Survival methods for the analysis of customer lifetime duration in insurance

Ana María Pérez Marín

ADVERTIMENT. La consulta d'aquesta tesi queda condicionada a l'acceptació de les següents condicions d'ús: La difusió d'aquesta tesi per mitjà del servei TDX (www.tesisenxarxa.net) ha estat autoritzada pels titulars dels drets de propietat intel·lectual únicament per a usos privats emmarcats en activitats d'investigació i docència. No s'autoritza la seva reproducció amb finalitats de lucre ni la seva difusió i posada a disposició des d'un lloc aliè al servei TDX. No s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX (framing). Aquesta reserva de drets afecta tant al resum de presentació de la tesi com als seus continguts. En la utilització o cita de parts de la tesi és obligat indicar el nom de la persona autora.

ADVERTENCIA. La consulta de esta tesis queda condicionada a la aceptación de las siguientes condiciones de uso: La difusión de esta tesis por medio del servicio TDR (www.tesisenred.net) ha sido autorizada por los titulares de los derechos de propiedad intelectual únicamente para usos privados enmarcados en actividades de investigación y docencia. No se autoriza su reproducción con finalidades de lucro ni su difusión y puesta a disposición desde un sitio ajeno al servicio TDR. No se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR (framing). Esta reserva de derechos afecta tanto al resumen de presentación de la tesis como a sus contenidos. En la utilización o cita de partes de la tesis es obligado indicar el nombre de la persona autora.

WARNING. On having consulted this thesis you're accepting the following use conditions: Spreading this thesis by the TDX (www.tesisenxarxa.net) service has been authorized by the titular of the intellectual property rights only for private uses placed in investigation and teaching activities. Reproduction with lucrative aims is not authorized neither its spreading and availability from a site foreign to the TDX service. Introducing its content in a window or frame foreign to the TDX service is not authorized (framing). This rights affect to the presentation summary of the thesis as well as to its contents. In the using or citation of parts of the thesis it's obliged to indicate the name of the author.

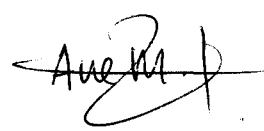
Departamento de Econometría, Estadística y Economía Española

UNIVERSIDAD DE BARCELONA

Survival methods for the analysis of customer lifetime duration in insurance

Ana María Pérez Marín

Diciembre 2005



Tesis doctoral para optar al Título de Doctora
Directores: Dra. Montserrat Guillén y Dr. Jens Perch Nielsen
Doctorado en Estudios Empresariales
Especialidad en Ciencias Actuariales y Financieras
Bienio 2001-2003

B.U.B. Secció d'Econòmiques
Diagonal, 690, 08034 Barcelona
Tel. 402 19 66

En primer lugar, agradezco a mis directores de tesis, la Dra. Montserrat Guillén y el Dr. Jens Perch Nielsen, toda la orientación y apoyo recibidos en la realización de esta tesis doctoral. Asimismo, agradezco a todos los miembros del Departamento de Econometría, Estadística y Economía Española y en especial a su director, el Dr. Miguel Ángel Sierra, y al Dr. Manuel Artís, la posibilidad que me han brindado de realizar esta tesis doctoral.

También doy las gracias a todos los integrantes del grupo de investigación *Risc en Finances i Assegurances* por el apoyo recibido durante todos estos años. Sin duda alguna, la posibilidad de compartir con ellos la labor de investigar ha hecho mucho más enriquecedora esta experiencia.

Muy en especial, agradezco a mis padres y a mi hermano el haberme apoyado y animado siempre a estudiar y a aprender. Por último, también doy las gracias a todos mis amigos y al resto de familiares, y a todos los que de algún modo han contribuido a que esta tesis sea una realidad.

A todos ellos mi más sincero agradecimiento.

A mis padres y a mi hermano

A Jiří

Contents

Chapter 1

Introduction and motivation

This thesis makes a contribution to the understanding of some elements in the assessment of the business risk in insurance companies. In order to measure operational risk[1], the insurer needs to account for the possibility of loosing a customer, i. e. to have a policy cancellation.

In this first chapter the problem of customer retention in insurance companies is presented. There are two agents in the problem addressed throughout this text: the insurer and the customer. The asymmetrical information phenomenon makes the insurer quite fragile in front of actions undertaken by the insured person. The one that is studied here is the decision to cancel the contract or contracts (if the customer has many underwritten policies). Therefore, cancellation may be viewed as an operational risk for the insurer, who may suddenly experience market share loss and/or portfolio distortion (meaning that some type of risks may be leaving the company making the portfolio overall risk different from the portfolio profile used for risk evaluation and actuarial calculations). Below we present, the current challengers in actuarial science regarding operational risk and business risk management are described. In the second and third sections of the current chapter, the customer characteristics and typical contracts in the insurance economic sector are briefly presented. The role of statistical survival analysis techniques in actuarial science

[1]We assume the classification of risks provided by Dhaene, Vanduffel, Tang, Goovaerts, Kaas & Vyncke (2004).

is introduced in the fourth section, where applications to both life and non-life insurance are illustrated with some examples.

The general and specific objectives of this thesis are presented in the final section, where the current perspective of data analysis in insurance is presented. This thesis is based on the following idea: rather than segmenting by line of business (which has been the traditional life, non-life dichotomy) one may see individual contracts as customers each one of whom is holding a micro-portfolio composed of a few contracts. The individual information framework and decision making schemes should not ignore correlation effects between policies underwritten by the same customer, which sometimes has traditionally been disregarded. This thesis is a little step forward in the multiline approach.

1.1 New challenges in actuarial science

Insurance companies provide cover against risks that citizens, corporations or organizations have to face. The importance of the insurance sector is not only derived from this significant fact, but also from their specific business activity. Insurance companies collect long-term savings of millions of citizens, and represent the largest institutional investor on EU stock exchanges (Linder & Ronkainen, 2004).

The EU solvency system, that is based on simple ratios representing percentages of risk exposure measures, was designed in a period which was completely different from current economic environment and insurance practices. Nowadays, the economic reality incorporates increasing competition, convergence between financial sectors and an international dependence.

One of the most important current challenges both in actuarial science and insurance practices, is driven by the implementation of the "Solvency II" project. In 2000, the European Commission initiated the"Solvency I" project in order to change the solvency system directives. This first project increased the capital requirements for the most volatile classes of business and also introduced improvements regarding early supervisors' intervention powers. Nevertheless, it was necessary to examine

the fundamentals of the EU insurance supervisory system much more in detail, and therefore the "Solvency II" project was meant to provide for this review.

The new system should include a framework that appropriately reflects insurance risks. Additionally, it should incorporate incentives for companies to assess and to manage these risks. Finally, it should be in line with international developments in solvency, risk management and accounting. In the new framework, operational risks, that has been traditionally forced into the background, has been recognized as one of the major sources of instability. Banks, for example, are required to meet a regulatory capital threshold to cover operational risks.

Operational risks are those that cannot be classified as either asset or liability risks[2]. They are subdivided in business risks, such as a production lower than expected, and event risks, like system failure (see Dhaene, Vanduffel, Tang, Goovaerts, Kaas & Vyncke, 2004). Traditionally, the lack of data has been one of the difficulties that should be faced in order to measure operational risks. Nevertheless, during the last years, insurance companies have been collecting more and more statistical information (see Gustafsson, Guillén, Nielsen & Pritchard, 2005). This has been partly motivated by the extremely competitive economic environment, where business risk management has become an important issue and a key factor for improving efficiency.

Loosing customers is part of the operational risk assumed by insurers, though very little has been said on how to measure and handle this kind of risk.

Not all customers have the same characteristics and not all policy cancellations are expected to have the same influence on the overall business risk. This influence is directly related to the lifetime value of the customer in the insurance company. While the estimation of the probability of a policy cancellation can be easily addressed, the assessment of the lifetime value of the customer in an insurance company incorporates several difficulties (Jackson, 1989).

The main contribution of this thesis is to provide a new methodology, in the

[2]Other definitions are possible, see for example the report "A global framework for insurer solvency assessment", IAA Documents, available at www.actuaries.org.

field of survival analysis, to estimate important elements of the lifetime value of a customer: the time that he/she will stay in the insurance company and some of the changes that presumably are going to occur in his/her behaviour along this time (his/her lifecycle).

An application of this methodology to real data has been carried out in order to analyse a particular period in the customer's lifetime. This research provides new insights and some conclusions that are a contribution to the understanding of the problem of policy cancellations and, therefore, to the assessment of the business risk. In consequence, this thesis aims to provide useful results for both the marketing and the risk manager in insurance companies.

1.2 The insurance customer

Schlesinger & Schulenburg (1993) claimed that many customers were unaware that there were price differences among insurance companies. Additionally, they remarked that comparative price shopping was very difficult since price differences for comparable coverages were not available in printed form. The same authors stressed two features of the automobile insurance market in 1993: substantial differences in price even though the contracts being sold were considered to be very homogeneous.

The reasons for this phenomenon were mainly informational. Firstly, the insurance product is much more that the contract itself. Apart from the insurance contract, the reputation of the insurance company, the marketing strategies and the claims handling procedure are influencing the insurance relationship[3], and therefore are part of the insurance product as a whole. Nevertheless, the information about most of these additional elements is not available for the customer at the time when the product is purchased.

Definitely, many things have changed during the recent years. Nowadays customers who want to switch insurers incur in much lower information search costs than before. One of the main reasons is the spread use of internet inside the infor-

[3]The insurance relationship is the relationship between the customer and the insurer.

mation society.

A very clear evidence of this change can be found in Denmark. The Danish insurance industry has developed an electronic system for comparing the three most common classes of customer insurance in terms of price and cover. These are household/contents insurance (which covers goods inside the house), buildings insurance (that covers the building itself) and motor insurance. The web site www.ForsikringsLuppen.dk presents the products in an easily understandable way next to each other and compared against a common standard. Moreover, the presentation includes a comparison of prices inclusive and exclusive of group discounts and comprises 14 of the largest insurers in the market. Surely, this is one of the reasons why the Danish insurance market is one of the most competitive among the European insurance markets.

For all these reasons, during the last years the customer is playing a much more important role in the insurance relationship than before. As a consequence, insurance companies have to face a new problem that was absolutely secondary some years ago: the retention and recruitment of customers, in order to keep or increase the market share, and the management of the business risk.

1.3 Insurance contracts and loyalty

Customer loyalty is becoming an important issue both in the actuarial science and insurance practice. It is important to remark that the specific features of the insurance sector should be taken into account when implementing any customer loyalty strategy or business risk management policy.

As mentioned before, insurance companies provide a complex product that includes much more elements that the contract itself (ranging from legal contrains to contract duration, financial management of funds and claims compensation). Additionally, the insurance relationship is also unique. The insured pays a premium in advance to be covered against the risk that a particular event would occur. The insurer has the commitment to provide the corresponding economic compensation

in case that this loss event would occur.

Apart from this, it is frequently the case that common risks affecting the members of the same household are covered by the same insurance company. The so-called cross-buying behaviour (the same household has different insurance contracts with the same insurer) is very common in the insurance market.

Therefore, the insurance relationship should be understood by taking simultaneously into account all single insurance contracts the customer has with the insurer in order to have some understanding of the overall relationship of the customer with the insurer.

The different business lines in insurance companies have been traditionally managed independently. It is frequently not an easy task to combine all information about a particular customer in different lines of business in the same insurance company and to have an overall picture of the insurance relationship (the so-called multiline approach).

Nevertheless, in the recent years, information systems and more sophisticated statistical tools have partly contributed to make information transfer and analysis much more efficient. That is why, in order to guarantee an overall view of the insurance relationship the multiline approach should be applied in the marketing and business risk management.

1.4 Survival analysis in insurance

Survival analysis is unavoidably part of the actuarial science. Most of the applications of statistical methods for mortality analysis are found in life insurance. The reason is obvious, the measurement of the time to death or the estimation of the risk of death in mortality studies should be necessarily addressed by using these techniques (see for example Guillen, Nielsen & Perez-Marin, 2006). An actuarial survey of statistical models for survival data can be found in MacDonald (1996).

Apart from classical mortality studies, recent research has extended the application of these methods to long-term-care insurance (Czado & Rudolph, 2002), where

a proportional hazards model is applied to estimate transition probabilities between care levels. More sophisticated models, such as random effects survival models, also called frailty models (Keiding, Andersen & Klein, 1997, and Hougaard, 1995) are nowadays widely extended in actuarial studies (Albers, 1999; Haberman & Pitacco, 1996 and Olivieri, 2003).

Nevertheless, survival analysis methods can also provide the solution to many new actuarial problems in non-life insurance. For example, Beirlant, Derveaux, De Meyer, Goovaerts, Labie & Maenhoudt (1991) used an accelerated failure time model when trying to explain claim size in the statistical risk evaluation applied to Belgian car insurance.

Another example can be found in Herbst (1999), who presented an application of randomly truncated data models in reserving IBNR claims, where the total size of the IBNR claims can be reduced to determining the joint probability distribution of the delay variable (time elapsed between the occurrence and the notification of a claim) and the claim size variable under a model of random truncation.

As a conclusion, survival analysis techniques have a great number of applications in actuarial science. The estimation of any time-to-event variable would require the use of such methods. In this thesis, a new non-parametric survival analysis technique is presented and it is applied to the estimation of customer lifetime duration in the non-life insurance lines of business. This new method can be used for improving the efficiency of estimations in survival studies with ritgh-censored data and so, it has a great number of potential additional applications in actuarial science.

1.5 Objectives

The general objective of this thesis is the identification of factors influencying customer loyalty as a contribution to the understanding and measurement of business risk in insurance companies. This contribution is focused both on the methodology and the application to real data.

Regarding the methodology, the objective is to provide a new technique in the

field of survival analysis methods with better efficiency performance than other standard methods. Additionally, the application to real data would provide the identification of factors influencing customer loyalty and the analysis of customer survival time in insurance companies.

The specific objectives are summarized as follows in nine items:

- Definition of a new methodology for the estimation of customer lifetime duration. This methodology would have a number of applications both in actuarial science and survival analysis.
- Estimation of the probability that a customer with several policies in the same insurance company would cancel all of them simultaneously, the so-called total cancellation.
- Determination of the factors associated to a higher risk of a total cancellation.
- Application of the new methodology to the analysis of a specific period of customer lifecycle: from the first cancellation of a policy to the moment when all the remaining policies would be cancelled.
- Analysis of survival beyond the first cancellation, with a specific attention to both the lifetime duration and the survival probabilities.
- Determination of the factors associated to a higher risk of cancelling all the remaining policies (shorter residual lifetime), given that one policy has already been cancelled.
- Comparison between the proposed methodology and other standard methods frequently used in marketing, such as the Tobit model.
- Extensions to the case where effects of covariates are allowed to vary over time in regression models for survival analysis.
- Conclusions about what should be taken into account when dealing with the insurance business risk management related to policy cancellations.

Chapter 2

Nonparametric methods for survival analysis

In this chapter the notation and functions describing the time-to-event variable in survival analysis are introduced. The basic elements of counting process theory that will be applied in Chapter 3 are presented in the second section. Two general approaches can be used in survival studies: parametric and nonparametric methods. The differences between them are extensively discussed, and the most widely used models are briefly presented. Regression methods for modelling the survival experience of a heterogeneous population are described in section 2.4. A main issue in survival analysis is the changing effect of covariates over time. The models that incorporate these time-varying effects are presented in the last section of this chapter.

2.1 Notation and functions

We will denote by X the random variable that measures time until some specified event. In classical survival analysis methods this event may be, for example, the death or the development of a certain disease. Duration data has a peculiar feature: our possibility to measure the time-to-event variable may be limited by our particular observation period and other characteristics of the phenomena being studied. This is the reason why specific methods in survival analysis and statistics have been

developed. Firstly, these methods have to deal with problems arising due to some of the properties of time-to-event variables, namely, censoring and truncation.

Generally speaking, censoring occurs when some lifetimes are known to have occurred only within certain intervals. There are three different types of censoring, namely, right censoring, left censoring and interval censoring. Right censoring corresponds to the case when the event is observed only if it occurs prior to some pre-specified time. If the event of interest has already occurred before the individual is observed in the study then the corresponding observation is said to be left censored. Finally, interval censoring can be considered a more general type of censoring, it occurs when individuals in the study have a periodic follow-up and the event time is only known to fall in the corresponding temporal interval.

Truncation occurs when the only individuals observed by the researcher are those who experience some event. This event may be some condition that occurs prior to the event of interest, and in this case the main event of interest is said to be left truncated. Right truncation occurs when individuals who have experienced the event are the only included in the study, while those individuals who have not experienced the event are not considered. More details about these definitions can be found in Klein & Moeschberger (1997).

We will consider three functions to characterize the distribution of X, such that, by knowing any of them, the other two can be determined uniquely. These functions are the survival function, which is the probability of surviving beyond a certain moment in time x, the probability density function, which is the probability of the event occurring at time x, and the hazard rate function, which can be interpreted as the chance that an individual of age x has to experience the event in the next instant.

The survival function is the probability of an individual to experience the event after time x, and is defined by

$$S(x) = \Pr(X > x).$$

If X is a continuous random variable then the survival function is the complement of the distribution function

$$S(x) = 1 - F(x),$$

where $F(x) = \Pr(X \leq x)$. The survival function also equals the integral of the probability density function $f(x)$

$$S(x) = \Pr(X > x) = \int_x^\infty f(u)du.$$

Among the basic properties of survival curves we point out that they are monotone, non-increasing functions equal to one at zero and equal to zero as the time (x) tends to infinity. The survival function is a widely-used tool to describe survival and to compare two or more mortality experiences.

The probability density function $f(x)$ can be written in terms of the survival function $S(x)$ according to the following relationship:

$$f(x) = -\frac{dS(x)}{dx}.$$

In order to have an interpretation of the previous function, we should note that $f(x)\Delta x$ is an approximation of the probability that the event occurs in $(x, x + \Delta x)$. It is also important to remark that $f(x)$ is a nonnegative function with the area under $f(x)$ being equal to one.

The hazard rate function (or simply, hazard function) is defined by

$$\alpha(x) = \lim_{\Delta x \to 0} \frac{P\left[x \leq X < x + \Delta x | X \geq x\right]}{\Delta x}.$$

In order to have an interpretation of the hazard rate, one can see that $\alpha(x)\Delta x$ corresponds to an approximation of the probability that an individual of age x experiences the event in the next instant $(x, x + \Delta x)$. In other words the individual has to be alive at the beginning of the interval. The hazard function corresponds to the force of mortality in demography and the intensity function in stochastic processes. If X is a continuous random variable, then

$$\alpha(x) = \frac{f(x)}{S(x)} = \frac{-d\ln[S(x)]}{dx}.$$

The hazard rate can be increasing, decreasing or constant, there is no specific shape describing the failure pattern. The only requirement is that the hazard rate function must be a non-negative function, $\alpha(x) \geq 0$ for all $x \geq 0$.

A very important related quantity is the cumulative hazard function $\widetilde{\Lambda}(x)$ defined by

$$\widetilde{\Lambda}(x) = \int_0^x \alpha(u)du = -\ln[S(x)],$$

and in the case of continuous lifetimes the following equality holds:

$$S(x) = \exp[-\widetilde{\Lambda}(x)] = \exp[-\int_0^x \alpha(u)du].$$

2.2 Counting process theory

Counting process methodology results from a combination of stochastic integration, continuous time margingale theory and counting process theory that leads to the development of inference techniques for censored and truncated survival data.

We assume in this section the same notation and formulation as in Klein & Moeschberger (1997).

A counting process $N(t)$, $t \geq 0$, is a stochastic process with the following properties: $N(0)$ is zero, $N(t) < \infty$, with probability one, and the sample paths of $N(t)$ are right-continuous and piecewise constant with jumps of size $+1$.

Let us consider a right censored sample of n individuals. For any individual in the study we assume that there is a lifetime X and a censoring time C_i, known but possibly different for each individual. We also assume that the $X's$ are independent and identically distributed with probability density function $f(x)$ and survival function $S(x)$. Then the exact lifetime X of an individual can be known if and only if X is less than or equal to C_i. If X is greater than C_i the event time is censored at C_i. We can conveniently represent survival data for individual i from this sample by a pair of random variables (T_i, δ_i), where the censoring indicator δ_i is equal to 1 when the lifetime X correspods to an event, and is equal to 0 when the observation is censored. T_i is equal to X if the lifetime is observed, and T_i is equal to C_i if it is

censored, thus $T_i = \min(X, C_i)$.

For a right-censored sample, we have that the process $N_i(t) = I[T_i \leq t, \delta_i = 1]$, where I represents an indicator function, is a counting process. Note that this process is zero until the individual i dies and then jumps to one. Similarly, we have that the process that counts the number of deaths in the sample at or prior to time t, namely, $N(t) = \Sigma_{i=1}^{n} N_i(t) = \Sigma_{t_i \leq t} \delta_i$, is also a counting process.

The history or filtration of the counting process at time t, $\mathcal{F}_t$, is the accumulated knowledge about the process up to time t. This knowledge includes information about when events occur, and additionally, in the case of right censored data, includes knowledge of which individuals have been censored prior to time t. In a causal model framework this knowledge about the process includes information about values for fixed or time-dependent covariates. We will require that $\mathcal{F}_s \subset \mathcal{F}_t$ for $s \leq t$, because as time progresses it is natural to expect that we accumulate more and more information about the sample. Additionally, we will denote the history at an instant just prior to time t by $\mathcal{F}_{t-}$.

We define $dN(t)$ as the change in the counting process $N(t)$ over a short time interval $[t, t + dt)$, thus $dN(t) = N[(t + dt)^-] - N(t^-)$, where t^- represents a time just prior to t. Additionally we define $Y(t)$ as the number of individuals with a study time $T_i \geq t$, thus provide us with the number of individuals at risk at a given time t. Then it can be proved that $E[dN(t)|\mathcal{F}_{t-}] = Y(t)\alpha(t)dt$, where the process $\lambda(t) = Y(t)\alpha(t)$ is called the intensity process of the counting process.

We define the cumulative intensity process $\Lambda(t) = \int_0^t \lambda(s)ds$, $t \geq 0$, that has the property that $E[N(t)|\mathcal{F}_{t-}] = E[\Lambda(t)|\mathcal{F}_{t-}] = \Lambda(t)$, that is derived from the fact that once we know the history just prior to t, the value of $Y(t)$ is fixed and thus $\Lambda(t)$ is not random. The stochastic process $M(t) = N(t) - \Lambda(t)$ is called the counting process martingale. This process has the property $E[dM(t)|\mathcal{F}_{t-}] = 0$, i.e. the increments of this process have an expected value, given the strict past $\mathcal{F}_{t-}$, that are zero. It can be proved that this property is equivalent to $E[M(t)|\mathcal{F}_s] = M(s)$ for $s < t$, that is the property that characterized the so-called martingale stochastic processes. Note that the counting process martingale $M(t) = N(t) - \Lambda(t)$ consist of two parts: $N(t)$

that is a nondecreasing step function and $\Lambda(t)$, that is called compensator of the counting process, that is a smooth process which is predictable, because its value at time t is fixed just prior to time t.

Another basic quantity in the theory of counting processes is the predictable variation process of $M(t)$, that is denoted by $\langle M \rangle (t)$. This process is defined as the compensator of the process $M^2(t)$, i.e. the predictable process needed to be substracted from $M^2(t)$ to produce a martingale. The label "predictable variation process" comes from the fact that, for a martingale $M(t)$, it can be proved that $var(dM(t)|\mathcal{F}_{t-}) = d\langle M \rangle (t)$.

In Chapter 3 of this thesis we will deal with a number of statistics that are essentially stochastic integrals of the martingale discussed in this section. To introduce this notion, let us assume that $K(t)$ is a predictable stochastic process, i.e. its value is known given the history just prior to time t, $\mathcal{F}_{t-}$. Over the interval 0 to t, the stochastic integral of such a process, with respect to a martingale, is denoted by $\int_0^t K(u)dM(u)$. These stochastic integrals have the property that they themselves are martingales as a function of t and their predictable variation process can be obtained from the predictable variation process of the original marginale,

$$\left\langle \int_0^t K(u)dM(u) \right\rangle = \int_0^t K(u)^2 d\langle M \rangle (u).$$

In Andersen, Borgan, Gill & Keiding (1993) and Klein & Moeschberger (1997) a more detailed discussion about stochastic integrals can be found.

2.3 Parametric versus non parametric methods

Two general approaches can be used in survival studies: parametric and nonparametric methods. Parametric methods assume that the time-to-event variable comes from a specific distributional family, while nonparametric methods make no assumption about the distribution of the time-to-event variable. More sophisticated methods are derived from the combination of these two approaches, the so-called semiparametric methods.

Nonparametric and semiparametric methods are extensively applied in survival studies. They will be widely used in subsequent sections of this thesis. Nevertheless, parametric models are very popular among researchers. Bowers, Gerber, Hickman, Jones, & Nesbitt (1997) summarize the justifications for postulating an analytic form for mortality basically according to both philosophic and practical arguments. Firstly, many phenomena analysed in physics have been explained efficiently by simple formulas, therefore using biological arguments, some authors have suggested that human survival is governed by an equally simple law. Secondly, it is easy to estimate a few parameters of the function from mortality data and besides it is also easier to provide a function with few parameters than a life table with may be 100 mortality probabilities.

Apart from these two arguments, it is important to remark that the main advantage of parametric methods is that in some cases there is a direct interpretation of some of these parameters in terms of specific functions describing the survival pattern.

Among parametric models or mortality laws formulated in the context of actuarial science and biometrics the Gomperz, Makeham and Weibull laws of mortality are specially relevant because of they are widely used.

Gompertz (1825) discovered a pattern in human mortality. He found that the probability of dying is high at birth but then declined until sexual maturity. Afterwards, it increases at an exponential rate. Thus, according to Gompertz's law of mortality the hazard rate has an exponential form $\alpha(x) = Bc^x$, where $B > 0$ and $c > 1$ and the survival function is $S(x) = \exp[-m(c^x - 1)]$. Note that Gompertz's law can also be expressed in the linear form $\ln \alpha(x) = \ln B + x \ln c$. The basic characteristic of this parametric model is that the hazard function is increasing but the relative increase is constant, i.e. $\alpha'(x)/\alpha(x) = \ln c$. Makeham (1860) suggested that Gompertz's law could be improved by adding a constant term so that $\alpha(x) = A + Bc^x$, where $B > 0$, $A \geq -B$ and $c > 1$. Note that the Gompertz's law is a special case of Makeham's law with $A = 0$. This constant A has been interpeted as capturing accident hazard and the term Bc^x as capturing the hazard of aging.

It has been found to fit adult populations well, with variations in the parameters allowing for differences between populations.

Before describing the Weibull law of mortality it is convenient to introduce the exponential distribution, that is one of the most significant parametric models because of its mathematical simplicity and important properties. The survival function is $S(x) = exp(-\lambda x)$, $\lambda > 0$, $x \geq 0$, and the hazard function is constant $\alpha(x) = \lambda$. One basic property of the exponential distribution is the so-called lack of memory property, that is given by $P(X \geq x+z|X \geq x) = P(X \geq z)$. Despite the mathematical tractability that result from this property, it is also reducing its applicability to many realistic situations. Another property is that the so-called mean residual life is constant. This property can be expressed like $E(X - x|X > x) = E(X) = 1/\lambda$, and is directly derived from the lack of memory property.

According to the Weibull law of mortality (Weibull, 1951), the distribution function is given by $S(x) = exp(-\lambda x^{\alpha})$, where $\lambda > 0$ is called the scale parameter and $\alpha > 0$ is called the shape parameter. The hazard function has a very flexible form $\alpha(x) = \lambda\alpha x^{\alpha-1}$. This flexibility allows for decreasing ($\alpha < 1$), increasing ($\alpha > 1$) and constant hazard rate ($\alpha = 1$). This flexibility and the model's simple formulation have made it a very popular parametric model.

Apart from these parametric models that have been formulated in the context of actuarial science and biometrics, other well known probability distributions are used in the context of survival analysis. This is the case of the gamma and the lognormal distribution.

The gamma distribution has similar properties to the Weibull distribution, but is not as mathematically tractable. Its density function is given by $f(x) = \lambda^{\beta}x^{\beta-1}$ $\exp(-\lambda x)/\Gamma(\beta)$, where $\lambda > 0$, $\beta > 0$, $x \geq 0$ and $\Gamma(\cdot)$ is the gamma function. Due to its similarity to the Weibull distribution, λ is called the scale parameter and β is the shape parameter. The hazard function for the gamma distribution is monotone increasing for $\beta > 1$, with $\alpha(0) = 0$ and $\alpha(x) \to \lambda$ as $x \to \infty$, and monotone decreasing for $\beta < 1$, with $\alpha(0) = \infty$ and $\alpha(x) \to \lambda$ as $x \to \infty$.

The lognormal distribution is also very used when modelling time-to-event data,

not only because of its relationship with the well known normal distribution, but also because some authors have observed that the lognormal distribution approximates survival times or ages at the onset of certain diseases. The survival funciton is given by $S(x) = 1 - \Phi[(\ln x - \mu)/\sigma]$ where $\Phi[\cdot]$ represents the standard normal distribution function. The hazard rate of the lognormal distribution is hump-shaped, i.e. its value at 0 is 0, it increases to a maximum and then decreases to 0 as x approaches infinity. Main critics to this models are based on the decreasing shape of the hazard function for large x, which seems not very realistic. Nevertheless the model fit very well in certain cases when large values of x are not of interest.

As claimed in Bowers, Gerber, Hickman, Jones & Nesbitt (1997), the support for simple analytic survival functions has declined in recent years, not only because many researchers have the feeling that to believe in universal laws is naive, but also because high-speed computers have made possible to develop sophisticated nonparametric and semiparametric methods that have provided new approaches to deal with advanced issued in mortality studies.

As an example of nonparametric methods, we mention because of its significance, the Product-Limit estimator, introduced by Kaplan & Meier (1958). This estimator is defined as follows

$$\widehat{S}(t) = \begin{cases} 1 & \text{if } t < t_1 \\ \prod_{t_i \leq t} \left[1 - \frac{d_i}{Y_i}\right] & \text{if } t \geq t_1 \end{cases}$$

where t_i represents the different values of t "where there is data", d_i is the number of events at time t_i and Y_i is the number of individuals at risk at time t_i.

Estimations of the survival function, and therefore of the cumulative hazard rate, for right-censored data can be easily obtained by using the Product-Limit estimator. An alternative method to estimate the cumulative hazard rate with better small-sample-size properties than the one based on the Product-Limit estimator is the so-called Nelson-Aalen estimator (Nelson, 1969; Nelson, 1972 and Aalen, 1978), that will be discussed in detail in Chapter 3.

2.4 Causal models in survival analysis

The methods discussed until this point have dealt with modelling the survival experience of a homogeneous population, but sometimes when analyzing survival data it is necessary to adjust the survival function to account for additional information that identifies heterogeneous populations. This information is normally given by the so-called independent variables, explanatory variables or covariates. In this section we will make an introduction to regression models in survival analysis, where this additional information is used to differenciate the survival experience of heterogeneous populations.

We will consider a time-to-event variable T and a $p-$ dimensional vector $Z = (Z_1, ..., Z_p)$ of covariates associated with the variable T. The vector Z can include quantitative variables, qualitative variables, and/or time-dependent covariates in which case we have that $Z(t) = [Z_1(t), ..., Z_p(t)]$. The matter of interest in this framework is to analyse the relationship between one or more explanatory covariates and the time-to-event variable.

Two approaches can be found in the statistical literature for modeling the covariate effects on the survival. The first approach is similar to the classical linear regression approach, and is called the accelerated failure time model. The natural logarith of the survival time $Y = \ln T$ is modelled linearly

$$Y = \mu + \gamma^t Z + \sigma W$$

where γ^t the transpose vector of regression coefficients, i.e. $\gamma^t = (\gamma_1, ..., \gamma_p)$, and W is the error distribution, that is commonly modelled by using a standard normal distribution, which yields a lognormal distribution. Other possibilities include the extreme value distribution that yields a Weibull regression model or a logistic distribution which yields a log logistic regression model. The estimation of the regression coefficients is performed by using maximum likelihood methods.

The reason why this model is called the accelerated failure-time model is derived from the following equality $\Pr[T > t|Z] = S_0[t \exp(-\gamma^t Z)]$ where $S_0(t)$ denote the

survival function of $T = e^Y$ when Z is zero, that is, $S_0(t)$ is the survival function of $\exp[\mu + \sigma W]$. Then the effect of the explanatory variables in the original time scale is to change the time scale by a factor $\exp(-\gamma^t Z)$. Therefore, the sign of $\gamma^t Z$ has the effect of either an acceleration or a degradation of the time by a constant factor. In this model the hazard rate of an individual with covariate value Z is related to a baseline hazard rate α_0 by $\alpha(t|Z) = \alpha_0[t\exp(-\gamma^t Z)]\exp(-\gamma^t Z)$.

The accelerated failure time model represents a direct extension of the classical linear model, but its use is restricted by the choice of the error distributions. Therefore, other approaches have been proposed, for example those arising from modelling the conditional hazard rate as a function of covariates. As mentioned in Klein & Moeschberger (1997), the easiest survival parameter to model is the hazard rate which reports how quickly individuals of a particular age are experiencing the event of interest. In this second approach two classes of models have been used to relate covariate effects to survival, namely the multiplicative hazard (rate) models and the additive hazard rate models.

Multiplicative hazard rate models specify the conditional hazard rate of an individual with covariate vector z as the product of a baseline hazard rate $\alpha_0(t)$ and a non-negative function of the covariates $c(\beta^t z)$, i.e.

$$\alpha(t|z) = \alpha_0(t)c(\beta^t z)$$

In applications of this type of models, $\alpha_0(t)$ may have a specific parametric from or may be estimated nonparametrically directly from the data. Regarding $c(\cdot)$, any nonnegative function can be used. Most applications use the Cox (1972) model (also called proportional hazards regression model), where $c(\cdot)$ is assume to have an exponential form. The main feature of multiplicative hazards models is that when all covariates are fixed at time 0 the hazard rates of two individuals with distinct values of z are proportional. It can be easily showed by considering two individuals with covariate values z_1 and z_2, then we have that

$$\frac{\alpha(t|z_1)}{\alpha(t|z_2)} = \frac{\alpha_0(t)c(\beta^t z_1)}{\alpha_0(t)c(\beta^t z_2)} = \frac{c(\beta^t z_1)}{c(\beta^t z_2)}.$$

For the class of additive hazard rate models, the conditional hazard function is modelled according to this specification

$$\alpha(t|z) = \alpha_0(t) + \sum_{j=1}^{p} z_j(t)\beta_j(t). \tag{2.1}$$

Note that the regression coefficients for these models are functions of time. Therefore, the effect of a given covariate on survival varies over time. The values of the p regression functions may be positive or negative, but they are constrained because (2.1) must be positive.

2.5 Time varying coefficients

The changing effect of covariates over time in a causal model is a main issue in survival analysis. Even when the model seems to provide a proper description of the covariate effect it is convenient to carry out some procedure to investigate whether or not the effect of covariate changes over time.

In Andersen, Borgan, Gill & Keiding (1993) we can find a summary of the approaches traditionally used for this purpose in the case of the proportional hazards regression model. According to this model, the intensity is specified as follows

$$\lambda_i(t) = Y_i(t)\alpha_0(t)\exp(\beta^t Z_i)$$

where $Y_i(t)$ is an indicator equal to 1 if the subject is at risk and zero otherwise, $\alpha_0(t)$ is the baseline hazard, $Z_i = (Z_{i1}, ..., Z_{ip})$ is the $p-$ dimensional vector of covariates (which may also be time-dependent) and β is the $p-$ dimensional vector of unknown regression parameters.

One method traditionally used to check whether or not the effect of covariates changes over time consist on investigating whether the baseline hazards for each strata in the data are proportional. Strata are defined according to the value of categorical covariates. In the case of a huge number of categorical covariates in the data set, the number of strata could be so large that it would not be possible to perform this procedure efficiently. In the case of continuous covariates strata can

ve arbitrarily defined based on the covariate values, but then the model is different than the original one.

Another approach consists on investigating whether or not a time-dependency with a pre-specified parametric form is present. For example, the following model can be consider (Scheike & Martinussen, 2004)

$$\lambda_i(t) = Y_i(t)\alpha_0(t)\exp(\beta^t Z_i + Z_{i1}\theta f(t))$$

where $f(x)$ is a pre-specified function. Then, it should be tested whether or not θ is equal to 0. With this procedure evidences of departures close to $f(t)$ can be found, but it has the disadvantage that it is necessary to have a precise idea about the type of departure from proportionality, which rarely occurs. This is even more difficult in the case of an extended version of this model where more than one dimension are considered.

These two traditional approaches can be used to have some rough evidence of departures from the model, but both of them have the disadvantage of considering the covariates one at a time in a model where constant effects are assumed for the rest of covariates. Scheike and Martinussen (2004) claim that it is preferable to make test for one component at a time, starting with the model where all effects are allowed to be time-varying and then gradually simplifying the model appropriately.

Alternative approaches based on functionals of martingale residuals have been proposed, for example Lin, Wei & Ying (1993) and Schoenfeld (1982), who introduced the so-called Schoenfeld residuals. Grambsch & Therneau (1994) (GT) estimated time-varying regression coefficients by smoothing Schoenfeld residuals (see Martinussen, Scheike & Skovgaard, 2002, for detailed comments about this procedure). Despite the fact that the GT method is widely used, two potential drawbacks of this procedure are mentioned in Scheike & Martinussen (2004): firstly, the method results in a one-step estimator based on Cox's estimate and cannot be proved to have good properties, such as consistency if the true model contains time-varying effects; secondly, confidence intervals are computed assuming the Cox model and are only valid under this model. Additionally, Scheike & Martinussen (2004) remark the

difficulties arising when the estimation of the possible time-dependent regression coefficient $\beta(t)$ is used for hypothesis testing. Many authors argue that cumulative regression functions, $B(t) = \int_0^t \beta(s)ds$, are preferable when hypothesis testing about $\beta(t)$ is the issue, see Murphy (1993) and Murphy & Sen (1991).

The following extension of the Cox model have been studied by a number of authors, Murphy & Sen (1991) and Grambsch & Thearneau (1994) among many others

$$\lambda_i(t) = Y_i(t) \exp\left[\beta(t)^t Z_i(t)\right] \tag{2.2}$$

where $Z_i(t)$ are p-dimensional covariates and $\beta(t)$ denote the associated regression coefficients. When the first covariate is constant and equal to one $Z_{i1}(t) = 1$ the model contains a baseline $\alpha_0(t)$ that is parametrized as $\exp[\beta_1(t)]$. Martinussen, Scheike & Skovgaard (2002) generalized the previous model to allow that some covariates have constant effects, therefore they formulate the following model

$$\lambda_i(t) = Y_i(t) \exp\left[\beta(t)^t Z_i(t) + \gamma X_i(t)\right] \tag{2.3}$$

where $Z_i(t)$ and $X_i(t)$ are covariates of dimension p_1 and p_2, respectively, and $\beta(t)$ and γ denote the associated regression coefficients. Martinussen, Scheike & Skovgaard (2002) remark that some effects may not depend on time and should therefore not be fitted as general non-parametric regression functions. The same authors recommend to start with the model where all the covariates are allowed to have time-varying effects, and provide tests to decide if these effects are in fact time-varying (see Martinussen, Scheike & Skovgaard, 2002, and Scheike & Martinussen, 2004). The corresponding test statistics are based on the asymptotic analysis of the cumulative regression functions in model (2.3). These tests are easy to implement, but the main contribution of the authors is that they propose a simple new procedure where they test if a specific covariate has a time-constant effect using a model that allows the other covariates to have combinations of time-varying and constant effects. The simulation techniques suggested by Lin, Wei & Ying (1993) and Lin, Fleming & Wei (1994) are used by the authors for the calculation of $p-$values and

the practical implementation of this procedure (for a more detailed description see Scheike & Martinussen, 2004).

Chapter 3

Improving the efficiency of the Nelson-Aalen estimator: the naive local constant estimator

3.1 Introduction

The Nelson-Aalen estimator, devised by Nelson (1969), Nelson (1972) and Aalen (1978), is well known to be an asymptotically efficient estimator of the cumulative hazard function, see Andersen, Borgan, Gill & Keiding (1993) among many others. In this chapter[1], we show that the efficiency of the Nelson-Aalen estimator can be considerably improved by using more information in the estimation process than the traditional Nelson-Aalen estimator uses. While our approach results in a biased estimator, the variance improvement is substantial. By carefully optimizing the balance between the bias loss and the variance improvement, we obtain results on the efficiency gain. Several examples for known failure time distributions are used to illustrate these ideas. In the following chapters, the proposed non-parametric estimator will be used in survival models. Here, for sake of simplicity, a small application of the proposed estimator is used in a classical context. A well-known survival data set on cancer research is used for illustrative pruposes and a comparison

[1]Some parts of this chapter are also part of the paper: Guillén, M., Nielsen, J.P. & Perez-Marin, A.M. (2005), "Imporving the efficiency of the Nelson-Aalen estimator: the naive local constant estimator," submitted for publication.

with the traditional Nelson-Aalen estimator is discussed.

3.2 The naive local constant estimator

There is a comprehensive knowledge of the main statistical properties of the basic nonparametric estimators of the survival function, the hazard rate, the density function and the distribution function. Azzalini (1981), Reiss (1981) and Falk (1983) can be mentioned as examples of contributions to the knowledge of the theoretical properties of the kernel distribution function estimator introduced by Nadaraya (1964).

Bandwidth selection is essential in non-parametric estimation. A number of methods have been proposed for selecting the smoothing parameter in kernel density estimation, see for example Rudemo (1982) and Bowman (1984). Wand & Jones (1995, Ch. 3) give a thorough discussion of those methods. Sarda (1993) and Altman & Leger (1995), studied bandwidth selection for estimating distribution functions. Falk (1983) also gave relative efficiencies of kernel type estimators of distribution functions.

Bowman, Hall & Prvan (1998) discussed the performance of optimal, data-based methods of bandwidth choices for distribution functions leading to results which do not have analogues in the context of density estimation. In the discussion they pointed out that "care should be taken in cases where the largest survival times are censored. A further issue arises from the fact that survival times are usually greater than zero. This 'edge effect' will also require special attention". In this chapter we will use counting process theory for non-parametric estimation of the cumulative hazard rate function of a duration variable in the context of censored data. Moreover, the behaviour of durations near zero ('edge effect') is studied in detail. Both problems were mentioned in Bowman, Hall & Prvan (1998) but have not been addressed before.

The Nelson-Aalen estimator has proved useful for a number of applications in fields including actuarial science, biostatistics, finance and reliability theory. The

Nelson-Aalen estimator is well known to be an asymptotically efficient estimator of the cumulative hazard, see Andersen, Borgan, Gill & Keiding (1993, section IV.1) among many others. In this chapter we challenge the more or less accepted point of view that the Nelson-Aalen estimator is also close to efficiency when it comes to the small samples encountered in real life. Consider the following quite simple heuristics: when the Nelson-Aalen estimator is estimated at a point t, indicating duration, then it only uses the information available in the interval $[0, t]$. One might be able to improve the Nelson-Aalen estimator using some information just to the right of t, let us say, in $[t, t + b]$, where b is small and depends on t. This will certainly improve on the variance, since more information is included. It does, however, introduce some bias. The more information we include, i.e. the bigger b is, the bigger the variance improvement and the bigger the bias. It seems natural to expect that there could exist some kind of trade-off, an optimal b, that gives the optimal balance between improved variance and penalty of bias.

We introduce the simplest possible estimator of the cumulative hazard that employs information to the right of the point of interest. Since our approach is based on the assumption that the hazard is locally constant around the point of interest, we name our estimator the naive local constant estimator. The term 'naive' is taken from Silverman (1986) who used it for a kernel density estimator, where the kernel equalled the uniform distribution. Our mathematical analysis shows that the variance improves for increasing b and a non trivial optimal b exists that has an improved performance compared to the Nelson-Aalen estimator. This improvement is quite substantial with efficiency gains of up to 60%. Not surprisingly, we normally obtain the biggest efficiency gains for small t's, where the variance improvement is more important than for big t's, and we normally see that the efficiency gain is a falling function of t. Exceptions to this simple and intuitive rule do however exist. If we consider a Gamma distribution on the positive axis, then the efficiency gain has a U-shape as a function of time, with the smallest efficiency gain obtained around the central quantiles. The naive local constant estimator can be seen as a second order approximation to the Nelson-Aalen estimator. Our efficiency considerations

are therefore most important for small sample sizes. This parallels other second order approximations in the statistical literature, for example the Bartlett correction (Bartlett, 1937 and Lawley, 1956). Even though the derivation of the Bartlett correction is based on asymptotic theory, it is designed to work well for smaller sample sizes, where the first order correction is not sufficiently accurate.

The ideas presented here can be extended with few modifications to the Kaplan-Meier estimator (Kaplan & Meier, 1958) or even to the analysis of conditional survival functions (Tsai, Jewell & Wang, 1987, and Van der Laan, Jewell & Peterson, 1997).

We adapt the model formulation of Andersen, Borgan, Gill & Keiding (1993, p. 176) with an infinite terminal point $\tau = \infty$ and where this terminal point is not included in the considered interval, namely $[0, \infty[$. We are interested in the asymptotic distribution of an estimator of the cumulated hazard for some $t \in [0, \infty[$. We consider a measurable space (Ω, F), equipped with a filtration $(F_t, t \in [0, \infty[)$ satisfying the usual conditions except for possible completeness, see Andersen, Borgan, Gill & Keiding (1993, p. 60), for each member of a family P of probability measures. Defined on (Ω, F) and adapted to the filtration, we have a multivariate counting process $N = \{N_1(t), \ldots, N_n(t)\}$, where $t \in [0, \infty[$ and n is the number of counting processes, satisfying Aalen's multiplicative intensity model, i.e., its $(P, F_t)-$intensity process is $\lambda_i(t) = \alpha(t)Y_i(t)$, where Y_i is an observable predictable process taking values in $\{0, 1\}$, indicating, by the value 1, when the ith individual is under risk. We assume in the following that the hazard function α does not depend on i, is twice continuously differentiable, $\int_0^t \alpha(s)ds < \infty$, $\int_0^t \alpha(s)ds \neq 0$ and $\alpha'(t) \neq 0$ for all $t \in [0, \infty[$. When studing the large sample properties in section 3.2, we consider the limit $n \to \infty$, and we also assume that the usual Lindeberg type of conditions, that make Rebolledo's martingale theorem apply, hold. Thus, we assume that conditions 4.1.12, 4.1.13 and 4.1.14 of Andersen, Borgan, Gill & Keiding (1993, p. 191) hold.

Our efficiency comparison of the classical Nelson-Aalen estimator and our naive local constant estimator of the cumulative hazard is based on the above standard assumptions.

If $Y = \sum_{i=1}^{n} Y_i$ is the aggregated exposure, the sum of individual processes indicating that the unit is under risk, and

$$\widetilde{\Lambda}(t) = \int_0^t \alpha(s) I_{\{Y(s)>0\}} ds,$$

is the cumulative hazard, where $I_{\{Y(s)>0\}}$ denotes the indicator for $Y(s) > 0$. Then the classical Nelson-Aalen estimator equals

$$\widehat{\Lambda}_{NA}(t) = \sum_{i=1}^{n} \int_0^t \frac{1}{Y(s)} dN_i(s).$$

We now get that

$$\widehat{\Lambda}_{NA}(t) - \widetilde{\Lambda}(t) = \sum_{i=1}^{n} \int_0^t \frac{1}{Y(s)} dM_i(s)$$

is a martingale since $Y(s)$ is predictable, where $M_i = N_i - \Lambda_i$ is the counting process martingale and $\Lambda_i(t) = \int_0^t \alpha(s) Y_i(s) ds$ is the compensator of N_i with respect to the considered filtration.

The original standard Nelson-Aalen estimator is changed to incorporate some information to the right of t in order to reduce the variance and obtain a more efficient estimator. To be precise, we include information from the interval $[t, t+b]$, where b is a bandwidth. It is clear that including some information above t implies that this estimator introduces some bias and that this bias increases with b, just as the improvement in variance increases with b. So, given t, we therefore expect some optimal b to exist, where we would get a good variance reduction without being penalized too much with respect to bias. Below we define our naive local constant estimator and show that with respect to mean square error an optimal non trivial b exists. That the optimal bandwidth is non trivial is defined to mean that it is above zero. This implies that our naive local constant estimator improves efficiency compared to the classical Nelson-Aalen estimator.

Minimizing the mean square error is the classical method applied to get an optimal bandwidth. A number of references can be mentioned in that sense. Bowman, Hall & Prvan (1998) propose a crossvalidation procedure consisting on minimizing

an unbiased estimator of the mean integrated squared error curve to get the optimal bandwidth parameter in kernel estimation of distribution functions, but censoring was not considered in their investigation.

Our approach is the following, we keep the classical Nelson-Aalen estimator on the interval $[0, t-b]$, but while estimating the part of the cumulative hazard belonging to the interval $[t-b, t]$, we use information from the interval $[t-b, t+b]$. We do not use a smooth kernel or any other kind of smoothing while including this extra information. This is the reason for proposing the term naive when naming the estimator. This parallels the way Silverman (1986) uses this notion when it comes to kernel density estimators. Our point of view is that this is the simplest possible way we can include the extra information on the interval $[t, t+b]$ and that this estimator has to be analysed and understood before other more advanced estimators are proposed.

Now consider a counting process martingale in T defined as

$$\sum_{i=1}^{n} \int_0^T h_i(s) dM_i(s),$$

where each h_i is a predictable function with respect to filtration $(F_s, s \in [0, T[)$. Then

$$\sum_{i=1}^{n} \int_{a_1}^{a_2} h_i(s) dM_i(s) = \sum_{i=1}^{n} \int_0^T I_{\{s \in [a_1, a_2]\}} h_i(s) dM_i(s) \tag{3.1}$$

is a martingale in T.

In the rest of the paper we will call an expression of form (3.1) a martingale. Often we will have $a_2 = t+b$, where b is a bandwidth that might itself depend on t. The expression (3.1) is not a martingale in t in this case, but a martingale in some T bigger than a_2.

The naive local constant estimator is defined as follows

$$\widehat{\Lambda}_{NLC}(t) = \int_0^\infty w(s,t) d\widehat{\Lambda}_{NA}(s), \text{ for all } t$$

where $w(s,t)$ is some weight function, which in this paper we assume to be the

"naive" function

$$w(s,t) = I_{\{s \leq t-b\}} + \frac{1}{\gamma_{t,b}} I_{\{s \in (t-b,t+b)\}},$$

where

$$\gamma_{t,b} = \frac{t+b-\max(t-b,0)}{t-\max(t-b,0)},$$

or equivalently

$$\gamma_{t,b} = \begin{cases} 2 & if \quad t \geq b \\ \frac{t+b}{t} & if \quad t < b \end{cases}, \tag{3.2}$$

for some $b > 0$. Note that when considering a "naive" weight function, this estimator can be expressed as follows

$$\begin{aligned} \widehat{\Lambda}_{NLC}(t) &= \sum_{i=1}^{n} \int_0^{\max(t-b,0)} \frac{1}{Y(s)} dN_i(s) + \sum_{i=1}^{n} \int_{\max(t-b,0)}^{t+b} \frac{1}{\gamma_{t,b} Y(s)} dN_i(s) \\ &= \widehat{\Lambda}_{NA}\{\max(t-b,0)\} + \frac{1}{\gamma_{t,b}} \left[\widehat{\Lambda}_{NA}(t+b) - \widehat{\Lambda}_{NA}\{\max(t-b,0)\}\right]. \end{aligned}$$

One could of course extend the same ideas to the local linear estimator or even the multiplicative bias correction method, see Jones, Linton & Nielsen (1995) for the density equivalent.

3.3 Efficiency considerations and bandwidth selection

In this section we consider the mean square error of the naive local constant estimator. Based on this calculation we can develop an optimal local bandwidth for the naive local constant estimator of the cumulative hazard. We divide the analysis in two parts according to whether we are estimating at a point t belonging to the boundary region or whether we are estimating at a point t belonging to the interior of the interval. We assume that a T bigger than $t+b$ and a positive continuous function y exist such that

$$\sup_{s \in [0,T]} \left| \frac{Y(s)}{n} - y(s) \right| \to 0$$

in probability. We also assume that $Y(s)$ is strictly positive for $s \in [0,T]$. We do this for notational convenience.

In the Appendix A we have derived the following expansion of our naive local constant estimator

$$\begin{aligned}\widehat{\Lambda}_{NLC} - \widetilde{\Lambda}(t) &= \sum_{i=1}^{n} \int_0^{\max(t-b,0)} \frac{1}{Y(s)} dM_i(s) + \sum_{i=1}^{n} \int_{\max(t-b,0)}^{t+b} \frac{1}{\gamma_{t,b} Y(s)} dM_i(s) \\ &\quad + \int_{\max(t-b,0)}^{t+b} \frac{1}{\gamma_{t,b}} \alpha(s) ds - \int_{\max(t-b,0)}^{t} \alpha(s) ds \\ &= V_t + B_t,\end{aligned}$$

where the variable term V_t asymptotic variance is defined as

$$V_t = \sum_{i=1}^{n} \int_0^{\max(t-b,0)} \frac{1}{Y(s)} dM_i(s) + \sum_{i=1}^{n} \int_{\max(t-b,0)}^{t+b} \frac{1}{\gamma_{t,b} Y(s)} dM_i(s),$$

and the stable term B_t asymptotic bias is defined as

$$\begin{aligned}B_t &= \int_{\max(t-b,0)}^{t+b} \frac{1}{\gamma_{t,b}} \alpha(s) ds - \int_{\max(t-b,0)}^{t} \alpha(s) ds \\ &= \int_{\max(t-b,0)}^{t} \left(\frac{1}{\gamma_{t,b}} - 1 \right) \alpha(s) ds + \int_{t}^{t+b} \frac{1}{\gamma_{t,b}} \alpha(s) ds.\end{aligned}$$

The predictable variation of the variable part process is

$$\langle V_t \rangle = \int_0^{\max(t-b,0)} \frac{\alpha(s)}{Y(s)} ds + \int_{\max(t-b,0)}^{t+b} \frac{\alpha(s)}{\gamma_{t,b}^2 Y(s)} ds.$$

Once we have obtained the general expression for the bias and the variance, in order to calculate the relative efficiency of the naive local constant estimator we have found it convenient to consider the standard case, $t \geq b$, and the boundary case, $t < b$, separately.

3.3.1 Relative efficiency if $t \geq b$

First we consider the analysis, where we are at the interior of the interval, $t \geq b$. The bias expression is

$$\begin{aligned} B_t &= -\int_{t-b}^{t} \frac{1}{2}\alpha(s)ds + \int_{t}^{t+b} \frac{1}{2}\alpha(s)ds \\ &= -\int_{t}^{t+b} \frac{1}{2}\alpha(s-b)ds + \int_{t}^{t+b} \frac{1}{2}\alpha(s)ds \\ &= \frac{1}{2}\int_{t}^{t+b} \{\alpha(s) - \alpha(s-b)\}ds = \frac{1}{2}b^2\alpha'(t) + o_P(b^2), \end{aligned}$$

and the predictable variation of the variable part is

$$\langle V_t \rangle = \int_0^{t-b} \frac{\alpha(s)}{Y(s)}ds + \int_{t-b}^{t+b} \frac{\alpha(s)}{4Y(s)}ds.$$

The variance gain with respect to the original Nelson-Aalen estimator is

$$\begin{aligned} \int_0^t \frac{\alpha(s)}{Y(s)}ds - \left\{\int_0^{t-b} \frac{\alpha(s)}{Y(s)}ds + \int_{t-b}^{t+b} \frac{\alpha(s)}{4Y(s)}ds\right\} &= \int_{t-b}^{t} \frac{\alpha(s)}{Y(s)}ds - \int_{t-b}^{t+b} \frac{\alpha(s)}{4Y(s)}ds \\ &= \frac{1}{2}b\frac{\alpha(t)}{Y(t)} + o_P(bn^{-1}). \end{aligned}$$

Then the total gain in terms of mean square error is

$$Q_0(b) = \frac{1}{2}b\frac{\alpha(t)}{Y(t)} - \left\{\frac{1}{2}b^2\alpha'(t)\right\}^2 + o_P(bn^{-1}) + o_P(b^4).$$

When $\alpha'(t)$ is positive this defines the leading term of the total gain as

$$Q(b) = \frac{1}{2}b\frac{\alpha(t)}{Y(t)} - \left\{\frac{1}{2}b^2\alpha'(t)\right\}^2.$$

The optimal bandwidth b is found by differentiating the total gain,

$$Q'(b_{opt}) = \frac{1}{2}\frac{\alpha(t)}{Y(t)} - b_{opt}^3\alpha'(t)^2 = 0,$$

so that the optimal bandwidth is

$$b_{opt} = \left\{\frac{\alpha(t)}{2Y(t)\alpha'(t)^2}\right\}^{\frac{1}{3}}. \tag{3.3}$$

The efficiency gain at the optimal bandwidth b can be simplified as

$$Q(b_{opt}) = \frac{3}{8} b_{opt} \frac{\alpha(t)}{Y(t)}.$$

The ratio between the total gain obtained for the naive local constant estimator and the variance of the Nelson-Aalen estimator, the relative efficiency gain, equals

$$\varepsilon(t) = \frac{Q(b_{opt})}{\int_0^t \frac{\alpha(s)}{Y(s)} ds} = \frac{3}{8} \frac{\alpha(t)}{\int_0^t \alpha(s) ds} b_{opt}.$$

The relative efficiency gain is therefore of the quite significant order of magnitude $n^{-1/3}$.

The relative efficiency gain does, however, also depend on t, and we see that it will go to 0 for t going to infinity for any bounded α. However, many hazards are unbounded and for some of those, a significant relative efficiency gain remains for $t \to \infty$. In section 3.4, where the relative efficiency gain is analysed for a number of realistic hazards, we see that the gain is far from negligible, with values up to 60% and mostly above 10% for most of the relevant t's considered.

3.3.2 Relative efficiency if $t < b$

Now we investigate the boundary case, where $t < b$. The bias is

$$\begin{aligned} B_t &= \int_0^t \left(\frac{t}{t+b} - 1 \right) \alpha(s) ds + \int_t^{t+b} \frac{t}{t+b} \alpha(s) ds \\ &= \int_0^{t+b} \frac{t}{t+b} \alpha(s) ds - \int_0^t \alpha(s) ds \\ &= \int_0^t \left\{ \alpha\left(\frac{t+b}{t} s \right) - \alpha(s) \right\} ds = \frac{1}{2} bt\alpha'(t) + o_P(bt) \end{aligned}$$

and the predictable variation process is

$$\langle V_t \rangle = \int_0^{t+b} \left(\frac{t}{t+b} \right)^2 \frac{\alpha(s)}{Y(s)} ds.$$

The variance gain is

$$\int_0^t \frac{\alpha(s)}{Y(s)} ds - \int_0^{t+b} \left(\frac{t}{t+b} \right)^2 \frac{\alpha(s)}{Y(s)} ds = \frac{tb}{t+b} \frac{\alpha(t)}{Y(t)} + o_P(\frac{tb}{t+b}).$$

The total gain is

$$Q_0(b) = \frac{tb}{t+b}\frac{\alpha(t)}{Y(t)} - \left\{\frac{1}{2}bt\alpha'(t)\right\}^2 + o_P\left(\frac{tb}{t+b}\right) + o_P(bt).$$

So, we define the leading term of the total gain as

$$Q(b) = \frac{\alpha(t)}{Y(t)}\frac{tb}{t+b} - \left\{\frac{1}{2}bt\alpha'(t)\right\}^2 = \alpha'(t)^2\left(2b_{opt}^3\frac{tb}{t+b} - \frac{1}{4}b^2t^2\right),$$

where $b_{opt} = [\alpha(t)/\{2Y(t)\alpha'(t)^2\}]^{1/3}$ is the optimal bandwidth we would have got in t, if t had not been in the boundary region.

In order to find the optimal bandwidth in the boundary case defined by $t < b_{opt}$, we normalise our optimization problem and express b as some portion of b_{opt}. Thus, $b = q_1 b_{opt}$ for $q_1 \in (0, q_2)$, where $q_2 = t/b_{opt}$ for $q_2 \in (0,1)$.

Note that the boundary bandwidth b can not be bigger than t. The variance gain in the boundary case can be expressed as

$$Q(q_1) = \alpha'(t)^2 b_{opt}^4\left(2\frac{q_1q_2}{q_1+q_2} - \frac{1}{4}q_1^2q_2^2\right), \text{ for } q_1 \in (0,q_2) \text{ and } q_2 \in (0,1).$$

The slope of $Q(q_1)$ is strictly greater than 0 and the maximum is therefore obtained for $q_1 = q_2$ and we get the optimal boundary bandwidth $b_{opt,b} = t$.

The total efficiency gain at the optimal boundary bandwidth $b_{opt,b}$ is

$$Q(b_{opt,b}) = \frac{t}{2}\left\{\frac{\alpha(t)}{Y(t)} - \frac{t^3}{2}\alpha'(t)^2\right\}$$

and the relative efficiency gain is

$$\varepsilon(t) = \frac{Q(b_{opt,b})}{\int_0^t \frac{\alpha(s)}{Y(s)}ds} = \frac{\frac{t}{2}\left\{\alpha(t) - \frac{Y(t)t^3}{2}\alpha'(t)^2\right\}}{\int_0^t \alpha(s)ds}.$$

This is always positive since $b_{opt} > t$. This condition is equivalent to $Y(t)t^3\alpha'(t)^2 < \alpha(t)/2$, and then $Y(t)t^3\alpha'(t)^2/2 < \alpha(t)/4$.

A general description of the relative efficiency gain is

$$\varepsilon(t) = \begin{cases} \frac{3}{8}\frac{\alpha(t)}{\int_0^t \alpha(s)ds}b_{opt} & \text{if } b_{opt} \le t \\ \frac{\frac{t}{2}\left\{\alpha(t) - \frac{Y(t)t^3}{2}\alpha'(t)^2\right\}}{\int_0^t \alpha(s)ds} & \text{if } b_{opt} > t. \end{cases} \tag{3.4}$$

Note that the relative efficiency gain is continuous in b_{opt} since

$$\lim_{t \to b_{opt}^-} \frac{\frac{t}{2}\left\{\alpha(t) - \frac{Y(t)t^3}{2}\alpha'(t)^2\right\}}{\int_0^t \alpha(s)ds} = \lim_{t \to b_{opt}^+} \frac{3}{8}\frac{\alpha(t)}{\int_0^t \alpha(s)ds} b_{opt},$$

and the optimal bandwidth b_{opt} is a function of t.

3.4 Relative efficiency gain

In this section we present the relative efficiency gain curve which compares, at every given t, the relative efficiency gain of the naive local constant estimator to the Nelson-Aalen estimator. Several typical distributions of a random variable measuring time-to-event are considered in this section. In the calculations $Y(t)$, the function that indicates the exposure risk, is assumed to be equal to $100S(t)$, where $S(t)$ is the survival function. For each distribution function considered here, the survival function, the hazard function and the relative efficiency gain curve are shown (see, Figures 3.1 and 3.2). The hazard and the relative efficiency gain are plotted as a function of the percentile of t.

Uniform distribution

Let us assume that the true distribution for the time-to-event variable is the uniform distribution. So, the density function $f(t)$ equals $1/\theta$ for $0 \leq t \leq \theta$, where θ represents the maximum value for the random variable measuring the time-to-event. The survival function is then a straight line with a constant negative slope and the hazard is an increasing function defined by $\alpha(t) = 1/(\theta - t)$ for $0 \leq t \leq \theta$.

The relative efficiency gain of the naive local constant estimator with respect to the Nelson-Aalen estimator has the following simple expression

$$\varepsilon(t) = \begin{cases} \frac{3}{-8\log(1-\frac{t}{\theta})\{2Y(t)\}^{1/3}} & \text{if } b_{opt} \leq t \\ \frac{2t(\theta-t)^3 - Y(t)t^4}{-4(\theta-t)^4\log(1-\frac{t}{\theta})} & \text{if } b_{opt} > t \end{cases}$$

It is easy to prove positiveness for $b_{opt} \leq t$, but when $b_{opt} > t$, the boundary case, we have to take into account that $t^3 < (t-\theta)^3/\{2Y(t)\}$, which is the boundary

condition for the uniform distribution. Therefore, the naive local constant estimator is more efficient than the Nelson-Aalen estimator in terms of the mean square error, and its relative efficiency gain depends on t, θ, and $Y(t)$ which is now supposed to be equal to $100S(t)$. Figure 3.1 shows the survival function, the hazard and the relative efficiency gain curve for three given values of θ.

On the one hand, for a given value for θ and $Y(t)$, the relative efficiency gain decreases as t increases, as expected. It can be shown that

$$\lim_{t\to\theta}\varepsilon(t) = \lim_{t\to\theta}\frac{3}{-8\log(1-\frac{t}{\theta})\left\{2Y(t)\right\}^{1/3}} = 0.$$

On the other hand the limit of $\varepsilon(t)$ when $t \to 0$ is

$$\lim_{t\to 0}\varepsilon(t) = \lim_{t\to 0}\frac{2t(\theta-t)^3 - Y(t)t^4}{-4(\theta-t)^4\log(1-\frac{t}{\theta})} = \frac{1}{2}.$$

The maximum relative efficiency gain is not obtained for $t \to 0$, because $\varepsilon(t)$ takes values greater than $1/2$ for some $0 < t < b_{opt}$. As shown in Figure 3.1, the relative efficiency gain is well above 20% in the first quartile, with values around 50% in the first decile.

Weibull distribution

If we assume that the time-to-event random variable follows a Weibull distribution, with density $f(t)$ equal to $\alpha\lambda t^{\alpha-1}\exp(-\lambda t^{\alpha})$ for $\lambda > 0, \alpha > 0$ and $t \geq 0$, then the hazard is $\alpha(t) = \alpha\lambda t^{\alpha-1}$and the relative efficiency gain is then

$$\varepsilon(t) = \begin{cases} \frac{3}{8}\frac{\alpha}{\left\{2Y(t)\lambda\alpha(\alpha-1)^2t^{\alpha}\right\}^{1/3}} & \text{if } b_{opt} \leq t \\ \frac{1}{2}\alpha\left\{1 - Y(t)\frac{\alpha\lambda t^{\alpha}(\alpha-1)^2}{2}\right\} & \text{if } b_{opt} > t. \end{cases}$$

In Figure 3.1, we have also presented a graph showing the survival function, the hazard function and the relative efficiency gain curve for some parameter values of the Weibull distribution.

Maximum relative efficiency gain is obtained for $t \to 0$, and its value is $\alpha/2$. Notice that the minimum relative efficiency gain is not neglectable, as one can see in Figure 3.1 that there is a relative efficiency gain above 30% for any value of t.

Gamma distribution

We will now assume that the distribution for the random variable indicating time to event is the Gamma distribution. In this situation, the density is $f(t) = \{\beta^{\alpha}\Gamma(\alpha)\}^{-1} t^{\alpha-1}\exp(-t/\beta)$, where $\alpha > 0, \beta > 0$ and $t \geq 0$. The hazard function is

$$\alpha(t) = \frac{f(t)}{1-F(t)} = \frac{\frac{1}{\beta^{\alpha}\Gamma(\alpha)} t^{\alpha-1}\exp(-t/\beta)}{1-\int_0^t \frac{1}{\beta^{\alpha}\Gamma(\alpha)} s^{\alpha-1}\exp(-s/\beta)ds} = \frac{\frac{1}{\beta^{\alpha}\Gamma(\alpha)} t^{\alpha-1}\exp(-t/\beta)}{\int_t^{\infty} \frac{1}{\beta^{\alpha}\Gamma(\alpha)} s^{\alpha-1}\exp(-s/\beta)ds},$$

where $f(\cdot)$ and $F(\cdot)$ represent the Gamma density function and distribution function, respectively. The relative efficiency gain is

$$\varepsilon(t) = \begin{cases} \frac{3}{8}\frac{f(t)^{4/3}}{-\log\{1-F(t)\}\{2Y(t)\}^{1/3}\left[f'(t)\{1-F(t)\}+\{f(t)\}^2\right]^{2/3}} & \text{if } b_{opt} \leq t \\ \frac{tf(t)\{1-F(t)\}^3-\frac{Y(t)t^4}{2}\left[f'(t)\{1-F(t)\}+\{f(t)\}^2\right]^2}{-2\log\{1-F(t)\}\{1-F(t)\}^4} & \text{if } b_{opt} > t. \end{cases}$$

In Figure 3.1, we have shown an example of the survival function, the hazard and the relative efficiency gain curve for this type of distribution.

When the duration variable follows a Gamma distribution, one can notice that the best gain is obtained at the lower and higher distribution quantiles. For the values of the parameters that were used for illustration, we see that the relative efficiency gain is in general above 20%, but it may reach almost 60% for the highest quantiles.

Lognormal distribution

Let us now assume that the random variable which measures time to event follows a Lognormal distribution, so

$$f(t) = \frac{1}{(2\pi)^{1/2}\sigma t}\exp\left[-\frac{1}{2}\left\{\frac{\log(t)-\mu}{\sigma}\right\}^2\right] \quad \text{for } \sigma > 0, \text{ and } t \geq 0.$$

The hazard function is

$$\alpha(t) = \frac{f(t)}{1-F(t)} = \frac{\frac{1}{(2\pi)^{1/2}\sigma t}\exp\left[-\frac{1}{2}\left\{\frac{\log(t)-\mu}{\sigma}\right\}^2\right]}{1-\Phi\left\{\frac{\log(t)-\mu}{\sigma}\right\}},$$

where $f(\cdot)$ and $F(\cdot)$ represent the Lognormal density function and distribution function, respectively, and $\Phi(\cdot)$ is the standard Normal distribution function.

The relative efficiency gain is

$$\varepsilon(t) = \begin{cases} \frac{3}{8} \frac{f(t)^{4/3}}{-\log\left[1-\Phi\left\{\frac{\log(t)-\mu}{\sigma}\right\}\right]\{2Y(t)\}^{1/3}\left(f'(t)\left[1-\Phi\left\{\frac{\log(t)-\mu}{\sigma}\right\}\right]+\{f(t)\}^2\right)^{2/3}} & \text{if } b_{opt} \leq t \\ \frac{tf(t)\left[1-\Phi\left\{\frac{\log(t)-\mu}{\sigma}\right\}\right]^3-\frac{Y(t)t^4}{2}\left(f'(t)\left[1-\Phi\left\{\frac{\log(t)-\mu}{\sigma}\right\}\right]+\{f(t)\}^2\right)^2}{-2\log\left[1-\Phi\left\{\frac{\log(t)-\mu}{\sigma}\right\}\right]\left[1-\Phi\left\{\frac{\log(t)-\mu}{\sigma}\right\}\right]^4} & \text{if } b_{opt} > t. \end{cases}$$

Figure 3.1 shows an example for the survival function, the hazard and the relative efficiency gain curve in the Lognormal situation. The shape of the relative efficiency gain curve is different to the previous ones for this kind of distribution. While lying above about 20% in all the domain, the relative efficiency gain may reach about 70% in the central part of the distribution domain.

Log-logistic distribution

Let us assume that the distribution for the time-to-event variable is the Log-logistic distribution. Then, the density is $f(t) = \alpha\lambda t^{\alpha-1}\left(1+\lambda t^{\alpha}\right)^{-2}$ for $\alpha > 0, \lambda > 0$ and $t \geq 0$. The hazard function is now $\alpha(t) = \alpha\lambda t^{\alpha-1}(1+\lambda t^{\alpha})^{-1}$.

The relative efficiency gain is

$$\varepsilon(t) = \begin{cases} \frac{3}{8} \frac{\left(\alpha\lambda t^{\alpha-1}\right)^{4/3}}{\log(1+\lambda t^{\alpha})\{2Y(t)\}^{1/3}\{\alpha(\alpha-1)\lambda t^{\alpha-2}(1+\lambda t^{\alpha})-(\lambda\alpha t^{\alpha-1})^2\}^{2/3}} & \text{if } b_{opt} \leq t \\ \frac{\alpha t^{\alpha}\lambda(1+\lambda t^{\alpha})^3-\frac{Y(t)t^4}{2}\{\alpha(\alpha-1)\lambda t^{\alpha-2}(1+\lambda t^{\alpha})-(\alpha\lambda t^{\alpha-1})^2\}^2}{2(1+\lambda t^{\alpha})^4\log(1+\lambda t^{\alpha})} & \text{if } b_{opt} > t. \end{cases}$$

In Figure 3.2 we graph the survival function, the hazard and the relative efficiency gain curve for a Log-logistic distribution.

When looking at the plot of the relative efficiency gain curve, we see that the larger gains, of about 50%, are reached at the lowest quantiles.

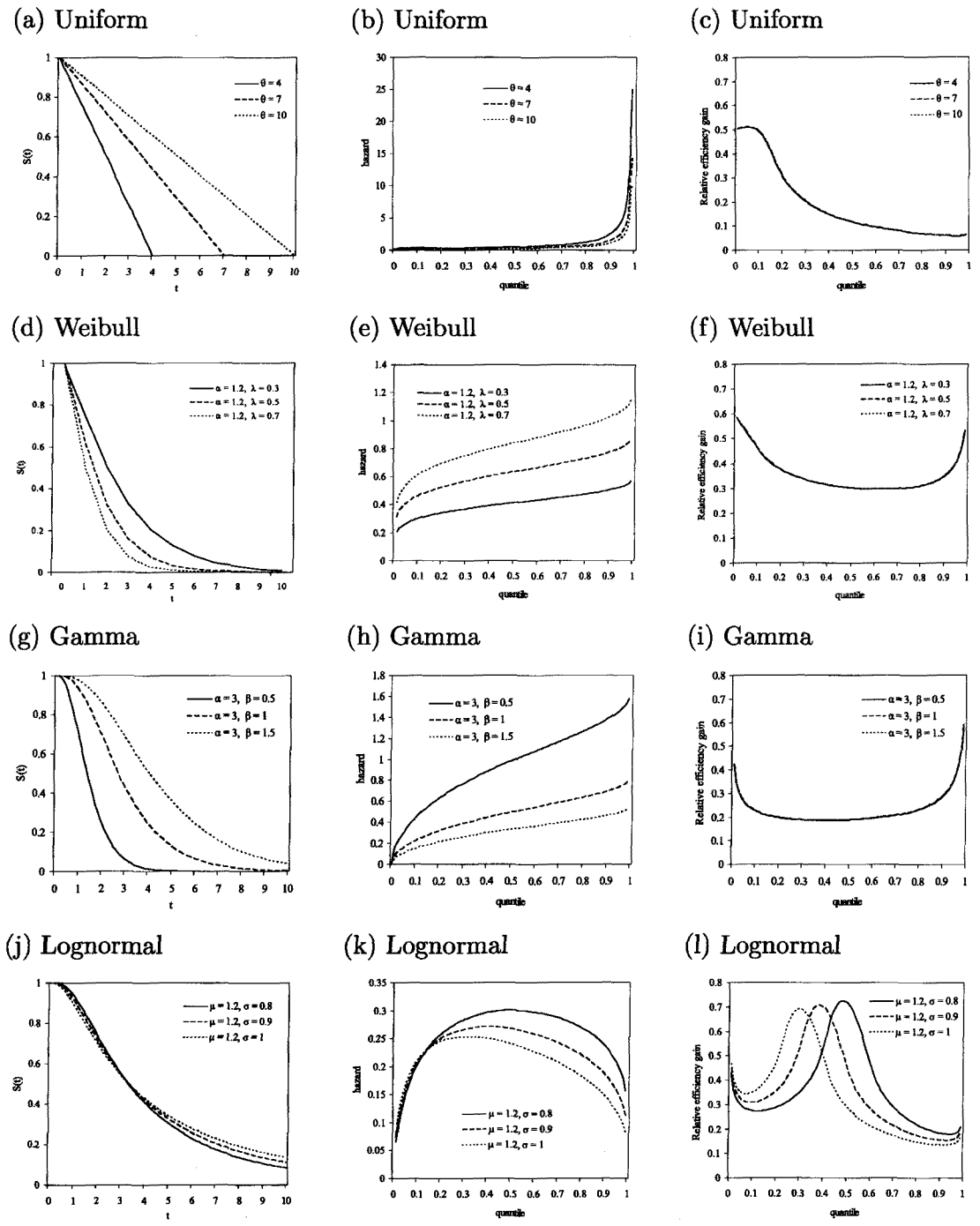


Figure 3.1. Survival function, hazard function and relative efficiency gain curve for some known distributions: Uniform, Weibull, Gamma and Lognormal. (a), (d), (g) and (j) survival function, (b), (e), (h) and (k) hazard function and (c), (f), (i) and (l) relative efficiency gain curve.

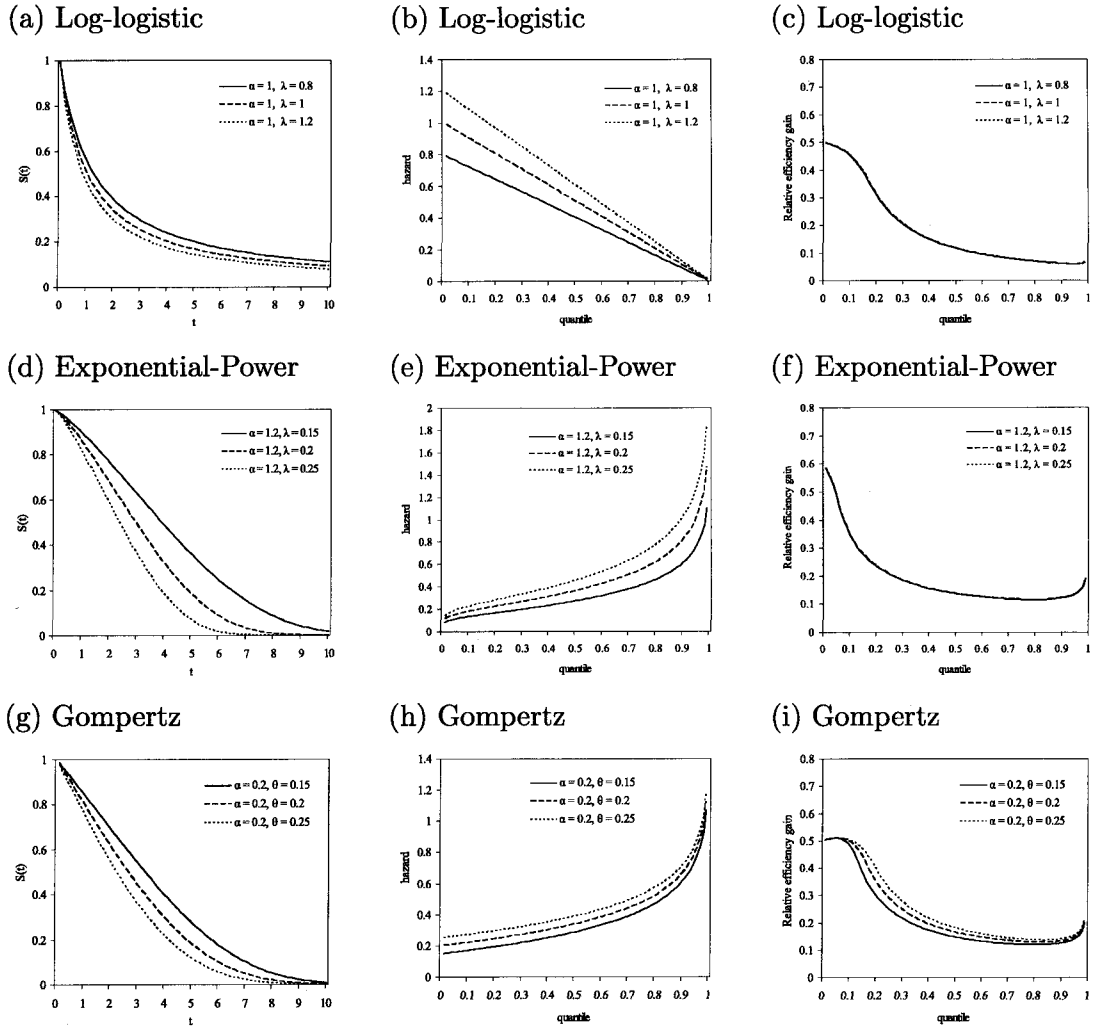


Figure 3.2. Survival function, hazard function and relative efficiency gain curve for some known distributions: Log-logistic, Exponential-Power and Gompertz. (a), (d) and (g) survival function, (b), (e) and (h) hazard function and (c), (f) and (i) relative efficiency gain curve.

Exponential Power distribution

If the distribution for the time-to-event variable is now the Exponential Power distribution, the density is $f(t) = \alpha\lambda^{\alpha}t^{\alpha-1}\exp\left[1+(\lambda t)^{\alpha}-\exp\left\{(\lambda t)^{\alpha}\right\}\right]$ for $\alpha, \lambda > 0$ and $t \geq 0$. The hazard function is $\alpha(t) = \alpha\lambda^{\alpha}t^{\alpha-1}\exp\left\{(\lambda t)^{\alpha}\right\}$ and the relative efficiency gain is then

$$\varepsilon(t) = \begin{cases} \frac{3}{8}\frac{\left[\alpha\lambda^{\alpha}t^{\alpha-1}\exp\{(\lambda t)^{\alpha}\}\right]^{4/3}}{\{2Y(t)\}^{1/3}[\exp\{(\lambda t)^{\alpha}\}-1][\alpha\lambda^{\alpha}t^{\alpha-2}\exp\{(\lambda t)^{\alpha}\}\{\alpha-1+\alpha(t\lambda)^{\alpha}\}]^{2/3}} & \text{if } b_{opt} \leq t \\ \frac{\alpha\lambda^{\alpha}t^{\alpha}\exp\{(\lambda t)^{\alpha}\}-\frac{Y(t)}{2}t^{4}\left[\alpha\lambda^{\alpha}t^{\alpha-2}\exp\{(\lambda t)^{\alpha}\}\{\alpha-1+\alpha(t\lambda)^{\alpha}\}\right]^{2}}{2[\exp\{(\lambda t)^{\alpha}\}-1]} & \text{if } b_{opt} > t. \end{cases}$$

In Figure 3.2 the survival function, the hazard and the relative efficiency gain curve are shown for the Exponential Power distribution. The maximum relative efficiency gain is obtained at the lower quantiles.

Gompertz distribution

Let us assume that the distribution for the time-to-event variable is the Gompertz distribution, so the density is $f(t) = \theta\exp(\alpha t)\exp\left[(\theta/\alpha)\left\{1-\exp(\alpha t)\right\}\right]$ for $\theta, \alpha > 0$, and $t \geq 0$. The hazard is $\alpha(t) = \theta\exp(\alpha t)$. The relative efficiency gain is

$$\varepsilon(t) = \begin{cases} \frac{3}{8}\frac{\alpha^{1/3}\exp\left(\frac{2}{3}\alpha t\right)}{\{2\theta Y(t)\}^{1/3}\{\exp(\alpha t)-1\}} & \text{if } b_{opt} \leq t \\ \frac{t\theta e^{\alpha t}-\frac{Y(t)t^{4}}{2}\{\theta\alpha\exp(\alpha t)\}^{2}}{2\frac{\theta}{\alpha}\{\exp(\alpha t)-1\}} & \text{if } b_{opt} > t. \end{cases}$$

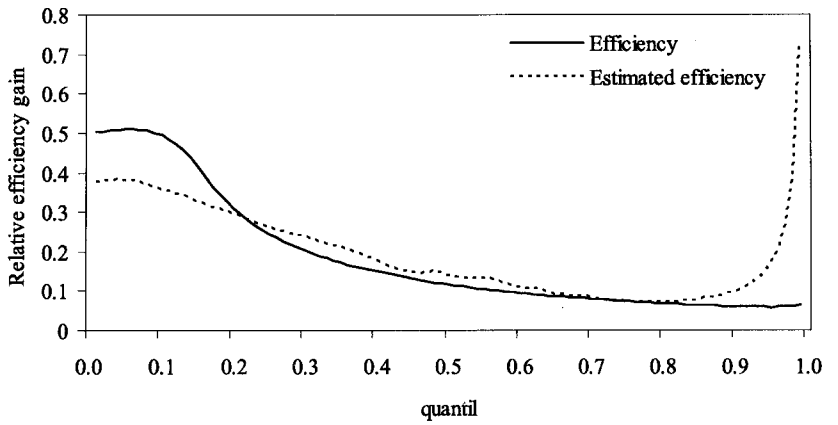
In Figure 3.2, we graph the survival function, the hazard and the relative efficiency gain curve. The best relative efficiency gains have been obtained at the lowest quantiles for this type of distribution, and again one can see gains reaching 50%.

3.5 The efficiency curve after estimating b

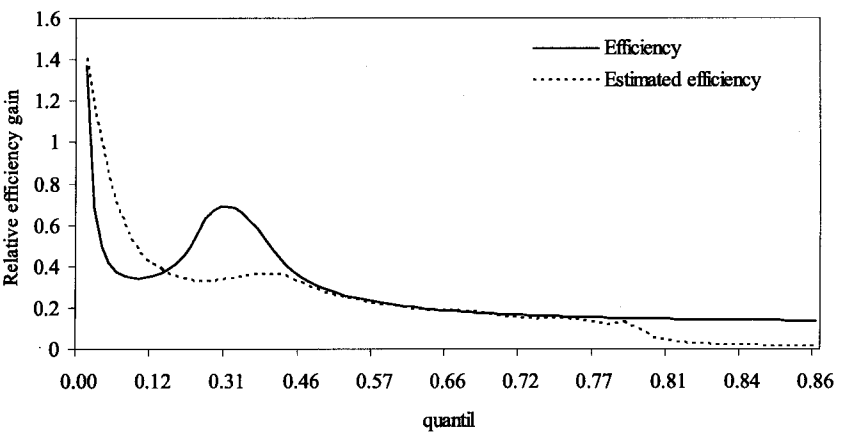
The theoretical efficiency curves shown in Figure 3.1 and 3.2 are based on knowing the optimal b, thus knowing $\alpha(t)$ and $\alpha'(t)$. Therefore, they do not adjust for the effect of plugging-in an estimation of b when calculating the efficiency.

A simulation study has been performed in order to account for the estimation of the optimal b (see Figure 3.3).

(a) Uniform, $\theta = 10$.



(b) Lognormal, $\mu = 1.2, \sigma = 1$.



(c) Gamma, $\alpha = 3, \beta = 1.5$.

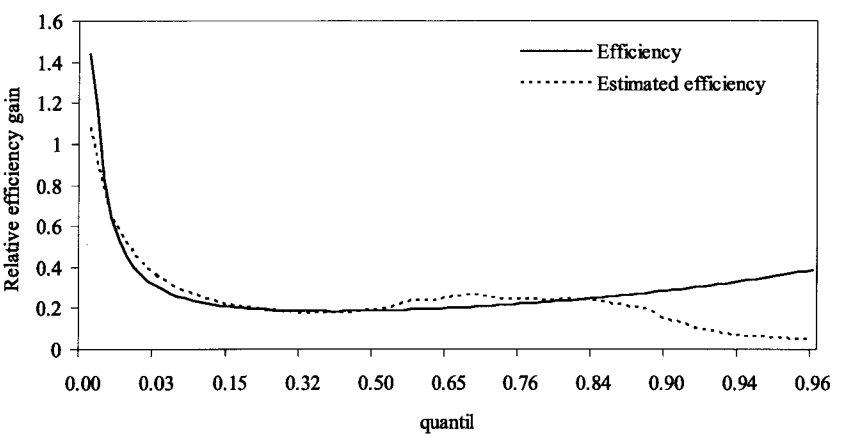


Figure 3.3. Adjustment for the effect of estimating b in the efficiency gain.

In the central quantiles, efficiencies resulting from the plug-in procedure are very similar to the theoretical ones. More advanced methods can be proposed to get better estimations for lower and higher quantiles. These methods can be based on using more sophisticated versions of the naive local constant estimator, where non-uniform weight functions could be considered. Nevertheless, we advocate for a detailed knowledge of the naive local constant estimator (the simplest possible estimator of the cumulative hazard that employs information to the right of the point of interest) before more advanced estimators would be investigated.

3.6 Implementation

As the optimal bandwidth b at t depends on $\alpha(t)$ and $\alpha'(t)^2$ both of them should be estimated first. We start with a local linear estimator of $\alpha(t)$ and $\alpha'(t)$. Nielsen & Tanggaard (2001) introduced local linear hazard estimation by transferring the well known principles from nonparametric regression and kernel density estimation (Wand & Jones, 1995; Fan & Gijbels, 1996 and Jones, 1993).

While the classical kernel hazard estimator of Ramlau-Hansen (1983) can be interpreted as a local linear estimator, we prefer the local linear estimator corresponding to a natural weighting defined as $\widehat{\alpha}_2(t) = \widehat{\Theta}_0$, according to

$$\begin{pmatrix} \widehat{\Theta}_0 \\ \widehat{\Theta}_1 \end{pmatrix} = \arg_{\Theta_0,\Theta_1} \min \sum_{i=1}^{n} \int_0^\infty \left\{\Delta N_i(s) - \Theta_0 - \Theta_1 (t-s)\right\}^2 K_b(t-s) Y_i(s) ds,$$

where $K(\cdot)$ is a probability density function with support $[-1,1]$ which is symmetric around zero and $K_b(\cdot) = b^{-1}K(\cdot/b)$ for a bandwidth b. This estimator has better robustness properties, and it is the direct analogy to the local linear estimator known from nonparametric regression.

Let

$$a_j(t) = \int_0^\infty K_b(t-s)(t-s)^j Y(s) ds, \quad \text{for } j = 0, 1 \text{ and } 2$$

and

$$G_j(t) = \sum_{i=1}^{n} \int_0^\infty K_b(t-s)(t-s)^j dN_i(s), \text{ for } j = 0 \text{ and } 1.$$

Then, to find $\widehat{\Theta}_0$ and $\widehat{\Theta}_1$ the following equations have to be solved

$$G_0(t) = \Theta_0 a_0(t) + \Theta_1 a_1(t) \tag{3.5}$$

$$G_1(t) = \Theta_0 a_1(t) + \Theta_1 a_2(t). \tag{3.6}$$

And this results in the local linear estimator $\widehat{\Theta}_0 = \widehat{\alpha}_2(t)$, where

$$\widehat{\alpha}_2(t) = \sum_{i=1}^{n} \int_0^\infty \overline{K}_{t,b}(t-s) dN_i(s),$$

and

$$\overline{K}_{t,b}(t-s) = \frac{a_2(t)K_b(t-s) - a_1(t)K_b(t-s)(t-s)}{a_0(t)a_2(t) - \{a_1(t)\}^2}.$$

An estimation of the first derivative $\alpha'(t)$ is provided by $-\widehat{\Theta}_1$, which can be directly obtained from (3.5) or (3.6) once Θ_0 has been estimated. Let us call $\widehat{g}_1(t)$ the estimator that we get for $\alpha'(t)^2$ based on $-\widehat{\Theta}_1$. For more details on local linear kernel hazard estimation see Nielsen & Tanggaard (2001).

Notice, that we do not wish our estimator of $\alpha'(t)^2$ to become too close to zero. Then we almost divide by zero while calculating our optimal b, that might become very big. The worst thing that can happen in our estimation process is that the bandwidth b becomes so big that the naive local constant estimator does not perform better than the classical Nelson-Aalen estimator. If the estimation procedure results in a bandwidth smaller than the optimal, we only lose some of the efficiency gain, but not all of it. The consequence of these considerations is that we define a robustified estimator $\widehat{g}_2(t)$ of $\alpha'(t)^2$ by smoothing one more time. Thus, in practice we will use

$$\widehat{g}_2(t) = \left\{ \int_0^\infty K(t-s) ds \right\}^{-1} \int_0^\infty K(t-s) \widehat{g}_1(s) ds.$$

3.7 An application to survival data

In the period from 1962 to 1977, 79 male and 126 female patients with malignant melanoma, cancer of the skin, had radical operations performed at the Department of Plastic Surgery, University Hospital of Odense, Denmark. The tumor was completely

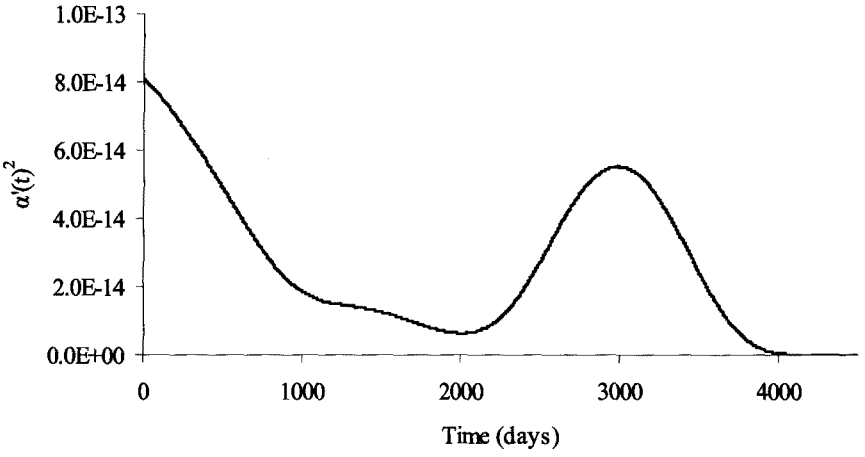
removed together with the skin within a distance of about 2.5 cm around it. All patients were followed until the end of 1977 and it was noted if and when any of the patients died, as well as the cause of death. Of the 79 male patients, 29 were observed to die from the disease, and of the 126 female patients, 28 were observed to die from the disease, while 14 died from other causes. The rest of them were alive at the end of 1977. The objective of this historically prospective clinical study was to asses the effect of risk factors on survival. The most important time variable is time since operation. Other factors were screened such as gender, age at operation and several variables related to the characteristics of the tumor.

Andersen, Borgan, Gill & Keiding (1993, example IV.1.2) present Nelson-Aalen estimates for these male and female patients where the survival time is measured since the time of operation. We will now compare their results with those corresponding to the naive local constant estimator.

As we mentioned before, in order to get the optimal bandwidth b both $\alpha(t)$ and $\alpha'(t)^2$ should be estimated first. According to the methodology described in section 5 these estimations can be obtained by using the local linear estimator. In that application a suitable probability function $K_b(\cdot)$ is the biweight kernel $K_b(\cdot) = \frac{15}{16}\{1-(\cdot/b)^2\}^2$ where $b = 800$ for both male and female. The same biweight kernel with the same b has been used to smooth $\alpha'(t)^2$ one more time (see Appendix B for details about the calculations). Thus, a more robustified estimator for $\alpha'(t)^2$ have been obtained, see Figure 3.4.

Once estimations of $\alpha(t)$ and $\alpha'(t)^2$ have been obtained the optimal b can easily be calculated according to (3.3). In Figure 3.5 optimal b as a function of t is shown both for male and female. Note that for small $t's$, in the boundary case, (3.3) provides bigger $b's$ than the corresponding t, so the optimal solution in that case is $b = t$, as we showed in section 3.2. At a certain point b equals zero, and $\widehat{\Theta}_0$ and $\widehat{\alpha}_2(t)$ too.

(a) Male



(b) Female.

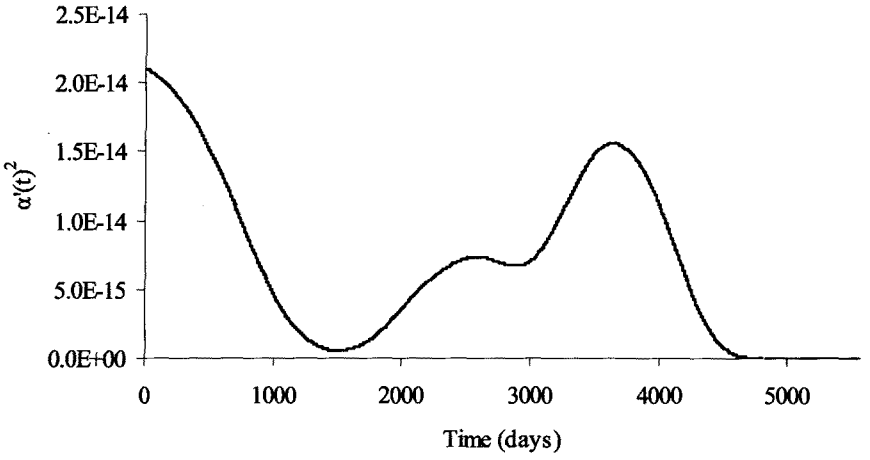


Figure 3.4. Estimation of $\alpha'(t)^2$.

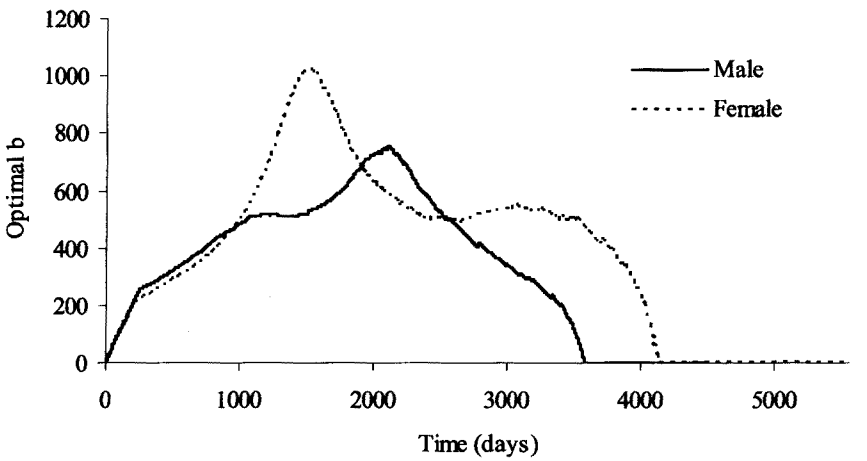
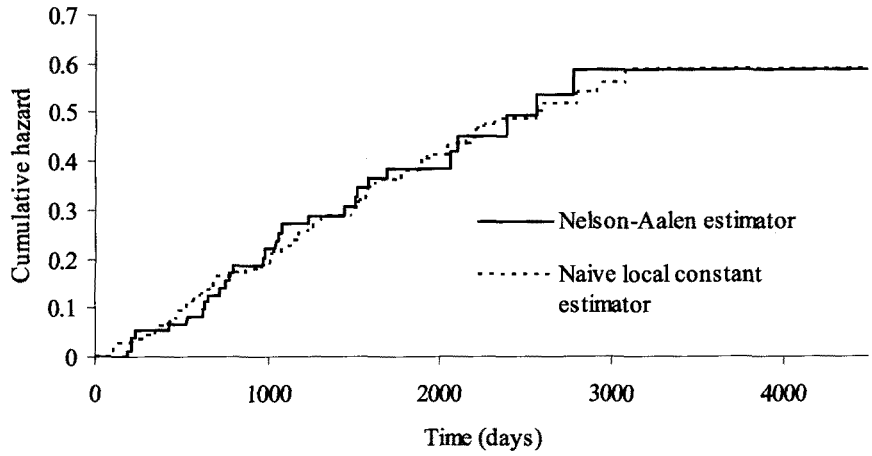


Figure 3.5. Optimal b used in the naive local constant estimator for male and female.

In Figure 3.6 we show the Nelson-Aalen and the naive local constant estimates of the cumulative hazard for both male and female. The most relevant feature of the curves being compared is that the naive local constant estimator provides a smoother curve than the Nelson-Aalen estimator. For any given t, the estimation of the cumulative hazard provided by the naive local constant estimator is taking into account all occurrences that took place in some period $[t-b, t+b]$ around t. For example, note that for both male and female the most important increase in the number of deaths occurs approximately around the end of the second year after operation or the beginning of the third year, approximately when $t = 621$ days for male and $t = 817$ days for female. This fact is reflected in the corresponding estimates of the naive local constant estimator prior to these time points, providing larger estimates than the Nelson-Aalen estimator.

The ratio between the naive local constant and the Nelson-Aalen estimator, see Figure 3.7, can be used for comparative purposes. Note that the ratio is quite large for small $t's$ but it decreases for larger $t's$. This ratio becomes very close to one after day 2000. After day 3000, approximately, the estimates seem to be equal, thus the ratio is 1.

(a) Males.



(b) Females.

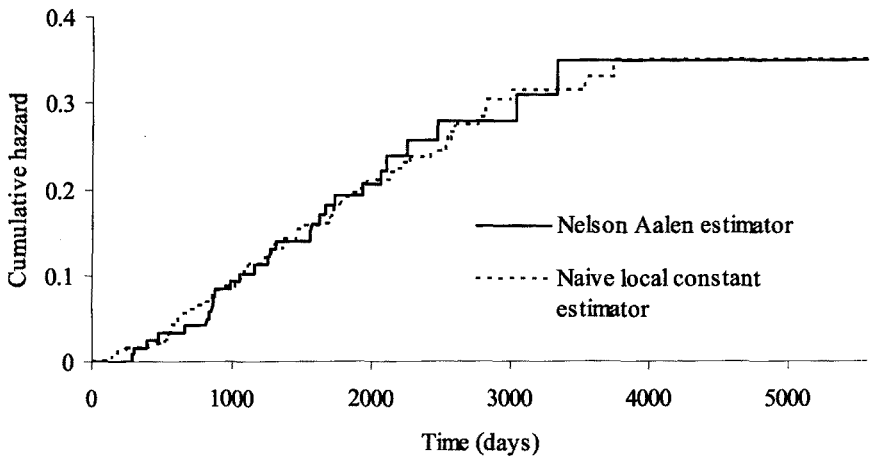
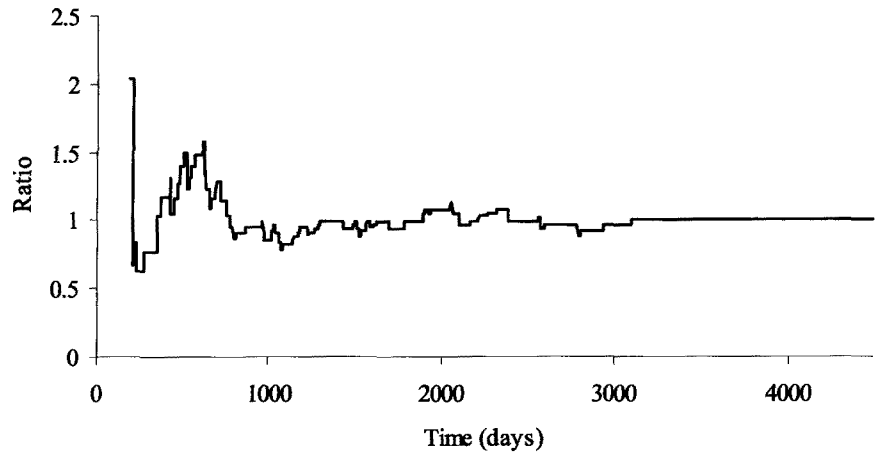


Figure 3.6. Comparison between the Nelson-Aalen and the naive local constant estimator for both male and female.

(a) Males.



(b) Females.

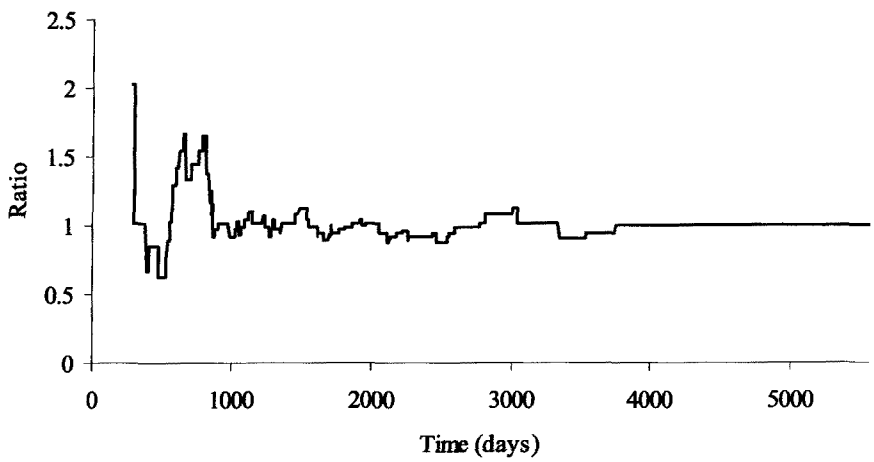


Figure 3.7. Comparison between the Nelson-Aalen and the naive local constant estimator for both male and female.

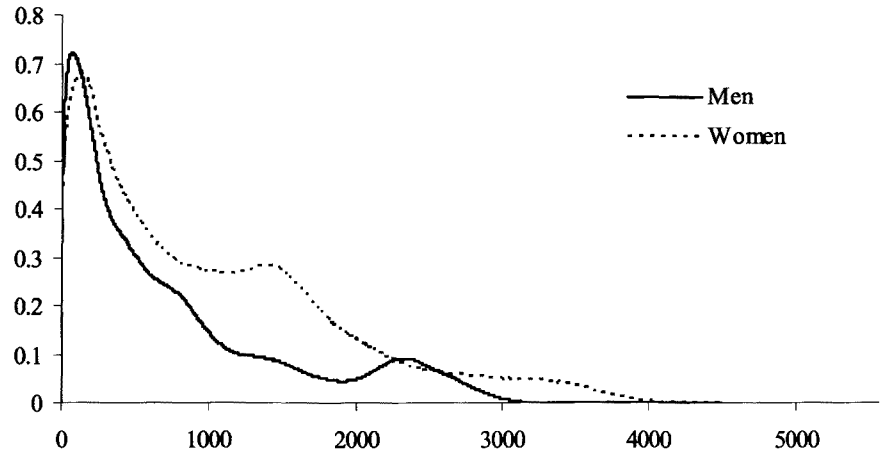


Figure 3.8. Relative efficiency gain curve of the naive local constant with respect to the Nelson-Aalen estimator.

For males, when $t < 779$ days, the naive local constant estimator provides larger estimates than the Nelson-Aalen estimator, except during a short period between day 210 and day 351. The reason is that it includes some information about what is going to happen after the time point being considered, so the substantial increase in the number of deaths occurring between day 621 and day 793 is taken into account. During a second period between day 779 and day 1892, the naive local constant estimator provides smaller estimates than the Nelson-Aalen, because it takes into consideration that no sudden increases in the mortality occur in subsequent periods. After day 1892, both estimates are close together, but the cumulative hazard curves do not become equal until day 3091.

A similar pattern is observed for women when comparing both estimates. During the period for $t < 872$ the naive local constant estimator provides larger estimates than the Nelson-Aalen estimator, except between day 386 and day 555. Again the naive local constant estimator captures the great increase that is going to occur in the number of deaths between day 817 and day 872. From $t = 872$ to $t = 2062$, the difference between the naive local constant and the Nelson-Aalen estimates is not so large, but after day 2062 the naive local constant estimator provides smaller esti-

mates than Nelson-Aalen, except during a short period around day 3000. Estimates become equal at day 3745.

Finally, in Figure 3.8 we plot the relative efficiency gain curve (3.4) for male and female. For males, the maximum value is 0.55 at $t = 146$ days. For females, the maximum value is 0.77 at $t = 67$ days. Relative efficiency gains range from 40% to 55%, for $t < 373$ for males. For females, relative efficiency gains for $t < 443$ go from 40% up to 70%.

Chapter 4

Customer lifetime duration

In this chapter[1] we introduce customer lifetime duration analysis in insurace companies. In the first section, the models traditionally used in this field of marketing are presented. Secondly, we describe the empirical and conceptual framework of customer lifetime duration analysis with a specific remark concerning the most important empirical studies in the field of customer loyalty in insurance companies. Finally, we present the empirical application to a real household dataset. The hypothesis and methodology applied in this empirical study are detailed in the last section of this chapter.

4.1 Models for customer lifetime duration

It was in the 50's when firms started to be interested in the reasons why customers are choosing a particular product or brand. The behavioural concept of loyalty was introduced by Brown (1952). According to his definition, customer loyalty is a tendency to buy one brand and it is directly related to the frequency of purchase.

Nevertheless, many authours were not satisfied with a pure behavoural concept of loyalty and they included a positive attitude towards the brand in the definition of loyalty (Day, 1969, and Jacoby & Chestnut, 1978). This was a second step towards

[1]Most parts of this chapter are also part of the paper: Brockett, P.L., Golden, L.L., Guillen, M., Nielsen, J.P., Parner, J. & Perez-Marin, A.M. (2005), "Household multiple policy retention effects of first policy cancellation: how much time do you have to stop total customer defection?," submitted for publication.

the modern idea of loyalty.

Nowadays, the idea that customer loyalty has both a behavioural and attitudinal component is widely accepted. Moreover, in recent years new factors such as sensitivity (Kapferer & Laurent, 1983), emotions towards the brand (Fourier & Yao, 1997) and stochastic elements (Uncles & Laurent, 1997) have been considered.

It is also well accepted that people grow into loyal customers by following a step-by-step progression[2] (Griffin, 2004). It has been proved that high levels of loyalty result in an increase in the customer average value (Riechheld, 1996). Mittal & Kamakura (2001) found out that keeping current customers is cheaper than recruiting new ones. Other important contributions to the analysis of customer loyalty are provided by Levitt (1988), Fornell (1992) and Bon & Tissier-Desbordes (2000) among many others.

Reinartz & Kumar (2003) give a brief review of the major findings of studies concerned with customer lifetime duration modelling. Firstly, the authors stress the limitations of several empirical studies (Allenby, Leone & Jen, 1999; Bolton, 1998; Dwyer, 1997 and Schmittlein & Peterson, 1994) due to the general lack of customer purchase history data. Nevertheless, during last years there is an increasing availability of longitudinal customer databases and researchers have started to take a longitudinal perspective in their work. Therefore, nowadays studies are mainly focused on the empirical measurement and modelization of the customer's relationship with the firm (Reinartz & Kumar, 2000).

Regarding the methodology, in some of these studies survival analysis techniques are used, namely the proportional regression model (Li, 1995 and Bolton, 1998). Helsen & Schmittlein (1993) supported the superiority of these methods when handling duration type data. Other methodologies are also applied, such as, for example, the Tobit regression model (Thomas, 2001) and Bayes models of customer interpurchase time (Allenby, Leone & Jen, 1999).

[2]Murray (1988) was the first to introduce a scale, and he proposed five levels of loyalty: prospects, shoppers, customers, clients and advocates.

Regarding the data sets used in these empirical studies, they are concerned about financial brokerage services, cellular or long-distance telephone service among many others. These results have provided several key results. The model proposed by Li (1995) identified variables (usage, marketing, demographics,...) that affect the lenght of customer subscription and made it possible to build profile of customers with high and low lifetimes in the long-distance telephone service.

Bolton (1998) found out that customer satisfaction is related positively to subscription duration in cellular phone service, but prior cumulative satisfaction is weighted more heavily than recent satisfaction in the decision on whether to continue or not. The Bayes model proposed by Allenby, Leone & Jen (1999) allows managers to recognize when a customer is changing his or her purchase patterns in financial brokerage services.

Very few applications to the insurance market can be mentioned. Reinartz & Kumar (2003) specially remark the contritution of Crosby & Stephens (1987) to the modelization of satisfaction with the service provider in life insurance. Their results suggest that nonlapsing customers report higher satisfaction than lapsed customers, but insureds were followed for 13 months only. The contribution of the rest of empirical studies about the insurance market will be quoted in the following sections, but in general they should be classified as studies related to purchasing behaviour.

4.2 Empirical and conceptual framework

It has long been recognized that insurance operates in a marketplace. Yet, the focus of the vast majority of research on this marketplace has traditionally been supply-side, emphasizing study of financial and actuarial elements. Very little has been published concerning the dynamics behind customer demand for insurance products. Demand side influences have been addressed, but to a lesser extent than have supply side considerations. For example, habit formation and the demand for insurance has been studied (e.g. Ben-Arab, Brys & Schlesinger, 1996), consumer

perceptions of service quality (Wells & Stafford, 1995 and Stafford, Stafford & Wells, 1998), individual portfolio decisions and demand (Mayers & Smith, 1983), household characteristics (Showers & Shotick, 1994), and demand in the presence of other risks (Doherty & Schlesinger, 1983; Schlesinger & Doherty, 1985; Gollier & Scharmure, 1994).

This chapter expands perspectives on the marketplace to the study of the behaviours of customers, who are the ultimate reason for the existence of insurance in the first place. Only by understanding the whole of the marketplace, both supply-side factors and demand-side factors, can insurance firms more optimally manage their operations in the marketplace.

The importance of demand-side analysis and money for product exchange

Without the ultimate sale of the insurance product, in the form of an exchange between a buyer and the insurer seller, there is no need for actuarial estimation or financial analysis, as there is no customer to insure. The firm must generate sales to survive, and just as actuarial estimations and financial analyses are critical for an insurer's long-term business survival, so too are sales and customer relationship management.

The firm's sales are not only a function of how many new customers are attracted, but are also a function of how many existing customers are retained. By retaining existing customers and attracting new customers profitably, the insurer can grow the business and potentially increase market share. Managing customer growth and market share requires an incorporation of consideration of demand side marketplace dynamics to better understand customer behaviours and responses, in order that customer relationships may be developed and matured. Insurance has long focused on compensation and marketing techniques, including a comparison of distribution systems, to study methods for providing sales incentives and better customer relationship management (although not necessarily discussed in those terms). Examples of this type of research are: Barrese, Doerpinghaus & Nelson (1995) and Gravelle (1994).

Persistency studies and the importance of the first lapse signal

Persistency studies have been conducted in life insurance to determine factors causing policy lapses (Kuo, Tsai & Chen, 2003). These studies often provide demand-side information, but are generally conducted for a single policy. Persistency studies, however, have not tended to be conducted for property and casualty insurance, even for single policies.

In spite of the lack of attention given to persistency in property and casualty, policy lapses are likely to be similarly important for the property and casualty area. They are a customer behaviour that can signal purchasing decisions in-process or on the customer's decision-making horizon. And, often a household will buy multiple policies from the same insurer, such that a lapse in one may be only the beginning of the customer's defection to a competitor.

Lapses signal the customer's brand switching behaviour: moving from one insurer/company to another indicates the customer's defection to the competitor in a demand-side market analysis. Customer retention is the opposite of customer defection-one firm's gain is another firm's loss. And, customer retention becomes increasingly important in a multiple product situation, where the company sells multiple policies to the same person or household. Losing one policy is likely to be the first step in a customer decision-making process resulting in lapses in all policies sold to a household. And, whether comprised of one individual or several, the household is an appropriate unit of analysis, because insurance purchases are often made as a bundle of products serving multiple risk management needs of the household operating as a decision-making unit.

Losing and gaining customers through brand switching is a major concern for firms in the insurance industry. This concern is well-founded, as customer marketplace purchasing is dynamic. For example, Schlesinger & Schulenberg (1993) found that 30.1% of customers interviewed had switched automobile insurance carriers at some point in time.

Insurance companies focus on both retention of existing customers and attraction of new customers (Cooley, 2002). And, retention is particularly important, because the most undesired consequence of losing a good customer is the possibility that they will be replaced by a not so good customer. The quality of the customer portfolio is essential to profitable survival in the insurance business, because pricing is based on an estimation of the quality and quantity of the risk being covered.

The Concept of Policy Lapse Versus Changing Customer Needs

In the insurance setting, when a policy is ended two basic situations are possible: (1) the risk is going to be covered by another insurance company (e.g., an automobile insurance policy is taken out with another insurance company), or (2) the risk does not exist any more for the policy holder (e.g., a car being sold). Any investigation into policy lapse must explicitly take the difference between these two situations into account empirically. The first type of policy termination, brand switching via the customer purchasing from another company, is the termination of interest to understanding demand-side market dynamics and customer relationships.

In keeping with the importance of distinguishing policy terminations from policy lapses, this research does not view all terminations as cancellations or lapses. The rule that we have applied to determine whether a termination is regarded as a cancellation or lapse, and thus immediately relevant as a signal of impending customer defection for other policies purchased from the same firm, is whether the risk still exists at the time the contract is ended.

It is important to note here differences that exist between the European processes of insurance contract renewal and those used in other countries, for example the United States of America. In Europe, the source of the longitudinal database used in this research, a policy will be automatically continued, via a bank account debit, unless the customer takes an explicit action to terminate it. Thus, a termination, because of a brand switch or lack of need, requires explicit action in Europe.

In the United States of America, the policy will automatically lapse if it is not explicitly renewed. The two processes are essentially reversed: Europe requires

explicit action to terminate (or a bank account debit will automatically occur and the policy will renew) and, conversely, the United States requires explicit action (i.e., payment) for renewal or the policy will automatically terminate after thirty days with non-payment. Europe is automatic renewal and the United States is automatic termination.

While this international difference is important to note, it makes the European database particularly relevant for study, even to gain insights into brand switch in any country. Customers in any country might notify their insurer of an intent to switch firms prior to non-payment, but customers in Europe (either the new firm or the customer) must notify the current insurer in order to switch brands. Thus, European insurers automatically have information of an impending brand switch without having to wait to realize a lost sale/customer. And, while the brand switch intention signal may come to the company in a non-European country at a later point in time (i.e., after thirty days of non-payment), the methodology and customer relationship management implications of this research are still relevant to guide further customer retention investigations in any country.

Customer Lapse Example: Policy Coverage Data versus Customer Notification Data

Here we present an example of customer lapse behaviour to illustrate the nature and scope of the investigation. Our focus will be on three types of insurance contracts: contents, house, and automobile. Contents insurance covers the items inside the house subject to loss, such as furniture, silver and gold, paintings, clothes, and audiovisual equipment. House insurance covers the building itself (roughly speaking, the bricks) from damage generally caused by phenomena such as fire, storm, or water. Automobile insurance covers bodily injury and property liability.

All notifications of cancellation in the house, contents, and automobile lines of business occur close to the renewal date. As mentioned previously, before a policy is cancelled there has to be a notification from the customer. If a customer does not want to renew a particular policy on the day it is due, the notification has to be made a minimum of one month before the renewal date. If the notification is made

during the last month before the renewal date, the policy will not be cancelled in the next renewal but will be cancelled in the following one, thirteen months after notification, i.e. the year after. Therefore, there is a period of time that can take up to 13 months from when the cancellation is notified until it is actually made effective via policy termination.

Thus, if several cancellation notices are made by the customer, the first to be actually cancelled is not necessarily the first policy the customer notified the insurer about canceling. It depends on the corresponding renewal dates. Figure 4.1 considers the number of policies that a particular customer has of a certain type (contents, house, and automobile) at each moment in time.

The customer shown in Figure 4.1 first purchases a house policy, and shortly afterwards an automobile policy, followed by a contents policy purchase. In this example, the first policy the customer cancels is the automobile policy.

Compare Figure 4.2 which represents the same information and adds a dashed line for the period of time when the risk is covered by the corresponding policy, even though the insurer has been notified of the cancellation. Note that the first notification corresponds to the house policy, while the first risk to be out of coverage corresponds to the automobile policy, because its renewal date comes first.

From the perspective of understanding the customer's intention and predicting the length of time the customer is likely to remain with an insurer after cancellation of a policy, the type of first notice of cancellation contains relevant information. Notification data properly reflects the customer point of view with respect to the insurance relationship, while policy lapse date information only describes the risk being covered. Therefore, the customer's intent is properly established by the type and moment in time of the first notification of cancellation (not the first policy to actually lapse).

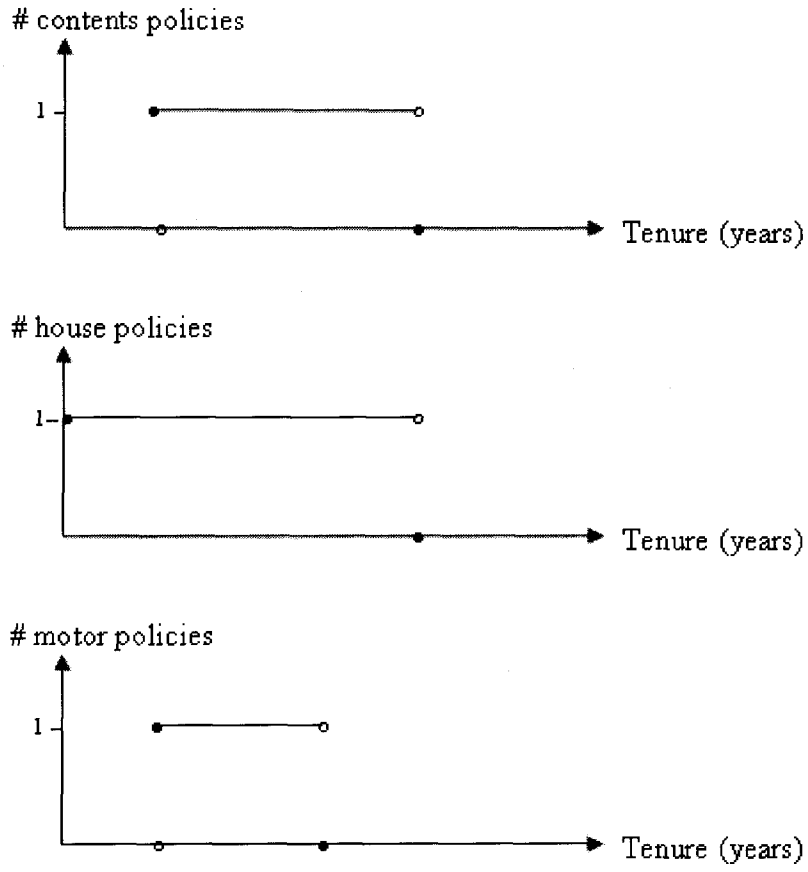


Figure 4.1. Number of contents, house and motor policies a customer has at a particular moment in time (policy coverage data).

When the first risk out of coverage is taken to define the first lapse then there is a problem of misclassification that clearly affects both the type and the time of the first lapse.

This distinction between lapse and notification is important, as the notification data define the demand side characteristics which are the source of the ultimate lapse. This issue only arises for multiple policies underwritten by the same insurer, as is considered here. The brand switch signal must be taken from the time of the first cancellation notification (and not the time of the first lapse in coverage).

This research measures lapse from the time of first policy cancellation notification to the insurer (not first policy to actually lapse in coverage).

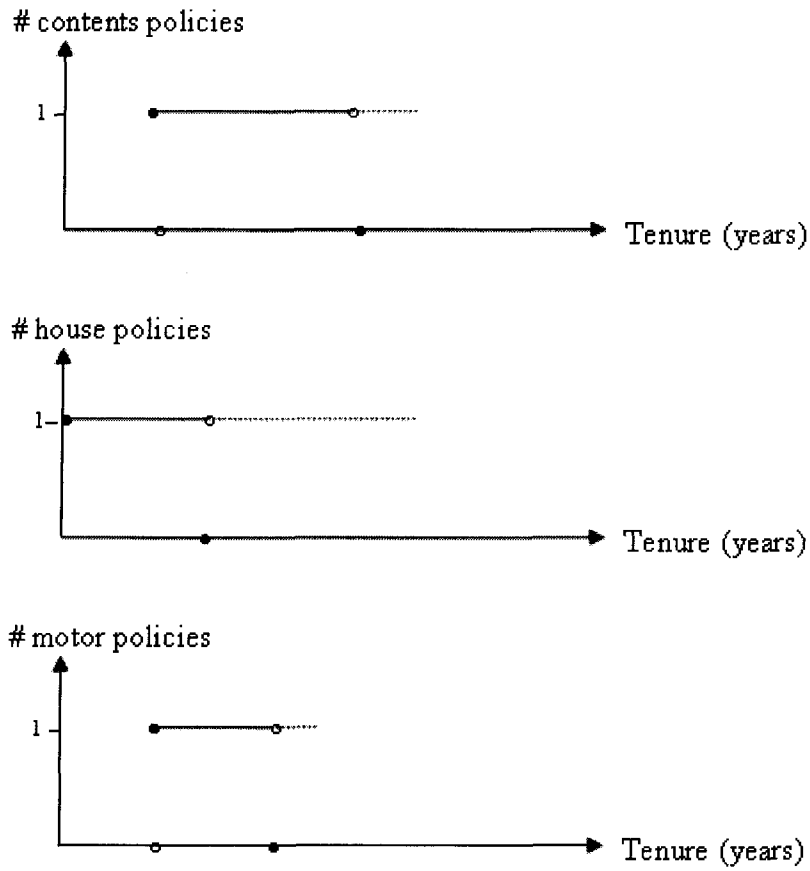


Figure 4.2. Number of contents, house and motor policies a customer has at a particular moment in time. Dashed lines represent the period when the risk is covered even though the cancellation has been notified.

Thus, we capture the brand switch signal so as to be able to analyze the response time available to the insurer before the customer is lost, as well as the probability that the customer household will subsequently cancel additional policies.

4.3 The household data set

The dataset used in this research consists of 151290 households possessing multiple insurance policies, who sent notification of cancellation of their first policy to a particular major Danish insurer between January 1, 1997 and June 1, 2001. The information was collected according to the time frame shown in Figure 4.3.

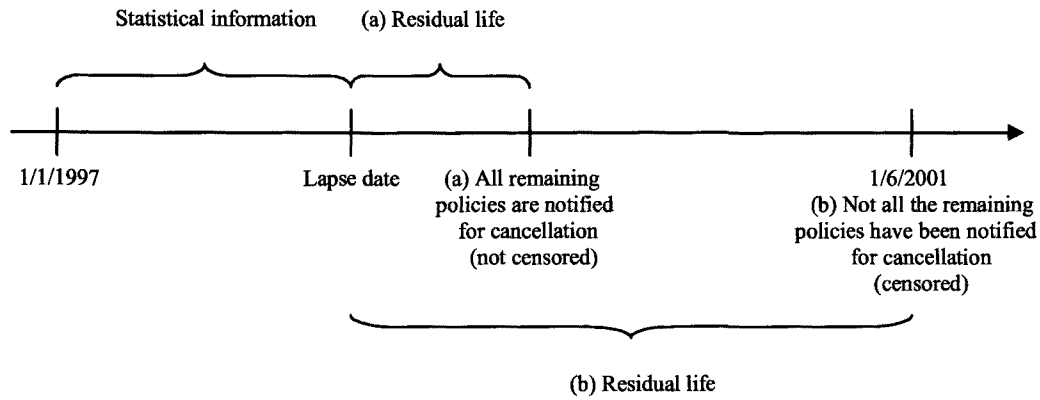


Figure 4.3. Time frame.

Some of the household covariates refer to the occurrence of an event (a claim, a premium increase, or a change of address) from January 1, 1997 until the date of the first lapse while other covariates are measured at the time of first policy cancellation (for example, the tenure or the age of the policy holder). Once the first policy cancellation occurs, the residual household customer lifetime is measured by the number of days until all remaining policies are notified for cancellation or until the end of the study, June 1, 2001, whichever comes first (some policyholders will cancel one policy but keep others).

In situation (a) in Figure 4.3, all the remaining policies are cancelled before June 1, 2001, so the household customer residual life is the time from the first lapse date until total cancellation of all other policies occurs. In situation Figure 4.3 (b), at the end of the study, we only know that the residual life is greater than the time from the first lapse until June 1, 2001. In this case, the residual life is the listed as the time elapsed from first policy cancellation until June 1, 2001, but note that the observation is right censored.

Table 4.1 lists the variables in the database and the label given to teach. Some of the variables that may not be immediately self-explanatory are explained further in the text.

Table 4.1. Variables in the Household Data Set

Age of named policyholder at the date of the first lapse (Age)
Gender of named policyholder (Gender)
Time from notification of first policy cancellation until the actual first cancellation (Notice)
Tenure of household with insurer (Tenure)
Core customer status =1 if two policies in addition to contents insurance (Corecust)
Change of address prior to cancellation (Change of Address, broken into six subcategories)
First Cancellation notice furnished by external company A (External Company A)
First Cancellation notice furnished by external company B (External Company B)
First Cancellation notice furnished by external company C (External Company C)
First Cancellation notice furnished by external company D (External Company D)
First Cancellation notice furnished by another known external company (Another Known External Company)
Claims history: Time since last claim (Claims, broken into six subcategories)
Contents insurance prior to cancellation (Contents0)
Contents after cancellation of first policy(Contents1)
House insurance prior to cancellation (House0
House after cancellation of first policy (House1)
Automobile insurance prior to cancellation (Motor0)
Automobile after cancellation of first policy (Motor1)
Indicator of household having underwritten the first contents policy within the 12 months previous to the date of the first lapse (Newcontents)
Indicator of if household has underwritten the first house policy within the 12 months previous to the date of the first lapse (Newhouse)
Indicator of if household has underwritten the first automobile policy within the 12 months previous to the date of the first lapse (Newmotor)
Premium increase (Pruning, broken into three subcategories)

Tenure is the number of years the household has been a customer of the company calculated as the number of years from the first policy issued to the policy holder, within the types of policies considered here, until the date of the first lapse. Notice is the time interval from notification of the first policy cancellation until the actual occurrence of the corresponding cancellation.

Since the types of policies held by the household could conceivably affect the retention attributes of the client with respect to the insurer, the following dummy

variables were developed: contents0, house0 and motor0. They indicate whether the household has contents, house, or automobile policies respectively before the first lapse. Contents1, house1 and motor1 indicate whether the household has contents, house, or automobile policies respectively after the first lapse. Newcontents, newhouse and newmotor indicate whether the household has underwritten the first contents, house, or automobile policies, respectively, within the 12 months prior to the date of the first lapse.

Corecust indicates whether the customer has a core customer status. A core customer is a customer that has a contents policy and at least two other types of policies (they could be automobile, house, or others like life insurance) with the insurer. In the insurance company that has been analyzed here, core customers have lower premiums, bonuses, and special advantages. From a marketing perspective core customers having multiple policies tend to be more profitable and, hence, deserve special consideration.

Information on whether a change of address has occurred was included, as it can affect the probability of house and contents cancellations. Six categories were developed for this variable: no change of address, change of address less than 2 months before the date of the first lapse, between 2 and 6 months before the date of the first lapse, between 6 and 12 months before the date of the first lapse, between 12 and 24 months before the date of the first lapse, and more than 2 years before the date of the first lapse.

Since premium increases might impact customer retention, information was included on whether the time period included a substantial increase in premium of 20 to 50%. Such premium increases are commonly termed pruning, since the insurer wants to persuade the customer to lapse, possibly due to a very bad claims history. Three categories were developed: no pruning, pruning within the past 12 months, and pruning more than one year before the date of the first lapse.

The data included information about recency of claims, as they can also affect the probability of lapse. The six categories for claims developed were: no claim, claim less than 2 months prior to the first lapse, between 2 and 6 months prior to

the first lapse, between 6 and 12 months prior to the first lapse, between 12 and 24 months prior to the first lapse, and more than 2 years prior to the first lapse.

Finally, considering the competitive nature of the marketplace, and the marketing dynamics of alternative brands in a brand switching model, we have also included information on whether there was any external company involved in the cancellation notice. The customer has a choice of notifying the current insurer him/herself of cancellation or of having the new insurer notify the current insurer. It is clear that when the new insurer does the notification, that a brand switch has already occurred and, at least for that policy, the customer is entrenched with the new insurer for at least the next year. It is likely, also, that the new insurer will wait until the last moment to signal their competitor of the upcoming brand switch, lest the competitor take measures to try to retain their customer. Further, the new insurer will likely be discussing other insurance policy needs with their newly acquired customer, so subsequent policy cancellations are likely.

We considered the four most important competitors, coded as A, B, C and D and developed six categories for this variable: no external company (notification by the customer himself), company A, company B, company C, company D, and finally another known external company. We considered a competitor to be involved if the notification was communicated by an insurance company on behalf of the customer.

Table 4.2 presents a description of the policy portfolio state before versus the state after the first lapse, thus comparing the types of policies the household had before and after the occurrence of the first lapse. This information is represented with a string of three characters of 0's and 1's where 1 (0) indicates that the household had (had not) one particular type of policy. The sequence order is contents - house - automobile. For example, if the state before the first lapse is 011 and the state after the first lapse is 010, then the household had house and automobile policies before the first lapse, but no automobile policy after the lapse.

As shown in Table 4.2 above, the number of households with only a contents policy, 34998, is slightly smaller than the number of households with the three types of policies, 37103.

Table 4.2. State before vs. after first lapse. Number of policies.

	State after[a]								
State before[a]	000	100	010	001	110	101	011	111	Total
000	0	0	0	0	0	0	0	0	0
100	34998	0	0	0	0	0	0	0	34998
010	10757	0	0	0	0	0	0	0	10757
001	28198	0	0	0	0	0	0	0	28198
110	2690	3060	4090	1	0	0	0	0	9841
101	3397	13764	1	10613	0	0	0	0	27775
011	166	1	1409	1042	0	0	0	0	2618
111	4389	471	2488	4535	12488	5957	6775	0	37103
Total	84595	17296	7988	16191	12488	5957	6775	0	151290

[a]States represented by a string of three characters of 0's and 1's where 1 (0) indicates that the household had (had not) one particular type of policy. Sequence order: contents - house - motor.

The most frequent state before the first lapse is 111, where the customer has contents, house, and automobile policies, while the most frequent state after the first lapse is 000, since many households have just one policy.

Focusing now on households with more than one type of policy (Table 4.2), initially we observe that for those with just contents and house policies (110), the most frequent state after the first lapse is 010, so the one being canceled is the contents policy. However, if the household initially has contents and automobile policies (101), or house, and automobile policies (011), or all three of contents, house, and automobile policies (111), then the automobile policy is the most likely to be cancelled.

Some useful additional observations can be made by examining the results for simple descriptive statistics. Amongst the 151290 households who notified their first cancellation during the analyzed period, 20740 had not cancelled all their remaining policies by June 1, 2001 (the end of the period), so the frequency of censored observations is 13.71%. The average age of the customers is 45.92 years (with standard deviation 17.22), and the average tenure is 9.03 years with the company (standard deviation of 10.19).

The whole data set consisted of those households who notified their first cancellation during the period between January 1, 1997 and June 1, 2001. This research focuses on that part of the dataset with more than one policy before the first lapse occurs, which totals 77337 households. As mentioned previously, some of these customers cancel all their policies at the same time, so the insurer does not have time to react once the first lapse (and total cancellation) has occurred, but some customers cancel sequentially, leaving the insurer time to avoid completely losing them. Our analysis focuses on the subset of multiple policy holders who cancel sequentially.

Table 4.3 shows the expectation of the residual life in days depending on the state before and after the first lapse, for those households with more than one policy at the beginning of the period that do not cancel all the policies simultaneously, i.e. the state after the first lapse is not 000.

Estimated residual lifetimes have been obtained using the Nelson-Aalen estimator, devised by Nelson (1969, 1972) and Aalen (1978). Transitions have been classified into three subsets: from two initial policies to one policy, from three initial policies to one policy and from three initial policies to two policies.

Table 4.3. Average residual life

Transition	State before[a]	State after[a]	n	Average (days)
From 2 to 1 policy	110	100	3060	644.809
	110	010	4090	306.711
	101	100	13764	600.681
	101	001	10613	357.765
	011	010	1409	513.986
	011	001	1042	516.909
From 3 to 1 policy	111	100	471	74.006
	111	010	2488	120.675
	111	001	4535	157.319
From 3 to 2 policies	111	110	12488	589.648
	111	101	5957	702.472
	111	011	6775	379.851

Those households having three policies at the beginning of the period who retain two policies after the first lapse has the largest overall average residual life, about 558 days. Further, those customers that started with three policies who end up with one policy after the first lapse have a smaller overall average residual life (140 days) than those who had two policies that end up with one policy (481 days). The estimated lifetime difference (341 days) is substantial and may be indicative of customer dissatisfaction motivating policy change. The number of days the insurer has to respond is important in itself.

Table 4.4 presents the estimated expected residual life for different types of households, depending on whether they have had any claim, change of address, or a substantial rise in the premium. Expectations have been obtained using the Nelson-Aalen estimator.

Table 4.4. Average residual life.

			Average
Factor	Status	n	(days)
Claims	None	32500	522.541
	At least one	34195	412.753
Change of address	None	45284	463.476
	At least one	21411	482.802
Pruning	None	64761	477.242
	At least one	1934	259.967

Those households that had at least one claim have a smaller average residual life (413 days) than those who did not have any claim (523 days). Schlesinger & Schulenburg (1993) found that in the German automobile market 14.3% of the switchers who filed a claim with their previous insurer had received an indemnity of less than 75% of the total insured damages, whereas only 5.4% of non-switchers who filed a claim with their current insurer received less than 75% of damages. They also found that for switchers, 52.5% of claims filed with previous insurers took three weeks or longer to get paid, while only 29.6% of those customers who filed a claim with current insurer had to wait that long. This may be in part why claims has the

impact it does on the expected time the client remains with the insurer (satisfaction level with the claims payments of the insurer).

According to the results in Table 4.4, households that suffered a substantial rise in their premium have an average residual life (260 days), smaller than those corresponding with no price increase (477 days). This result was expected, as households who experienced the premium increase probably tried to find a cheaper product in another company. A lower premium has been shown to be the most important reason for choosing a particular insurer for the German automobile insurance market (Schlesinger & Schulenburg, 1993).

The expected time until final cancellation for those who have had a change of address (483 days) is slightly longer than the corresponding lifetime for those who did not move (463 days). This result suggests that although a contents or a house policy cancellation is more likely to happen when a family moves, this does not seem to affect household behaviour regarding remaining policies.

4.4 The hypothesis and the methods

Our modelling process includes two stages (see Figure 4.4). Firstly, we consider those households with more than one policy in the insurance company. Some of them would cancell all their policies simultaneously and some of them would make a partial cancellation (they cancel some of their policies but not all of them). For those households making a total cancellation the insurer has no time to react after these first cancellation, the remaining lifetime is zero. For those households who make a partial cancellation the insurer can estimate the remaining lifetime, i. e. the time between the first cancellation and the moment when all the remaining policies would be cancelled.

Therefore, the modelling process includes a first step where the probability of a total cancellation is estimated for those households with more than one policy in the insurance company. A logistic regression model will be used to estimate this probability as a function of some explanatory covariates.

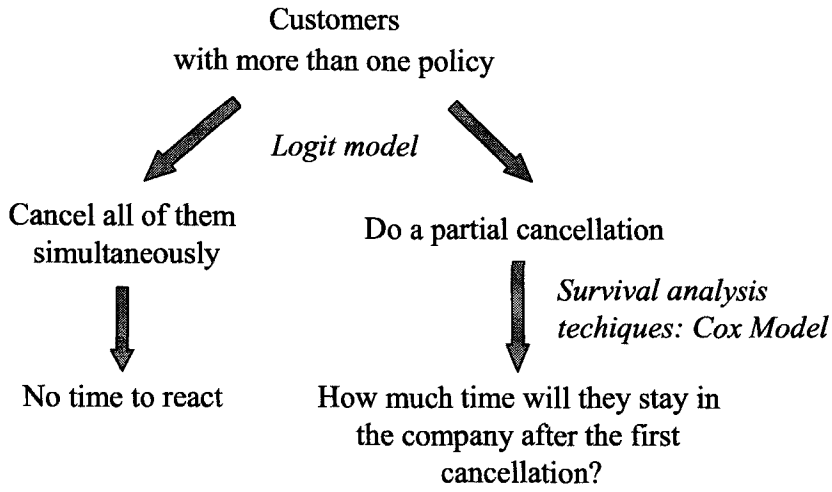


Figure 4.4. Modelling strategy.

In the second stage we focus on those households who made a partial cancellation. The risk that all the remaining policies would be cancelled (therefore, the insurer loose the customer) and the customer lifetime is estimated as a function of some covariates by using survival analysis techniques. The methodological contribution takes place in this second stage. The naive local constant estimator formulation is adapted in order to be used for the estimation of the non parametric part of the proportional hazards regression model.

Therefore, with this methodology the insurer is able to identify those households with a high risk of a total cancellation. At the same time, the logistic regression model let us know the effect of each covariate on that risk. Additionally, in the second stage an estimation of the risk of cancelling all the remaining policies for those with a partial cancellation is provided together with the effect of each covariate on that risk. Estimations of the remaining customer lifetime duration can be easily obtained by using this method.

Nevertheless, the scope of this research is the customer lifecycle being studied. The first stage of our research is focused on those customers who make a first cancellation, and the second stage is focused on those households who do not cancel all their policies simultaneously in their first cancellation. Therefore, conclusion about only these two groups of customers can be drawn.

Chapter 5

The risk of non-renewal

In this chapter the prediction of the risk of non-renewal is carried out by applying a logistic regression model. This method is briefly introduced in the first section. Secondly, the estimation results obtained for our household dataset are presented and the effect of each covariate in the risk of a total cancellation is discussed. Finally, we identify different groups of customers with different probabilities of a total cancellation and we estimate them.

5.1 Logistic regression for choice prediction

As mentioned before, in the first stage of this research we use logistic regression to determine the probability, based upon known covariates of the insured, that a household originally having more than one policy will cancel all the policies simultaneously. The same methodology was used by Guillen, Parner, Densgsoe & Perez-Marin (2003) in order to predict the probability of a policy cancellation in a three-month period.

For household i, $i = 1, ..., n$ we assume that

$$\Pr(R_i = 1|x_i) = \frac{\exp(\beta' x_i)}{1 + \exp(\beta' x_i)} \tag{5.1}$$

where $R_i = 0$ for a partial cancellation and $R_i = 1$ for a total cancellation, x_i is a vector of the observed explanatory variables, β is a vector of unknown parameters. Consistent and asymptotically efficient estimates of the parameters in the

logistic regression model (5.1) are obtainable using the conditional maximum likelihood method (Snell & Cox, 1989 and Agresti, 1990) implemented in many common statistical packages. In this manner we are able to look at the effect of household characteristics (covariates) on the likelihood of total simultaneous cancellation.

5.2 Estimation results

Predicting this probability is assessed using a logistic regression model where the covariates described in Table 4.1 are used, except for *contents1*, *house1* and *motor1* which are, of course, all zero after a total cancellation has occurred. The data set used in the estimation of the model consists of 74969 households (a few observations were eliminated due to missing values on some covariates). Among those 74969 observations, 10317 simultaneously effected a total cancellation of all policies with the insurer.

The overall statistical test of no covariate effect provided a likelihood ratio statistic of $LR = 9178.68$ with 28 degrees of freedom ($p < .001$). Household characteristics significantly affect the probability that a customer will totally simultaneously cancel all policies. Individual parameter estimates are shown in Table 5.1.

All of the parameters in the model are significant, except for having added a new house policy in the last 12 months (*newhouse*) and having had a premium increase more than one year prior to the household first giving a cancellation notice (*pruning more than one year past*). Thus, the potential customer repelling effects of premium increases seems to wear out after 12 months. Nevertheless, the overall test of significant effect for the risk factor pruning let us refuse the null hypothesis ($p-$ value 0.0251).

The individual parameter tests indicate that the change of address within the first 6 months or more than two years prior to the first lapse, the occurrence of a claim and the existence of a premium increase (*pruning within past* 12 *months*) are three factors that influence the probability of a total cancellation.

Table 5.1. Logistic regression model estimates.

Parameter	Estimate	Stand. Error	OR	p-value
Constant	-2.201	0.112	-	<0.001
Change of address, less 2 m. ago	-0.596	0.060	0.551	<0.001
Change of address, 2 - 6 m. ago	-0.122	0.052	0.885	0.019
Change of address, 6 - 12 m. ago	-0.095	0.048	0.909	0.049
Change of address, 12 - 24 m. ago	0.229	0.041	1.258	<0.001
Change of address, more 24 m. ago	0.531	0.044	1.701	<0.001
Tenure	-0.011	0.001	0.989	<0.001
Claims, less 2 months ago	0.230	0.040	1.259	<0.001
Claims, 2 and 6 months ago	0.324	0.035	1.383	<0.001
Claims, 6 and 12 months ago	0.440	0.035	1.553	<0.001
Claims, 12 and 24 months ago	0.469	0.037	1.598	<0.001
Claims, more 2 years ago	0.546	0.054	1.727	<0.001
Contents0	0.277	0.087	1.319	0.001
Corecust	0.109	0.025	1.116	<0.001
Age	0.004	0.001	1.004	<0.001
External company A	2.548	0.041	12.779	<0.001
External company B	2.165	0.046	8.718	<0.001
External company C	1.893	0.048	6.637	<0.001
External company D	2.270	0.047	8.834	<0.001
Another known external company	1.686	0.035	9.680	<0.001
Gender (male)	0.099	0.028	1.104	<0.001
House0	-0.657	0.030	0.518	<0.001
Motor0	-1.253	0.033	0.286	<0.001
Newcontents	-0.113	0.043	0.893	0.008
Newhouse	0.073	0.060	1.076	0.225
Newmotor	-0.208	0.050	0.813	<0.001
Notice	-0.002	<0.001	0.998	<0.001
Pruning within past 12 months	-0.187	0.072	0.829	0.009
Pruning more than one year ago	0.086	0.111	1.089	0.442

By looking at the odds ratios, we see that external companies, claims, change of address more than one year ago, and contents policy are the most relevant factors influencing the probability that a total cancellation occurs. Far and away the most important determinant of the probability of a total cancellation is, however, the intervention of an external company (competitive effects). Among external companies, the one coded as A is the one with the largest odds ratio, which identifies

company A as an aggressive competitor that frequently captures all household contracts simultaneously.

Both claims occurrence and change of address increase the probability of a total cancellation as the time since the corresponding event has occurred increases.

A surprising result was that core customer status increases the probability of a total cancellation. A possible explanation is that core customers may be more likely to be solicited by competitors, and they may receive very persuasive offers from competitors.

If this is so, the company should increase the efforts to retain them as these are the precise customers which the insurer would like to keep, and, unfortunately, for which the insurer has the least amount of notice (residual time) to recapture the defecting client. Their risk characteristics may make them desirable customers to all insurers and highly sought after.

We will now address the description of the model's ability to discriminate between partial and total first cancellations. In Figure 5.1 the absolute frequencies of the predicted probabilities for the observed total and partial cancellations are shown.

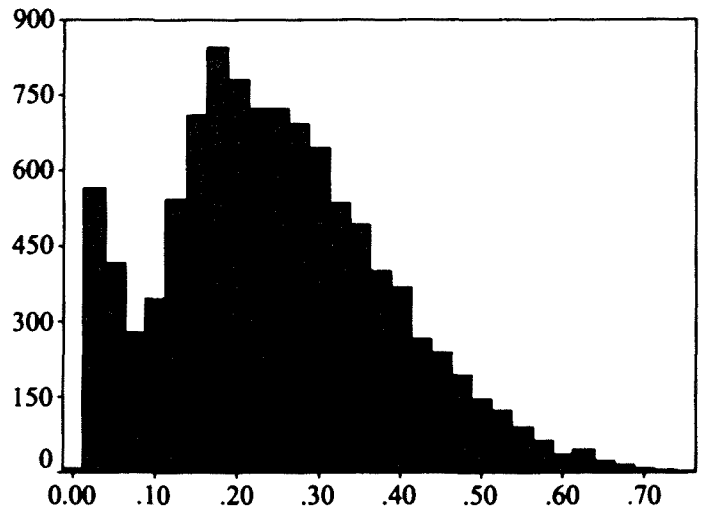
It is possible to compare what the model discriminates with the real observed results, i.e, whether or not customers actually do a total cancellation.

We will consider that customers with a predicted probability greater than a given threshold value p as customers for whom the model predicts that he/she will do a total cancellation.

For any given threshold probability p one can calculate *sensitivity*, *specificity*, *predictive positive value* (PV_{pos}) and *predictive negative value* (PV_{neg}). *Sensitivity* versus 1− *specificity* is represented in the ROC curve, dotted line in Figure 5.2 (a). The figure also illustrates the identity (solid line), meaning that the model has no discrimination ability.

The model data set consisting of 74969 customers is scored for selected threshold levels. The results are shown in Figure 5.2 (b) and Table 5.2. For a probability level of 16.5% *sensitivity* approximately equals *specificity*.

(a) Total cancellations.



(b) Partial cancellations.

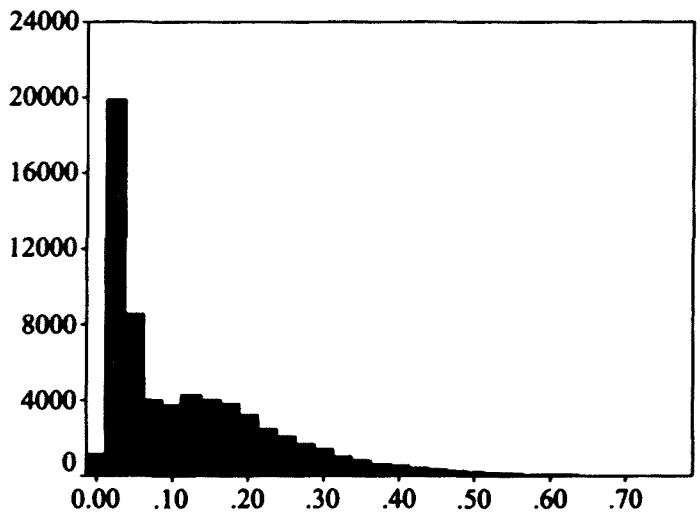
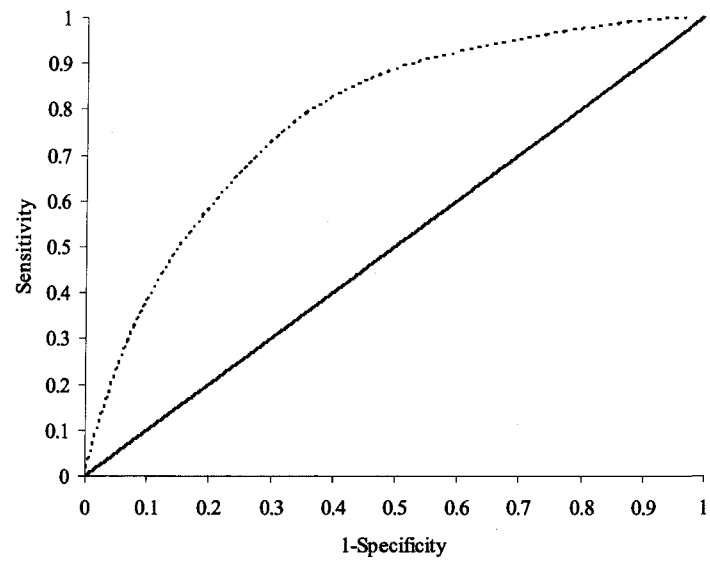


Figure 5.1. Predicted probabilities for observed total and partial cancellations.

(a) ROC curve



(b) Sensitivity, specificity, PV_{pos} and PV_{neg} (b).

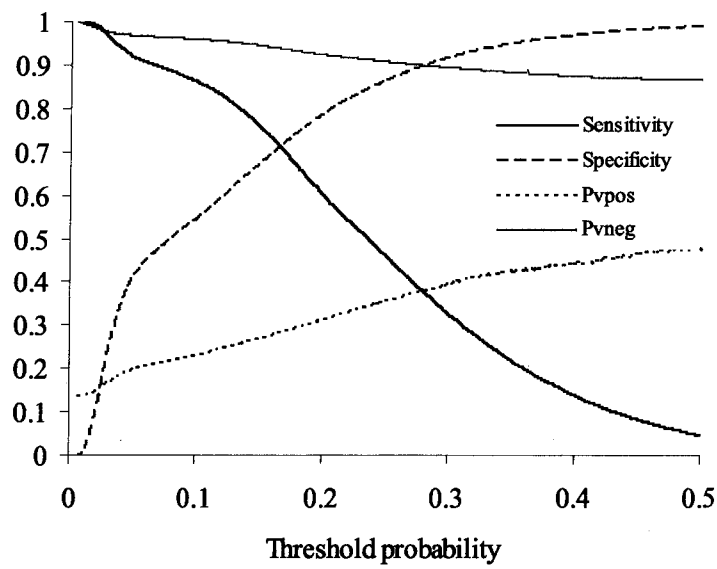


Figure 5.2. ROC curve and Sensitivity, specificity, PV_{pos} and PV_{neg}.

Table 5.1. Classification results for different probability thresholds.

Threshold prob	Pred. Total canc.	Obs. Total canc.	Sensitivity	Specificity	PV_{pos}	PV_{neg}
12.0%	34565	8564	83.01%	59.78%	24.78%	95.66%
13.5%	31690	8225	79.72%	63.78%	26.00%	95.17%
15.0%	28836	7813	75.73%	67.48%	27.09%	94.57%
16.5%	25926	7366	71.40%	71.29%	28.41%	93.98%
18.0%	23211	6869	66.58%	74.72%	29.59%	93.34%
19.5%	20622	6391	61.95%	77.99%	30.99%	92.78%
21.0%	18276	5894	57.13%	80.85%	32.25%	92.20%

This criteria is usually applied to choose the probability threshold. Nevertheless, if costs associated to each possible misclassification are known they could be used to get better results. In our case, if we assume a probability level of 16.5% we will detect approximately 71% of actual total cancellations (*sensitivity*) and 71% of actual partial lapses (*specificity*). Therefore, for this threshold probability we have a reasonably good discrimination ability of the model.

5.3 Analysing different types of customers in separate

For illustration purposes, let us consider a 35 year-old male customer, with 5 years of seniority, 90 days of notice before renewal and no new policies within last 12 months. The estimated probability of a total cancellation if he has house and motor policies is 1.6% (by setting the remaining covariates equal to 0). This probability is 4.1% if he has contents and motor policies and 7.1% if he has contents and house policies. In case that he has the three types of policies this probability equals 2.1%. It is important to remark that additional policies are always associated with a lower probability of a total cancellation except for the case of the contents policy, for which the effect is the oposite.

If we now take as a standard customer the same 35 year-old male customer, with 5 years of seniority, 90 days of notice before renewal, no new policies within

last 12 months and with the three types of policy contracts, we can compare the correspoding probability of a total cancellation (2.1%) when additional risk factors are considered. For example, his probability increases to 2.7% if he had a change of address within 12 and 24 months ago. In the case of a claim within 6 and 12 months ago the probability of the standard customer increases to 3.3%. Finally, this probability is equal to 2.4% if he has a core customer status. The opposite effect would have a substantial increase in the premium within last year, as it reduces the probability of a total cancellation for the standard customer to 1.8%.

As mentioned before, the factor with the most dramatic impact on the risk of a total cancellation is external companies. If we consider the same standard customer but we assume that external company A is involved in the first cancellation, then the probability of a total cancellation increases to 22%. As a final example, if we consider the standard customer with a claim within 6 and 12 months ago, change of address within 12 and 24 months ago, core customer status and external company involved in the first cancellation, then the probability of a total cancellation increases to 38%.

Therefore, with this type of analysis we can have an estimation of the probability of a total cancellation for each particular customer. Additionally, we have also identified factors having the largest effect on increasing the risk of a total cancellation. Firstly, according to the types of policies we can identify two types of customers, those with contents policy and those with only house and/or motor policies. The first group will always have a higher risk of a total cancellation than the second one. In any group, the probability will always be increased by the presence of risk factors such as claims, change of addres more than one year ago, core customer status and, most important, external companies. The rest of covariates in our analysis would also have their specific contribution when identifying more precisely the risk group each customer belongs to. This information can be the basis for segmentation procedures that could result in segment-specific marketing strategies.

Chapter 6

A comparison of alternative models for customer lifetime duration

In this chapter[1] we address alternatives procedures aimed at predicting the duration of a customer, once a notice for cancellation has been reported to the company. One possible statistic that can be used to compare these methods is the expected remaining lifetime, another one could be the probability that the insured stays in the company three more months. We wonder which of these measures is more useful to capture the information on the customer and to furher implement retention policies. Several methods provide useful tools to characterise the individuals and to predict their behaviour.

In this chapter we firstly introduce the proportional hazards regression model and the Tobit model. These two models are frequently used in marketing for modelling customer lifetime duration. Secondly, the comparison of these two methods is performed. Finally, the extension to the case where parameters in the proportional hazards regression model are time-dependent is presented.

[1]Most parts of this chapter are also part of the paper: Brockett, P.L., Golden, L.L., Guillen, M., Nielsen, J.P., Parner, J. & Perez-Marin, A.M. (2005), "Household multiple policy retention effects of first policy cancellation: how much time do you have to stop total customer defection?," submitted for publication.

6.1 Proportional hazards regression model

We use a proportional hazards model to concentrate on those households that do not cancel all their policies simultaneously. For these households, we estimate the length of time from the initial consumer cancellation of the first policy, until cancellation of the last policy held by the household with the insurer. This proportional hazards analysis originates in the biostatistical literature and a brief explanation appears below (Cox, 1972).

In the case of sequential withdrawal of the customer, we model the time between first cancellation notification and the final complete withdrawal by assuming that there is a baseline (stochastic) distribution for the time a customer will take for defection, and that the relative risk of an individual customer defecting completely changes from this baseline according to their particular set of individual household covariates.

The instantaneous probability of total defection at time t given survival (partial defection) up to time t is called the hazard function at time t. The actual time which the household stays with the company, T, is a random variable, and the proportional hazards regression model (Cox, 1972) specifies that the hazard function for a random survival time T is given by

$$\alpha(t) = \lim_{dt \to 0} \frac{\Pr(t \leq T < t + dt | T \geq t)}{dt},$$

which is the product of a baseline hazard $\alpha_0(t)$ and a specific covariate dependent factor,

$$\alpha(t|z_i) = \alpha_0(t) \exp(\beta' z_i)$$

where z_i is the p dimensional observed covariate column vector for individual i and β is the unknown regression coefficient column vector.

This model is called the proportional hazards model since if we look at two individuals with covariate values z_0 and z_1, the ratio of their hazard functions is constant ($\exp[\beta'(z_0 - z_1)]$) over time. The hazard function $\alpha(t)$ can be used for determining the survival function $S(t) = 1 - F(t)$, where $F(t)$ is the distribution

function, on the basis of the relationship $S(t) = \exp\left(-\int_0^t \alpha(s)ds\right)$. Expectation of the time to total withdrawal can be obtained by integrating the survival function.

The parameters β can be estimated even without pre-specifying the baseline survival curve or hazard function. Efron (1977) has shown that the partial likelihood

$$L(\beta) = \prod_{j=1}^{D} \frac{\exp(\beta' s_j)}{\prod_{l=1}^{d_j} \exp(\beta' z_k) - \frac{l-1}{d_j} \sum_{k \in \Delta_j} \exp(\beta' z_k)},$$

can be maximized independently of the unknown baseline hazard function, $\alpha_0(t)$ to yield estimations of β. Here $t_1 < t_2 < ... < t_D$ denotes the D distinct ordered event times, d_j denotes the number of total cancellations that occur at t_j, Δ_j denotes the set of all households who cancel all the policies at time t_j, and R_j is the set of all households at risk of canceling their policy just prior to t_j. This regression model is semiparametric as the baseline hazard function has to be obtained using non parametric methods.

Most of the covariates in our application are binary and can be understood as indicators of the presence of a risk factor (for example, a change of address or a claim). The sign of the parameter estimate can be interpreted as the effect of the corresponding covariate on the expected time to final withdrawal from the company.

When the parameter estimate is positive, we conclude that the hazard for the household with the associated covariate is larger than in the absence of the indicator of this covariate. On the basis of proportionality, the corresponding resulting survival function is also steeper. Thus, a positive parameter estimate is associated to a shorter time to total withdrawal for those households that have the risk factor signaled by the covariate, compared to those without the risk factor. Parameters with positive and significant coefficients thus signal to the insurer that they have a shorter time to react to the initial lapse in order to forestall total household withdrawal than would be the situation otherwise.

For estimating the baseline hazard $a_0(t)$ we use a modification of the standard Nelson-Aalen estimator (Aalen, 1978; Nelson, 1969 and Nelson, 1972), the naive local constant estimator, introduced in Chapter 3. By adapting the formulation

of this new estimator to the estimation of the baseline cumulative hazard in the proportional hazards regression model, we obtain the following estimator

$$\widehat{\Lambda}_{NLC}(t) = \sum_{t_j=0}^{\max(t-b,0)} \frac{d_j}{\sum_{l\in R_j} \exp(\beta' z_l)} + \sum_{t_j=\max(t-b,0)}^{t+b} \frac{d_j}{\sum_{l\in R_j} \gamma_{t,j} \exp(\beta' z_l)}, \quad (6.1)$$

where b is the bandwith and $\gamma_{t,b}$ is a normming constant defined in (3.2).

6.2 The Tobit model

Another method that can be applied in that context is the standard Tobit model (Tobin, 1958). Applying that model to the customer lifetime analysis results in the following specification

$$\begin{aligned} y_i^* &= \beta' z_i + \varepsilon_i \\ y_i &= \begin{cases} 0 & \text{if } y_i^* \leq 0 \\ c_i & \text{if } y_i^* \geq c_i \\ y_i^* & \text{otherwise} \end{cases} \end{aligned}$$

for $i = 1, ..., n$, where y_i^* is the so-called index variable, y_i is the lenght of the customer i's lifetime, c_i is the censoring point for customer i (the customer's maximum observable lifetime), z_i is the vector of covariates affecting the lenght of the customer's lifetime, and $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$. Parameter estimates are obtained by maximizing the likelihood function. With an estimation of β and σ_ε^2 one can estimate the probabilities concerning the customer residual lifetime.

In this model framework, it can be proved that the expectation of y for a particular customer with covariate vector z_i is given by

$$\begin{aligned} E[y|z_i] &= c_i \left(1 - \Phi\left(\frac{c_i - \beta' z_i}{\sigma_\varepsilon}\right)\right) + \left(\Phi\left(\frac{c_i - \beta' z_i}{\sigma_\varepsilon}\right) - \Phi\left(\frac{-\beta' z_i}{\sigma_\varepsilon}\right)\right) \beta' z_i + \\ &\quad + \sigma_\varepsilon \int_{\alpha_0}^{\alpha_{c_i}} \left(\frac{\varepsilon}{\sigma_\varepsilon}\right) \phi\left(\frac{\varepsilon}{\sigma_\varepsilon}\right) d\left(\frac{\varepsilon}{\sigma_\varepsilon}\right), \end{aligned}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal density and distribution function, respectively, and $\alpha_j = (j - \beta' z_i)/\sigma_\varepsilon$ for $j = 0, c_i$.

In the case of the Tobit model when the parameter estimate is significant and negative the customer residual lifetime duration is shorter than in the absence of the indicator of the corresponding covariate. Parameters with positive and significant coefficients let us identify factors associated to a longer residual lifetime duration.

6.3 Comparison of the methods

For those households with more than one policy that do not cancel all their policies at the same time, we analyze expected amount of time between the first cancellation and the final termination of all policies with the company (called the residual lifetime of the household with the insurer) by using the proportional hazards regression model and a Tobit model described earlier.

Table 6.1 displays parameter estimates for the proportional hazards regression model. The likelihood ratio test for the overall significance of the Cox regression model is high. The corresponding LR tests statistics for the Cox model is 32623.9, which are chi-squared distributed with 31 degrees of freedom ($p - value < .001$). Therefore, this result indicates that the covariates have a significant effect.

Parameter estimates for the Tobit model are displayed in Table 6.2. The intercept estimate is equal to 666.13 and its standard error is 14.58[2]. The scale parameter is 396.06 and its standard error is 1.37. The log likelihood equals -340533.761.

According to parameters in Tables 6.1 and 6.2 covariates with significant effects in both models have the same influence on the risk (or the residual lifetime). Therefore, generally speaking, the same conclusions about the direction of the effect of significant risk factors are obtained for both models.

For the proportional hazards regression model, having made a change of address between six and twelve months prior to the first lapse is not significantly related to the expected time between the first lapse and the final termination of all policies, however all other parameters associated in the model are significant at the 5% level, except the indicator for having experienced a premium increase more than one year

[2]The intercept is significant (chi-squared statistic equal to 2087.91, $p - value < .0001$).

Table 6.1.

Cox regression model results.

Parameter	Estimate	Standard Error	Hazard Rate	p-value
Change of address less 2 m. ago	-0.245	0.019	0.783	<0.001
Change of address 2 - 6 m. ago	-0.083	0.020	0.920	<0.001
Change of address 6 - 12 m. ago	-0.023	0.020	0.978	0.252
Change of address 12 - 24 m. ago	0.044	0.019	1.045	0.021
Change of address more 24 m. ago	0.157	0.025	1.170	<0.001
Tenure	-0.003	0.001	0.997	<0.001
Claims, less 2 m. ago	0.096	0.016	1.100	<0.001
Claims, 2 - 6 m. ago	0.161	0.015	1.175	<0.001
Claims, 6 - 12 m. ago	0.185	0.015	1.203	<0.001
Claims, 12 - 24 m. ago	0.228	0.017	1.256	<0.001
Claims, more 24 m. ago	0.257	0.027	1.294	<0.001
Contents0	0.681	0.029	1.975	<0.001
Contents1	-0.869	0.016	0.419	<0.001
Corecust	-0.042	0.011	0.959	<0.001
Age	-0.003	<0.001	0.998	<0.001
External company A	1.727	0.018	5.625	<0.001
External company B	1.528	0.020	4.611	<0.001
External company C	1.652	0.019	5.217	<0.001
External company D	1.778	0.020	5.919	<0.001
Another known external company	1.643	0.012	5.170	<0.001
Gender (male)	0.103	0.012	1.109	<0.001
House0	0.222	0.017	1.249	<0.001
House1	-0.559	0.017	0.571	<0.001
Motor0	0.415	0.021	1.515	<0.001
Motor1	-0.529	0.018	0.589	<0.001
Newcontents	-0.059	0.016	0.942	<0.001
Newhouse	-0.109	0.024	0.897	<0.001
Newmotor	-0.076	0.017	0.927	<0.001
Notice[a]	>-0.001	<0.001	1.000	<0.001
Pruning within past 12 months	0.188	0.029	1.207	<0.001
Pruning more than 1 year ago	0.095	0.053	1.100	0.075

[a]Negative parameter estimate.

Table 6.2.

Tobit regression model results.

Parameter	Estimate	Standard Error	Chi-Square	p-value
Change of address less 2 m. ago	70.198	6.803	106.49	<0.001
Change of address 2 - 6 m. ago	6.754	7.168	0.89	0.346
Change of address 6 - 12 m. ago	-4.130	7.119	0.34	0.562
Change of address 12 - 24 m. ago	-29.116	6.926	17.67	<0.001
Change of address more 24 m. ago	-82.058	8.909	84.84	<0.001
Tenure	1.942	0.210	85.14	<0.001
Claims, less 2 m. ago	-61.178	5.838	109.81	<0.001
Claims, 2 - 6 m. ago	-77.610	5.397	206.75	<0.001
Claims, 6 - 12 m. ago	-90.527	5.505	270.43	<0.001
Claims, 12 - 24 m. ago	-109.482	6.138	318.18	<0.001
Claims, more 24 m. ago	-151.214	9.798	238.18	<0.001
Contents0	-174.219	10.693	265.45	<0.001
Contents1	266.768	6.133	1892.11	<0.001
Corecust	32.199	3.903	68.05	<0.001
Age	1.904	0.132	206.82	<0.001
External company A	-562.644	6.948	6558.26	<0.001
External company B	-528.710	7.583	4861.87	<0.001
External company C	-580.379	7.302	6316.73	<0.001
External company D	-597.857	7.605	6180.00	<0.001
Another known external company	-565.852	4.231	17888.20	<0.001
Gender (male)	-39.388	4.117	91.55	<0.001
House0	-12.742	6.523	3.82	0.051
House1	133.892	6.433	433.21	<0.001
Motor0	-112.112	7.849	204.02	<0.001
Motor1	133.336	6.973	365.60	<0.001
Newcontents	-6.088	5.779	1.11	0.292
Newhouse	62.253	8.552	52.99	<0.001
Newmotor	11.626	6.215	3.50	0.061
Notice	0.056	0.015	13.18	<0.001
Pruning within past 12 months	-40.515	11.214	13.05	<0.001
Pruning more than 1 year ago	-34.307	19.686	3.04	0.081

before the first lapse (pruning more than one year ago) which is marginally significant (not significant at the 5% level of significance, but significant at the 10% level).

Change of address, claims, external company, and pruning are the strongest factors contributing to reducing the expected residual life. For the Tobit regression model, change of address between 2 and 6 months ago and between 6 and 12 months ago do not have a significant effect on the residual lifetime. The same occurs for newcontents. The effect of house0, newmotor and pruning within more than one year ago are not significant at the 5% level, but are significant at the 10% level.

For covariates having a significant effect on the risk (or residual lifetime), the following comments can be made on the basis of any of the two models (see Table 6.1 and 6.2). When looking at all the parameters related to the change of address one can see that the effect is different depending on the moment when this event took place. A recent change of address slightly increases the expected time between the first policy lapse and the final termination of the last household policy. But, as the time elapsed since the household has moved increases, the contribution of this factor on the expected time until final termination (residual life) has the opposite effect: it reduces residual life.

The parameter associated with tenure is significant and negative (in the case of the Cox model, positive for the Tobit model) suggesting the length of time that members of a household have been with the insurer the less likely they are to switch brands—the longer it will take them to switch. This finding underscores the importance of customer loyalty. Female customers have also a slightly longer expected residual life than do male customers. Beyond the actuarial difference, others studies have also found meaningful purchasing behavioural differences between men and women [3].

The presence of claims reduces residual life, and the effect becomes more remarkable as the time since the claim has occurred increases. This later effect can be due to some delay in the compensation of claims, so the assessment of the claims

[3]For example, Gandolfi & Miners (1996) found gender differences in life insurance ownership.

handling process by the household is delayed.

An expected result is that the core customer status increases residual life. As we mentioned before, these households have some advantages, and these special features seem to be effective for customer retention, or at least to dissuade customers from completely leaving the company, although it does also increase their attractiveness for competitors. The parameter estimate corresponding to the covariate age is significant and negative sign in the Cox model (positive in the Tobit model) indicating that as the age of the policyholder increases there is a slight increase in the expected residual life.

As with the probability of total cancellation discussed earlier, the competitive effects of the market are very important. Furthermore, this is the factor most significantly related to a reduction in residual life of the client household with the insurer. This is particularly true for the external companies coded as A and D.

When an external company is involved in a particular cancellation, it seems to attract the new customer toward other lines of business. It is also easier for the customer to get information about the premium costs and the characteristics of other products in the new insurance company. Competitors may provide information that the current company does not to existing customers, making switching more attractive, possibly because of customer ignorance of product opportunities with their current insurer.

Tables 6.1 and 6.2 also shows that variables measuring new business of the household with the insurer within past 12 months (newcontents, newhouse, newmotor) have a significant and negative impact. This suggests that recent business is contributing to an increase residual life. Again, this may be related to customer contact and education. The new business, makes the biggest contribution is buying a new contents policy.

The parameter associated with notice is also significant and negative, and as the time since notification until renewal increases, the expectation of residual life increases slightly, also. And a substantial rise in premiums is associated with a reduced residual life, but the effect is different depending on the moment when this

event took place. A substantial rise in the premium during last year is associated with a shorter residual lifetime duration compared to a premium rise that took place more than one year ago.

For variables relating to the insurance portfolio of the household before the first lapse, one causing less reduction in residual life is house0, followed by, motor0 and finally contents0. The fact that the initial state of the insurance portfolio has a significant effect on residual life can be tested, for example in the case of the Cox model, using Wald's statistic (value 306.606, chi-squared with 2 degrees of freedom, $p - value < .001$).

Concerning the insurance portfolio of the household after the first lapse, the one associated with the largest increase in the expected residual life is contents1, followed by motor1 and finally house1. Again, a Wald statistic test confirms that the state of the household's insurance portfolio following the first lapse also significantly impacts the residual time until final total policy cancellation (test statistic equal to 941.135, chi-squared distributed with 2 degrees of freedom, $p - value < .001$).

Examples of customer survival functions and expectations

This section develops an illustration of the application of the proportional hazards regression model and the Tobit model for understanding customer retention. The estimated parameters can be used to obtain the survival function to model retention for any given customer, which is potentially useful strategic demand side information. To do this, in the case of the proportional hazards regression model, we first need to estimate the baseline hazard rate using non parametric techniques. In our case, we use the naive local constant estimator (6.1) and we proceed in the same way as explained in section 3.5. By integrating the resulting baseline cumulative hazard[4] we obtain the baseline survival curve plotted in Figure 6.1. Regarding the Tobit model, the details about the calculations are shown in Appendix C.

[4]A suitable probability function $K_b(\cdot)$ is the biweight kernel $K_b(\cdot) = \frac{15}{16}\{1 - (\cdot/b)^2\}^2$ where $b = 400$. The same biweight kernel with the same b has been used to smooth $\alpha'(t)^2$ one more time.

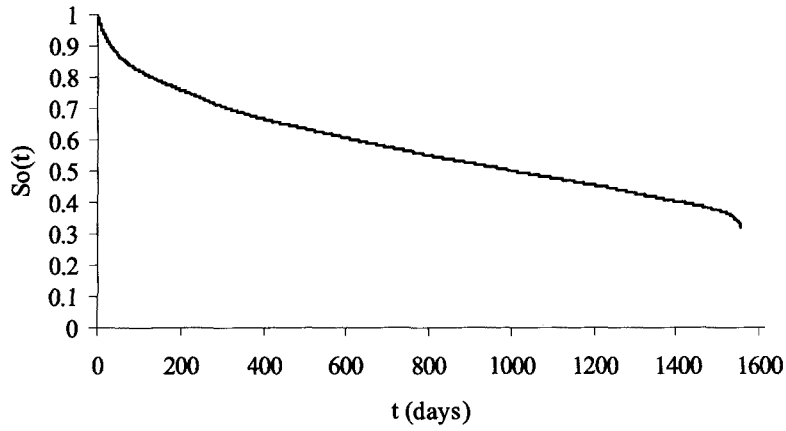


Figure 6.1. Baseline survival function.

Figures 6.2 and 6.6 show the survival function for a 55 year-old male customer, with ten years of tenure with the insurer, no change of address within the last two years, a claim between two and six months ago, and no external company involvement in the notification, giving 150 days of notice before renewal, no new business with the insurer within the past twelve months, no core customer status, and no pruning. Assume also that the customer has contents, house, and automobile policies before the first lapse. Survival curves and expectations are shown for both models depending on the first policy being cancelled.

It can be observed that the survival curve with the steepest slope is the one corresponding to those households who first cancel the contents policy for whom we observe a shorter residual life (of about 779 days in the case of the Cox model and 658 in the case of the Tobit model) before final expected exit from the company. The largest expected residual life corresponds to the case in which the automobile policy is the first to be notified for cancellation.

Figures 6.3 and 6.7 consider the same customer as in the previous example, but with contents and automobile policies before the first lapse and only an automobile policy after the first lapse. In this case, we compare the survival function and the residual life depending on whether or not any external company was involved in the cancellation.

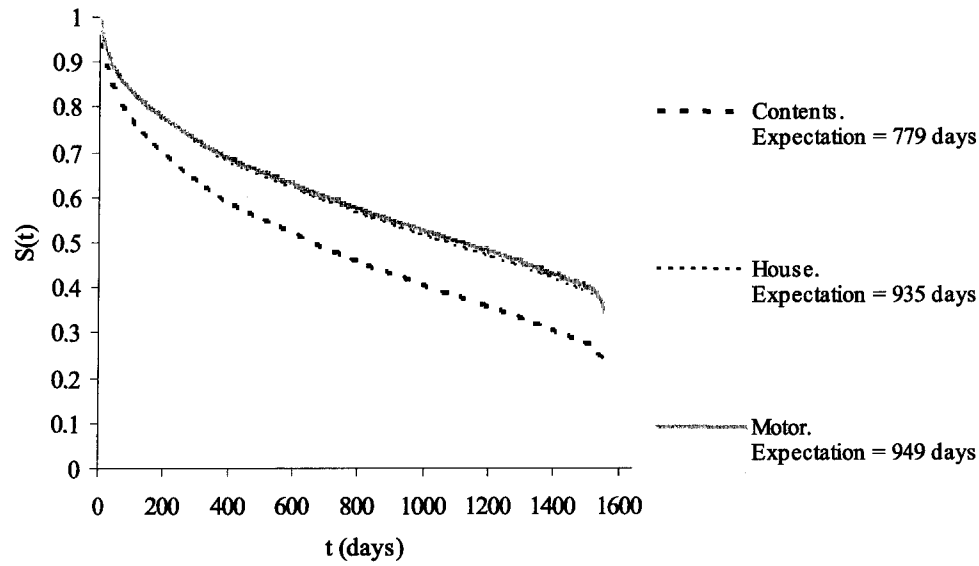


Figure 6.2. Survival function depending on the first policy being cancelled. Cox model.

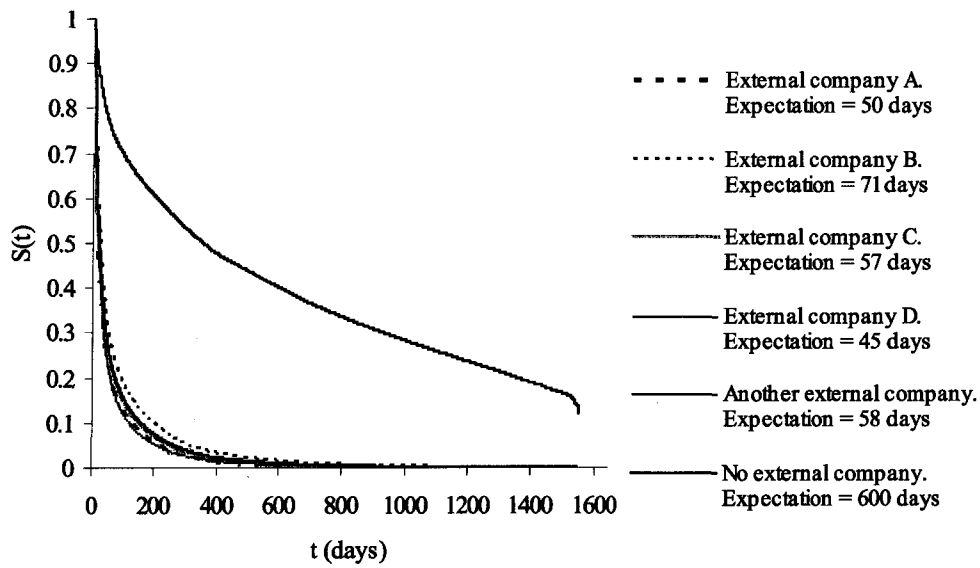


Figure 6.3. Survival functions depending on external companies. Cox model.

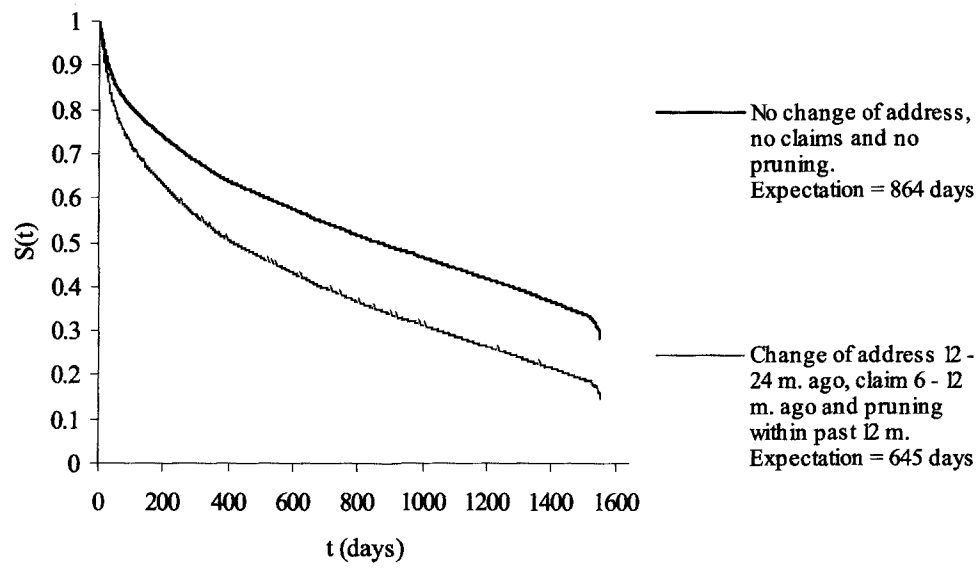


Figure 6.4. Survival function depending on change of address, claims and pruning. Cox model.

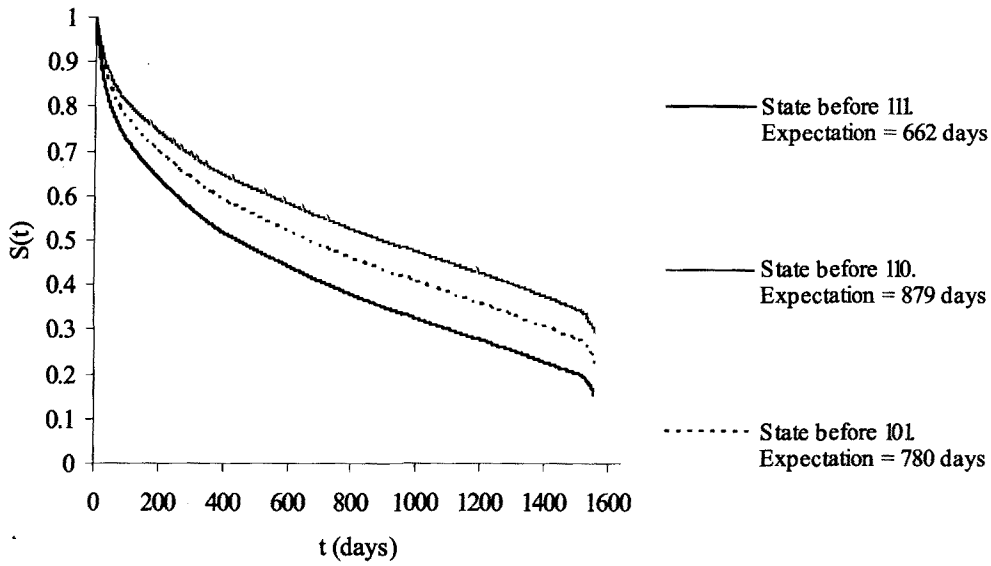


Figure 6.5. Survival function for a households with only the contents policy after the first lapse depending on the types of policies they had before the first lapse. Cox model.

The external competitor coded as company D is the one causing the most reduced residual life (of 45 days for the Cox regression model and 126 days for the Tobit model) while, on the other hand, the residual life is substantially larger when no external company is involved in the first cancellation (600 days for the Cox model and 545 days for the Tobit model).

Figures 6.4 and 6.8 show the survival function and the estimated expected residual life for a thirty year-old female customer with five years of tenure with the insurer, with both contents and automobile policies before the first lapse, but only a contents policy after the first lapse, with no external company involved in the notification, ninety days of notice of cancellation given to the insurer before renewal, no new business with the insurer within the past twelve months, and no core customer status. Results are compared depending on whether or not the customer has had a change of address, claims, and if there has been pruning or not. As would be expected, when none of these events have occurred, residual life (864 days for the Cox model and 723 for the Tobit model) is larger than when the three of them have occurred (645 days for the Cox model and 572 days for the Tobit model).

We should now illustrate the importance of the particular insurance policies owned by the household before the first lapse. We consider the same customer as in Figure 6.2, but with only the contents policy remaining after the first lapse. Figures 6.5 and 6.9 compare the survival functions for the portfolio before the first lapse. The survival function with the steepest slope, as well as the shortest residual life, corresponds to the household that has all three types of policies before the first lapse (662 days for the Cox model and 657 for the Tobit model).

The examples above illustrate how dramatic the impact of the covariates can be on the time the insurer has to respond to the first cancellation, before subsequent cancellations. These covariates are measurable characteristics of customers in the marketplace and can be employed to help signal customer defection to the competition.

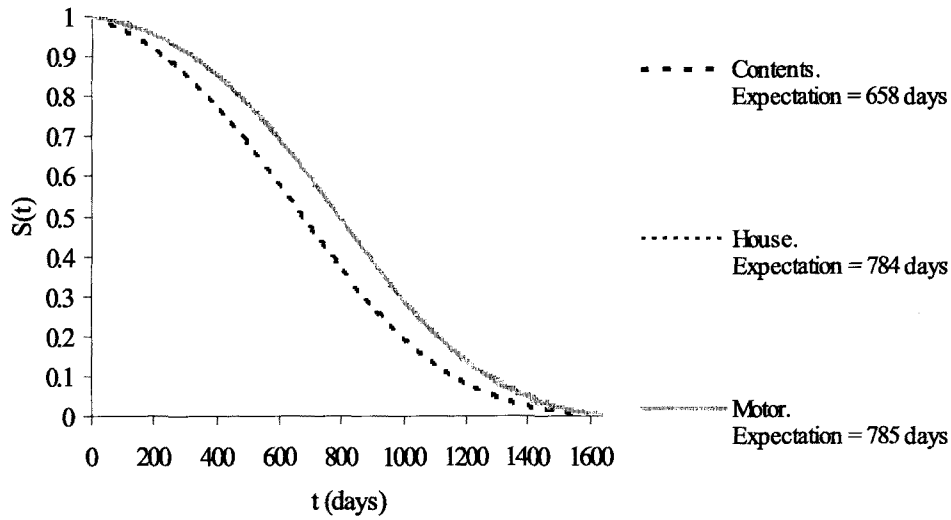


Figure 6.6.Survival function depending on the first policy being cancelled. Tobit model.

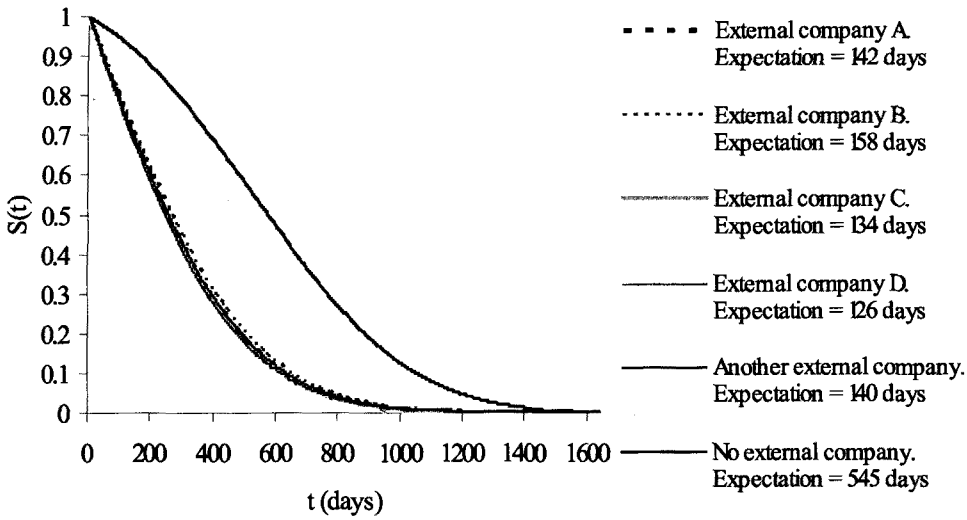


Figure 6.7. Survival functions depending on external companies. Tobit model.

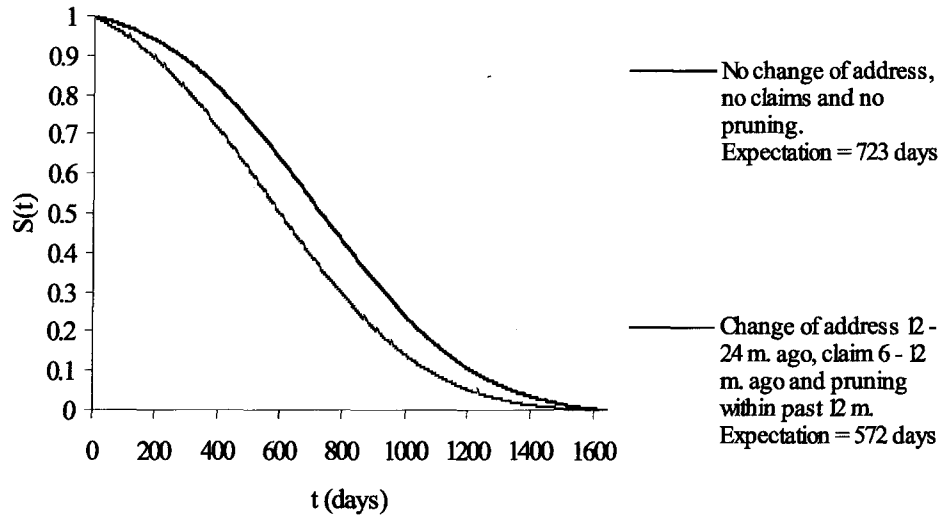


Figure 6.8. Survival function depending on change of address, claims and pruning. Tobit model.

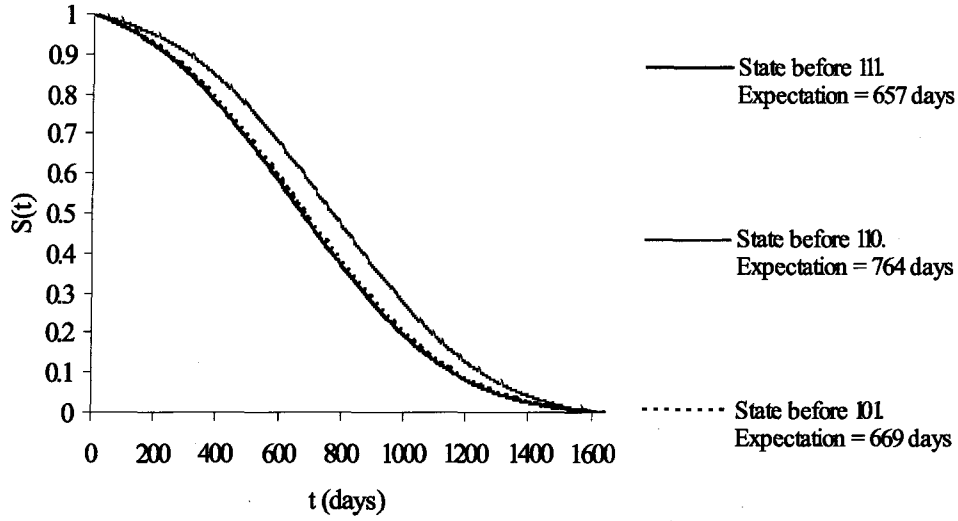


Figure 6.9. Survival function for a households with only the contents policy after the first lapse depending on the types of policies they had before the first lapse. Tobit model.

The risk to loose a customer within a given period of time

Now we compare both model's ability to detect customers with a high risk to compleately leave the company within different periods of time, namely 3 months, 6 months and 12 months. In order to do that, we estimate the probabilities of cancelling all the remaining policies in the given period of time and we combine this information with the actually observed lifetime.

Results for the 3-month period case are summarized in Tables 6.3 and 6.4 for the Cox and Tobit model, respectively. On the one hand, among those who actually experience a total cancellation within a 3-month period, 28548 customers, 3642 have an estimated probability by using the Tobit model higher or equal to 0.75 and 1346 between 0.5 and 0.75. Therefore only 17.47% of those customers have an estimated probability higher or equal to 0.5. This frequency increases to 74.7% in the case of the Cox model. On the other hand, in the case of the Tobit model, among those who actually survive beyond a 3-months period, 32377 customers, 95.34% of them have an estimated probability of a total cancellation within the reference period lower than 0.5, while this probability is 74.4% in the case of the Cox model.

Table 6.3. Expected probabilities for the Cox Model for a 3-month time period.

	Observed residual lifetime		
p =P(residual life<3 month)	Residual life>3 month	Residual life$\leq$3 month	
$p < 0.25$	19929	4559	24488
$0.25 \leq p < 0.5$	4162	2662	6824
$0.5 \leq p < 0.75$	6433	13099	19532
$p \geq 0.75$	1853	8228	10081
	32377	28548	60925

Table 6.4. Expected probabilities for the Tobit Model for a 3-month time period

	Observed residual lifetime		
p =P(residual life<3 month)	Residual life>3 month	Residual life$\leq$3 month	
$p < 0.25$	28147	19967	48114
$0.25 \leq p < 0.5$	2720	3593	6313
$0.5 \leq p < 0.75$	947	1346	2293
$p \geq 0.75$	563	3642	4205
	32377	28548	60925

Table 6.5. Expected probabilities for the Cox Model for a 6-month time period

	Observed residual lifetime		
p =P(residual life<6 month)	Residual life>6 month	Residual life$\leq$ 6 month	
$p < 0.25$	11664	3606	15270
$0.25 \leq p < 0.5$	8328	6197	14525
$0.5 \leq p < 0.75$	2946	7720	10666
$p \geq 0.75$	3129	17335	20464
	26067	34858	60925

Table 6.6. Expected probabilities for the Tobit Model for a 6-month time period

	Observed residual lifetime		
p =P(residual life<6 month)	Residual life>6 month	Residual life$\leq$ 6 month	
$p < 0.25$	17809	8180	25989
$0.25 \leq p < 0.5$	5950	15507	21457
$0.5 \leq p < 0.75$	1389	2552	3941
$p \geq 0.75$	919	8619	9538
	26067	34858	60925

The same probabilities can be estimated for a 6-month period. Results are shown in Tables 6.5 and 6.6 for the Cox and Tobit model respectively. Among those who actually cancel all their remaining policies within a 6-month period, 34858 customers, 8619 have an estimated probability by using the Tobit model higher or equal to 0.75 and 2552 between 0.5 and 0.75. Thus, 32.05% of them have an estimated probability higher or equal to 0.5. In the case of the Cox model, this frequency is much higher, 71.88% . For those who actually survive beyond a 6-months period, 26067 customers, 91.15% of them have an estimated probability of a total cancellation within the reference period lower than 0.5 if we consider the Tobit model. For the Cox model, this probability is 76.7%.

We finally present the probability of a total cancellation within a 12-month period, see Tables 6.7 and 6.8. In that case, among those who actually experience a total cancellation within a 12-month period, 42922 customers, 78.66% % of them have an estimated probability higher or equal to 0.5 (if we consider the Tobit model). This frequency is 69.44% in the case of the Cox model. For the Tobit model, among

Table 6.7. Expected probabilities for the Cox Model for a 12-month time period

	Observed residual lifetime		
p =P(residual life<12 month)	Residual life>12 m.	Residual life$\leq$ 12 m.	
$p < 0.25$	2541	966	3507
$0.25 \leq p < 0.5$	11695	12151	23846
$0.5 \leq p < 0.75$	953	2718	3671
$p \geq 0.75$	2814	27087	29901
	18003	42922	60925

Table 6.8. Expected probabilities for the Tobit Model for a 12-month time period

	Observed residual lifetime		
p =P(residual life<12 month)	Residual life>12 m.	Residual life$\leq$ 12 m.	
$p < 0.25$	8203	2827	11030
$0.25 \leq p < 0.5$	5446	6336	11783
$0.5 \leq p < 0.75$	3386	15892	19278
$p \geq 0.75$	968	17867	18835
	18003	42922	60925

those who actually survive beyond a 12-months period, 18003 customers, 75.81% of them have an estimated probability of a total cancellation within the reference period lower than 0.5, while this probability is 79.04% in the case of the Cox model.

Therefore, when the detection of customers with a high risk of a total cancellation within a short time period is the issue, the Cox regression model let us better identify them than the Tobit model. When dealing with the detection of customers who are going to survive beyond short time periods the Tobit regression model seems to have a better performance than the Cox regression model. Nevertheless, in both cases the Cox regression model always provides reasonably good estimations, while the Tobit model clearly fails to detect total cancellations in short time periods.

Additionally, in order to choose the optimal method for detecting these customers, the cost of the potential under-estimation should be considered. In that case, the cost of not detecting a customer that is going to make a total cancellation within a 3 or 6 month-period is supposed is be higher than the cost of not detecting those who are going to survive beyond this time period. For all these reasons, the Cox regression model is preferable.

When considering longer periods of time, namely one year, the performance of both models seem to be opposite. The Cox model provides better estimated probabilities than the Tobit model for those who survive beyond the one year time period, while the Tobit model seems to be preferable when estimating probabilities for those who do a total cancellation within a one year time period. Nevertheless, the performances of the two models are not so different like in the short-time period, therefore as an overall conclusion we advocate for the use of the Cox regression model.

6.4 Time-varying covariate effect in the survival model

We now investigate whether or not the effects of covariates in the Cox regression model change over time. Methods described in section 2.4 have been widely used in survival studies in medicine or biology, where sample sizes are normally not very large and the number of covariates or risk factors is small. On the other hand, insurance companies have large portfolios, therefore in actuarial studies we frequently have to face the problem to deal with massive data sets. In this section, we present two applications of methodology described in section 2.4 to our data set.

Application 1.

In this first example, we randomly selected 2069 customers among those who do a first partial cancellation (approximately 3.4% of the total). The number of covariates has also been limited to the following 10 selected risk factors: *address* (1 if a change of address has been registered, 0 otherwise), *claim* (1 if a claim has been registered, 0 otherwise), *contents0*, *contents1*, *corecust*, *extc* (1 if any external company has been involved in the first cancellation, 0 otherwise), *house0*, *house1*, *motor0* and *motor1*. Parameter estimates of the Cox model for this subsample are shown in Table 6.9.

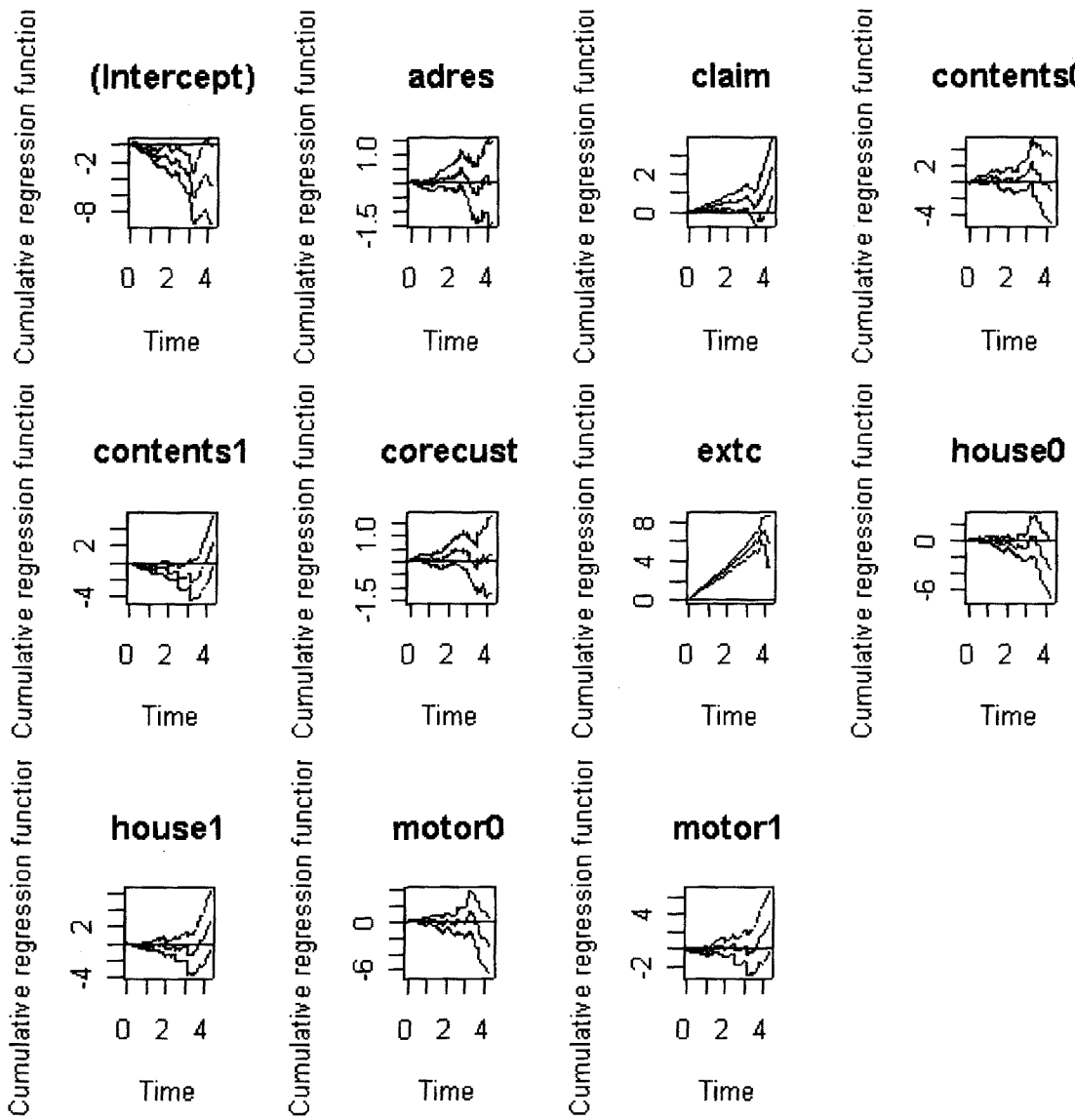


Figure 6.10. Cumulative regression functions (time measured in years).

Table 6.9. Cox regression model fitting information.

Parameter	Estimate	Standard Error	Hazard Rate	p-value
address	-0.088	0.059	0.916	0.136
claim	0.237	0.053	1.267	<.001
contents0	0.581	0.169	1.787	<.001
contents1	-0.895	0.089	0.409	<.001
corecust	0.134	0.058	1.143	0.021
extc	1.740	0.059	5.697	<.001
house0	0.074	0.091	1.077	0.416
house1	-0.586	0.093	0.557	<.001
motor0	0.206	0.111	1.229	0.064
motor1	-0.423	0.100	0.655	<.001

Table 6.10. Test for time-dependent effects.

Parameter	Test statistic	p-value
(intercept)	2.863	0.148
address	0.523	0.686
claim	1.616	0.008
contents0	3.360	0.026
contents1	4.079	0.000
corecust	0.622	0.526
extc	1.977	0.130
house0	3.434	0.002
house1	3.483	0.014
motor0	3.634	0.010
motor1	2.966	0.028

All parameters in the model are significant except for *address* and *motor0*. The likelihood ratio test for the overall significance of the Cox regression model is high at LR test statistic equal to 1122.866, which is chi-squared distributed with 10 degrees of freedom ($p < .0001$), indicating that the covariates have a significant effect on the hazard.

We now investigate for potential time-dependent covariates effects in that model. In order to do that, we assume model (2.2) and apply the *timereg* R-package (available in http://www.biostat.ku.dk/~ts/timereg.html, see Appendix D) and the methodology devised by Scheike & Martinussen (2004). In Figure 6.10 the cumulative regression functions together with the corresponding 95% pointwise confidence bands are shown (see Scheike & Martinussen, 2004, for details about the calculations).

It is important to remark that in the case of constant effects a straight line (resulting from integrating a constant parameter) should be observed. A departure from this pattern is an indicator of a time-varying effect for the corresponding covariate. Additionally, several tests of time-varying effects can be used, like the one based on the Kolmogorov-Smirnov type statistic (described in Scheike & Martinussen, 2004). Results for time-varying effects testing provided by this package are summarized in Table 6.10.

According to these results, the null hypothesis of time-constant effects is rejected at the 5% level of significance for *claim*, *contents0*, *contents1*, *house0*, *house1*, *motor0* and *motor1*. Therefore, influence of all these covariates varies with time. Generally speaking, for all of them the main change in the effect of the covariate occurs approximately at $t = 2.5$ years.

The overall effect of *claim* on the increase of the risk of a total cancellation is much more remarkable after this time point. Regarding *contents0*, there seems to be a change in the direction of the effect after $t = 2.5$. The overall effect before this time point seems to be positive (producing an increase of the risk of a total cancellation) while after $t = 2.5$ years is negative (reducing the risk of a total cancellation).

On the other hand, *contents1* is contributing to a lower risk of a total cancellation before $t = 2.5$, but afterwards its effect is the opposite. For *house0* and *motor0* a remarkable change in the effect of the covariate occurs after this time point, where a clear contribution to reducing the risk of a total cancellation is observed. The same occurs for *house1* and *motor1* after this time point, but in the opposite direction.

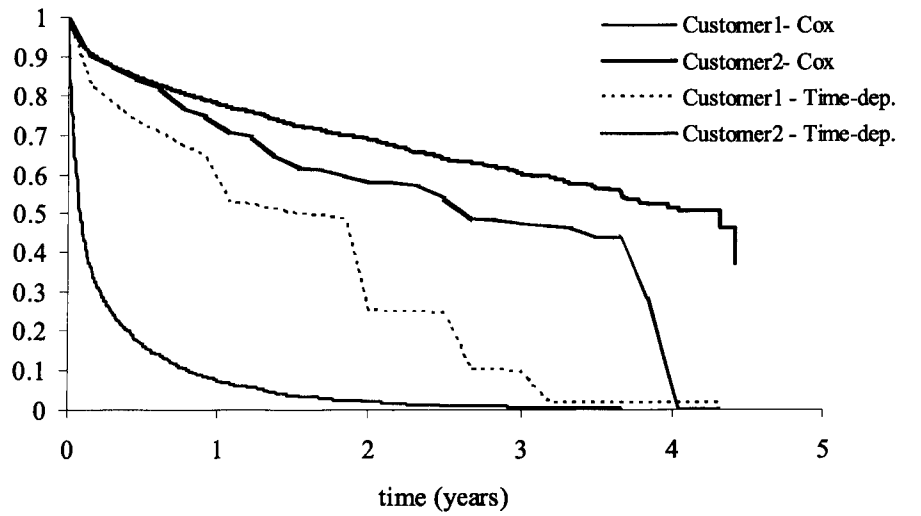


Figure 6.11. Survival curves.

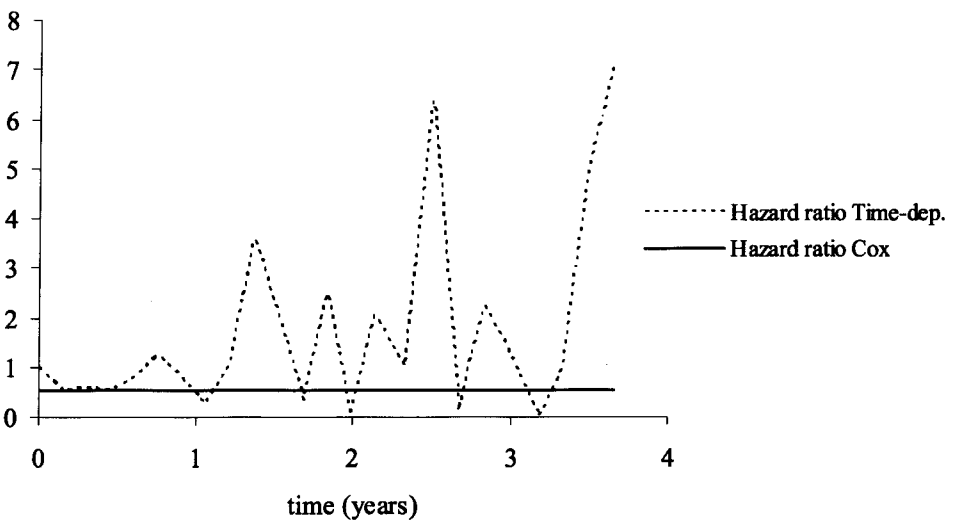


Figure 6.12. Hazard rates.

Therefore, a direct implication of this results is that the proportionality assumption in the Cox model does not hold in that case.

In order to illustrate this idea, let us consider two customers. The first one (*customer*1) has the three types of policies and cancels both the house and the motor policies, has a change of address, a claim, does not have a core customer status in the company and there is an external company involved in the cancellation.

The second one (*customer*2) has exactly the same characteristics except for the fact that he only cancels the motor policy. The corresponding survival curves for both customers obtained by using both the standard Cox model and the generalized Cox model with time dependent parameters are shown in Figure 6.11.

Both models let us conclude that *customer*2 would have a longer expected residual lifetime after first cancellation than *customer*1, but important differences in the shape of the survival function due to the changing effects of covariates over time are only captured by the model including time-dependent effects in the parameters.

Additionally, differences in the survival curves of both individuals seem to be very remarkable according to the standard Cox regression model, while the extended model with time-varying coefficients shows us that the difference between both customers is not so extensive as reported by the Cox model, specially before $t = 2$ years.

Therefore, expectations for the residual life calculated on the basis of curves obtained by using the standard Cox model would provide a too much short residual lifetime for *customer*1 and a too much long residual lifetime for *customer2.*

The completely different pattern of survival curves captured by the extended Cox model with time-dependent parameters is an evidence of the non proportionality of the hazards of both customers. This can be better illustrated in Figure 6.12, where the hazard rate of *customer*2 with respect to *customer*1 is represented. The horizontal line represents the hazard rate in the case where proportionality is assumed, i. e. the standard Cox model.

Application 2.

In this second application we focus on a very important subset of customers, those who have the three types of policies before the first cancellation and simultaneously cancel two of them. For these customers, quick marketing actions should be addressed in order to retain them, because their expected remaining lifetime is considerably short.

We randomly select 600 customers among 7381 who simultaneously cancel two policies (8.13%).

We investigate potential time-varying effects of three selected covariates considered in this analysis, *contents1*, *house1* and *extc* (1 if any external company has been involved in the first cancellation, 0 otherwise).

By applying the same methodology as in the previous application (see Scheike & Martinussen, 2004) the cumulative regression functions and 95% pointwise confidence bands can be obtained (see Figure 6.13). Results for time-varying effects testing provided by *timereg* are summarized in Table 6.11.

Evidences of time-varying effects are found in the case of *house1*. For *contents1* the null hypothesis of constant effects is rejected at the 6% level of significance, but not at the 5% level.

Results are very clear in the case of *extc*, this covariate has a constant effect on the risk of cancelling all the remaining policies.

Table 6.11. Test for time-dependent effects.

Parameter	Test statistic	p-value
(intercept)	384.276	0.000
contents1	693.626	0.058
house1	631.527	0.014
extc	96.000	0.936

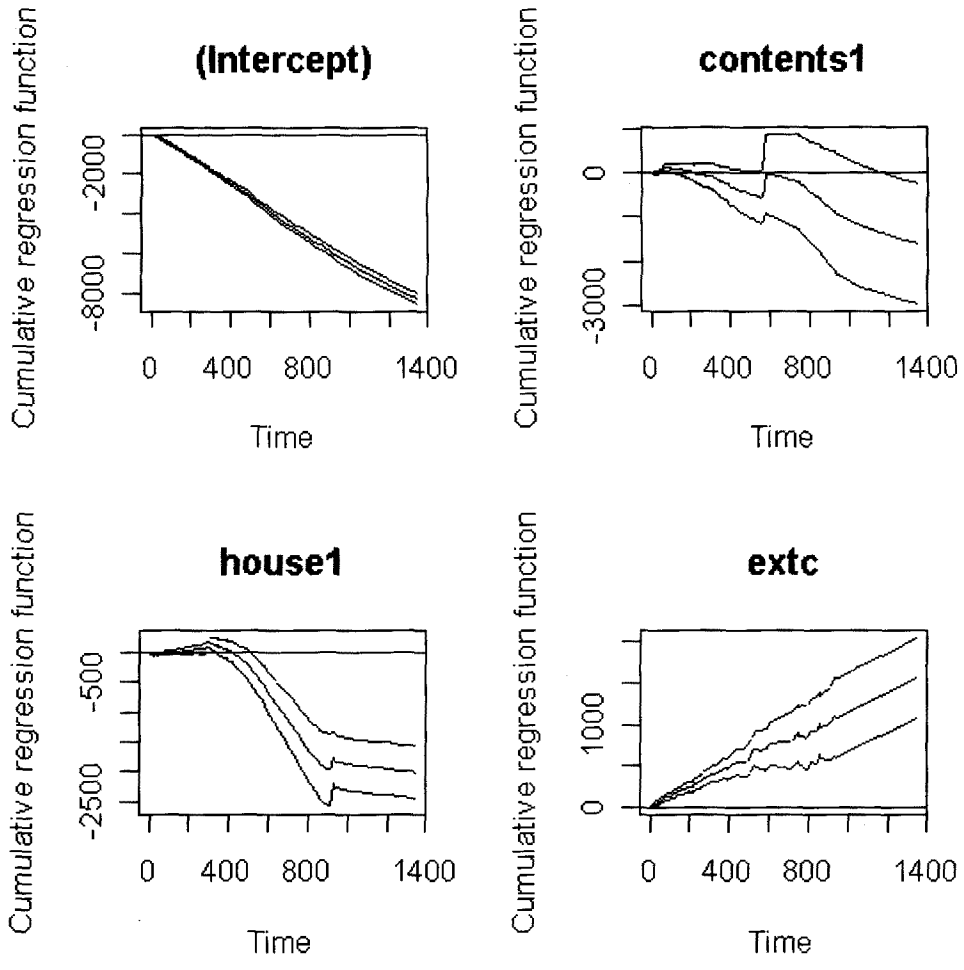


Figure 6.13. Cumulative regression functions (time measured in days).

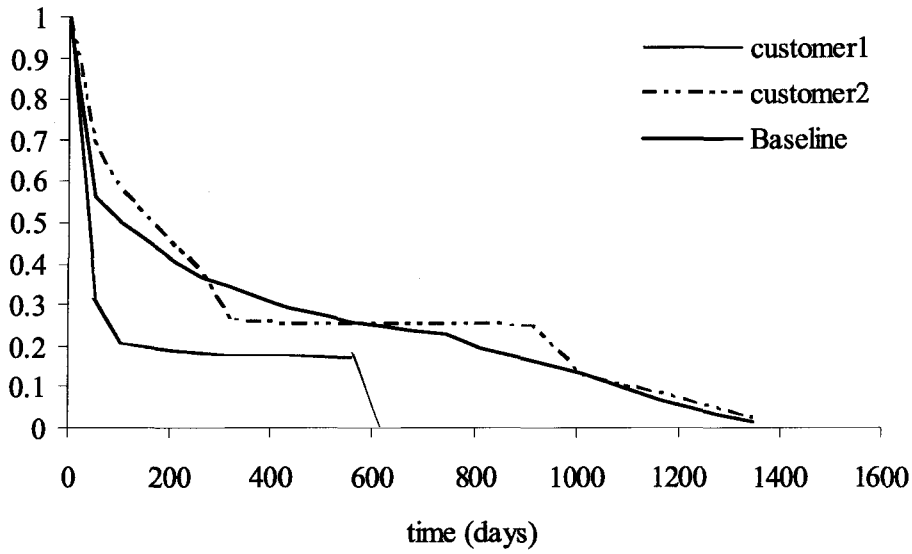


Figure 6.14. Survival curves.

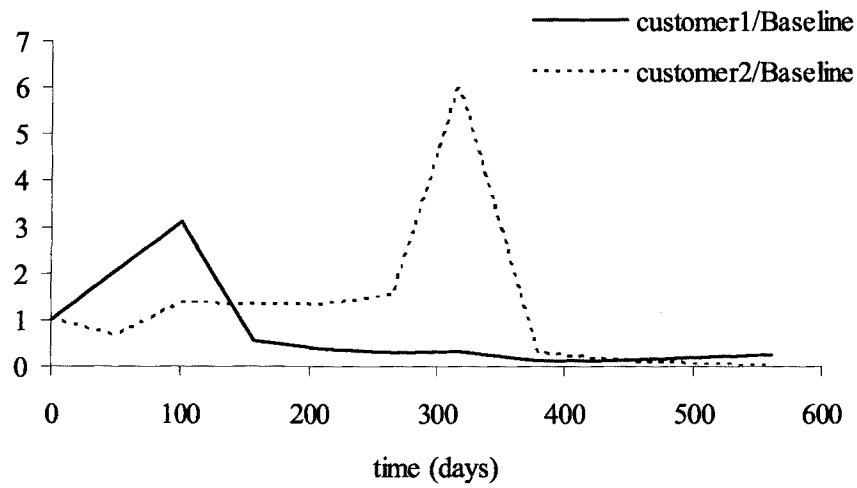
As we can see in Figure 6.13, *house1* has a positive effect (increase in the risk) during the first year approximately. This means that the risk of cancelling all the remaining policies during the first year is higher for those customers with just the house policy after the first cancellation than for those with just the motor policy (baseline group).

After the first year, the effect of this covariate is the opposite, it contributes to reduce the risk (except during a short period of time around $t = 900$ days approximately).

This can be shown in Figure 6.14, where *customer*1, *customer*2 and baseline represents a customer with just the contents, the house and the motor policy respectively after the first cancellation.

The fact that the survival curves for the baseline case and *customer*2 cross each other is an evidence of the non proportionality of hazards.

(a) hazard rates, $t <= 613$.



(b) hazard rates, $t > 613$.

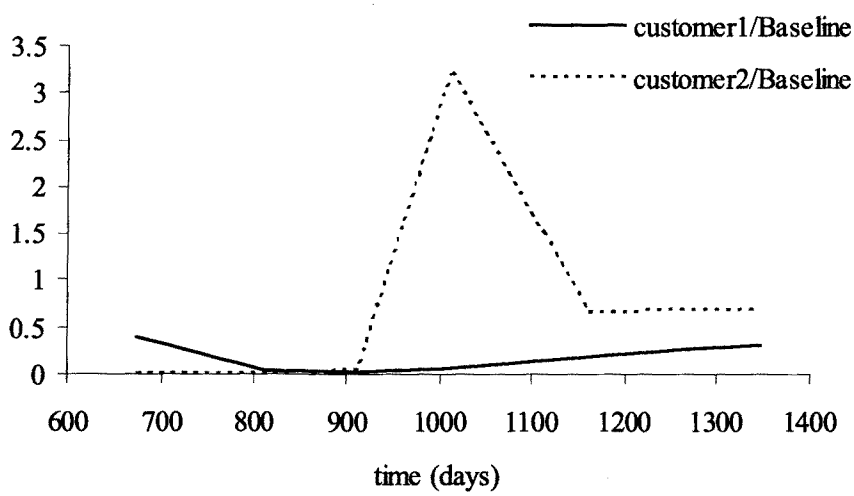


Figure 6.15. Hazard rates.

This is illustrated in Figure 6.15, where the hazard rates of *customer*1 and *customer*2 with respect to the baseline case are represented (in that case, two plots are used in order to avoid the representation of the hazard rates corresponding to $t = 613$ because of the dramatic increase of the hazard function at this point in time for *customer*1).

Chapter 7

Conclusions

In this last chapter, final remarks and conclusions are presented in a systematic way. In order to organize the argumentation, we firstly discuss the methodological contribution of this thesis and secondly, the results obtained in the empirical applications. Finally some extensions of this research are outlined.

7.1 About the methodology

The methodological contribution of this research is directly related to the first and fifth specific objectives presented in section 1.5. Conclusions are therefore presented according to the following two general topics:

Definition of a new methodology for the estimation of customer lifetime duration.

The definition of a new non parametric estimator in the field of survival analysis is the main methodological contribution of this thesis. Namely, we have introduced a new estimator of the cumulative hazard function, the naive local constant estimator, with improved bias versus variance properties compared to the traditional Nelson-Aalen estimator.

While arguing for the superiority of the proposed naive local constant estimator, we have also introduced a set-up for evaluating an estimator of the cumulative hazard function. In particular such a set-up must be able to cope with the behaviour of

the boundary region, namely the "edge effect" mentioned in Bowman, Hall & Prvan (1998). This is extensively discussed in section 3.2, where the final expressions of the optimal bandwidth and efficiency gain are functions of t. These results provide valuable properties of bandwidth selection in the context of survival analysis where the largest times are censored.

We also quantify the efficiency gain for a number of well-known distribution functions of a random variable measuring the time-to-event. In most of the cases considered here, the best gain is obtained at the lower and higher distribution quantiles, except for the Lognormal distribution, for which the shape of the curve is different from the rest of them. Most of the efficiency gains obtained are well above 20%, except for the Log-logistic, Exponential-Power and Gompertz distributions in higher quantiles. The highest efficiency gains are approximately between 50% and 70% and they correspond to the case of the Lognormal distribution.

The theoretical efficiency gain curves are obtained for a given distribution function, which is usually unknown in practice. In real applications the efficiency gain depends on the estimation of the optimal bandwidth parameter. Therefore, we carried out a little simulation study in order to adjust for the effect of plugging-in an estimation of the bandwidth in the efficiency calculations. Our results are shown in section 3.4.

Generally speaking, differences between theoretical efficiencies and those obtained by using the plug-in procedure are not dramatic, specially for the central quantiles. In any case, even after estimating the optimal bandwidth parameter, the estimated efficiency gain curve is capturing reasonably well the efficiency gain performance of the new estimator.

We give practical notes about the implementation of the new estimator in section 3.5. We suggest to use the local linear estimator (Nielsen & Tanggaard, 2001) in order to have an approximation of the optimal bandwidth that could be used in the application of the new estimator to real data.

If the estimated squared first derivative of the hazard function is very close to zero, the optimal bandwidth will be so much large that the new estimator will not

perform better than the Nelson-Aalen estimator. In this case, our suggestion is to use a more robustified estimator of the squared first derivative of the hazard function by smoothing it one more time.

The practical insights provided in section 3.5 are used in the application of the new proposed estimator to the analysis of survival with malignant melanoma (data set in Example I.3.1, Andersen, Borgan, Gill & Keiding, 1993). The results let us compare estimations of the cumulative hazard obtained by using the Nelson-Aalen estimator and the naive local constant estimator. We confirm the improved efficiency performance of the new estimator with respect to the Nelson-Aalen estimator. The efficiency gains are specially remarkable for low quantiles.

Application of the new methodology to the analysis of the remaining customer lifetime duration after the first cancellation.

In the empirical application, the general formulation of the new estimator is adapted to the estimation of the non parametric component of the proportional hazards regression model.

Therefore, this empirical study is at the same time illustrating one of the many possible applications of the naive local constant estimator when addressing new challenges in actuarial science. Apart from this, the new estimator can be directly applied to classical survival studies in life insurance statistics and survival analysis.

The methodology applied in our empirical application consist of two stages that conveniently fits the problem of analysing customer lifetime duration after the first policy cancellation. Even though it only analyses the time elapsed between two particular moments in the customer lifecycle, they have been chosen in order to provide a reasonable/right understanding of the relevant factors influencing customer residual lifetime duration after the first cancellation.

The first policy cancellation is clearly indicating a change in the insurance relationship from the customer point of view. Because of this fact, we chose this moment as the starting point of the customer lifecycle period we wanted to analyse.

Subsequent cancellations are of course important moments of the remaining customer lifecycle. Nevertheless, our empirical study has proved that, in many cases, the remaining policies are cancelled after a very short period of time.

For these customers it is probably more practical to have an estimation of the total remaining lifetime duration (time until all the remaining policies are cancelled) than only until the following cancellation. This was one of the reason for choosing the moment when all the remaining policies are cancelled as the terminal point of the customer lifecycle period we wanted to analyse. Other reasons are derived from the lack of availability of historical datasets about policy cancellation in a multiline context.

It is clear that incorporating information about subsequent cancellations would improve our understanding of the insurance customer behaviour, specially in the case of insureds with a long remaining lifetime duration.

Nevertheless, the estimation of the total remaining lifetime duration can be used in order to design an initial retention strategy (for a particular group of customers who have just made their first cancellation) and the information provided by prospective events in the customer lifecycle can always be incorporated in order to conveniently reconduct the initial retention strategy.

The proposed methodology lets us analyse the two most important elements of period going from the first cancellation to the total cancellation of all policies: a) the risk of cancelling all policies simultaneously (logistic regression model) and b) the remaining lifetime duration in case of a partial cancellation (proportional hazards regression model).

Our empirical results let us conclude that the propose methodology provides a reasonably good understanding the insurance customer lifecycle. We conclude that there is a significant methodological contribution in this thesis, because the customer behaviour in the insurance market has not been studied from this perspective before.

7.2 About the empirical application

Insurance purchasing is a complex process surrounding an intangible product[1] about which consumers are likely to be ignorant (Gravelle, 1994; Showers & Shotick, 1994 and Schlesinger & Schulenburg, 1993). Due to all the intangibility of the product, service, and purchasing process uncertainties, customers are likely to be particularly vulnerable to competitive influences that provide some certainty, for example, knowledge that the competitor offers a lower premium. Information provided by insurers to existing and potential customers may be particularly important to the brand switching potential of customers. And, the relationships formed by the insurer and its agents or representatives, will also be particularly important for customer retention (Crosby & Stephens, 1987).

The empirical study focuses on demand-side dynamics by analyzing household customer behaviour for a bundle of insurance products purchased from a single insurer. Many times a household will concentrate its policy portfolio with a single company making a policy lapse all the more important. Specifically, we focus on households for whom at least one policy has lapsed and investigate the effects for the lapse rates of other policies owned by the same household. Multiple risks, such as are reflected in multiple policies, and household decision-making have been shown to be important for understanding customer behaviours (cf., Bonato & Zweifel, 2002 and Dionne, Gourieroux & Vanasse, 1997).

In this manner, we are able to provide useful managerial information on the response time that the firm has from the first lapse signal to total customer defection (lapse of all policies purchased from the firm). This time frame is a window of opportunity for the firm to be able to stop the customer defection through customer relationship management techniques. From a pure financial perspective, it will be

[1]While it is common to speak about insurance as a product, conceptually from a marketing perspective, it is also a service. All products have facilitating services and all services have facilitating products, sometimes blurring the distinction between the product and the service for the customer. The agent may often be viewed as the service provider (to the customer), with the insurer being perceived as the producer of the product. This distinction may be very important to customer relationship management.

less costly for the firm to retain an existing customer, when possible, than it is to attract a new customer or win-back a customer who completed a total defection (lapse of all policies owned from a single customer or household).

This research also provides information on the customer variables and experiences that relate to cancellation behaviour. Specifically, we investigate customer demographics, customer history and firm experiences, the manner in which the cancellation is made, and household portfolio dynamics. This information allows the firm to be able to segment the customer market on the basis of demographics (and other variables), so as to better predict the likely time until defection from the first cancellation by customer characteristics[2].

We summarize the main conclusion in the same order as the objectives presented in the first chapter. We have reorganized the items in five general topics:

Estimation of the probability of a total cancellation. Determination of the factors associated to a higher risk of a total cancellation.

We have analyzed the behaviour of households having more than one policy in the same company (but not necessarily of the same type) that make a first cancellation. It can be either a total or a partial cancellation.

A logistic regression model has been used to estimate the probability of a total cancellation for a sample of customer from a Danish insurance company. Our results support the overall significance of the model. Additionally, its discrimination ability is reasonably good (it identifies around 71% of total policy cancellations).

The main advantage of this method is that it can be easily implemented in order to identify groups of households with a higher risk of simultaneous total cancellation of all their policies. This information can be used to design specific customer retention strategies (for another application in the Spanish market see Pujol, 2004)

The logistic regression results indicated that having an external company notify

[2] Jill Griffin discusses the general concepts of customer loyalty, retention and win-back and their relative financial costs in her two books, Customer Loyalty (Second Edition, Jossey-Bass, New York, 2004) and Customer Win-back (Josey-Bass, New York, 2003)

the customer cancellation is probably the most relevant factor associated to a higher risk of simultaneous total cancellation of all household policies.

Additionally, the occurrence of a claim is always associated to a higher risk of a total cancellation. Actually, both claims occurrence and change of address increase the probability of total cancellation, at this probability even increases further as time goes by.

We also found that customer retention begets customer retention: the longer the customer was with the firm, the less likely they were to cancel a policy. And, if one policy is cancelled, the longer the customer can be expected to remain with the firm after first cancellation. Thus, the already loyal customers seems to want to remain loyal, and be relatively reluctant to switch. These customers are critical and should not be ignored, i.e. a relationship with them needs to be a strategic focus. Loyal customers are often ignored in the question of market share expansion.

A warning note from this research is that core customers, those with multiple policies that receive a special treatment, are among the most likely to switch all policies simultaneously to an external company. These are the most valuable customers from a risk perspective, and those are the most likely to be recruited away, as other firms are likely to see them as valuable, also.

The occurrence of a premium increase within past 12 months is associated to a lower risk of a total cancellation. When the customer is paying a high price for an insurance contract is less likely to announce a total policy cancellation (at least in the short term) because probably he/she is waiting to take some profit from these high premium. In the long term the effect of this covariate is the opposite (this will be discussed in the following item).

The more in advance the customer is announcing the cancellation the lower is the risk of a total cancellation. Definitely, the marketing manager should take the advantage of the time elapsed between the notification and the moment when the risk is not covered any more in order to retain the customer.

In the Danish company dataset, those customers who have a contents policy have a higher risk to make a total cancellation than those who only have house

or motor policies. Therefore, special attention should be paid to customers with a contents policy in the company because in case that they decide to move some of their policies (not necessarily the contents policy) to another insurer, then the rest of policies are very likely to be moved as well.

Analysis of survival beyond the first cancellation. Determination of the factors associated to a higher risk of cancelling all the remaining policies (shorter residual lifetime).

When a partial cancellation occurs, this is the first evidence of a reconsideration of the insurance relationship from the customer point of view, and the insurer may have time to take action to retain the customer for other policies held by the household (multiple types of coverage). For the analysed data set, if policies of more than one type are cancelled, the expected remaining lifetime of the customer with the firm (before all policies are cancelled) is substantially smaller than when only one line of business is cancelled.

The analysis of customer residual lifetime duration after the first cancellation has been carried out by using two models: the proportional hazards regression model and the Tobit model. Generally speaking, the effect of each covariate on the customer remaining lifetime duration is the same for both models.

Our main conclusion, for the analysed data, is that the amount of time that the insurer has to retain the customer is dependent upon the type of customer. Our results let us conclude that the policy type first cancelled is one of the key factors in explaining the expected residual life, or the estimated time left until the customer-insurer relationship is terminated.

The knowledge of the specific estimated residual life corresponding to a particular customer will allow the insurer to more effectively tailor their marketing strategy to increase customer retention. For example, an informational campaign strategy, which may be effective if the expected residual life is long, may be ineffective if the expected customer lifetime duration with the firm is short. The procedures presented in this paper will allow the insurer to make such customer market segmentation

decisions, so as to more effectively manage the customer relationship.

In the real dataset that is studied in this thesis, we have seen that: External companies, change of address, claims and pruning are the strongest factors contributing to reducing the expected residual life. Nevertheless, a recent change of address slightly increases the expected time between the first policy lapse and the final termination of the last household policy.

The effect of claims in the reduction of the residual life is more remarkable as the time since the claim has occurred increases. A substantial rise in premiums is associated with a reduced residual life, but this effect is less remarkable the more time has passed since the premium increase.

The longer the customer was with the firm, the longer residual customer lifetime duration he/she has. Core customer status is also increasing the residual lifetime duration, even though it is associated to a higher probability of a total cancellation. Therefore, for those core customers who do not cancel all their policies simultaneously, the residual lifetime duration is longer than for non core customers.

Finally, the composition of the insurance portfolio before and after the first cancellation (the type of first cancellation) is a relevant factor explaining the remaining customer lifetime duration. Concerning the insurance portfolio of the household before the first lapse, the contribution of the house type of policy in the reduction of the residual lifetime duration is more remarkable than the one corresponding to the contents and motor types of policies. Regarding the insurance portfolio after the first lapse, the contribution of the contents type of policy in the increase of the residual lifetime duration is higher than the one corresponding to the motor and house types of policies.

Comparison between the proposed methodology and the Tobit model.

The proportional hazards regression model and the Tobit model have been used to obtain an estimation of customer residual lifetime durations and survival probabilities of different types of insureds.

Estimations about lifetime durations are a very useful information for the insurer in order to design retention strategies. One way to compare both methodologies is in terms of survival probabilities. In the case of the proportional hazards regression model, lifetime durations are obtained in some indirect way by integrating the corresponding survival curve. In the case of the Tobit model, residual lifetime duration is directly the dependent variable. In any case, these estimations depend on the censoring times. For example, in the case of the proportional hazards regression model, the higher the risk of cancelling all the remaining policies is the better the approximation to the real residual lifetime duration.

Therefore, we compared these two methods in terms of survival probabilities to given time thresholds. On the one hand, our results let us conclude that the proportional hazards regression model provides a better detection of customers with a high risk of a total cancellation within a short period of time than the Tobit model does. On the other hand, the Tobit model seems to better identify customers with a high probability of a total cancellation within longer periods of time. Nevertheless, in any case the overall performance of the proportional hazards regression model is preferable than the one corresponding to the Tobit model, specially if we take into account the cost of the potential under(over)-estimations of both models.

Extensions to the case where effects of covariates are allowed to vary over time in regression models in survival analysis.

Time-varying coefficients are included in the proportional hazards regression model in two empirical applications to our insurance data set. We have considered two randomly selected samples and specific covariates in each application.

Our results in any case are indicating that there is a clear evidence for the time-dependency of the effect of some risk factors on the probability to cancel all the remaining policies. Generally speaking, the overall effect of claims and the composition of the customer portfolio is dramatically changing at around the time point $t = 2.5$ years. Factors with a constant effect over time are change of address, core customer status and external companies.

Insurance business risk management.

The methodology applied to the empirical application provides an important tool for the design of marketing strategies and business risk management guidelines. The knowledge of the risk of a policy cancellation for each particular customer can be used in order to measure the overall business risk of the portfolio.

Probably, the most important contribution of this research is derived from the formulation of the problem itself. Our results support our initial intuition about how to set the problem out and provide useful insights applicable to marketing and business risk management in insurance.

The basic elements of the formulation of the problem are the conceptual framework (the concept of policy cancellation and the household as the individual in our study) and the multiline approach (multiple policies analysed simultaneously).

Marketing and business risk management in insurance should not consider the customer as the individual policy holder of a particular contract but a customer of the company as a whole. Our hypothesis is that any event in the customer lifecycle in one line of business is influencing the customer behaviour in the rest of them.

For the three types of insurance contracts considered in the applied part of this research (contents, house, and automobile), it may be the case that all adult household members participate in the decision to cancel (or to purchase). In this research we analyse the behaviour of households having more than one policy in the same company, but not necessarily of the same type. The marketing and business risk manager should link policy holders who are members of the same household, i.e. the same decision-making unit. The household is an appropriate unit of analysis for marketing multiple (or individual) insurance products to a decision-making unit[3].

[3]Even though linked as a household decision-making unit, the insurer can still recognize the different decision influences of various individuals (e.g., husband-dominate, wife-dominate, joint-equally, or delegated to one member exclusively). Future research can be conducted to determine the style of decision-making for various types of customers or policies held.

7.3 Extensions

This research has a number of relevant extensions both in its methodological and applied component. In this section they are briefly summarized.

About the methodology

We view the new estimator introduced in Chapter 3 as the simplest possible extension of the traditional Nelson-Aalen estimator, since it to some extent assumes the hazard to be constant in a neighbourhood of the point of interest, using an unweighted local approach for estimation. Therefore, our estimator has similarities to both the histogram estimator and the naive kernel density estimator, which leads to the choice of the name of the method: the naive local constant estimator. More sophisticated versions of the naive local constant estimator can be investigated, where non-uniform weight functions could be considered in order to improve its efficiency.

About the empirical application

The analysis of the complete customer lifecycle is the main extension of this research. The scarce availability of historical insurance data in a multiline context was our main obstacle. Nevertheless, as we mention in the introduction of this thesis, in the recent years, information systems and more advanced statistical packages have partly contributed to make information transfer and analysis much more efficient. Therefore, our intuition is that in a short period of time extension of the current investigation will be needed, in order to analyse longitudinal and multi-product information.

As we mention before, not all policy cancellations are the same. The incorporation to the analysis of information on premiums and claim compensations would let the insurer discriminate between policy cancellations made by bad customers and those made by good customers. This would also improve any estimation of the real

risk derived from the changing composition of the portfolio due to cancellations and new policy underwritings.

A more complete analysis of the time-varying effects of covariates in the survival model should be addressed, specially to incorporate all the explanatory variables in the model. This would provide a better understanding of the impact of each risk factor in customer lifetime duration. In order to do that, we may work with randomly selected small samples in order to overcome the problem of working with big data sets.

The implementation and follow-up of marketing strategies and business risk management guidelines is the final extension of the current research. Nevertheless, this research has provided a right understanding of the most relevant elements of the problem in order to carry out more advanced investigations within this field.

References

Aalen, O. O. (1978) "Nonparametric Inference for a Family of Counting Processes," *Annals of Statistics* 6, 701-726.

Agresti, A. (1990) *Categorical Data Analysis.* Wiley, New York.

Albers, W. (1999) "Stop-loss premiums under dependence," *Insurance: Mathematics and Economics* 24, 173-185.

Allenby, G. M., Leone, R. P. & Jen, L (1999) "A dynamic model of purchase timing with application to direct marketing," *Journal of the Americal Statistical Association* 94, 446, 365-374.

Altman, N. & Leger, C. (1995) "Bandwidth selection for kernel distribution function estimation," *Journal of Statistical Planning and Inference* 46, 195-214.

Andersen, P. K., Borgan, O., Gill, R. D. & Keiding, N. (1993) *Statistical models based on counting processes.* Springer-Verlag, New York.

Azzalini, A. (1981) "A note on the estimation of a distribution function and quantiles by a kernel method," *Biometrika* 68, 326-328.

Barrese, J., Doerpinghaus, H. & Nelson, J. (1995) "Do Independent Agent Insurers Provide Superior Service? The Insurance Marketing Puzzle," *Journal of Risk and Insurance* 62, 2, 297-308.

Bartlett, M. (1937) "Properties of sufficiency and statistical tests," *Proceedings of the Royal Society of London Series A* 160, 168-182.

Beirlant, J., Derveaux, V., De Meyer, A. M., Goovaerts, M.J., Labie, E. & Maenhoudt, B. (1991) "Statistical risk evaluation applied to (Belgian) car insurance," *Insurance: Mathematics and Economics* 10, 289-302.

Ben-Arab, M., Brys, E. & Schlesinger, H. (1996) "Habit Formation and the Demand for Insurance," *Journal of Risk and Insurance* 63, 1, 111-119.

Bolton, R. N. (1998) "A dynamic model of the duration of the customer's relationship with a continuous service provider: the role of satisfaction," *Marketing Science* 17, 1, 45-65.

Bon, J. & Tissier-Desbordes, E. (2000) "Fidéliser les clients? Oui, mais...," *Revue Française de Gestion* 127, 52-60.

Bonato, D. & Zweifel, P. (2002) "Information about Multiple Risks: The Case of Building and Contents Insurance," *Journal of Risk and Insurance* 69, 4, 469-487.

Bowers, N. L. JR., Gerber H. U., Hickman, J. C., Jones, D. A. & Nesbitt, C. J. (1997) *Actuarial Mathematics*, 2nd Edition. The Society of Actuaries, Schaumburg.

Bowman, A. (1984) "An alternative method of crossvalidation for the smoothing of density estimates," *Biometrika* 71, 353-360.

Bowman, A., Hall, P. & Prvan, T. (1998) "Bandwidth selection for the smoothing of distribution functions," *Biometrika* 85, 4, 799-808.

Brockett, P. L., Golden, L. L., Guillen, M., Nielsen, J. P., Parner, J. & Perez-Marin, A. M. (2005) "Household multiple policy retention effects of first policy cancellation: how much time do you have to stop total customer defection?," submitted for publication.

Brown, G. H. (1952) "Brand loyalty - fact or fiction?," *Advertising Age* 9, 53-55.

Cooley, S. (2002) "Loyalty Strategy Development Using Applied Member-Cohort Segmentation," *Journal of Consumer Marketing* 19, 7, 550-563.

Cox, D. R. (1972) "Regression Models and Life Tables," *Journal of the Royal Statistical Society B* 34, 187-220.

Crosby, L. A. & Stephens, N. (1987) "Effects of Relationship Marketing on Satisfaction, Retention, and Prices in the Life Insurance Industry," *Journal of Marketing Research* 24, 4, 404-411.

Czado, C. & Rudolph, F. (2002) "Application of survival analysis methods to long-term care insurance," *Insurance: Mathematics and Economics* 31, 396-413.

Day, G. S. (1969) "A two dimensional concept of brand loyalty," *Journal of Advertising Research* 9 (September), 29-36.

Dhaene, J., Vanduffel, S., Tang, Q. H., Goovaerts, M., Kaas, R. & Vyncke, D. (2004) "Capital requirements, risk measures and comonotonicity," *Belgian Actuarial Bulletin* 4, 53-61.

Dionne, G., Gourieroux, C. & Vanasse, C. (1997) *The Informational Content of Household Decisions with Application to Insurance under Adverse Selection.* Universite de Montreal, Montreal.

Doherty, N. A. & Schlesinger, H. (1983) "Optimal Insurance in Incomplete Markets," *Journal of Political Economy* 91, 1045-1054.

Dwyer, F. R. (1997) "Customer lifetime valuation to support marketing decision making," *Journal of Direct Marketing* 11 (Fall), 6-13.

Efron, B. (1977) "The Efficiency of the Cox's Likelihood Function for Censored Data," *Journal of the American Statistical Association* 72, 557-565.

Falk, M. (1983) "Relative efficiency and deficiency of kernel type estimators of smooth distribution functions," *Statistica Neerlandica* 37, 73-83.

Fan, J. & Gijbels, I. (1996) *Local Polynomial Regression.* Chapman and Hall, London.

Fornell, C. (1992) "National satisfaction barometer: the Swedish experience," *Journal of Marketing* 56 (January), 6-21.

Fourier, S. & Yao, J. L. (1997) "Reviving brand loyalty: a reconceptualization within the framework of customer-brand relationships," *International Journal of Research in Marketing*, 14, 5, 451-472.

Gandolfi, A. S. & Miners, L. (1996) "Gender-based Differences in Life Insurance Ownership," *Journal of Risk and Insurance* 63, 4, 683-693.

Gollier, C. & Scarmure, P. (1994) "The Spillover Effect of Compulsory Insurance," *Geneva Papers on Risk and Insurance Theory* 19, 1, 23-34.

Gompertz, B. (1825) "On the nature of the function expressive of the law of human mortality and on the new mode of determining the value of life contingencies," *Philosophical Transactions of the Royal Society of London* 115, 513-585.

Grambsch, P. M. & Therneau, T. M. (1994) "Proportional hazards tests and diagnostics based on weighted residuals," *Biometrika* 81, 515-526.

Gravelle, H. (1994) "Remunerating Information Providers: Commissions versus Fees in Life Insurance," *Journal of Risk and Insurance* 61, 3, 425-457.

Griffin, J. (2003) *Customer Win-Back.* Josey-Bass, New York.

Griffin, J. (2004) *Customer Loyalty*, Second Edition. Josey-Bass, New York.

Guillen, M., Nielsen, J. P. & Perez-Marin, A. M. (2005) "Improving the efficiency of the Nelson-Aalen estimator: the naive local constant estimator," submitted for publication.

Guillen, M., Nielsen, J. P. & Perez-Marin, A. M. (2006) "Multiplicative hazard models for studying the evolution of mortality," *Annals of Actuarial Science*, accepted for publication.

Guillen, M., Parner, J., Densgsoe, C. & Perez-Marin, A. M. (2003) "Using logistic regression models to predict and understand why customer leave an insurance company," in *Intelligent and other Computational Techniques in Insurance. Theory and Applications*, Lakhmi Jain and Arnold Shapiro eds. World Scientific, pp 465-490.

Gustafsson, J., Guillen, M., Nielsen, J. P. & Pritchard, P. (2005) "Using external data in the calculation of operational risk capital requirements with particular reference to under-reporting," Working paper University of Barcelona.

Haberman S. & Pitacco, E. (1996) *Actuarial models for disability insurance.* Chapman and Hall/CRC Press, Boca Raton.

Helsen, K. & Schmittlein, D. C. (1993) "Analysing duration times in marketing: evidence for the effectiveness of hazard rate models," *Marketing Science* 11 (Fall), 395-414.

Herbst, T. (1999) "An application of randomly truncated data models in reserving IBNR claims," *Insurance: Mathematics and Economics* 25, 123-131.

Hougaard, P. (1995) "Frailty models for survival data," *Lifetime Data Analysis* 1, 255-273.

Jackson, D. (1989) "Determining a Customer's Lifetime Value," *Direct Marketing* 52, 1, 24-32.

Jacoby, J. & Chesnuy, R. (1978) *Brand loyalty: measurement and management.* Willey, New York.

Jones, M. C. (1993) "Simple boundary correction for kernel density estimation," *Statistics and Computing* 3, 135-146.

Jones, C., Linton, O. & Nielsen, J. P. (1995) "A simple bias reduction method for density estimation," *Biometrika* 82, 2, 327-338.

Kapferer, J. N. & Laurent, G. (1983) *La sensibilité aux marques.* Foundation Jours de France, Paris.

Kaplan, E. L. & Meier, P. (1958) "Non-parametric estimation from incomplete observations," *Journal of the American Statistical Association* 53, 457-481, 562-563.

Keiding, N., Andersen, P. K. & Klein, J. P. (1997) "The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates," *Statistics in Medicine* 16, 215-224.

Klein, J. P. & Moeschberger, M. L. (1997) *Survival analysis techniques for censored and truncated data*, Springer Verlag, New York.

Kuo, W., Tsai, C. & Chen, W. -K. (2003) "An Empirical Study on the Lapse Rate: The Cointegration Approach," *Journal of Risk and Insurance* 70, 3, 489-501.

Lawley, D. N. (1956) "A general method for approximating the distribution of likelihood ratio criteria," *Biometrika* 43, 295-303.

Levitt, T. (1988) *La comercialización creativa.* Compañía Editorial Continental S. A., Mexico.

Li, S. (1995) "Survival analysis," *Marketing Research* 7 (Fall), 17-23.

Lin, D. Y., Fleming, T. & Wei, L. J. (1994) "Confidence bands for survival curves under the proportional hazads model," *Biometrika* 81, 73-81.

Lin, D. Y., Wei, L. J. & Ying, Z. (1993) "Checking the Cox model with cumulative sums of martingale-based residuals," *Biometrika* 80, 557-572.

Linder, U. & Ronkainen, V. (2004) "Towards a new insurance supervisory system in the EU," *Scandinavian Actuarial Journal* 6, 462-474.

Macdonald, A. S. (1996) "An actuarial survey of statistical models for decrement and transition data, II: competing risks, non-parametric and regression models," *British Actuarial Journal* 2, 2,429-448.

Makeham, W. M. (1860) "On the law of mortality and the construction of annuity tables," *Journal of the Institute of Actuaries* 8, 301-310.

Martinussen, T., Scheike, T. H. & Skovgaard, I. (2002) "Efficient estimation of fixed and time-varying covariate effects in multiplicative intensity models," *Scandinavian Journal of Statistics* 29, 57-75.

Mayers, D. & Smith, C. W. Jr. (1983) "The Interdependence of Individual Portfolio Decisions and the Demand for Insurance," *Journal of Political Economy* 91, 304-311.

Mittal, V. & Kamakura, W. A. (2001) "Satisfaction, repurchase intent, and repurchase behaviour: investigating the moderating effect of customers characteristics," *Journal of Marketing Research* 38, 131-142.

Murphy, S. A. (1993) "Testing for time dependent coefficient in Cox's regression model," *Scandinavian Journal of Statistics* 20, 35-50.

Murphy, S. A. & Sen, P. K. (1991) "Time-dependent coefficients in a Cox-type regression model," *Stochastic Processes and Their Applications* 39, 153-180.

Murray, R. (1988) "Up the loyalty ladder," *Direct Marketing*, December.

Nadaraya E. A. (1964) "Some new estimates for distribution functions," *Theory of Probability and its Applications* 9, 497-500.

Nelson, W. (1969) "Hazard plotting for incomplete failure data," *Journal of Quality Technology* 1, 27-52.

Nelson, W. (1972) "Theory and applications of hazard plotting for censored failure data," *Technometrics* 14, 945-965.

Nielsen, J. P. & Tanggaard, C. (2001) "Boundary and bias correction in kernel hazard estimation," *Scandinavian Journal of Statistics* 28, 675-698.

Olivieri, A. (2003) "Allowing for heterogeneity in life insurance reserving," *Proceedings of the 7th Insurance: Mathematics and Economics Congress*, Lyon, 2003.

Pujol, M. (2004) "El valor y la fidelidad de los asegurados en el ramo del automóvil," Tesis Doctoral, Universidad de Barcelona.

Ramlau-Hansen, H. (1983) "Smoothing counting process intensities by means of kernel functions," *Annals of Statistics* 11, 453-466.

Reinartz, W. J. & Kumar, V. (2000) "On the profitability of long-life customers in a noncontractual setting: an empirical investigation and implications for marketing," *Journal of Marketing* 64 (October), 17-35.

Reinartz, W. J. & Kumar, V. (2003) "The impact of customer relationship characteristics on profitable lifetime duration," *Journal of Marketing* 67, 77-99.

Reiss, R. -D. (1981) "Nonparametric estimation of smooth distribution functions," *Scandinavian Journal of Statistics* 8, 116-119.

Riechheld, F. (1996) *The loyalty effect.* Harvard Business School Press.

Rudemo, M. (1982) "Empirical choice of histograms and kernel density estimators," *Scandinavian Journal of Statistical* 9, 65-78.

Sarda, P. (1993) "Smoothing parameter selection for smooth distribution functions," *Journal of Statistical Planning and Inference* 35, 65-75.

Scheike, T. H. & Martinussen, T. (2004) "On estimation and tests of time-varying effects in the proportional hazards model," *Scandinavian Journal of Statistics* 31, 51-62.

Schlesinger, H. & Doherty, N. A. (1985) "Incomplete Markets for Insurance: An Overview," *Journal of Risk and Insurance* 52, 402-423.

Schlesinger, H. & Schulenburg, J. M. (1993) "Customer Information and Decisions to Switch Insurers," *Journal of Risk and Insurance* 60, 4, 591-615.

Schmittlein, D. C. & Peterson, R. A. (1994) "Customer base analysis: an industrial purchase process application," *Marketing Science* 13, 1, 41-67.

Schoenfeld, D. (1982) "Partial residuals for the proportional hazards regression model," *Biometrika* 69, 239-241.

Showers, V. & Shotick, J. (1994) "The Effect of Household Characteristics on Demand for Insurance: A Tobit Analysis," *Journal of Risk and Insurance* 61(3), 492-502.

Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

Snell, E. J. & Cox, D. R. (1989) *Analysis of Binary Data.* Chapman and Hall, London.

Stafford, M. R., Stafford, T. F. & Wells, B. P. (1998) "Determinants of Service Quality and Satisfaction in the Auto Casualty Claims Process," *Journal of Services Marketing* 12, 6, 426-40.

Thomas, J. S. (2001) "A methodology for linking customer acquisition to customer retention," *Journal of Marketing Research* 38, 2, 262-268.

Tobin, J. (1958) "Estimation of relationships for limited dependent variables," *Econometrica* 26, 24-36.

Tsai, W. -Y., Jewell, N. P. & Wang, M. -C. (1987) "A note on the product-limit estimator under right censoring and left truncation," *Biometrika* 74, 883-886.

Uncles, M. & Laurent, G. (1997) Editorial. *International Journal of Research in Marketing* 14, 5, 399-404.

Van der Laan, M. J., Jewell, N. P. & Peterson, D. R. (1997) "Efficient estimation of the lifetime and disease onset distribution," *Biometrika* 84, 539-554.

Wand, M. P. & Jones, M. C. (1995) *Kernel Smoothing.* Chapman and Hall, London.

Weibull, W. A. (1951) "A statistical distribution of wide applicability," *Journal of Applied Mechanics* 18, 293-297.

Wells, B. P. & Stafford, M. R. (1995) "Service Quality in the Insurance Industry. Customer Perception versus Regulatory Perceptions," *Journal of Insurance Regulation* 13, 4, 462-477.

Appendix A. Expansion of the naive local constant estimator when $Y(s) > 0$ for $s \in [0, t+b]$.

$$
\begin{aligned}
\hat{\Lambda}_{NLC}(t) - \tilde{\Lambda}(t) &= \sum_{i=1}^{n} \int_0^{\max(t-b,0)} \frac{1}{Y(s)} dN_i(s) + \sum_{i=1}^{n} \int_{\max(t-b,0)}^{t+b} \frac{1}{\gamma_{t,b} Y(s)} dN_i(s) \\
&\quad - \int_0^t \alpha(s) ds \\
&= \sum_{i=1}^{n} \int_0^{\max(t-b,0)} \frac{1}{Y(s)} dN_i(s) - \int_0^{\max(t-b,0)} \alpha(s) ds \\
&\quad + \sum_{i=1}^{n} \int_{\max(t-b,0)}^{t+b} \frac{1}{\gamma_{t,b} Y(s)} dN_i(s) - \int_{\max(t-b,0)}^{t} \alpha(s) ds \\
&= \sum_{i=1}^{n} \int_0^{\max(t-b,0)} \frac{1}{Y(s)} dM_i(s) + \sum_{i=1}^{n} \int_0^{\max(t-b,0)} \frac{1}{Y(s)} d\Lambda_i(s) \\
&\quad - \int_0^{\max(t-b,0)} \alpha(s) ds + \sum_{i=1}^{n} \int_{\max(t-b,0)}^{t+b} \frac{1}{\gamma_{t,b} Y(s)} dM_i(s) \\
&\quad + \sum_{i=1}^{n} \int_{\max(t-b,0)}^{t+b} \frac{1}{\gamma_{t,b} Y(s)} d\Lambda_i(s) - \int_{\max(t-b,0)}^{t} \alpha(s) ds \\
&= \sum_{i=1}^{n} \int_0^{\max(t-b,0)} \frac{1}{Y(s)} dM_i(s) + \int_0^{\max(t-b,0)} \frac{1}{Y(s)} \alpha(s) \sum_{i=1}^{n} Y_i(s) ds \\
&\quad - \int_0^{\max(t-b,0)} \alpha(s) ds + \sum_{i=1}^{n} \int_{\max(t-b,0)}^{t+b} \frac{1}{\gamma_{t,b} Y(s)} dM_i(s) \\
&\quad + \int_{\max(t-b,0)}^{t+b} \frac{1}{\gamma_{t,b} Y(s)} \alpha(s) \sum_{i=1}^{n} Y_i(s) ds - \int_{\max(t-b,0)}^{t} \alpha(s) ds
\end{aligned}
$$

$$= \sum_{i=1}^{n} \int_{0}^{\max(t-b,0)} \frac{1}{Y(s)} dM_i(s) + \sum_{i=1}^{n} \int_{\max(t-b,0)}^{t+b} \frac{1}{\gamma_{t,b} Y(s)} dM_i(s)$$
$$+ \int_{\max(t-b,0)}^{t+b} \frac{1}{\gamma_{t,b}} \alpha(s) ds - \int_{\max(t-b,0)}^{t} \alpha(s) ds.$$

Appendix B. SAS program for the estimation of the naive local constant estimator.

```
proc iml;
  start k(x);
  b=; /* BANDWIDTH PARAMETER */
  if (x>=-b) & (x<=b) then c=(15/16)*(1/b)*((1-(x/b)**2)**2);
  if (x<-b) then c=0;
  if (x>b) then c=0;
  return (c);
  finish k;
  r=; /* VECTOR OF INDIVIDUAL SURVIVAL TIMES */
  d=; /* VECTOR OF INDIVIDUAL CENSORING - EVENT INDICATOR */
  y=; /* VECTOR OF INDIVIDUAL EXPOSITIONS TO RISK */
  mil=; /*MAXIMUM SURVIVAL TIME */
  milu=mil+1;
  nou=mil-1;
  n=nrow(r);
  in1=j(milu,3,0);
  in1[i,]=; /* SURVIVAL TIMES */
  /* LOCAL LINEAR ESTIMATOR */
  do i=1 to n;
  ni=r[i]+1;
```

```
if ((milu-ni)>0) then vec1=j(ni,1,1)//j((milu-ni),1,0);
if ((milu-ni)=0) then vec1=j(ni,1,1);
if ((milu-ni)<0) then vec1=j(milu,1,1);
in1[,2]=in1[,2]+(vec1#y[i]);
if ((milu-ni)>0) then in1[ni,3]=in1[ni,3]+(1-d[i]);
if ((milu-ni)<=0) then in1[milu,3]=in1[milu,3]+(1-d[i]);
end;
sum2=0;
sum3=0;
sum4=0;
sum5=0;
sum6=0;
sum7=0;
aj2=(1:mil+1);
aj3=(1:mil+1);
aj4=(1:mil+1);
aj5=(1:mil+1);
aj6=(1:mil+1);
aj7=(1:mil+1);
do t1=1 to (mil+1);
do t2=2 to (mil+2);
dife=((t1-1)-(t2-2))/10;
sum2=sum2+k(dife)*1*in1[t2-1,2]*0.1;
sum3=sum3+(k(dife))*(dife)*in1[t2-1,2]*0.1;
sum4=sum4+(k(dife))*((dife)**2)*in1[t2-1,2]*0.1;
sum5=sum5+(k(dife))*1*in1[t2-1,3];
sum6=sum6+(k(dife))*((dife)**1)*in1[t2-1,3];
end;
aj2[t1]=sum2;
sum2=0;
```

```
aj3[t1]=sum3;
sum3=0;
aj4[t1]=sum4;
sum4=0;
aj5[t1]=sum5;
sum5=0;
aj6[t1]=sum6;
sum6=0;
end;
do t1=1 to (mil+1);
do t2=2 to (mil+2);
dife=((t1-1)-(t2-2))/10;
if ((((aj2[t1]*aj4[t1])-((aj3[t1])**2)))^=0) then kbar=(aj4[t1]*k(dife)-(aj3[t1]*
k(dife)*(dife)))/((aj2[t1]*aj4[t1])-((aj3[t1])**2));
if ((((aj2[t1]*aj4[t1])-((aj3[t1])**2)))^=0) then sum7=sum7+kbar*in1[(t2-1),3];
end;
aj7[t1]=sum7;
sum7=0;
end;
aj2=aj2';
aj3=aj3';
aj4=aj4';
aj5=aj5';
aj6=aj6';
alpa=aj7';
ac=(1:mil+1);
ac[1]=aj7[1]*0.1;
do ss=2 to (mil+1);
ac[ss]=ac[ss-1]+aj7[ss]*0.1;
end;
```

```
act=ac';
der=(1:mil+1);
do ff=1 to (mil+1);
der[ff]=(aj6[ff]-aj7[ff]*aj3[ff])/aj4[ff];
end;
dert=der';
acder=(1:(mil+1));
acder[1]=der[1];
do gg=2 to (mil+1);
acder[gg]=acder[gg-1]+der[gg]*0.1;
end;
acdert=acder';
der2=(1:mil+1);
do ff=1 to (mil+1);
if (aj3[ff]^=0) then der2[ff]=(aj5[ff]-aj7[ff]*aj2[ff])/aj3[ff];
end;
dert2=der2';
acder2=(1:mil+1);
acder2[1]=der2[1];
do gg=2 to (mil+1);
acder2[gg]=acder2[gg-1]+der2[gg]*0.1;
end;
acdert2=acder2';
time=(0:mil);
base=time'||in1[,2]||in1[,3]||aj2||aj3||aj4||aj5||aj6||alpa||act||dert||acdert||dert2||acdert2;
dera=j(milu,1,0);
do kk=1 to milu;
dera[kk]=base[kk,11]**2;
end;
sum1=0;
```

```
sum2=0;
pp=nrow(dera)-1;
vec1=(1:pp+1);
vec2=(1:pp+1);
do t1=1 to (pp+1);
do t2=2 to (pp+2);
sum1=sum1+k2((t1-1)-(t2-2))*dera[t2-1];
sum2=sum2+k2((t1-1)-(t2-2));
end;
vec1[t1]=sum1;
vec2[t1]=sum2;
sum1=0;
sum2=0;
end;
vec3=vec1/vec2;
basef=base||vec3';
bb=j(milu,1,0);
do i=1 to milu;
if ((basef[i,9]/(2*basef[i,2]*(basef[i,15])))>0 & (2*basef[i,2]*(basef[i,15]))>0) then
bb[i]=(((basef[i,9]/(2*basef[i,2]*(basef[i,15])))**(1/3)));
end;
bopt=j(milu,1,0);
efi=j(milu,1,0);
do i=1 to milu;
bopt[i]=(basef[i,9]>0)*(bb[i]<(basef[i,1]))*bb[i]+
(basef[i,9]>0)*(bb[i]>(basef[i,1]))*basef[i,1]+
(basef[i,9]<=0)*0;
if (basef[i,10]^=0 & basef[i,9]>0 & basef[i,10]>0) then efi[i]=(bopt[i]<=basef[i,1])*
(3/8)*bopt[i]*basef[i,9]/basef[i,10]+
(bopt[i]>basef[i,1])*(basef[i,1]/2)*(basef[i,9]-(basef[i,11]**2)*basef[i,2]*
```

```
(basef[i,1]**3)/2)/basef[i,10];;
end;
basefin=basef||bopt||efi;
create lc from base; /* RESULTS LOCAL LINEAR ESTIMATOR */
append from base;
close lc;
create b_opt from basef; /* OPTIMAL BANDWIDTH PARAMETER AND
EFFICIENCY ESTIMATION */
append from basef;
close b_opt;
run;
/* LOCLIN CONTAINS THE 14 VARIABLES OF LC, THE ESTIMATION
OF THE SQUARED HAZARD DERIVATIVE AND THE ESTIMATION
OF THE OPTIMAL B */
proc iml;
use res.loclin;
read all into loclin;
inte=1;
pp=nrow(loclin)-1;
naiv=(1:pp+1);
ind=(1:pp+1);
do i=1 to 1;
t=loclin[i,1];
b=int(loclin[i,16]);
mm1=max((t-b),0);
c=0;
do j=1 to mm1+1;
c=c+loclin[j,3]/loclin[j,2];
end;
r=0;
```

```
mm2=min(t+b,loclin[nrow(loclin),1]);
do s=1 to mm2+1;
r=r+loclin[s,3]/loclin[s,2];
end;
naiv[i]=c*0.5+0.5*r;
end;
do i=2 to (pp+1) ;
t=loclin[i,1];
b=int(loclin[i,16]);
mm1=max((t-b),0);
c=0;
do j=1 to mm1+1;
c=c+loclin[j,3]/loclin[j,2];
end;
r=0;
mm2=min(t+b,loclin[nrow(loclin),1]);
do s=1 to mm2+1;
r=r+loclin[s,3]/loclin[s,2];
end;
naiv[i]=c*0.5+0.5*r;
ind[i]=((c*0.5+0.5*r)>=naiv[i-1]);
end;
time=(0:pp);
base2=time'||naiv'||ind';
create res.nlc from base2;
append from base2;
close res.nlc;
```

Appendix C. SAS program for the calculation of expectations for the Tobit model.

```
proc iml;
   use res.outest; /* RES.OUTEST CONTAINS PARAMETER ESTIMATES
   OF THE TOBIT MODEL */
   read all var {_scale_} into sigma;
   use res.outest;
   read all var {
   /* INTERCEPT AND EXPLANATORY COVARIATES */
   } into b;
   n=; /* NUMBER OF DIFFERENT TYPES OF CUSTOMERS FOR WHOM
   WE WANT TO GET THE ESTIMATION */
   xt={
   }; /* VECTOR OF COVARIATES FOR THESE CUSTOMERS */
   print xt;
   print b;
   x=xt';
   baseini=j(n,3,0);
   do kk=1 to n;
   baseini[kk,1]=kk;
   end;
   baseini[,2]=(b*x)';
```

```
baseini[,3]=; /* CENSORING TIMES */
integ=j(n,2,0);
do i=1 to n;
alpha_a= (0-baseini[i,2])/sigma;
alpha_b= (baseini[i,3]-baseini[i,2])/sigma;
a = 0;
b = baseini[i,3];
inc1=(alpha_b-alpha_a)/2000;
inc2=(b-a)/2000;
sum1 = 0;
do j=1 to 2000;
sum1 = sum1 + (alpha_a+j*inc1)*pdf('NORMAL',(alpha_a+j*inc1))*inc1;
end;
integ[i,2]=sum1;
integ[i,1]=baseini[i,1];
end;
inte=integ||baseini[,2]||baseini[,3];
create res.merg from inte;
append from inte;
close res.merg;
run;
data res.merg;
set res.merg;
rename col1=numperf col2=integ1 col3=Xbeta col4=dif;
run;
data res.out1;
drop _scale_;
set res.merg;
a=0;
b=dif;
```

```
if _n_ eq 1 then set res.outest;
Predic1= a*cdf('NORMAL',(a-Xbeta)/_scale_)+b*(1-cdf('NORMAL',
(b-Xbeta)/_scale_))+(cdf('NORMAL',(b-Xbeta)/_scale_)-cdf('NORMAL',
(a-Xbeta)/_scale_))*Xbeta + _scale_*integ1;
label Xbeta='MEAN OF UNCENSORED VARIABLE'
Predic1 = 'MEAN OF CENSORED VARIABLE';
run;
```

Appendix D. Cox model with time-varying coefficients. R program with timereg library.

```
library(timereg,survival)
    library(survival)
    mostra<-_data set_
    out<-timecox(Surv(reslife,status)~_list of covariates_,mostra,
    max.time=,n.sim=,band.width=)
    summary(out)
    par(mfrow=c(3,4))
    plot(out)
```