the hedge funds literature. The hedge fund industry is growing at a fast pace, growth in assets under management being estimated to be 40% per annum. The growth and change of the hedge fund industry lead to the existence of different hedge fund tracking databases. The most well known are the Tremont TASS database and the Hedge Fund Research (HFR) database. Our data come from HFR database. HFR was established in 1992, and specializes in the areas of indexation and analysis of hedge funds. Liang [8] investigated the data quality of those two database and found that the two databases cover a small proportion of common funds and differ not only in the number of dissolved funds covered, but also in some other aspects. Several academic studies and researchers (see [47, 5, 37, 56, 57]) have used the TASS database to analyse survival time of hedge funds, however, from the best of our knowledge, very few researchers (see [12, 8]) have used the HFR database to analyze the survival time of hedge funds. Therefore, an important feature of our work is to study the relationship between hedge funds characteristics and the probability of hedge funds demise, and compare with those found in the TASS database.

The HFR database is structured into 5 tables related together by the variable *code* (the unique ID of each hedge fund):

1. The *Administrative* table is the main table and contains several variables, among which the *code ID* of hedge fund and *fund status*.

2. The *Performance* table is a secondary table and contains 5 variables, among which the *code ID*, the *date* when the performance was reported, and the *percentage of performance* of a fund.

3. The *Assets* table is a secondary table and contains 4 variables, among which the *code ID*, the *date* when the asset was reported, and the *assets* for a period in millions of currency units.

4. The *Regions* table is a secondary table and contains 5 variables, among which the *code ID* and the *Region of investment* (Canada, USA, Eastern Europe, etc.)

5. The *Instruments* table is a secondary table and contains 5 variables, among which the *code ID* and the *Instrument* invested (corporate bonds, equities, etc.)

Our data cover the period from September 1963 to June 2005. The *Administrative* table, *Performance* and *Assets* tables contained 7093 distinct hedge funds and *Regions* and *Instruments* tables contained less than 1046 hedge funds. So, to avoid losing a lot of data, we decided to drop off the last two tables. From the remaining 3 tables with 7093 distinct hedge funds, we eliminated 146 funds due to poor data quality or inconsistencies. As a result, the data studied in this thesis consisted of 6947 hedge funds. Several fund characteristics included in the *Administrative* table are considered as fixed covariates. Other characteristics included in the *Performance* and *Assets* tables are considered as time-dependent covariates.

4.2.1 Administrative table

The *Administrative* table contains a variety of information about each fund. Among them are the following variables: *strategy, minimum investment, fund assets, leverage, high watermark, fund offshore, annual audit, incentive fee, management fee, denomination.* To better understand and estimate the regression coefficients of our variables with the proportional hazard models, we classify them into categories.

The variable *strategy*

The variable *strategy* defines the type of strategy that a hedge fund uses. As we have mentioned previously, hedge funds use a variety of strategies to perform. The variable *strategy* in the original table identifies 30 different strategies, some strategies being highly used by some managers while others being very little used. Therefore, we classified these strategies in 7 major groups, namely: *Event Driven, Fund of funds, Tactical, Relative value, Long/Short, Location* and *Others* (see Figure 4.1). The strategy *Others* includes all the strategies that could not be classified.

The variable *denomination*

The variable *denomination* defines the currency base for the fund performance: US dollar (USD), Canadian dollar (CAD), Euro (Eur), and others. This variable contains 16 currencies in the original table. However, 88% of the hedge funds use the US dollar currency and 7% the Euro currency. Therefore, we grouped this variable in 3 major groups, namely: *USD, Eur* and *Others* (see Figure 4.2). The denomination *Others* includes all the 14 other denominations.

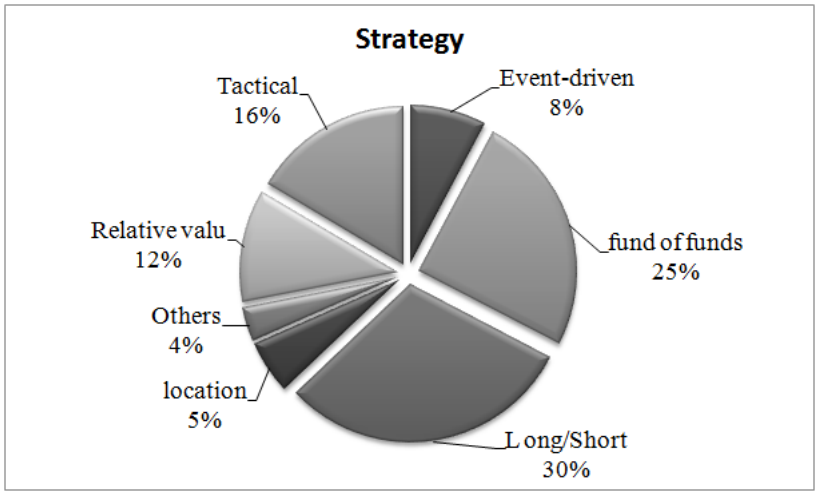

Figure 4.1: Pie chart defining the percentage of the 7 major strategies

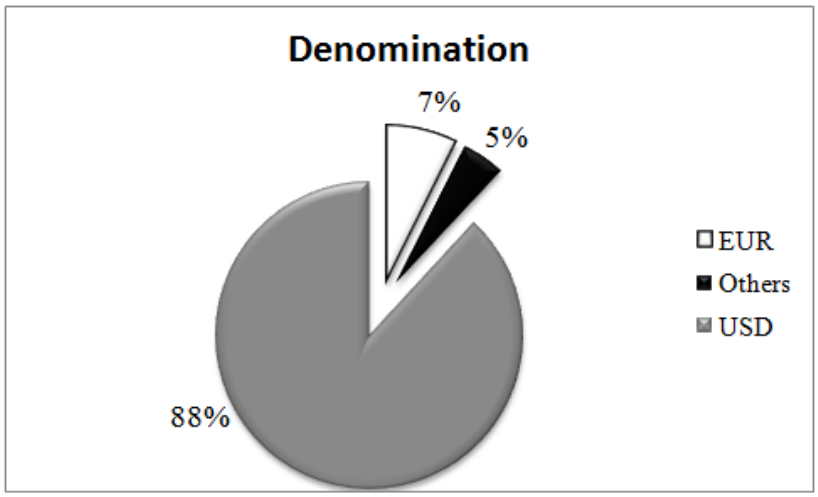

Figure 4.2: Pie chart defining the percentage of the 3 major denominations

The variables *incentive fee* and *management fee*

To facilitate our study, we decided to split the variable *incentive fee* in 4 classes (see Figure 4.3):

- Incentive fee less than 20% ,
- Incentive fee equal to 20%,

- Incentive fee greater than 20%,
- N/A.

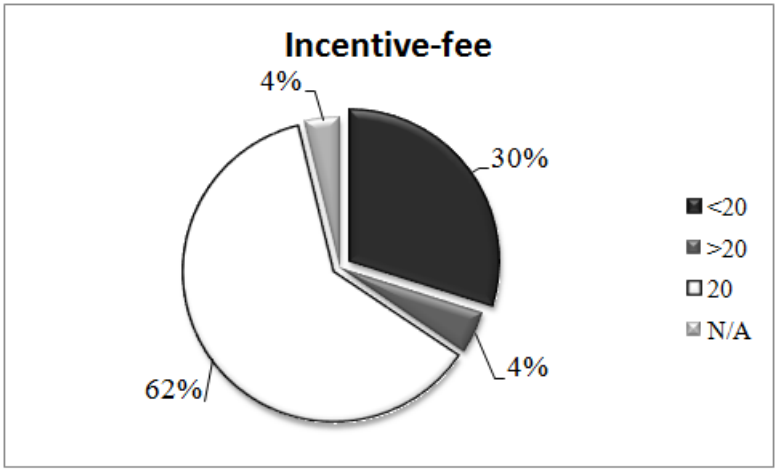

Figure 4.3: Pie chart defining the percentage of the 4 classes of the *incentive fee*

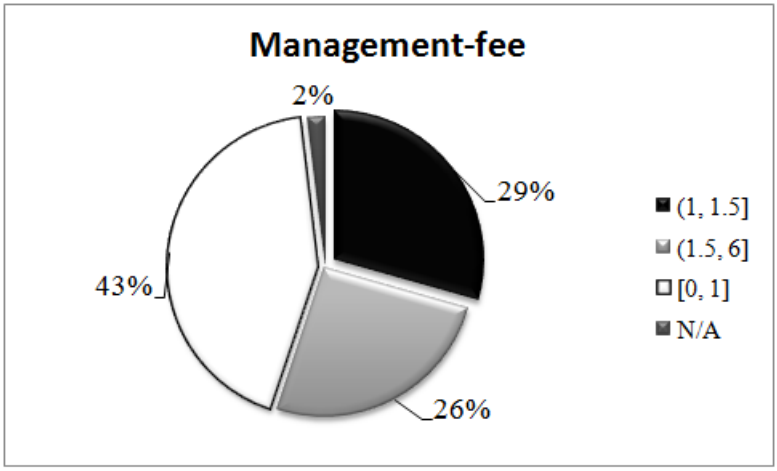

Figure 4.4: Pie chart defining the percentage of the 4 classes of the *management fee*

Similarly, for the variable *management fee*, we split the data into 4 classes (see Fig. 4.4):

- class $[0, 1]$ contains the funds whose management fee is less than or equal to 1%,

- class $(1, 1.5]$ contains the funds whose management fee is greater than 1% but less than or equal to 1.5%,
- class $(1.5, 6]$ contains the funds whose management fee is greater than 1.5% but less than or equal to 6%,
- N/A.

In these two variables, N/A corresponds to missing data.

The variable *age*

To calculate the age of a hedge fund, we use the participation time which is the time elapsed between the first date of reporting (beginning of study) and the last date of reporting (or the date of end of study). We created this variable in order to have the proportion of failure in each range. We classified *age* in three classes (see Figure 4.5):

- Less than 5 years old,
- between 5 to 10 years old,
- between 10 and 30 years old.

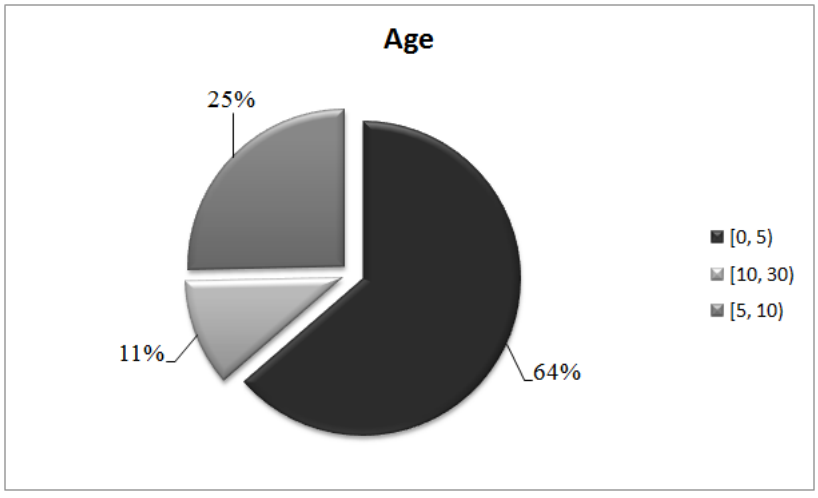

Figure 4.5: Pie chart defining the percentage of each *age*

The variable *size*

PerTrac [58] recently released an interesting study investigating the impact of size and age on hedge fund performance. They classified hedge fund sizes in three main categories:

- Small: Less than \$100 million (M) in assets under management (AUM)
- Mid-size: Between $\$100 - 500$ M in AUM
- Large: Greater than \$500 M in AUM

This classification is used to split our data in 3 classes as shown in Figure 4.6.

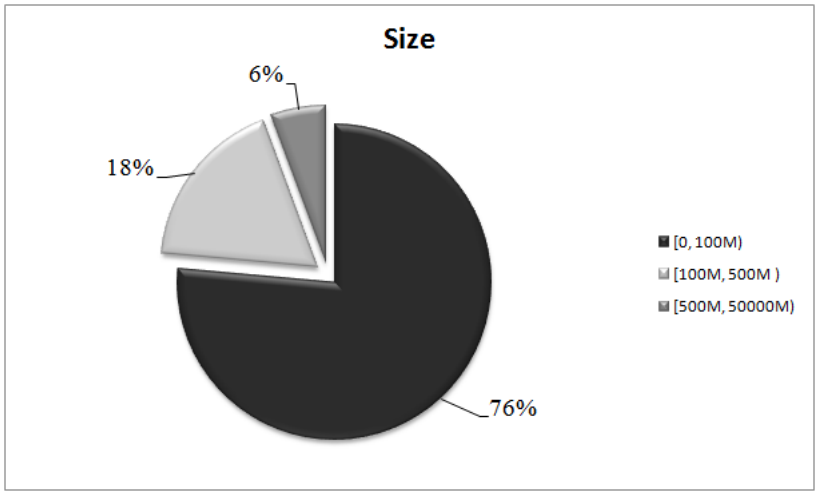

Figure 4.6: Pie chart defining the percentage of each *size*

The variable *minimum investment*

The minimum investment is a restriction imposed by a fund on new investors. To test the impact of minimum investment on the lifetime, we split the data into two groups: a *low minimum investment level group* consisting of funds having a minimum investment level of less than $\$250,000$ and a *high minimum investment level group* consisting of funds having a minimum investment level of $\$250,000$ or more (see Figure 4.7). This deviates from the approach commonly used in the literature where unadjusted minimum investment level is treated as a continuous variable. In our opinion, the minimum investment resembles more closely a categorical variable with a small number of levels.

Hence, treating it as a continuous variable will make its effect on performance extremely susceptible to outliers.

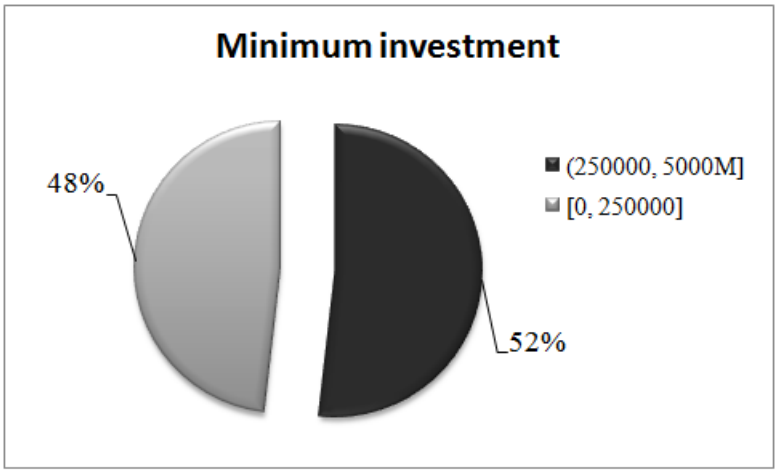

Figure 4.7: Pie chart defining the percentage of each class of the *minimum investment*

4.2.2 *Assets* and *Performance* tables

The *Assets* and *Performance* tables contain time dependent variables. In these tables, several hedge funds backfilling their old performances have been dropped. After a hedge fund is created, it may enter the database and begin to report its performance. However, other funds enter the database only several months after their *inception date* and refill their history with past performance. Sometimes they even report performances anterior to their *inception date*. These performances must be dropped off from the sample because the first date of report is prior to the date of entry in the database, but not to the *inception date*, which could bias our results.

4.3 Hedge funds lifetimes

To define the *lifetime* of a hedge fund, we used the variables *inception date* and *fund status*. We also used the 12 variables describing the dates in our data such as the minimum and maximum reporting dates for the *assets*, the *performance*, and the *allocations* (*allocation* is a variable of *Region* and *Instruments* table). There are also date of latest *audit* , date of latest *fund asset size*, date of latest *firm asset size*, and the date of

liquidation. Note that every hedge fund may have each of these 12 dates. The variable *inception date* represents the day on which a fund started trading. The variable *fund status* indicates the reason for which the hedge fund was removed from the database. In the HFR database, funds exit for three reasons: they are liquidated, they are closed to new investors or they simply stopped reporting to HFR. In our study, the starting point from which the hedge fund is observed is the first reporting date. This corresponds to the minimum $\min(T_i)_{i=1,2,...,12}$ of all the 12 dates reported in our database. The date of the end of study is June 2005. So, we are in the presence of a left truncation and a right censoring of type I.

We define $T_1, T_2, ..., T_{12}$ as the different dates of reporting of a hedge fund. More precisely, the truncation time, the participation time, the lifetime and the censoring time are defined as follows:

Truncation time

All the hedge funds for which the minimum reporting date $\min(T_i)_{i=1,2,...,12}$ is strictly greater than the inception time T_0 will be left truncated. So, if T_r represents the truncation time, then

$$T_r = \min(T_i)_{i=1,2,...,12} - T_0 \tag{4.1}$$

Participation time

Before defining the participation time, it is important to notice that all the hedge funds for which the status indicates *Dead* have their maximum reporting date (which is the date of death) strictly less than our date of end of study (June 2005), and the hedge funds without an indicated status may have a maximum reporting date strictly greater than June 2005. Therefore, the time elapsed between the first date of reporting and the last date of reporting (or the date of end of study T_f) is the *participation time* denoted by T_p which will be calculated as the difference between the maximum reporting date (or the date of end of study T_f) and the minimum reporting date.

$$T_p = \begin{cases} \max(T_i)_{i=1,2,...,12} - \min(T_i)_{i=1,2,...,12} & \text{if } \max(T_i)_{i=1,2,...,12} < T_f \\ T_f - \min(T_i)_{i=1,2,...,12} & \text{if } \max(T_i)_{i=1,2,...,12} \geq T_f \end{cases} \tag{4.2}$$

Lifetime

In our study, the *lifetime* of a hedge fund, denoted by X is defined as the time elapsed from the date of creation (*inception date*) until the release date of the database or the end of the study. So, it is the time of *truncation* plus the time of *participation* .

$$X = T_r + T_p \tag{4.3}$$

Figure 4.8 describes the lifetime of a hedge fund.

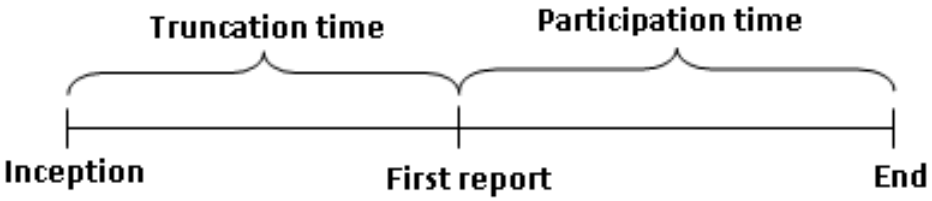

Figure 4.8: Description of the lifetime of a hedge fund

Censoring time

All the hedge funds for which the maximum reporting date is equal or greater than June 2005 will be censored. So, we have attributed the status "0" to all the censored hedge funds and the status "1" (i.e. dead) to all the hedge funds with a maximum date of reporting strictly less than June 2005.

$$\delta = \begin{cases} 1 & \text{if } \max(T_i)_{i=1,2,\ldots,12} < \text{June 2005} \\ 0 & \text{if } \max(T_i)_{i=1,2,\ldots,12} \geq \text{June 2005} \end{cases} . \tag{4.4}$$

Our final *Administrative* table contains the following variables: fund id, truncation time (Tr), participation time (Tobs), delta, strategy, size, age, fund denomination, leverage, HighWatermark, fundoffshore, annualaudit, incentivefee, managmentfee, minimum investment. Table 4.1 presentes a brief overview of this table.

4.4 Descriptive statistics

The data studied in this thesis comprise 6947 hedge funds, 2403 living funds and 4544 dead funds. Therefore, the mortality rate is estimated to be 65.41% during the period

Fund ID	**Tr(months)**	**Tobs(months)**	**Delta**	**Strategy**	**Size**	**Age**	**...**
4	114	195	1	event-driven	[100M, 500M)	[10, 30)	...
7	22	103	1	relative value	[100M, 500M)	[5, 10)	...
16	22	61	1	Others	[0, 100M)	[5, 10)	...
18	0	116	0	fund of funds	[100M, 500M)	[5, 10)	...
32	1	55	1	Long/short	[0, 100M)	[0, 5)	...
...	...	...	...	...	...	...	...

Table 4.1: The new Administrative table

from September 1963 to June 2005. Some variables greatly influence this mortality rate. In order to determine if a variable may be significant (variables that greatly affect the rate of mortality of hedge funds), we calculated the mortality rate of hedge funds for all the classes describing each variable. A variable may be significant if the rate of mortality differs significantly from one class to another. While such methods as the chi-squared or Fisher's exact test could be considered for further exploratory analysis, since they do not take into account truncation and censorship, it would not do much to improve our intuition in selecting important covariates. Therefore the preliminary analysis presented is considered sufficient to give a first idea of which variables could have a significant effect on the lifetime of our data.

Table 4.2 summarizes the variations of the mortality rate for each category for a given variable and specifies if the variable may be significant or not. For the variable *strategy* for example, the rate of mortality greatly depends on the type of strategy used to manage the hedge funds. The mortality rate for the *Event Driven* strategy is 55.64% whereas for the *Fund of funds* strategy is 73.64%. Therefore, due to this important variation, the variable *strategy* is a significant variable that will be taken into account in our study. As it can be observed in the table, all the variables have an effect on the lifetime. However, there are some that are strongly significant such as: *strategy, size, age, high watermark, annual audit, incentive fee, management fee.*

The *high watermark* describes the highest net asset value previously seen at the end of the fiscal year. It ensures that the manager does not get paid large sums for poor performance. Therefore, if the manager loses money over a period, he or she must get the fund above the *high watermark* before receiving a performance bonus. Our variable *high watermark* specifies if the fees are taken after a high-water mark or not. We note that the mortality rate of hedge funds without a high watermark (No) is 76.04%, versus

VARIABLES	CLASSES	MORTALITY RATE(%)	SIGNIFICANCE
strategy	Event driven	**55.64**	strong
	Long/Short	61.47	
	Relative value	60.82	
	Tactical	65.94	
	Location	67.72	
	Fund of funds	**73.64**	
	Others	71.84	
minimum investment	$[0, 250000]$	68.88	weak
	$(250000, 5000M]$	62.17	
age	$[0, 5)$	67.15	strong
	$[5, 10)$	64.66	
	$[10, 30)$	57.10	
denomination	USD	65.96	weak
	EUR	58.03	
	Others	66.66	
size	$[0, 100M)$	69.74	strong
	$[100M, 500M)$	53.08	
	$[500M, 50000M)$	45.92	
leverage	Yes	67.63	weak
	No	64.10	
high watermark	Yes	**63.15**	strong
	No	**76.04**	
fund offshore	Yes	64.03	weak
	No	67.23	
annual audit	Yes	63.74	strong
	No	**83.56**	
incentive fee	$< 20\%$	72.02	strong
	20%	61.08	
	$> 20\%$	69.16	
	N/A	79.38	
management fee	$[0, 1]$	68	strong
	$(1, 1.5]$	62.81	
	$(1.5, 6]$	62.41	
	N/A	87.69	

Table 4.2: Variations of the mortality rate for each category in a variable

63.15% for those with a *high watermark* (Yes). This means that when a hedge fund takes into account the *high watermark* before paying manager, it performs better. An annual audit is an annual review of the financial records of an organization, such as a hedge fund. In our data, the death rate among hedge funds for which annual audits were not performed was 83.56%.

Table 4.3 is a three dimensional contingency table with categorical variables *strategy, size* and *age*. It defines the mortality rate for each *size* and each *strategy* and for a given strategy and a given size. It also defines the mortality rate for each category of *age*. The mortality rate is calculated as the ratio between the number of deaths for a given category and the total number of hedge funds invested in the category. In addition, this table allows us to determine the mortality rate of each strategy for a given size. For example, the mortality rate for the *Event Driven* strategy is 61.58% for all hedge funds with a size between $[0, 100M)$ (see the lines *total*). The last column of the table summarizes the mortality rate by age of each strategy and also the mortality rate of each strategy (the intersection between this column and the lines *total*), and the last line summarizes the mortality rate by size.

This table shows that the mortality rate of large size funds which are less than 5 years old and which use a strategy classified in the category *Others* is 0%. This result appears like a perfect rate but it is a poor estimation due to the outlier because in this class, we only have one hedge fund whose lifetime extends from the beginning to the end of the study (see Figure 4.9). As it can be observed in the Figure 4.9, more than 80% of the hedge funds have small sizes. Therefore, this category will be not considered in our analysis. The same thing happens for the larger funds with age range in $[10, 30)$ using the *Location* strategy. Table 4.3 allows us to conclude that the highest mortality rates are obtained for:

- Hedge funds using fund of funds strategies with a mortality rate of 73.63%. For this strategy, regardless of the size and the age range, the mortality rate is high, as shown in Figure 4.10.
- The hedge funds with a small size (i.e smaller than $\$100M$) have a mortality rate of 69.74%.
- The mortality rate of hedge funds which are less than 5 years old is between 59.06% and 75.18%.

The lowest mortality rates shown in our contingency table are:

Strategy	**Age**	**Size**			**% TOTAL**
		$[0, 100M)$	$[100M, 500M)$	$[500M, 50000M)$	
Event-Driven	$[0, 5)$	64.13	49.25	**25.00**	**59.06**
	$[5, 10)$	55.67	45.45	50.00	52.77
	$[10, 30)$	60.60	34.37	41.66	46.75
	`total`	**61.58**	44.69	38.09	55.63
Fund of funds	$[0, 5)$	76.20	70.87	73.80	**75.18**
	$[5, 10)$	78.81	70.19	71.42	75.57
	$[10, 30)$	63.21	44.44	55.00	56.57
	`total`	75.79	67.32	69.36	**73.63**
Long/Short	$[0, 5)$	66.12	49.72	30.55	63.021
	$[5, 10)$	65.25	46.22	40.00	60.45
	$[10, 30)$	62.90	39.53	22.72	52.91
	`total`	65.68	47.30	30.76	61.47
Location	$[0, 5)$	78.94	45.71	37.50	72.53
	$[5, 10)$	68.49	**33.33**	46.15	58.18
	$[10, 30)$	70.96	50	0.00	65.78
	`total`	75.51	41.53	40.90	67.71
Relative value	$[0, 5)$	64.87	49.54	44.82	60.46
	$[5, 10)$	71.18	43.47	31.81	57.89
	$[10, 30)$	75.51	81.25	47.36	70.23
	`total`	67.22	50.00	41.42	60.82
Tactical	$[0, 5)$	70.31	38.46	40.00	67.19
	$[5, 10)$	74.8	56.36	**16.66**	68.42
	$[10, 30)$	65.48	52.27	40.54	57.73
	`total`	70.93	49.00	33.84	65.93
Others	$[0, 5)$	75.51	20.00	**0.00**	73.20
	$[5, 10)$	72.22	100.00	100.00	74.57
	$[10, 30)$	62.5	50.00	100.00	60.60
	`total`	73.33	50.00	75.00	71.83
TOTAL		**69.74**	53.08	45.91	**65.41**

Table 4.3: Three dimensional contingency table describing the relation between the variables *age*, *size* and *strategy*

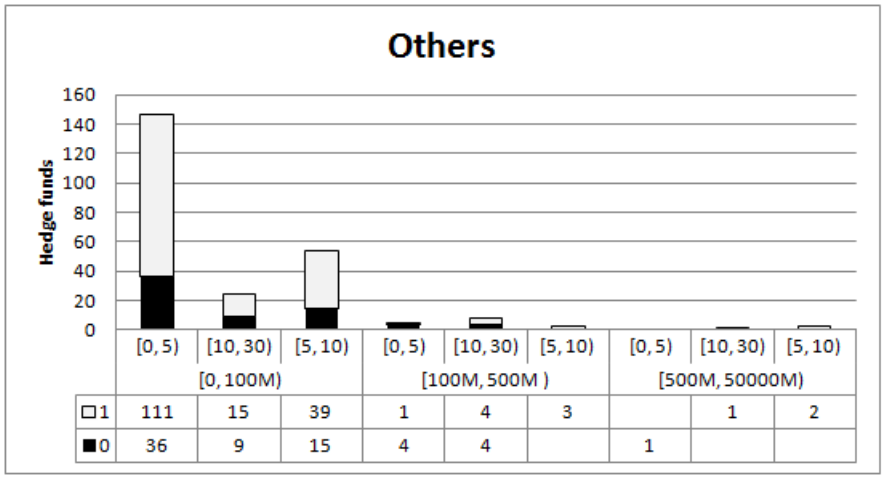

Figure 4.9: Distribution of dead and alive hedge funds in the strategy *Others* as a function of the *size* and *age*.

- The mortality rate of large fund with age in the range $[5, 10)$ using the *Tactical* strategy is 16.66%;
- The mortality rate of large and young (less than 5 years old) hedge funds using the *Event-Driven* strategy is 25%;
- The mortality rate of medium funds with age in the range $[5, 10)$ using the *Location* strategy is 33.33%.

These results are surprising compared to the mortality rate of all our data which is 65.41%.

To describe time varying covariates, we chose 2 hedge funds who reported their assets virtually at the same periods from 1994 to 2005. During this period, one died and the other was still alive at the end of study. Figure 4.11 allows us to visualize the trends over this period of assets. The line with the square symbols describes the trends of the dead fund, while the line with the diamond symbols describes the trends of the fund alive. The two hedge funds are medium funds with age range in $[10, 30)$, having their minimum investment in the large range $(250000, 5000M]$. They used leverage, had a high watermark and audited their funds during the fiscal year. However, while the dead fund used the strategy *Others* to manage, the alive fund used the strategy *Tactical* to manage. We see that for the alive fund, the trend appears to be increasing over a reasonably long period and attains its high watermark in 2004 before changing direction with a small slope

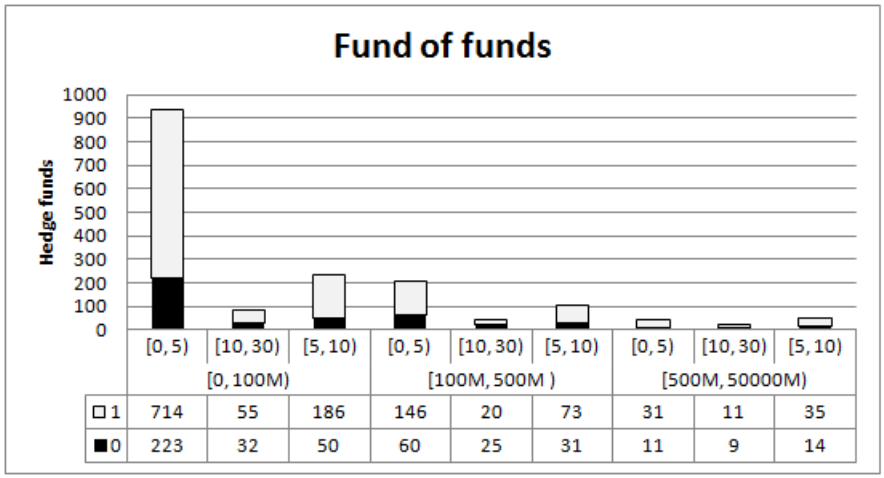

Figure 4.10: Distribution of dead and alive hedge funds in the strategy *Fund of funds* as a function of the *size* and *age*.

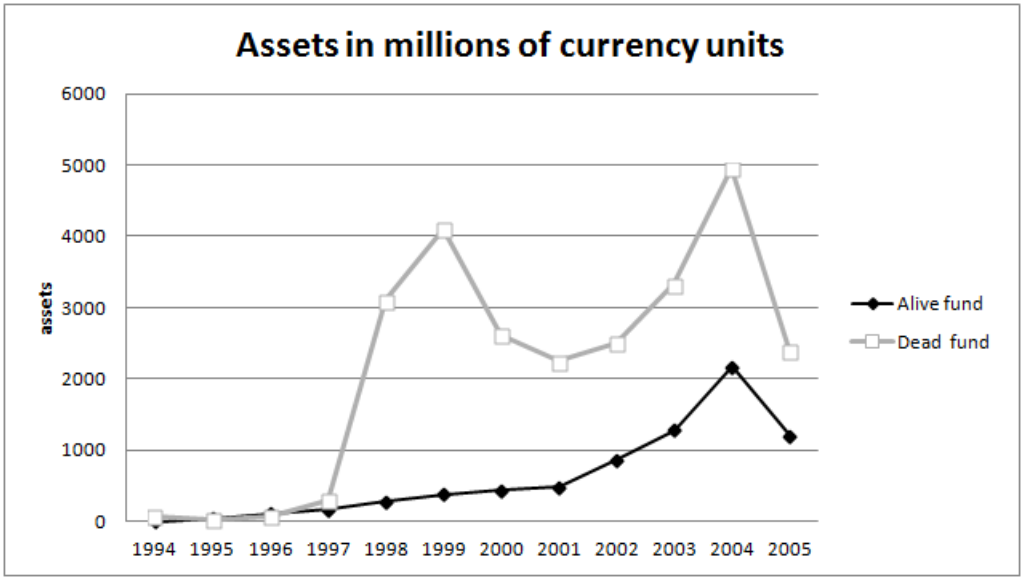

Figure 4.11: Trends over time of assets for two funds

down, while for the dead fund, the trend fluctuates during the observed period, attains its high watermark in 2004 and changes direction with a high slope down. Figure 4.12 presents the trends over this same period of performance of our two funds. It can be observed that the living fund has a better performance than the dead fund during the

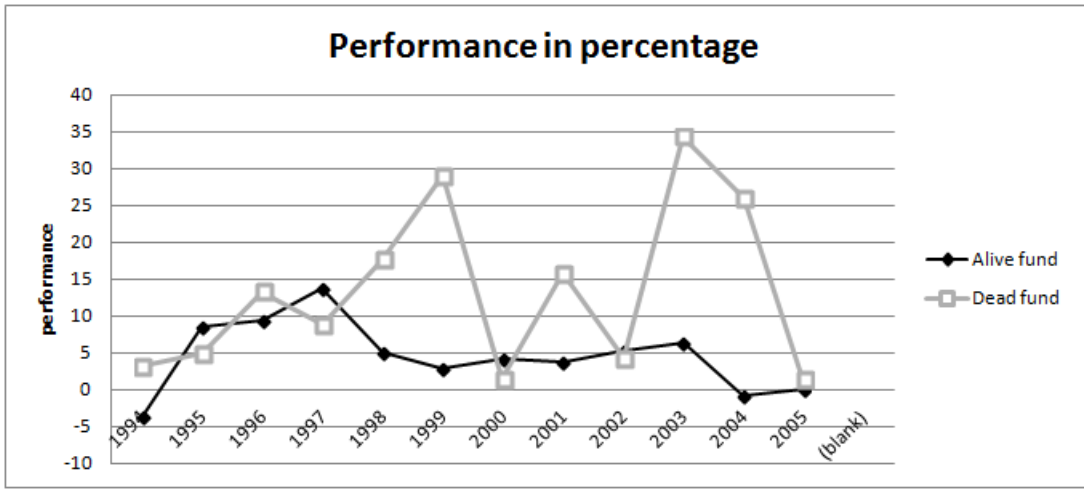

Figure 4.12: Trends over time of performance for two funds

observed period.

4.5 Summary

In this chapter, we presented an overview of our data and how we selected, organized and processed them, so that they are adapted to survival analysis and interpretation. In fact, using basic statistical tools, we described and identified the lifetime of hedge funds. After having constituted our three main tables (*Administration*, *Assets* and *Performance*) the study of the covariates allowed us to classify them into different categories. Therefore, the covariates such as *age, size, strategy, minimum investment, incentive fee* and *management fee* were created or reclassified. This chapter also allowed us to get a first insight on failure rate in each category using basic statistical tools. Consequently, the covariables such as age, size, strategy, annual audit and high watermark were identified as those which may strongly affect the lifetime of a hedge fund.

In the next chapter, we will use more complex statistical methods and the R programming language, to find a suitable model for the distribution of the lifetime of a hedge fund and identify among the hedge fund characteristics those which are likely to affect its lifetime.

Chapter 5

Methods and Results

5.1 Methods

In this chapter, we examine various survival analysis methods for analyzing our data. These methods are implemented using R. More precisely, we are interested in finding a suitable model for the lifetime distribution of a hedge fund and identify among the hedge funds characteristics those which affect this lifetime. Below is a list of methodologies which are presented in this chapter:

- we use the nonparametric methods of Kaplan-Meier and Nelson-Aalan to estimate the survival distribution of our data,
- we use the graphical displays to check the appropriateness of the Weibull model (i.e goodness of fit plots),
- we compare the Weibull proportional hazard (PH) model that incorporates only fixed covariates with the Cox PH model as follows:
    1. we estimate the regression coefficients of our data using the selected model
    2. we select and interpret the significant variables from our selected model
- we compare the extended Weibull proportional hazard (PH) model that incorporates only time dependent covariates with the extended Cox PH model as follows:
    1. we estimate the regression coefficients of our data using the selected model
    2. we select and interpret the significant variables from our selected model

- we check if there exists a significant interaction between some fixed covariates and time dependent covariates in the context of an extended Weibull model.

5.2 Non-parametric analysis

In this section, we review the nonparametric method of inference concerning the survival function $S(x) = P(X > x)$ to find the best fitting probability model for our data (x is the lifetime of a hedge fund given by equation (4.3)). We use the survival package in R to create the survival object and plot the Kaplan-Meier (K-M) estimate of S (see the code and table in Appendix B.1). Graphing the survival function is important because it provides valuable insight about the data (see Figure 5.1). The plot shows the K-M estimate with 95% confidence bands. These confidence bands are obtained using the Greenwood's formula (see equation 3.43) which gives the estimated $\widehat{S(x)}$. Note that

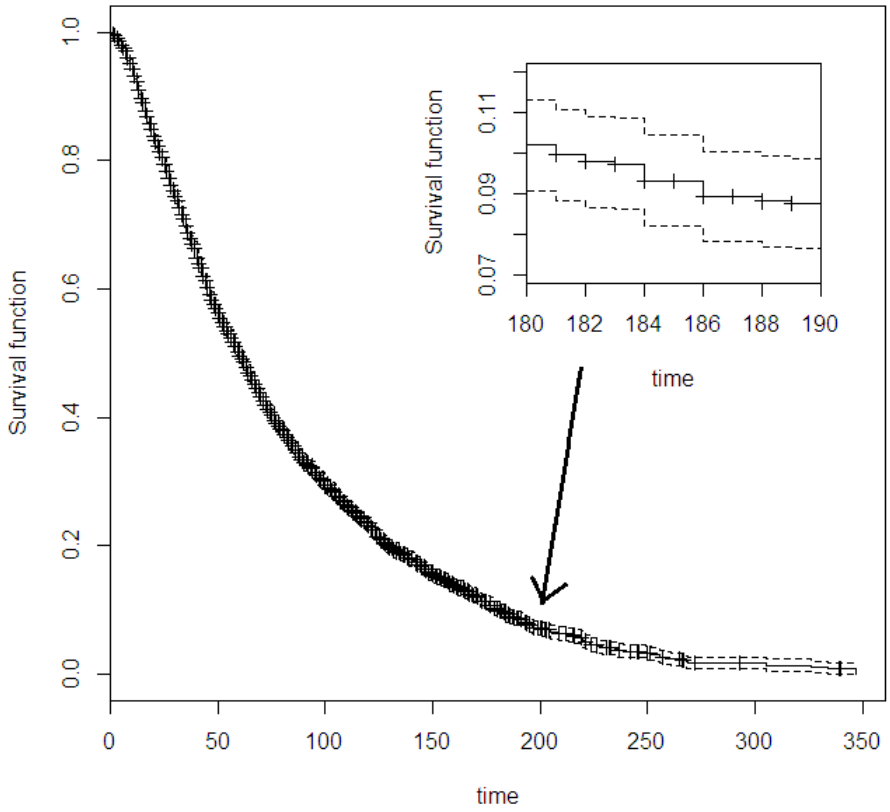

Figure 5.1: Kaplan-Meier estimate with 95% confidence bands.

the bands are quite narrow, due to the size of the sample, which would provide reliable estimation barring issues of sampling bias. The survival curve is drawn as a step function, so the proportion of surviving hedge funds remains unchanged between the events, even if there are some intermediate censored observations.

An estimate of the distribution mean is useful in many contexts, but for lifetime data an estimate of the quantile t_p of the distribution is usually of more interest. The median is often used as a *measure of location* or *central tendency.*(see equation 3.46 for the standard error formula)

	time (months)	**standard error**	**Interval**
1^{st} quartile	30	0.5667	(28.89, 31.11)
Median	60	1.0319	(57.98, 62.0226)
3^{rd} quartile	115	1.9449	(111.19, 118.81)

Table 5.1: The quartiles of survival estimate

Table 5.1 gives the quartiles of our hedge fund lifetime data (X) along with their standard error and an approximate 95% confidence interval. We conclude from this table that:

1. 25% of hedge funds died by the age range (28.89, 31.11),
2. 50% of hedge funds died by the age range (57.98, 62.0226) ,
3. 75% of hedge funds died by the age range (111.19, 118.81).

The Kaplan-Meier estimate that we computed in this section allow us to estimate the survival function $S(x)$ from our sample. Its advantage is that it does not depend on any parametric assumptions on the underlying probability distribution of the data. However, its disadvantage is that it does not take into account the covariates. In the next section, we use the parametric analysis to choose the best distribution model for our data.

5.3 Parametric analysis

To examine the underlying distribution of the data, the simplest model assessment procedure is to fit various parametric models to the data and compare visually how similar

the corresponding survival functions are to the Kaplan-Meier and the Nelson-Aalen estimates. However, the fact that our data has the right censoring and left truncation will limit the usefulness of such a comparison. In the literature, there is a strong tendency to use models that are mathematically or computationally convenient. To a large extent this accounts for the extensive use of models based on the log-normal, log-logistic and Weibull distributions. As the number and complexity of fixed covariates increase, the emphasis on the shape of the distribution is usually much reduced, the primary focus being on the location and dispersion aspects of X or $\log(X)$ (see [33] (pages 16-38) for more information).

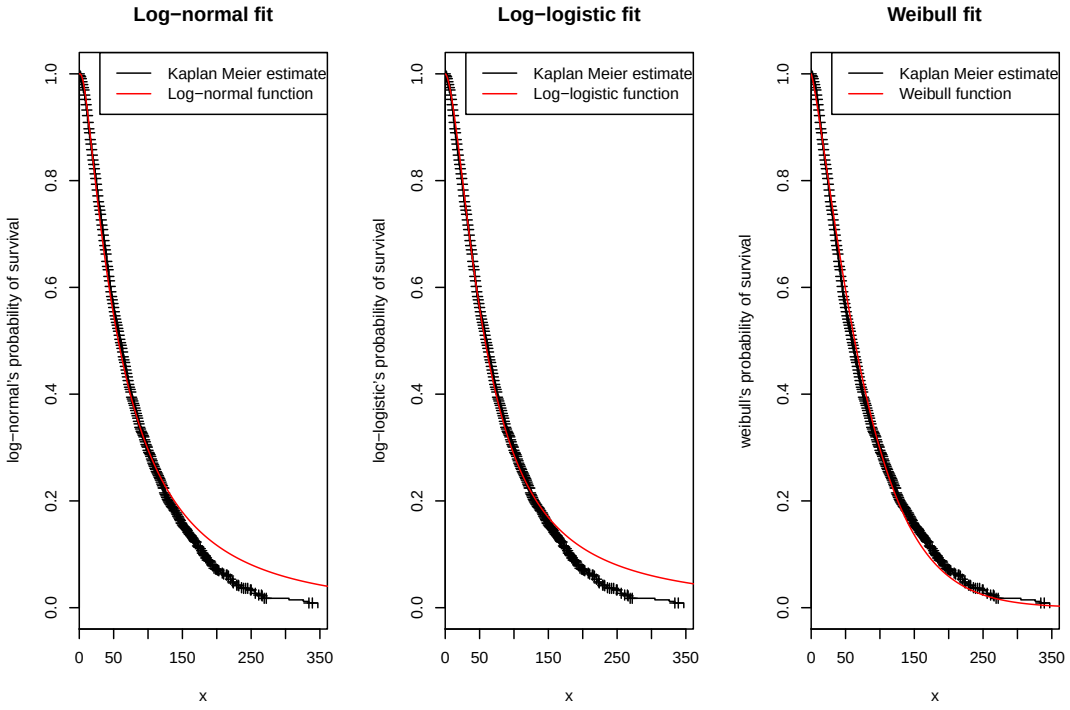

Figure 5.2: The Kaplan-Meier estimate versus some parametric estimates of the survival function

Figure 5.2 (see Appendix B.2 for the R code) shows the plots of the Kaplan-Meier estimate $\widehat{S(x)}$ of the survival function $S(x)$, together with the estimates $S(x;\widehat{\theta})$ of the survival function obtained using the log-normal, log-logistic and Weibull models. As it can be seen from the plots of this survival function, these three models fit well our data from the beginning of the study until 150 months, but not so well after this time. Weibull

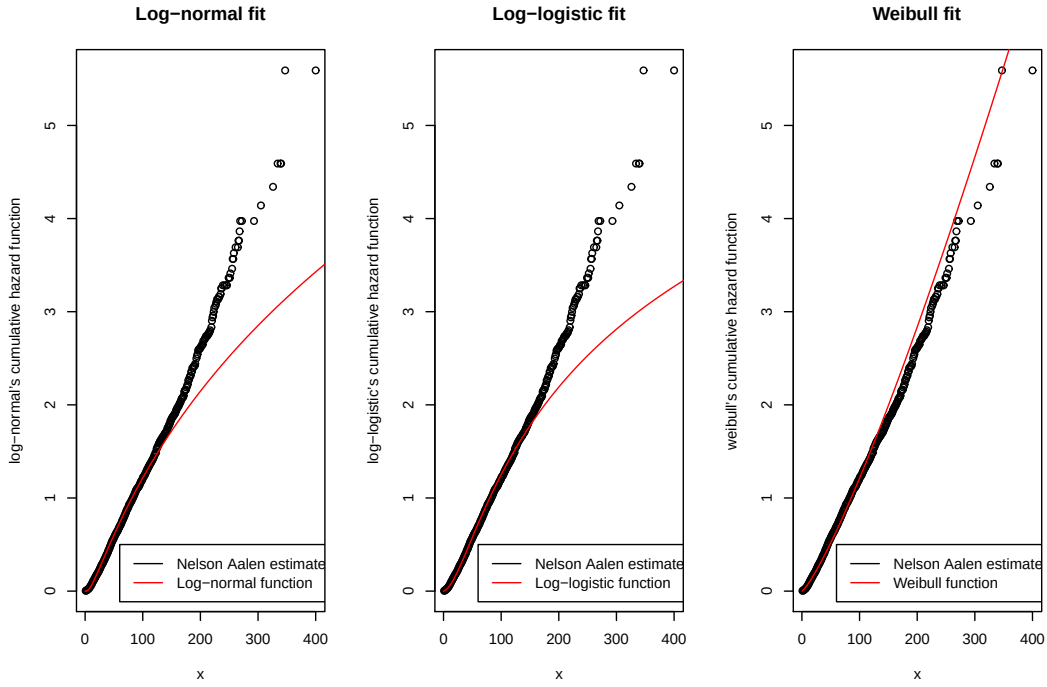

Figure 5.3: The Nelson-Aalen estimate versus some parametric estimates of the cumulative hazard function

appears to give the best fit compare to the two other models.

Although different survival functions can have the same basic shape, their hazard functions can differ dramatically. The hazard function usually gives more information about the underlying mechanism of failure than the survival function. For this reason, modelling the hazard function is an important method for analyzing survival data. Figure 5.3 shows the plot of the Nelson-Aalen estimate, together with the estimates $H(x, \widehat{\theta})$ of the cumulative hazard function obtained using the log-normal, log-logistic and Weibull models. As compared to the survival function plot (see Figure 5.2), this one allows us to visualize the large lag between the log-normal and log-logistic models. Weibull seems to fit well as compared to the two others and will be used to analyse our data. For both Figures 5.2 and 5.3, $\widehat{\theta}$ is an estimate of the parameters describing the model (see Table 5.2)

To assess the goodness-of-fit of the Weibull model, we used the graphical method. However, since we have a lot of categorical covariates with many groups, the best way to assess our model is to compare the distribution of the Cox-Snell residuals (based on

Models	$S(x)$	$\widehat{\theta}$	**Median(years)**
Log-normal	$1-\mathbf{\Phi}\left(\frac{\log(x)-\mu}{\sigma}\right)$	$\hat{\mu}=4.04$ $\hat{\sigma}=1.06$	4.73
Log-logistic	$(1+(x/\alpha)^{\beta})^{-1}$	$\hat{\beta}=1.67$ $\hat{\alpha}=57.92$	4.82
Weibull	$\exp(-(\lambda x)^{\gamma})$	$\hat{\gamma}=1.225$ $\hat{\lambda}=0.0117$	5.28

Table 5.2: Estimates of the model parameters

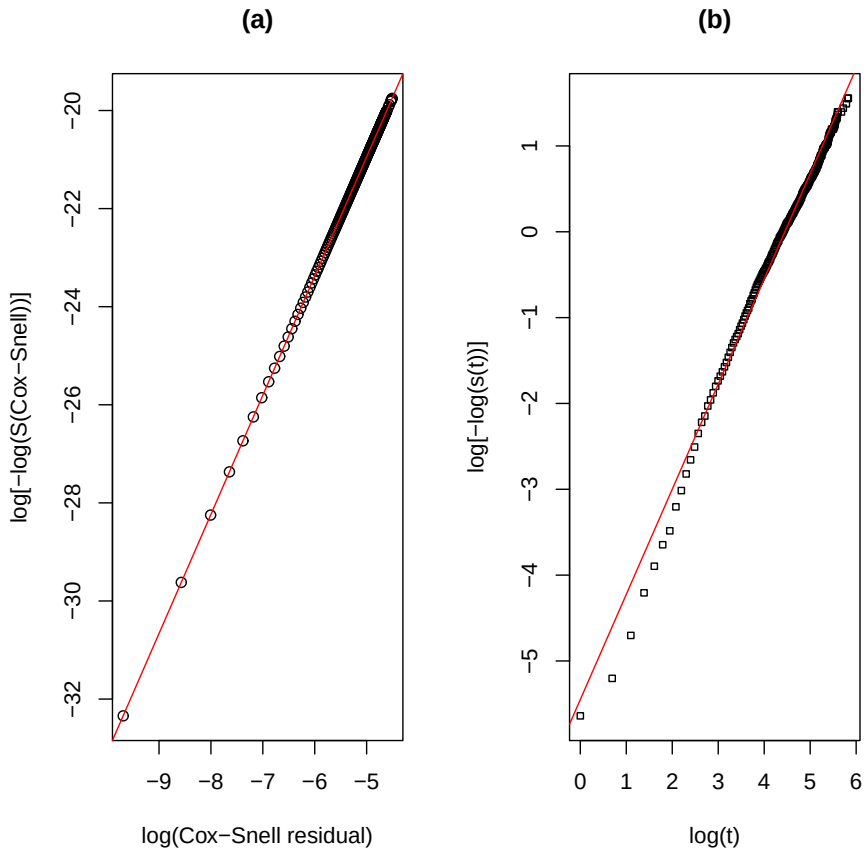

Figure 5.4: Goodness of fit plots, (a) Weibull probability plot of Cox-Snell residuals , (b) Weibull probability plot of hedge funds failure times

the *extreme value* distribution) with a unit exponential distribution. Figure 5.4 (see Appendix B.2 for the R code) shows the Weibull probability plot of the Cox-Snell residuals (rc) on graph (a) and the Weibull probability plot of the hedge funds failure times with a straight line (reference line) on graph (b). The plot of $\log[-\log \widehat{S}(Cox - Snell)]$ versus $\log(Cox - Snell\, residual)$ (graph (a)) is a straight line with unit slope through the origin. We conclude that the Weibull model fits the data well and thus may be considered as the model of our data. The graph (b) allows us to assess the proportionality hazard assumption for the Weibull model and indicates that $\log[-\log \widehat{S}(x)]$ versus $\log(x)$ is roughly linear. Thus, the Weibull distribution assumption seems reasonable.

The visual examination of the different fitted curves and the graphical goodness of fit allowed us to choose the Weibull model as the best model compared to the log-normal and log-logistic models.

5.4 Regression models with fixed covariates: Weibull PH and Cox PH

In this section, we explore the relationship between the lifetime of a hedge fund and the following explanatory variables: *strategy, size, age, minimum-investment, fund-denomination, fundoffshore, laverage, annual-audit, Incentivefee* and *managementfee*. As noted in Section 3.7, the Weibull regression model, (which includes the exponential model), is a special case of both the PH models and the AFT models. In this section, we estimate the regression coefficients of our data using the Weibull PH model and compare these estimates with those found using the Cox PH model. Finally, we will select the best model.

In a proportional hazard model, the unique effect of a unit increase in a covariate is multiplicative with respect to the hazard rate. In this class, we consider two broad classes of regression models: Weibull PH model and Cox PH model. Cox model is one of the finest techniques in identifying combined effects of several covariates on the relative risk (hazard). This model assumes that the hazards of the different strata formed by the levels of the covariates are proportional. As explained in Subsection 3.7.2, the Weibull PH model is a particular case of a PH regression model. Our aim is to compare the estimates of the slope of the covariate in the proportional hazards model using the parametric Weibull model and the semi-parametric Cox proportional hazards model.

Table 5.3 summarizes the results obtained using the Cox PH model and the Weibull

Variables		**Cox PH model**			**Weibull PH model**		
		Coef	**S.E.**	**Wald p**	**Coef**	**S.E.**	**Wald p**
Strategy	Event-Driven	0	(reference)		0	(reference)	
	fund of funds	0.417	0.076	0.000	0.374	0.076	0.000
	Long/Short	0.113	0.065	0.081	0.111	0.064	0.086
	location	0.227	0.087	0.009	0.225	0.087	0.010
	Others	0.186	0.097	0.056	0.176	0.097	0.071
	Relative value	0.127	0.074	0.088	0.137	0.074	0.065
	Tactical	0.057	0.071	0.423	0.085	0.071	0.227
Min Invest	[0;250000]	0	(reference)		0	(reference)	
	(250000;5000M]	0.033	0.032	0.308	0.023	0.032	0.474
Size	[0, 100M)	0	(reference)		0	(reference)	
	[100M, 500M)	-0.480	0.043	0.000	-0.458	0.043	0.000
	[500M, 50000M)	-0.777	0.078	0.000	-0.740	0.078	0.000
Age	[0, 5)	0	(reference)		0	(reference)	
	[5, 10)	**-18.955**	**156.481**	**0.904**	**-2.243**	**0.042**	**0.000**
	[10, 30)	**-36.925**	**223.474**	**0.869**	**-3.857**	**0.068**	**0.000**
Denomination	EUR	0	(reference)		0	(reference)	
	USD	0.012	0.064	0.852	0.018	0.064	0.779
	Others	0.275	0.091	0.002	0.280	0.091	0.002
Leverage	No	0	(reference)		0	(reference)	
	yes	-0.036	0.033	0.272	-0.022	0.033	0.509
HighWatermark	No	0	(reference)		0	(reference)	
	Yes	0.045	0.043	0.291	0.048	0.042	0.257
Fundoffshore	No	0	(reference)		0	(reference)	
	Yes	0.071	0.034	0.038	0.060	0.034	0.079
Annualaudit	No	0	(reference)		0	(reference)	
	Yes	-0.421	0.053	0.000	-0.389	0.053	0.000
Incentivefee	< 20	0	(reference)		0	(reference)	
	> 20	0.025	0.081	0.756	0.010	0.081	0.906
	20	0.028	0.050	0.571	-0.007	0.050	0.884
	N/A	-0.132	0.093	0.155	-0.102	0.095	0.282
Managementfee	[0, 1]	0	(reference)		0	(reference)	
	(1, 1.5]	0.017	0.037	0.651	0.023	0.037	0.533
	(1.5, 6]	0.033	0.040	0.405	0.037	0.040	0.350
	N/A	0.292	0.125	0.020	0.332	0.127	0.009

Table 5.3: Summary of fixed covariates using Cox PH model and Weibull PH model

PH model using the *coxreg()* and *weibreg()* functions in eha package (see Appendix B.3 for the R code). These first models include all the 10 fixed covariates. The *Coef*, *S.E* and *Wald p-value* column are respectively the estimates of the regression coefficient, the standard error of the estimate coefficient and the test of significance of the estimate coefficient using the Wald statistic for each covariate in each model. This table indicates that for both Cox PH model and Weibull PH model, the standard errors are the same and the coefficient estimates are almost the same, except for the variable *age* for which we notice a significant difference. In fact, both the coefficients and the standard errors in the Weibull model are considerably smaller, and these represent statistically significant regression coefficients. However, the estimates for the semiparametric Cox model are much larger in magnitude but have even larger standard errors, making the estimation unreliable. This may be a computational issue.

For the aim of the comparison between the Cox PH and Weibull PH models, we first use the graph of $\log[\hat{H}[Cox\text{-}Snell\,residual]]$ against $\log(Cox\text{-}Snell\,residual)$ for the two models. Figure 5.5 shows the log of the estimated cumulative hazard function of Cox-Snell residuals against the log of Cox-Snell residuals on graph (a) and the estimated cumulative hazard function of Cox-Snell residuals against the Cox-Snell residuals (see Appendix B.3 for the R code). From this figure, we see that the plot of $\log[\hat{H}(Cox\text{-}Snell\,residual)]$ against $\log(Cox\text{-}Snell\,residual)$ is not really a straight line plot with unit slope and zero intercept. Similarly, the plot of $\hat{H}[Cox\text{-}Snell\,residual]$ against $Cox\text{-}Snell\,residual$ is not really a straight line plot with unit slope and zero intercept. Therefore, the PH assumption seems to be violated. Comparing Figure 5.5 with Figure 5.4, we infer that the Cox PH model is not suitable for our data.

We now compare the two models discussed above using the *Akaike information criterion* (AIC) and the likelihood ratio test (LRT) given by equations (3.64) and (3.63) respectively . Table 5.4 summarizes the AIC and the LRT for the two models. From this table, we see that the Weibull model has the lowest LRT and AIC. So, it is the best regression model for our data. We will use it to select and interpret the fixed covariates (i.e the covariates which do not depend on time).

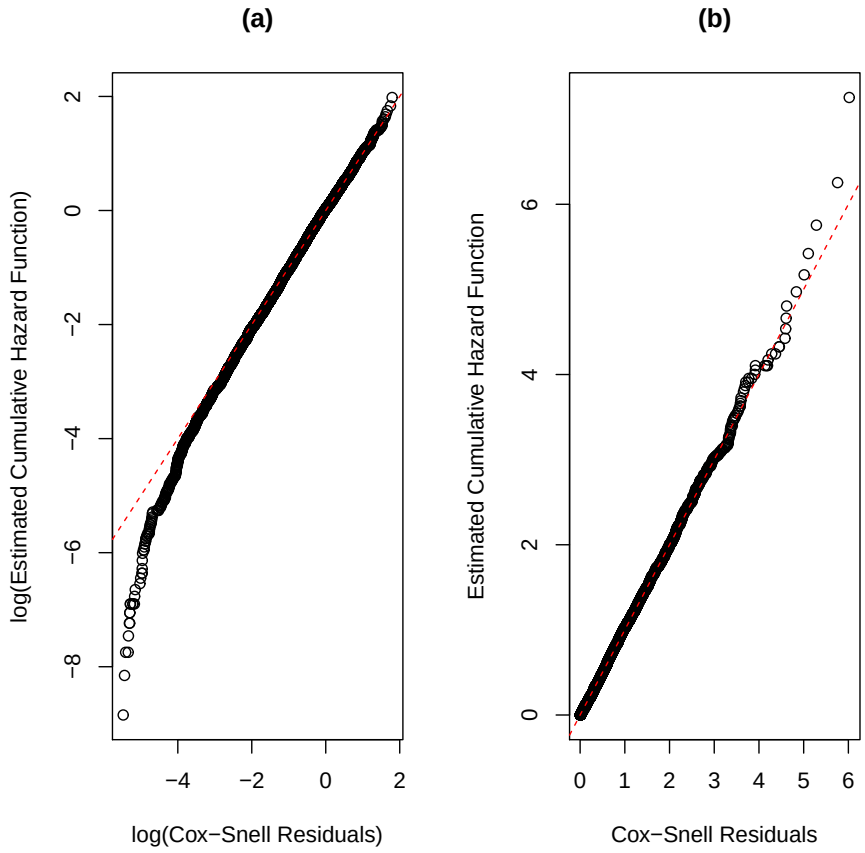

Figure 5.5: Goodness of fit plots: (a) The log of estimated cumulative hazard function of Cox-Snell residuals against the log of Cox-Snell residuals; (b) The estimated cumulative hazard function of Cox-Snell residuals against the Cox-Snell residuals.

5.5 Covariates selection and interpretation of fixed variables

Since there are many covariates in our data, to reduce the modeling bias, we have to select only those which have a significant effect on the lifetime. Covariate selection is a form of

Models	Degree of freedom	Log. likelihood	LRT	AIC
Cox PH model	23	-31700	8262	63445.33
Weibull PH model	23	-21694	5911	43434.37

Table 5.4: Comparison between the Cox PH model and the Weibull PH model

model selection in which the class of models under consideration is represented by a subset of the covariate components to be included in the analysis. An important and challenging task is to efficiently select this subset of significant variables upon which the hazard function depends. There are many variable selection techniques in linear regression models. Some of them have been extended to the context of censored survival data analysis, such as the stepwise deletion using the AIC. We use the *stepAIC()* function in R which consists of building a model by successively adding or removing variables, starting from the full model and deleting the least important variable each time. The procedure ends when no term can be deleted. This procedure allows us to delete the following variables: *incentive fee*, *leverage*, *minimum investment*, *high watermark*, *management fee* which were observed to be not significant in Table 5.3 as shown by the Wald p-value column. Table 5.5 represents the best fitting model for our data (see Appendix B.4 for the R code which was used for producing this table).

Interpreting the Weibull PH model involves examining the coefficients for each explanatory variable. A positive regression coefficient for an explanatory variable or a value of the hazard rate (HR) strictly greater than one means that the hazard is increasing; hence, the survival time is shortened, which gives a worse prognosis. Conversely, a negative regression coefficient or a value of the HR strictly less than one implies that the hazard rate is decreasing; hence, the survival time is lengthened, which gives a good prognosis for a hedge fund lifetime with higher values of that variable. According to the *Wald p-value* column in Table 5.5, *size*, *age* and *annualaudit* are the most significant variables with *p-value*$= 0$.

In the following subsections, we examine separately each explanatory variable.

5.5.1 The variable *size*

In Table 5.5, the smallest size is the reference for the variable *size*. Thus the medium size hazard ratio (HR) is :

$$HR = \frac{h_0(x)\exp(\hat{\boldsymbol{\beta}}_1 \times medium\ size)}{h_0(x)\exp(\hat{\boldsymbol{\beta}}_0 \times small\ size)} = \frac{\exp(\hat{\boldsymbol{\beta}}_1 \times medium\ size)}{\exp(\hat{\boldsymbol{\beta}}_0 \times small\ size)} = \exp(\hat{\boldsymbol{\beta}}_1) \qquad (5.1)$$

Note that the small size has a code 0 and the medium has a code 1. Therefore, $HR = \exp(-0.452) = 0.636$ as obtained in Table 5.5. That means that the risk of death is 0.636 times lower for hedge funds with the medium size when compared to the hedge funds with small size. Similarly, the risk of death is 0.481 lower for the hedge funds with large size when compared to hedge funds with small size. Thus, every $\$1M$ increase in

average AUM decreases the hazard by 36.4% (since $(0.636 - 1) \times 100\% = 36.4\%$) for the medium size funds and by 51.9% for large size funds. This result is similar to what was obtained in literature, see [6, 7, 8, 9, 10, 11, 12].

5.5.2 The variable *age*

The risk of death is 0.106 times lower for hedge funds with age between 5 and 10 years as compared to the hedge funds which are less than 5 years old. The risk of death is 0.021 times lower for hedge funds with age between 10 and 30 years as compared to the funds which are less than 5 years old. Thus, there is a decrease of 89.4% in the expected hazard relative to a one year increase in age for hedge funds with age in the range (5,10] and an 97.9% decrease in the expected hazard relative to one year increase in age for funds with age in the range (10, 30]. This is also similar to what we found in Table 4.2

Covariate		**Mean**	**Coef**	**Exp**	**S.E**	**Wald p**
`Strategy`	Event-Driven	0.088	0	1	(reference)	
	fund of funds	0.223	0.364	1.438	0.065	0.000
	Long/Short	0.290	0.106	1.111	0.064	0.100
	location	0.055	0.217	1.243	0.086	0.011
	Others	0.036	0.183	1.200	0.096	0.057
	Relative Value	0.115	0.134	1.144	0.074	0.068
	Tactical	0.194	0.086	1.089	0.069	0.214
`Size($M)`	[0, 100M)	0.690	0	1	(reference)	
	[100M, 500M)	0.220	-0.452	0.636	0.043	0.000
	[500M, 50000M)	0.090	-0.733	0.481	0.077	0.000
`Age(year)`	[0, 5)	0.320	0	1	(reference)	
	[5, 10)	0.374	-2.246	0.106	0.042	0.000
	[10, 30)	0.306	-3.865	0.021	0.067	0.000
`funddenomination`	EUR	0.049	0	1	(reference)	
	Others	0.028	0.273	1.313	0.090	0.003
	USD	0.923	0.021	1.021	0.063	0.744
`fundoffshore`	No	0.503	0	1	(reference)	
	Yes	0.497	0.058	1.059	0.033	0.080
`annualaudit`	No	0.084	0	1	(reference)	
	Yes	0.916	-0.403	0.668	0.050	0.000
`log(scale)`			3.594	36.370	0.041	0.000
`log(shape)`			0.887	2.427	0.012	0.000

Table 5.5: Regression with the Weibull PH model

and to what was obtained in literature, see [51, 49, 48, 6].

5.5.3 The variable *Annualaudit*

The risk of death is 0.668 times lower for funds who indicated that an annual audit is performed as compared to the funds who did not indicate that an annual audit was performed. Therefore, every performed audit decreases the hazard by 33.2%. This result is not surprising since it can be expected that hedge funds performing an annual audit are more controlled and should perform better. This conclusion is consistent with Table 4.2 where we saw that 83.56% of funds who did not perform the annual audit died but only 63.74% from those who performed the annual audit died.

5.5.4 The variable *Strategy*

All the categories of the variable *strategy* have their regression coefficients positive when compared to the strategy *Event-driven*. This means that they affect negatively the lifetime of funds. *Fund of fund* seems to be the worst strategy to use, since the risk of death using this strategy is 1.438 times higher than when using the *Event-driven* strategy. This result is consistent with what was found in Table 4.2 (i.e *Fund of fund* has the highest the mortality rate of $73,64\%$). The *location* strategy is the second worst with a risk of death 1.243 times higher than *Event-driven*. The other strategies have their p-values greater than 0.05, which means they do no affect negatively the lifetime of fund. In particular, the strategies *tactical* and *long/short* with 1.089%, respectively 1.111% higher than *Event-driven*.

5.5.5 The variable *Funddenomination*

In our data, 88% of hedge funds use the US dollar currency, 7% the Euro currency and 5% use another denomination. Using Table 5.5, we see that the risk of death for hedge funds using the US dollar currency is 1.021 times higher than for those with Euro currency.

5.5.6 The variable *Fundoffshore*

In our data, 57% of hedge funds are offshore funds (with a mortality rate of 64.03%) and 43% of them are non-offshore hedge funds (with mortality rate of 67.23%). From

Table 5.5, the risk of death for the offshore funds is 1.059 times higher than the non-offshore hedge funds. This result is not really surprising since when a hedge fund goes offshore, it has significant privacy benefits, its assets are not tracked by U.S. regulation and the investment is not subject to U.S taxation. This means that the gains are not taxed.

5.6 Regression models with time dependent covariates

In this section, we examine the combined effects of the time dependent covariates *performance* and *assets* using the extended Weibull PH and the extended Cox PH models and we choose the best estimation. As in the PH model, the extended model contains a baseline hazard function $h(x)$ which is multiplied by an exponential function. However, in the extended model, the exponential part contains the time dependent covariate denoted by $z(x)$ (see equation 3.62). The most important feature of the hazard ratio (HR) is that the proportional hazard assumption is no longer satisfied when using the extended model, since it is a function of time.

In R, it is easy to work with covariates which do not depend on time. However, working with time-dependent covariates in R is an exercise in organization, since there is no function in R that will directly accept time-dependent covariates. The most common way to solve the problem is to use the counting process format. To illustrate how this works, we consider the example of the hedge fund with *fund_id* 4 (see Table 5.6).

Hedge fund	**Start**	**Stop**	**Status**	**Assets**	**Performance**
4	167	168	0	0.112	0.0186
4	168	169	0	0.112	0.0129
4	169	170	0	0.112	0.0094
4	170	171	0	0.115	0.0049
4	171	172	0	0.115	0.0002
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.

Table 5.6: Example of time dependent covariate in data frame

For this hedge fund, the monthly report of *asset* and *performance* varies and is treated respectively as 0.112 and 0.0186 over the interval (167, 168], etc. The intervals are opened at left and closed at right, which means that the *asset* and *performance* are taken to be respectively 0.112 and 0.0186 at 168 months after the inception date. The variable *Status* describes whether or not each interval ended in an event. This explains how we use R to make time-dependent variables. When the data is in this format, we can use the function *coxph()* in R for the extended Cox PH model and the function *weibreg()* for the extended Weibull PH model. These functions interpret our data as left-truncated. Table 5.7 is a summary of time-dependent covariates using the extended Cox PH model and the extended Weibull PH model. The columns *Coef*, *Exp(Coef)*, *S.E.* and *Wald p* in this table give the estimates of the regression coefficients of *assets* and *performance*, the exponential of these coefficients, their standard errors and the p-values of the Wald test.

Variables	**Cox PH model**			**Weibull PH model**		
	Coef	**S.E.**	**Wald p**	**Coef**	**S.E.**	**Wald p**
`Performances(%)`	-0.034	0.002	0.000	-0.034	0.002	0.000
`Assets(bn)`	0.010	0.019	0.586	0.003	0.021	0.871
`log(scale)`				4.361	0.0127	0.000
`log(shape)`				0.195	0.012	0.000
`Max.log.likelihood`	-35304			-24139		
`LR test statistic`	172			169		
`AIC`	70611.26			48281.53		

Table 5.7: Summary of time-dependent covariates using the extended Cox PH model and the extended Weibull PH model

From Table 5.8, we see that the variable *performance* has the same parameter estimates in both extended Cox PH and extended Weibull models. The AIC criterion for the extended Weibull PH model is smaller than that for the extended Cox PH model, which means that the extended Weibull PH model is better than the extended Cox PH model. Therefore, we use the estimated coefficients from the extended Weibull PH model to interpret our data. With this model, the p-value of the *assets* coefficient is 0.871, which means the *assets* covariate is not significant. We use the *stepAIC()* function in R to select covariates and remove the *assets* covariate. Table 5.8 shows the final model (see Appendix B.5 for the R codes used to produce this Table).

Covariate	Mean	Coef	Exp(Coef)	S.E(Coef)	Wald p
`Performance per ten thousand`	0.978	-0.034	0.967	0.002	0.000
log(scale)		4.361	78.309	0.012	0.000
log(shape)		0.195	1.215	0.012	0.000
Max. log. likelihood		-24139			

Table 5.8: Summary of time-dependent covariates using the extended Weibull PH model

The estimated hazard ratio (HR) of the variable *performance* is defined as :

$$HR(x) = \frac{h_0(x)\exp(\hat{\boldsymbol{\beta}} \times z_1(x))}{h_0(x)\exp(\hat{\boldsymbol{\beta}} \times z_0(x))} = \frac{\exp(\hat{\boldsymbol{\beta}} \times z_1(x))}{\exp(\hat{\boldsymbol{\beta}} \times z_0(x))}. \tag{5.2}$$

Hence

$$HR(x) = \exp(\hat{\boldsymbol{\beta}} \times (z_1(x) - z_0(x))) = \exp(-0.034 \times (z_1(x) - z_0(x))), \tag{5.3}$$

where $h_0(x)$ is the baseline hazard function and $z_i(x)$ is the performance of hedge fund i at time x. The resulting formula for the hazard ratio time depends on x, since its value depends on performance at time x. Note that even though the values of the variable $z_i(x)$ may change over time, the estimated coefficient $\widehat{\beta}_i$ of this variable does not depend on time. Thus, this coefficient represents the *overall* effect of the corresponding time dependent variable, considering all times at which this variable has been measured in the study. The coefficient $\widehat{\beta}$ of the *performance* variable is equal to -3.394, which means that a good performance has a positive effect on the lifetime, decreasing the risk of death. Thus, every 1% increase in performance decreases the hazard by 3.3% (since $(0.967 - 1) \times 100\% = -3.3\%$).

Until now, we have studied the effect of each characteristic on hedge funds lifetime independently. In the next section, we will study the interaction between these variables.

5.7 Mixed Weibull PH model

In this section, we examine the interaction between the fixed and time-dependent covariates to see if the effect of one variable on the occurrence of the event (life fund death) depends on the risk of another variable. To determine the presence of interaction a product term is added to the regression model. In linear regression, the regression coefficient of the product term refers to interaction as a departure from additivity. However,

in a PH regression model it refers to interaction as a departure from multiplicativity. When we consider the interaction between variables, we find that there is a significant interaction between the variable *performance* and each of the variables *size*, *strategy*, *age* and *fundoffshore*, and also between the variable *assets* and *age*. The mixed Weibull PH model with the interaction covariates can be written as:

$$h(x) = h_0(x)\exp[(\hat{\beta}_1 \times S_1 + \hat{\beta}_2 \times S_2 + \hat{\beta}_3 \times A_1 + \hat{\beta}_1 \times F) \times P(x) + \hat{\beta}_4 A_1 \times A_2(x)] \quad (5.4)$$

where $h_0(x)$ is the baseline hazard function for the Weilbull PH model, S_1 represents the variable *strategy*, S_2 the *size*, A_1 the *age*, A_2 the *assets*, F the *fundoffshore* and P is the *performance*. The summary of this model is shown in Table 5.9 below (see Appendix B.6 for the R code which was used for producing this table).

From this table, the notation "Performance:Strategy" indicates respectively the interaction between *performance* and *strategy*. Similar notations are used for the other interactions. To interpret the interaction between these variables, we use the hazard ratio define as:

$$HR = HR_i \times HR_{ij} = \exp(\beta_i)\exp(\beta_{ij}) = \exp(\beta_i + \beta_{ij}) \quad (5.5)$$

where i represents either the *performance* or the *assets* covariates, and j is one of the variable *size*, *strategy*, *age* and *fundoffshore*. In the next subsections, we examine separately these interactions.

5.7.1 Interaction between *performance* and *size*

For the medium size and large size, the hazard ratio are respectively $HR = 0.964 \times 0.976 = 0.9408$ and $HR = 0.964 \times 0.974 = 0.9389$. Therefore, $(0.94 - 1) \times 100 = -6\%$ means that every 1% increase monthly in performance decreases the risk of death to 6% as compared to the smallest size where the risk of death decreases only by 3.6%. This result suggests that for hedge funds with size smaller than $\$100M$, the risk of failure is high, independently of his performance. However, for hedge funds whose size is larger than $\$100M$, the risk of failure decreases drastically as the performance increases.

5.7.2 Interaction between *performance* and *age*

For the hedge funds with age in the range [10, 30), the hazard ratio is $HR = 0.964 \times 1.021 = 0.984$. So, $(0.984 - 1) \times 100 = -1.5\%$ means that every 1% increase monthly in

Covariate	**Coef**	**Exp**	**S.E**	**Wald p**	**Coef**	**Exp**	**S.E**	**Wald p**
`strategy`					`strategy:performance`			
Event-Driven	0	1	(reference)		0	1	(reference)	
fund of funds	0.395	1.484	0.065	0.000	-0.024	0.977	0.015	0.119
Long/Short	0.099	1.104	0.064	0.122	0.020	1.020	0.014	0.142
location	0.225	1.252	0.086	0.009	0.044	1.045	0.015	0.004
Others	0.256	1.292	0.096	0.007	-0.012	0.988	0.020	0.562
Relative Value	0.161	1.175	0.074	0.028	-0.011	0.989	0.015	0.464
Tactical	0.157	1.170	0.069	0.021	0.037	1.038	0.014	0.008
`size`					`size:performance`			
$[0, 100M)$	0	1	(reference)		0	1	(reference)	
$[100M, 500M)$	-0.462	**0.630**	0.043	0.000	-0.024	**0.976**	0.009	0.008
$[500M, 50000M)$	-0.853	**0.426**	0.084	0.000	-0.027	**0.974**	0.017	0.121
`age`					`age:performance`			
$[0, 5)$	0	1	(reference)		0	1	(reference)	
$[5, 10)$	-2.193	**0.112**	0.043	0.000	-0.000	**1.000**	0.006	0.955
$[10, 30)$	-3.705	**0.025**	0.069	0.000	0.021	**1.021**	0.009	0.019
`fundoffshore`					`fundoffshore:performance`			
No	0	1	(reference)		0	1	(reference)	
Yes	0.009	**1.009**	0.032	0.787	-0.018	**0.983**	0.005	0.001
`age`					`age:assets`			
$[0, 5)$	0	1	(reference)		0	1	(reference)	
$[10, 30)$	-3.705	**0.025**	0.069	0.000	0.144	**1.155**	0.061	0.017
$[5, 10)$	-2.193	**0.112**	0.043	0.000	0.284	**1.329**	0.071	0.000
`performance`	-0.036	**0.964**	0.014	0.008				
`assets`	0.035	**1.036**	0.013	0.007				
log(scale)	3.733	41.786	0.026	0.000				
log(shape)	0.865	2.376	0.013	0.000				

Table 5.9: Summary of interaction between covariates using the extended Weibull PH model

performance decreases the hazard rate only to 1.5% as compared to the other range of age where the performance decreases the hazard rate by 3.6%. This result suggests that when a hedge fund reaches the age of 10 years, its risk of death increases independently of its performance.

5.7.3 Interaction between *performance* and *offshore*

For offshore hedge funds, the hazard ratio is $HR = 0.964 \times 0.983 = 0.947$. So, $(0.947 - 1) \times 100 = -5.2\%$ means that every 1% increase monthly in performance for the fund offshore decreases the risk of death by 5.2% as compared to the non fund offshore where the performance decreases the hazard rate by 3.6%.

5.7.4 Interaction between *performance* and *strategy*

For the *strategy* variable, only *location* and *tactical* strategies have a significant interaction with performance. For the *location* and *tactical* strategies, the hazard ratio are, respectively $HR = 0.964 \times 1.252 = 1.00738$ and $HR = 0.964 \times 1.038 = 1.000632$. So, $(1.00738 - 1) \times 100 = 0.7\%$ and $(1.000632 - 1) \times 100 = 0.06\%$ means that every 1% increase monthly in performance for the funds using *location* or *tactical* increases the risk of death respectively by 0.7% or 0.06% as compared to the *event-driven* strategy where the performance decreases the hazard rate by 3.6%. This result needs more attention, since one would expect that the increase in *performance* would decrease the risk of death.

5.7.5 Interaction between *assets* and *age*

The *assets* variable is significant in this model, but its hazard ratio is 1.036. This means that every one million monthly increase in assets increases the risk of death by 3.6%. This result is surprising; logically one would expect that the increase in assets would decrease the risk of death. So, this variable deserves more attention and a thorough study. We get the same result with its interaction with *age*. This interaction is significant but the estimated coefficients are positive. This means that the increase in assets affect negatively the lifetime in the range ages.

5.8 Summary

While Chapter 4 focused on the issue of how to describe and identify the lifetime of hedge funds, this chapter focuses on selecting and interpreting the hedge fund characteristics which affect mostly its lifetime. We applied non-parametric, parametric and semi-parametric survival methods to investigate the factors that affect the survival and mortality patterns of hedge funds. This study used the HFR database on which few works have been done. Most of the results obtain in this thesis are consistent with what

were obtain in the literature using the TASS database. For example, we found that large hedge funds survive longer than smaller and that 50% survival time is estimated to be around 5 years.

Chapter 6

Conclusion

A hedge fund is typically the part of the investment portfolio that looks for good returns via an active portfolio management. Most of the new money owing to hedge funds is from institutional investors. These institutions wish to invest into hedge funds on a long-term basis. So, they seek hedge funds likely to survive a long time and try to avoid liquidation. However, an undesirable outcome often associated with large capital losses appears. Therefore, the aim of survival analysis is to help investors select funds with good long-term prospects. Survival analysis is a class of statistical methods for studying the occurrence and timing of events. These methods are most often applied to the study of deaths or failure. And unlike ordinary regression models, survival methods correctly incorporate information from censored and truncated observations in estimating important model parameters.

This thesis has adapted and extended some existing methods in survival analysis to investigate the survival of hedge funds from the HFR database over September 1963 to June 2005. The main idea was to use the best survival model to analyze and identify among the hedge funds characteristics those which affect its lifetime and interpret them. However, having received more than $453,728$ data disordered and distributed in more than 100 tables, our first purpose was to identify, select, organize and process these data, so that they could be treated using survival analysis methods (see Chapter 4). In the same way, our database contains left-truncated, right-censored data and time-dependent covariates, so our second purpose was to identify some statistical models which are able to take into account these features (see Chapter 5).

The data studied in this thesis comprised 6947 hedge funds: 2403 live funds and 4544 dead funds. Therefore, the mortality rate has been estimated to be 65.41% during the

period from September 1963 to June 2005. *Kaplan-Meier* curve provided valuable insight into the behavior of the data and allowed us to choose the parametric model. We found that the *Weibull* model was a better parametric model compare to the *Log-normal* and *log-logistic* distributions. Akaike Information Criterion allowed us to choose the *Weibull PH* model as the best to estimate and interpret the covariates compared to the *Cox PH* model. After establishing the *Weibull PH* model with fixed covariates, as well as with time dependent covariates, under three specifications of the *Weibull PH* model (fixed model, time-varying model, and mixed model), we predicted that only the variables *strategy, size, age, fund denomination, annual-audit* and *performance* are those which affect mostly the lifetime of hedge funds during the study period. Similarly, we found a significant interaction between the variable *performance* and each of the variables *size*, *strategy*, *age* and *fundoffshore*, and also between the variable *assets* and *age*.

Some of our important results are: firstly, the risk of death is 0.668 times lower for funds who indicated that an annual audit is performed than the funds who did not indicated. In the same way, the risk of death for the offshore funds is 1.059 times higher than for the non-offshore hedge funds. Secondly, about 1% increase monthly in performance decreases the hazard by 3.3%. The interaction between *performance* and *size* shows that the large and medium funds perform better than the small funds. Most important, the interaction between *performance* and *age* reveals that when a hedge fund reaches the age of 10 years, its risk of death increases independently of its performance. That is why most hedge funds usually close their funds when they reaches a certain age to start a new hedge fund. On the other hand, we found some unexpected results such as the fact that the monthly increase in performance for hedge funds using *location* or *tactical*, increases the risk of death and also the fact that every one million dollars monthly increase in assets increases the risk of death by 3.6%. These may be due to the fact that in our sample data, we considered hedge funds with the same reporting date for assets and performance (the data of the hedge funds with the different reporting date for assets and performance were dropped off). These results need more attention in future works. The results presented in the thesis should be interpreted with a certain caution. The models used for the data analysis, while relatively complex, did not take into account potential features of the data such as backlog bias, calendar-time trends in the market, and potential correlation between hedge funds. Also, the data predate the global financial crisis of 2007-2008, and should not be taken as representative of hedge fund lifetimes since then. All these aspects should be considered as potential developments in future research.

Appendix A

Using Excel for basic data summary

A pivot table is a special type of summary table that is unique to *Excel*. Pivot tables are great for summarizing values in a table because they can be created without using any formula to perform the calculations. Pivot tables also let you play around with the arrangement of the summarized data. It is this capability of changing the arrangement of the summarized data simply by rotating row and column headings that gives the pivot table its name.

Follow these steps to create a pivot table in *Excel 2007* :

1. Open the worksheet that contains the table you want to summary with a pivot table and select any cell in the table. Ensure that the table has no *blank rows* or *columns* and that each column has a header.

2. Click the *PivotTable* button in the Tables group on the *Insert table* tab. Click the top portion of the button; if you click the arrow, click *PivotTable* in the drop-down menu. *Excel* opens the *Create PivotTable* dialog box and selects all the table data, as indicated by a marquee around the cell range.

3. If necessary, adjust the range in the Table/Range text box under the *Select a Table* or *Range* option button.

4. Select the location for the pivot table. By default, *Excel* builds the pivot table on a new worksheet and adds it to the workbook. If you want the pivot table to appear on the same worksheet, click the existing Worksheet option button and then indicate the location of the first cell of the new table in the Location text box.

5. Click OK. *Excel* adds a blank grid for the new pivot table and displays a *PivotTable Field List* task pane on the right side of the worksheet area. The *PivotTable Field List* task pane is divided into two areas: the *Choose fields to add to Report* list box with the names of all the fields in the source data for the pivot table and an area divided into four drop zones (*Report Filter, Column Labels, Row Labels*, and *Values*).

6. To complete the pivot table, assign the fields in the *PivotTable Field List* task pane to the various parts of the table. You do this by dragging a field name from the *Choose fields to add to report* list box and dropping it in one of the four areas below, called drop zones:

   - *Report Filter*: This area contains the fields that enable you to page through the data summaries shown in the actual pivot table by filtering out sets of data they act as the filters for the report. So, for example, if you designate the *Year Field* from a table as a *Report Filter*, you can display data summaries in the pivot table for individual years or for all years represented in the table.
   - *Column Labels*: This area contains the fields that determine the arrangement of data shown in the columns of the pivot table.
   - *Row Labels*: This area contains the fields that determine the arrangement of data shown in the rows of the pivot table.
   - *Values*: This area contains the fields that determine which data are presented in the cells of the pivot table they are the values that are summarized in its last column (totaled by default).

7. Continue to manipulate the pivot table as needed until the desired results appear.

To obtain our Table 4.3 describing the relation between the variables *age*, *size* and *strategy*, we put *age* and *strategy* at *Row Labels*, *delta* at the *Report Filter* and *Values*, and finally, we put *size* at the *Column Labels*. Figure A.1 presents an overview of this table in *Excel*.

To obtain all the pie charts to display the contribution of each category, we put the concerning variable at the *Row Labels* and *delta* at the *Values* and we click on *Insert* and chose pie chart in the chart board (see Figure A.2).

To obtain all the *Column charts* to compare values across categories, we put *age* and *size* at *Row Labels*, *strategy* at the *Report Filter* and *delta* at the *Values* and the *Column Labels* (see Figure A.3).

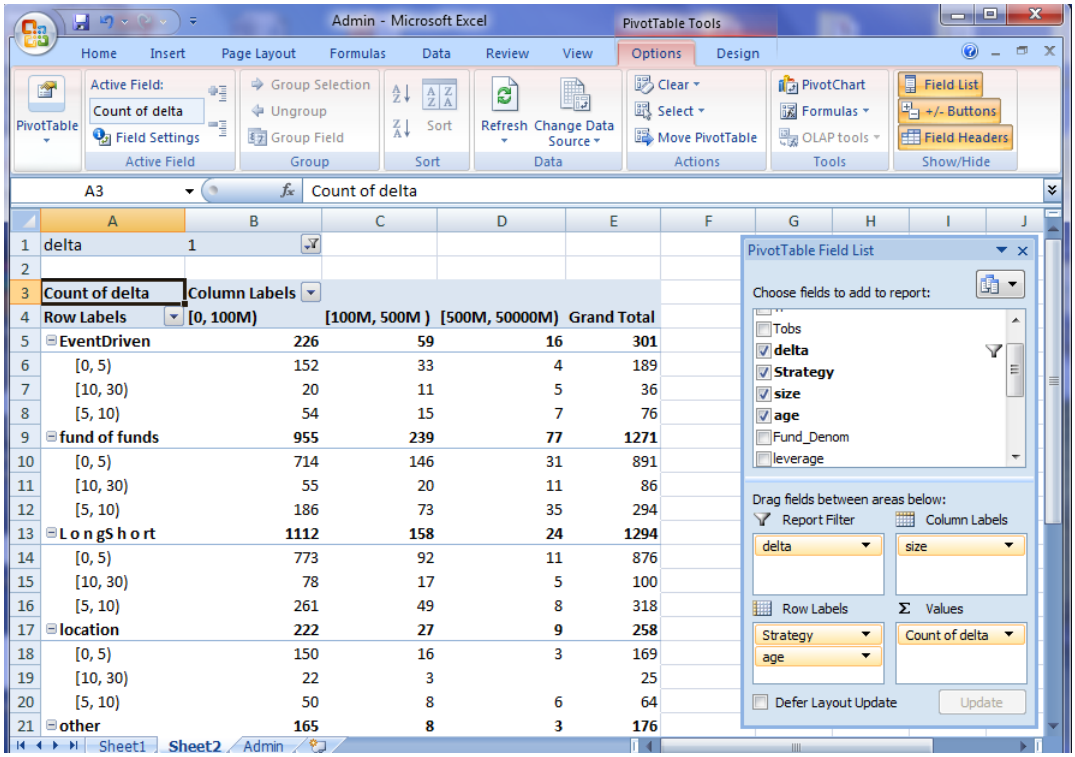

Figure A.1: An overview of a contingency table in *Excel*

To obtain all the *Line charts* to display trend over time, we put *years* at *Row Labels*, *performance* at the *Values* and *Fund-id* at the *Column Labels* (see Figure A.4).

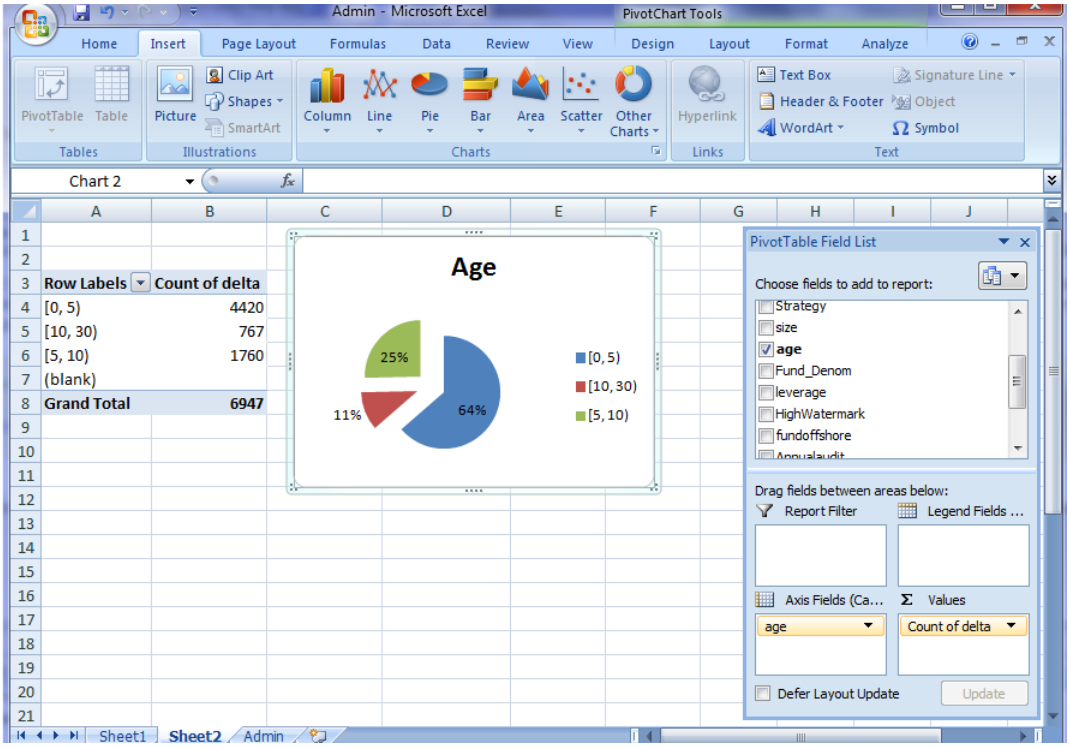

Figure A.2: An example of how to construct a pie chart in *Excel*

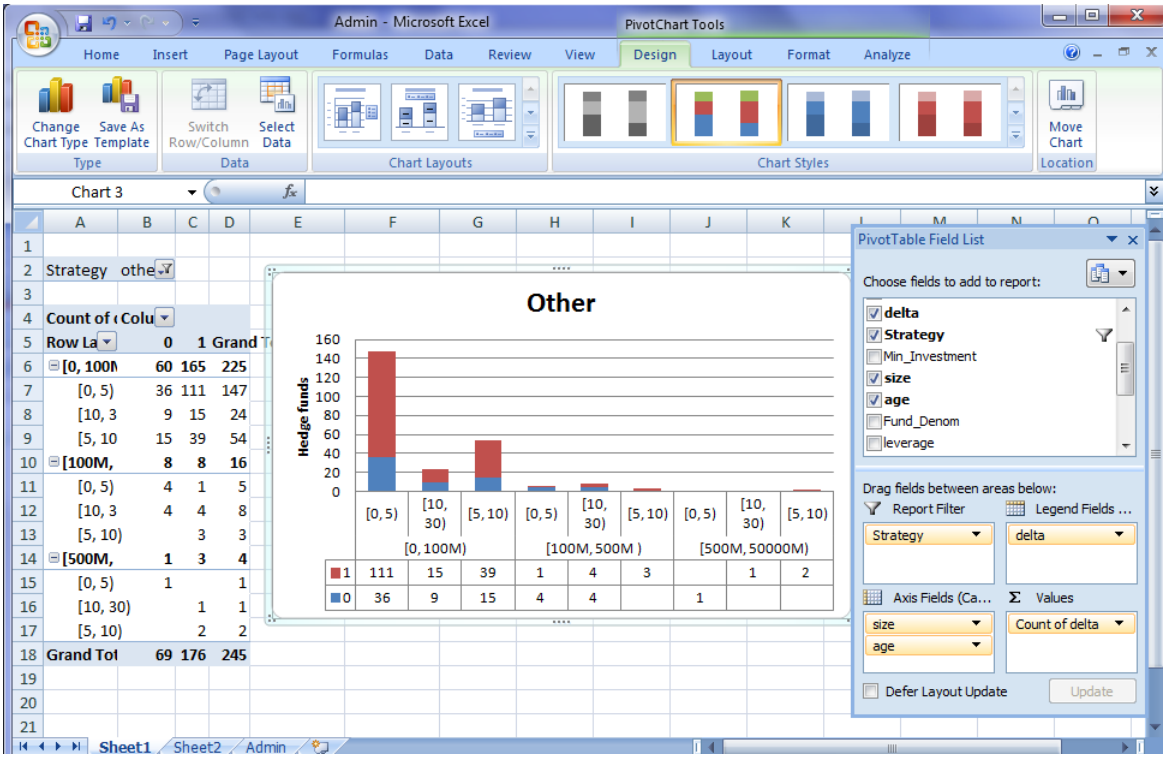

Figure A.3: An example of how to construct a column chart in *Excel*

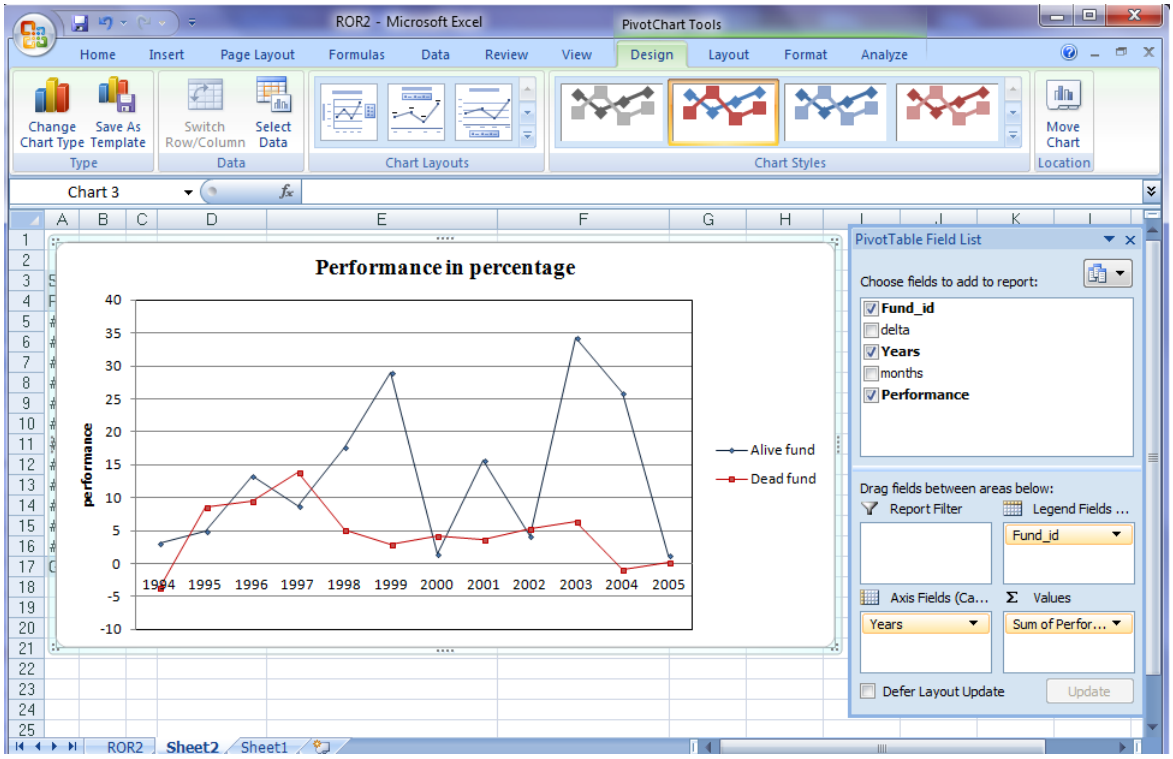

Figure A.4: An example of how to construct a line chart in *Excel*

Appendix B

The R code

B.1 Non-parametric analysis

Before using the survival analysis functions in R, we need to load the package *survival* and its library using the following code:

```
> install.packages("survival")
> library(survival)
```

Before complex functions may be performed, the data has to be put into the proper format, i.e. a survival object. In our data we have the left-truncated and right-censored time, so we use the *Surv()* function with parameters Tr (our variable truncation time), Tobs(the lifetime of the hedge funds), delta (our status variable) as follows:

```
> my.surv<- Surv(Tr, Tobs, delta, type="counting")
```

We use the *survfit()* function to estimate the survival function by the Kapan-Meier estimate and the *summary()* function to print the Kaplan-Meier estimator, its estimated variance given by the Greenwood's formula 3.43, and the 95% (Wald) confidence interval (using the type *plain*). We use the *plot()* function to visualize Kaplan-Meier curve, and the (point-wise) 95% confidence interval. The R-code of these are:

```
> my.fit <-  survfit(my.surv~1, conf.type="plain")
> summary(my.fit)
```

time	n.risk	n.event	entered	censored	survival	std err	lower 95%	upper 95%
1	6483	23	236	8	0.9965	0.000738	0.995005	0.9979
2	6688	13	67	12	0.9945	0.000912	0.992728	0.9963
3	6730	24	22	17	0.9910	0.001161	0.988694	0.9932
4	6711	39	9	14	0.9852	0.001476	0.982318	0.9881
.....more output omitted								

Table B.1: The object survfit created.

```
> plot(my.fit, xlab="time", ylab="Survival function")
```

Nelson-Aalen estimator may be constructed as follows:

```
> h.sort.of <-  my.fit$n.event[-1] / my.fit$n.risk[-1]
> H.tilde <-   cumsum(h.sort.of)
> H.tilde <-   c(H.tilde, tail(H.tilde, 1))
> plot(c(my.fit$time[-1], 400), H.tilde, xlab="time",
 ylab="Cumulative hazard" )
```

The R-code to obtain the Table 5.1 (the quartiles of survival estimate) is the following:

```
  # quantile
>quantile.km <- function(data, p, eps, z)
{
## data is survfit object, p is between 0 and 1
## eps is epsilon 0f 0.05 or bigger,
## z iz z-score for confidence coefficient
+ time <- summary(data)$time
+ ni <- summary(data)$n.risk
+ di <- summary(data)$n.event
+ surv <- summary(data)$surv
+ stderr <- summary(data)$std.err
+ qp <- min(time[surv <= 1 - p])
## The point estimate of pth-quantile
+ se.S.qp <- stderr[surv == max(surv[surv <= 1 - p])]
## S.qp is the standard error of the estimated survival
```

```
   ## probability at qp.
+ u.p <- max(time[surv >= 1 - p + eps])
# the largest time at which surv >= 1-p+eps
+ l.p <- min(time[surv <= 1 - p - eps])
# the smallest time at which surv <=1-p-eps
+ S.u.p <- surv[time == u.p] # survival probability at u.p
+ S.l.p <- surv[time == l.p] # survival probability at l.p
+ f.qp <- (S.u.p - S.l.p)/(l.p - u.p)
## estimated probability density at pth-quantile
+ se.qp <- se.S.qp/f.qp
## estimated standard error of the sample pth-quantile
+ LCL <- qp - z * se.qp
+ UCL <- qp + z * se.qp
+ out <- round(data.frame(qp, se.S.qp, f.qp, se.qp, LCL, UCL), 4)
+ print("summary")
+ print(out, invisible(1))
##print("An approximate 1-alpha confidence interval for
the true pth-quantile")
}
> quantile.km(my.fit,p=.5,eps=.05,z=1.96)
> quantile.km(my.fit,p=.25,eps=.05,z=1.96)
> quantile.km(my.fit,p=.75,eps=.05,z=1.96)
```

B.2 Parametric analysis

We need to load the package "eha" and its libraries using the following code:

```
> install.packages("eha")
> library(eha)
```

The R-code to obtain Figure 5.2 (Kaplan-Meier and parametric estimates of the survival function), and Table 5.2 are the following:

```
#Kaplan-Meier and parametric estimates of survival
#for weibull distribution
> weib <-  aftreg(Surv(Tr, Tobs, delta, type="counting")~1,
```

```
dist="weibull")

> g1 <- function(t){
+  s1 <- pweibull(t,shape=exp(weib$coefficients[2]),
scale=exp(weib$coefficients[1]),lower=FALSE)
+  return(s1)
  }
> gamma = exp(weib$coefficients[2]);
> lambda = 1/exp(weib$coefficients[1])

#for log-normal distribution
> norm <- aftreg(Surv(Tr, Tobs, delta, type="counting")~1,
dist="lognormal")
> g2 <- function(t){
+ s2 <- plnorm(t,norm$coefficients[1],
exp(-norm$coefficients[2]),lower=F) + return(s2)}
>  mu = norm$coefficients[1]$ ;
>  sigma = exp(-norm$coefficients[2]

#for logistic distribution

> logis <- aftreg(Surv(Tr, Tobs, delta, type="counting")~1,
 dist="loglogistic")
> g3 <- function(t){
+ s3 <- pllogis(t,exp(logis$coefficients[2]),
exp(logis$coefficients[1]), lower=FALSE)
+ return(s3)
}
> beta = exp(logis$coefficients[2]);
>  alpha = exp(logis$coefficients[1])

#summarize the three models with Kaplan-Meier estimate
> my.fit <- survfit(my.surv~1, conf.type="none")
> xx=seq(0, 2500, by=1)
> op <- par(mfrow=c(1,3))
```

```
> plot(my.fit, xlab="x", ylab="log-normal's probability
of survival",col="black", lty=c("solid"),
main="Log-normal fit")
> lines(xx, g2(xx), col=2red")
> legend(30, 1.04, c("Kaplan Meier estimate",
"Log-normal function"),
col=c("black", "red"), lty=c("solid","solid"), cex=1)
> plot(my.fit, xlab="x", ylab="log-logistic's
probability of survival",
col="black", lty=c("solid"), main="Log-logistic fit")
> lines(xx, g3(xx), col="red")
> legend( 30, 1.04, c("Kaplan Meier estimate",
"Log-logistic function"),
col=c("black", "red"), lty=c("solid", "solid"), cex=1)
> plot(my.fit, xlab="x", ylab="weibull's
probability of survival", col=2black",
lty=c("solid"), main="Weibull fit")
> lines(xx, g1(xx), col="red")
> legend(30,1.04, c("Kaplan Meier estimate","Weibull function"),
col=c("black", "red"), lty=c("solid","solid"),cex=1)
> par(op)
```

The R-code to obtain Figure 5.2 (the Nelson-Aalen estimate and parametric estimates of the cumulative hazard function) is the following:

```
> g11 <- function(t){
+ s1 <- -log(pweibull(t, shape=exp(weib$coefficients[2]),
scale=exp(weib$coefficients[1]), lower=FALSE))
+ return(s1)
}
>  g22 <- function(t){
+ s2 <- -log(plnorm(t,norm$coefficients[1],
exp(-norm$coefficients[2]), lower=FALSE))
+ return(s2)
```

```
}
> g33 <- function(t){
+ s3 <- -log(pllogis(t,exp(logis$coefficients[2]),
exp(logis$coefficients[1]), lower=FALSE))
+ return(s3)
}
>  op <- par(mfrow=c(1,3))
> plot(c(my.fit\$time[-1], 400), H.tilde, col="black",
lty=c("solid"), xlab="x", ylab="log-normal's
cumulative hazard function", main="Log-normal fit")
> lines(xx, g22(xx), col="red")
> legend(60, 0.25, c("Nelson-Aalen estimate",
"Log-normal function"), col=c("black", "red"),
 lty=c("solid", "solid"), cex=1)
> plot(c(my.fit$time[-1], 400), H.tilde, col="black",
 lty=c("solid"), xlab="x", ylab="log-logistic's
 cumulative hazard function",main="Log-logistic fit")
> lines(xx, g33(xx), col="red")
> legend(60, 0.25, c("Nelson-Aalen estimate",
"Log-logistic function"), col=c("black", "red"),
lty=c("solid","solid"), cex=1)
> plot(c(my.fit$time[-1], 400), H.tilde, col="black",
lty=c("solid"), xlab="x", ylab="weibull's
cumulative hazard function", main= "Weibull fit")
> lines(xx, g11(xx),col="red")
> legend(60, 0.25, c("Nelson-Aalen estimate",
"Weibull function"), col=c("black", "red"),
lty=c("solid","solid"), cex=1)
> par(op)
```

The following R-code is used to assess the goodness-of-fit of the Weibull model (see Figure 5.4):

```
> weib <- weibreg(Surv(Tr, Tobs, delta, type="counting")~Strategy
+MinInvestment+size+age+FundDenom+leverage+HighWatermark
```

```
+fundoffshore+Annualaudit+incentivefee+managmentfee)
> a1=exp(weib$coefficients[25]);
>  b1=scale=exp(weib$coefficients[24])
# Cox-Snell residual function for graph (a)
> rc <- function(x){
+ st$ <- -log(pweibull(x, a1, b1, lower=F))
+ return(st)
}

> weibhat <- function(t){
+ st <- log(-log(pweibull(t, a1, b1, lower=F)))
+ return(st)
}
#the code for graph (b):
> St <- log(-log(my.fit$surv))
> t <- log(my.fit$time)
> wei <- aftreg(Surv(Tr, Tobs, delta, type="counting")~1,
dist="weibull")
> a=exp(wei$coefficients[2]);
>  b=scale=exp(wei$coefficients[1])
> op <- par(mfrow=c(1,2))
> plot(log(rc(t)), weibhat(rc(t)), xlab="Cox-Snell residual",
ylab="log[-log(S(Cox-Snell))]", main="(a)")
> abline(-a1*log(b1), a1, col=2)
> plot(sort(t),sort(St), pch=22,lty=1,xlab="log(t)",
ylab="ln[-ln(s(t))]", cex=0.7, main="(b)")
> abline(-a*log(b), a, col=2)
```

B.3 Regression models with fixed covariates

We need to load package "bootStepAIC" and its libraries using the following code:

```
> install.packages("bootStepAIC")
> library(bootStepAIC)
```

The R-code used to obtain Table 5.3 and Table 5.4 is the following:

```
> weib <- weibreg(Surv(Tr, Tobs, delta,type="counting")~Strategy
+MinInvestment+size+age+FundDenom+leverage+HighWatermark
+fundoffshore+Annualaudit+incentivefee+managmentfee )
> summary(weib)
> extractAIC(weib,  k = 2)
> cox1 <- coxreg(Surv(Tr, Tobs, delta, type="counting")~ Strategy
+MinInvestment+size+age+FundDenom+leverage+HighWatermark
+fundoffshore+Annualaudit+incentivefee+managmentfee)
> summary(cox1)
> extractAIC(cox1, k=2)
```

The R-code used to obtain Figure 5.5 is the following:

```
#first get the Cox-Snell residuals for Coxph
> coxsnellres <- delta - resid(cox1,type="martingale")
#get the estimated of cumulative hazard function for residuals
> fitres <- survfit(coxph(Surv(coxsnellres, delta)~1,
method='breslow'),type='aalen')
> op <- par(mfrow = c(1,2))
> plot(log(fitres$time),log(-log(fitres$surv)),
xlab="log(Cox-Snell Residuals)", ylab="log(Estimated
Cumulative Hazard Function)", main="(a)")
> abline(0,1,col='red',lty=2)
> plot(fitres$time, -log(fitres$surv), xlab="Cox-Snell
Residuals", ylab="Estimated Cumulative Hazard Function",
main="(b)")
> abline(0,1,col="red", lty=2)
```

B.4 Covariates selection and interpretation of fixed variables

The R-code used to select covariable and create Table 5.5 is the following:

```
> stepAIC(weib, data=NEW-ADMIN,  direction = c("both") )
```

B.5 Regression models with time dependent covariates

The R-code used to create Table 5.8 is the following:

```
#regression with Weibull
> fit1 <- weibreg( Surv(Start, Stop, Delta)~Assets+ Performance)
#regression with Cox
> fit2 <- coxreg( Surv(Start, Stop, Delta)~Assets +Performance)
> extractAIC(fit1, k=2)
> extractAIC(fit2, k=2)
> stepAIC(fit1, data=NEW-ADMIN,  direction = c("both") )
```

B.6 Mixed Weibull PH model

The R-code used to create Table 5.9 is the following:

```
> fit1 <- weibreg( Surv(Start, Stop, Delta)~(Strategy + size + age
  + fundoffshore )*Performance + age*Assets)
```

Bibliography

[1] G. Connor and T. Lasarte. *An Introduction to Hedge Fund Strategies*. Working Paper, Financial Markets Group, London School of Economics, 2003.

[2] R.M. Stulz. Hedge funds: Past, present, and future. *Journal of Economic Perspectives*, 21:175–194, 2007.

[3] D. Tudor and B. Cao. The absolute returns of hedge funds. *Emerald Group Publishing Limited*, 38:280–302, 2012.

[4] M. Strömqvist. Hedge funds and financial crises. *Sveriges Riksbank Economic Review*, 2009.

[5] N. Baba and H. Goko. Survival analysis of hedge funds. *Bank of Japan Working Paper, Tokyo*, 6, 2006.

[6] G. Gregoriou. Hedge fund survival lifetime. *Journal of Asset Management*, 3:237–252, 2002.

[7] N.M. Boyson. Why do experienced hedge fund managers have lower returns. *Working Paper, Purdue University.*, 2003.

[8] B. Liang. Hedge funds: The living and the dead. *Journal of Financial and Quantitative Analysis*, 35:309–326, 2000.

[9] H. Baquero, J. Horst, and M. Verbeek . Survival, look-ahead bias, and persistence in hedge fund performance. *Journal of Financial and Quantitative Analysis*, 40:493–517, 2005.

[10] N. Chany, S. Haas, M. Getmansky, and Andrew. W. Lo. Systemic risk and hedge funds. *Working Paper, MIT Sloan School of Management*, 2005.

[11] G.N. Gregoriou and N.E. Duffy. Hedge funds: A summary of the literature. *Pensions: An International Journal*, 12:24–32, 2006.

[12] F. Rouah. Competing risks in hedge fund survival. *Social Science Research Network*, pages 1–52, 2006.

[13] F. Rouah. *Survival and Mortality of Hedge Funds.* McGill University, Montreal, Canada, 2005.

[14] D. Capocci. *Introduction aux hedge funds.* Economica, Paris, 2004.

[15] Jaeger R.A. *All About Hedge Funds: A hedge fund is an actively managed investment fund.* Mcgraw Hill, 2003.

[16] G.M. Henry. *Les hedge funds.* Eyrolles, 2008.

[17] J.C. Loomis. The jones nobody keeps up with. *Fortune*, pages 237–247, 1966.

[18] A. Eid and J. Minor. *Hedge Funds - Mathematics.* Investcorp Bank, Illinois, 2008.

[19] G. Connor and Woo M. *An Introduction to Hedge Funds.* working paper, Financial Markets Group, London School of Economics, 2003.

[20] lexinter.net. *Hedge Funds: Fonds d'Arbitrage.* http://www.lexinter.net/JF/hedge-funds.htm, 2012.

[21] HFR. *HFR strategy and regional classifications.* Hedge Funds Research Inc., Chicago, 2011.

[22] Investopedia. *Higher returns of hedge funds come at a price.* http://www.investopedia.com/articles/mutualfund/08/hedge-fund.asp, 2012.

[23] A. Ineichen. *Absolute Returns: the risks and opportunities of hedge fund investing.* John Wiley and Sons, New Jersey, 2002.

[24] FA Fund associates. Offshore hedge funds vs. onshore hedge funds. *A Fund Associates White Paper*, 2008.

[25] N.S. Nahman et al. Modification of the percutaneous approach to peritoneal dialysis catheter placement under peritoneoscopic visualization: Clinical results in 78 patients. *Journal of The American Society of Nephrology*, 3:103–107, 1992.

[26] B.W. Turnbull and L. Weiss. A likelihood ratio statistic for testing goodness of fit with randomly censored data. *Biometrics*, 34:367–375, 1978.

[27] H.C. Kraemer, B.A. Hamburg, and W. Jahnke. A hierarchy of drug use in adolescence behavioral and attitudinal correlates of substantial drug use. *American Journal of Psychiatry*, 132:1155–1163, 1975.

[28] Center for Human Resource Research. *National Longitudinal Survey of Youth.* The Ohio State University, Columbus, Ohio, 1995.

[29] J.T. Wassell, M.D. Keller, J.M. Ichida, and L.W. Ayers. Evaluation of protocol change in burn-care management using the cox proportional hazards model with time-dependent covariates. *Statistics in Medicine*, 12:301–310, 1993.

[30] S. Come, C. Henderson, G.F. Beadle, B. Silver, and S.A.H. Hellman. The effect of adjuvant chemotherapy on the cosmetic results after primary radiation treatment for early stage breast cancer. *International Journal of Radiation Oncology*, 10:2131–2137, 1984.

[31] B. Silver, J. R. Harris, G.F. Beadle, L. Botnick, and S.A.H. Hellman. Cosmetic results following primary radiation therapy for early breast cancer. *Cancer*, 54:2911–2918, 1984.

[32] J.P. Klein and M.L. Moeschberger. *SURVIVAL ANALYSIS: Techniques for Censored and Truncated Data*. Springer-Verlag, New York, second edition, 2003.

[33] J.F. Lawless. *Statistical Models and Methods for Lifetime Data*. John Wiley and sons, Hoboken, New Jersey, second edition, 2003.

[34] A.C. Cohen and B.J. Whitten. *Parameter Estimation in Reliability and Life Span Models.* Marcel Dekker, New York, 1988.

[35] Wikipédia l'encyclopdie libre. *Waloddi Weibull*. Wikipédia en français, http://fr.wikipedia.org/wiki/Waloddi_Weibull, 2012.

[36] M. Debanjan. *Likelihood inference for left truncated and right censored lifetime data*. Open Access Dissertations and Theses, Paper 7599, 2013.

[37] J.P. Florens, S. Darolles, and G. Simon. *Nonparametric Analysis of Hedge Funds Lifetimes*. Institut d'Économie Industrielle, Toulouse France, 2010.

[38] E.L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.*, 53:457–481, 1958.

[39] P. Armitage and T. Colton. *Encyclopedia of Biostatistics*. John Wiley and Sons, Ltd, New Jersey, 2005.

[40] D.R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B*, 34:187–220, 1972.

[41] D. Schoenfeld. Residuals for the proportional hazards regression model. *Biometrika*, 69:239–241, 1982.

[42] J. Patel, D. Hendricks, and R. Zeckhauser. The j-shape of performance persistence given survivorship bias. *Review of Economics and Statistics, MIT Press*, 79:161–166, 1997.

[43] M. Njavro. *Asia-Focused Hedge Funds: Analysis of Performance, Performance Persistence and Survival*. Dissertation, University of St. Gallen, 2012.

[44] M. Getmansky. The life cycle of hedge funds: Fund flows, size and performance. *working paper, MIT Laboratory for Financial Engineering*, 1708:1–64, 2005.

[45] B. Liang and H. Park. Predicting hedge fund failure: A comparison of risk measures. *Journal of Financial and Quantitative Analysis*, 45:199–222, 2010.

[46] G.N. Gregoriou and F. Rouah. Large versus small hedge funds. the journal of alternative investments. *Institutional Investor Journals*, 5:75–77, 2002.

[47] S. Brown, W. Goetzmann, and J. Park. Careers and survival: competition and risk in the hedge fund and cta's industry. *Journal of Finance*, 56:1869–1886, 2001.

[48] R. Gibson, P.A. Barès, and H. Gyger. Style consistency and survival probability in the hedge fund industry. *Working Paper, Swiss Federal Institute of Technology*, pages 1–25, 2001.

[49] G. Amin and H. Kat. Welcome to the dark side: hedge fund attrition and survivorship bias over the period 1994-2001. *Journal of Alternative Investments*, 6:57–73, 2003.

[50] Staff Report. *Implications of the Growth of Hedge Funds*. United States Securities and Exchange Commission, Washington, 2003.

[51] S. Brown, W. Goetzmann, and J. Park. Conditions for survival: changing risk and the performance of hedge fund managers and cta'ś. *Social Science Research Network*, pages 1–28, 1997.

[52] C. De Souza and S. Gokcan. A quantitative approach to hedge fund manager selection and de-selection. *Journal of Wealth Management,*, 6,:52–73, 2004.

[53] F. Lhabitant, G.N. Gregoriou, and F.D. Rouah. The survival of exchange-listed hedge funds. *Journal of Applied Research in Accounting and Finance*, 4:2–11, 2009.

[54] M.J. Howell. Fund age and performance. *Journal of Alternative Investments*, 4:57–60, 2001.

[55] G. Amin and H. Kat. Hedge fund attrition and survivorship bias over the period 1994-2001. *Working Paper, Cass Business School Research Centre*, 2002.

[56] M. Ammann and P. Moerth. *Impact of Fund Size on Hedge Fund Performance*. Swiss Institute of Banking and Finance University of St. Gallen, Switzerland, 2005.

[57] P. Gagliardiniz, S. Darolles, and C. Gouriroux. Survival of hedge funds: Frailty vs contagion. *Social Science Research Network*, to appear, 2013.

[58] PerTrac. *Impact of size and age on hedge fund.* http://www.pertrac.com/, 2012.